BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 7 2001
Pages 664–668

T-REX: reconstructing and visualizing phylogenetic trees and reticulation networks

Vladimir Makarenkov

Département de Sciences Biologiques, Université de Montréal, CP 6128, succ. Centre-ville, Montréal, Québec H3C 3J7, Canada and Institute of Control Sciences, 65 Profsoyuznaya, Moscow 117806, Russia

Received on January 10, 2001; accepted on April 3, 2001

ABSTRACT

Summary: T-REX (tree and reticulogram reconstruction) is an application to reconstruct phylogenetic trees and reticulation networks from distance matrices. The application includes a number of tree fitting methods like NJ, UNJ or ADDTREE which have been very popular in phylogenetic analysis. At the same time, the software comprises several new methods of phylogenetic analysis such as: tree reconstruction using weights, tree inference from incomplete distance matrices or modeling a reticulation network for a collection of objects or species. T-REX also allows the user to visualize obtained tree or network structures using Hierarchical, Radial or Axial types of tree drawing and manipulate them interactively.

Availability: T-REX is a freeware package available online at: http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex

Contact: makarenv@magellan.umontreal.ca or casgrain@magellan.umontreal.ca

INTRODUCTION

A tree is a formal structure used for the representation of the process of evolution. The leaves represent the species under study, the interior nodes represent virtual ancestors and the branches represent evolutionary events. In biology, trees are called *phylogenetic trees*, or *additive trees* if tree branches have valuations. The principal goal of phylogenetic reconstruction is to infer a phylogenetic tree from imperfect contemporary data, which do not correspond directly to any tree topology. Consequently, one should utilize available fitting methods to obtain the data corresponding to phylogenetic tree. Five fitting algorithms for inferring a phylogenetic tree from distance matrices are provided by T-REX.

In phylogenetics, techniques for generating distances (e.g. DNA hybridization) are subject to measurement errors, so that the experimental values do not always satisfy the mathematical properties of distance matrices. For example, the famous *four-point condition* (Zaretskii, 1965; Buneman, 1971):

$$d(i,j)+d(k,l) \leqslant \max\{d(i,k)+d(j,l); d(i,l)+d(j,k)\},$$
for any four objects i, j, k, and l,

enabling one to infer a unique phylogenetic tree associated with the given distance measure d, is never met when raw empirical data are considered. As a matter of fact, even the *triangular inequality*:

$$d(i,j) \leqslant d(i,k)+d(j,k), \text{ for any three objects } i, j, \text{ and } k,$$

is not always satisfied for empirical matrices of evolutionary distances. Moreover, some of the entries of an empirical distance matrix can be unknown. Recently a number of methods for reconstruction of phylogenetic trees from partial distance or dissimilarity data have been proposed. T-REX allows the user to carry out four of these recently developed methods.

From the biological point of view a reticulogram (or a reticulation network) represents an evolutionary structure in which the species may be related in a non-unique way to a common ancestor. A phylogenetic tree cannot represent such a structure. Reticulate patterns have been 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 (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. When biological data are analyzed the reticulation branches linking the species or their ancestors can be interpreted as mutation or hybridization events that have occurred during the evolution pro-

cess. The reticulation branches can also represent the phenomenon of homoplasy or parallel evolution that might have taken place in the past. Applications of the reticulogram reconstruction algorithm implemented in T-REX can be found in Makarenkov and Legendre (2000, 2001), or in the Special Section of the *Journal of Classification* (see Lapointe *et al.*, 2000) dedicated to the reticulate evolution.

SYSTEMS AND METHODS

T-REX carries out *five methods of fitting a phylogenetic tree* to a given distance or dissimilarity matrix between objects, namely:

(1) *ADDTREE* by Sattath and Tversky (1977).

(2) *Neighbor-Joining* (NJ) by Saitou and Nei (1987).

(3) *Unweighted Neighbor-Joining* (UNJ) by Gascuel (1997).

(4) *Circular order reconstruction* by Makarenkov and Leclerc (1997), and Yushmanov (1984).

(5) *Weighted least-squares MW* by Makarenkov and Leclerc (1999).

The first two methods, ADDTREE and NJ, are the most frequently used methods for inferring phylogenetic trees from evolutionary distances. They reconstruct a phylogenetic tree structure starting from a star tree that contains n leaves associated with the objects and $n - 1$ branches. The star tree is repeatedly developed by adding new internal nodes to it until a binary tree structure consisting of $2n - 2$ nodes and $2n - 3$ branches is obtained. The third method, called UNJ, uses at each step the same as NJ selection criterion, but more general estimation and reduction formulae than NJ to infer the tree. Such a strategy usually enables one to obtain better results using UNJ than NJ without increasing the time complexity (Gascuel, 1997). The fourth method reconstructs a phylogenetic tree using circular orders of objects associated with a given dissimilarity. This fitting method, discussed in Makarenkov and Leclerc (1997), was inspired by Yushmanov's (1984) paper which introduced the notion of circular orders of objects corresponding to the circular (say, clockwise) scanning of leaves of a tree drawn in the plane. The fifth method, called Method of Weights (MW), looks for the best phylogenetic tree with respect to dissimilarity and weight matrices supplied by the user. This method allows the usage of an arbitrary matrix of weights, which may be chosen according to one of the classical weighting models existing in the literature. The MW method can be carried out without taking into account the matrix of weights if the power value appearing in the tree reconstruction dialogue box is set to 0 (default option).

T-REX also performs four methods of *fitting a phylogenetic tree to a given distance or dissimilarity matrix containing missing entries*, namely:

(1) *Triangle method* by Guénoche and Grandcolas (1999).

(2) *Ultrametric procedure* for estimation of missing values by De Soete (1984) and Landry *et al.* (1996) followed by MW method.

(3) *Additive procedure* for estimation of missing values by Landry *et al.* (1996) followed by MW method.

(4) *Weighted least-squares method MW* by Makarenkov and Leclerc (1999), performed with the option giving weight 1 to the existing entries and weight 0 to the missing ones.

The first method, called the Triangle method, reconstructs a tree from a given partial data table in a sequential way; the objects are added one by one to the growing tree according to the order defined by the algorithm. The second method, called Ultrametric estimation + MW, proceeds by carrying out two algorithms. The first one consists of applying the ultrametric inequality to assess all the missing values in the dissimilarity matrix (see De Soete, 1984 or Landry *et al.*, 1996 for more details). It is followed by the MW algorithm by Makarenkov and Leclerc (1999) on the completed dissimilarity matrix. The third method available is named Additive estimation + MW. This is the same as the second method, except the first step is replaced by an additive estimation, in which the four-point condition is applied to assess the missing values in the dissimilarity matrix (see Landry *et al.*, 1996, for more details). The fourth method available for reconstruction of trees from partial matrices is a modified version of the method of weights (MW) in which all missing entries of the dissimilarity matrix receive weight 0, whereas all present entries receive weight 1. In a tree reconstruction process, such a strategy enables one to take into account only existing dissimilarity entries.

The *reticulation network* (e.g. *reticulogram or reticulated cladogram*) *reconstruction* algorithm available in T-REX works by starting with a phylogenetic tree, which provides the initial fit for the given distance matrix. The algorithm then adds new branches into the growing reticulogram. To add a new branch, the algorithm minimizes the least-squares loss function computed as the sum of the quadratic differences between the original dissimilarities and the associated reticulogram distances. Two statistical criteria are incorporated to measure the gain in fit when new branches are added. The minimum of each of these criteria may be used as a stopping rule for addition of new (reticulation) branches. Thus, the user can either select an appropriate criterion to stop the procedure of adding reticulation branches or can indicate an exact number of

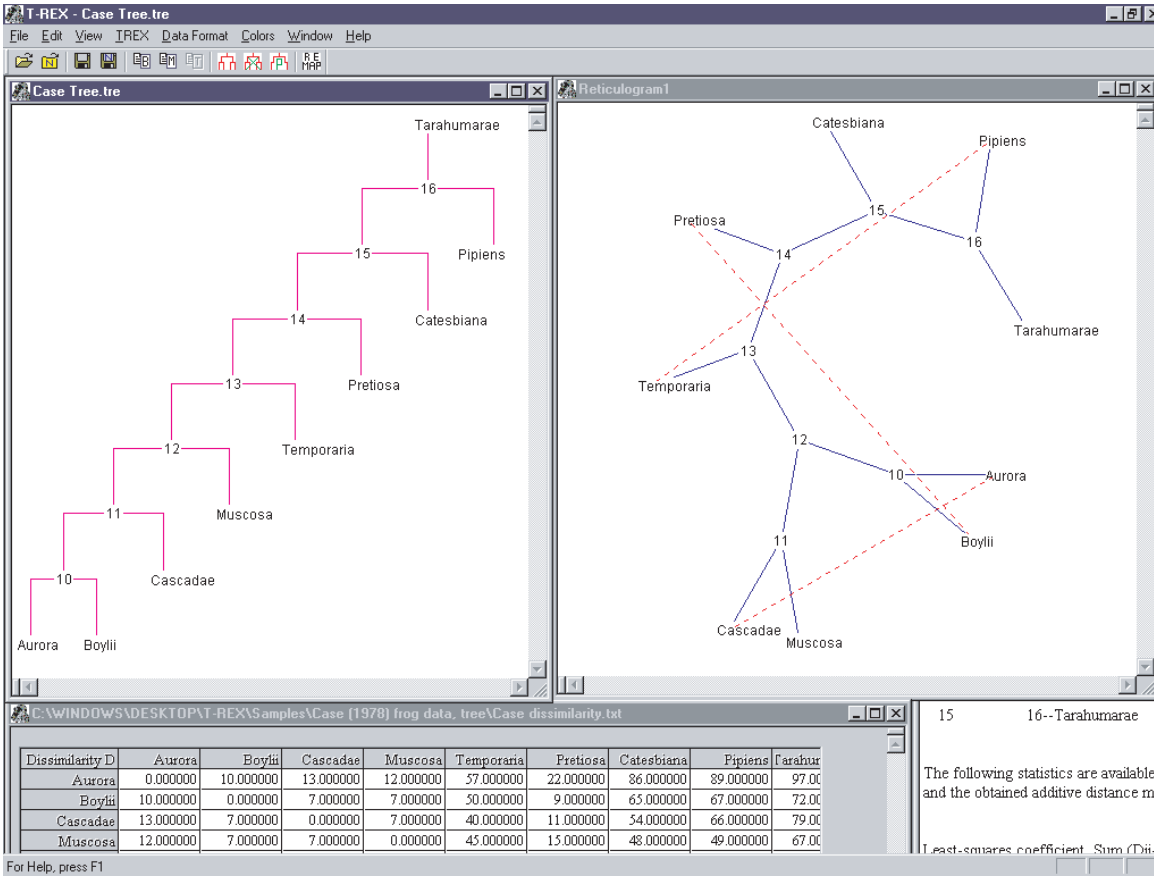

Fig. 1. T-REX screenshot, showing a phylogenetic tree (left part) and reticulation network (right part) inferred from a dissimilarity matrix among nine species of frogs (see Case, 1978 for more details).

reticulation branches to be placed into the reticulogram. For a detailed description of the reticulogram reconstruction method, see the papers by Makarenkov and Legendre (2000) or Makarenkov and Legendre (2001).

MAIN FEATURES OF THE WINDOWS 2.0.1 VERSION

Input and output data format

Several input and output formats have been adopted in T-REX, including Newick, Text (ASCII) and T-REX data formats. The *distance matrix menu* offers the commands, which enable the user to define the input format of his/her distance matrix. The input dissimilarity matrix can be presented as a square matrix, a lower triangular matrix without diagonal or an upper triangular matrix without diagonal. The names of the objects can be supplied or not. T-REX also allows the user to open tree files in a standard Newick format also used in PHYLIP by Felsenstein (1989), PAUP by Swofford (1998) and other well-known phylogenetic packages. Obtained tree or network structures can be saved either in Newick or in a special T-REX format. The fitting statistics can be saved in T-REX format.

Edit menu

This menu allows the user to copy tree or network map or fitting statistics from the document onto the clipboard in the Plain Text (ASCII), Bitmap or Windows Enhanced Metafile formats.

T-REX menu

The T-REX menu offers the commands which enable the user to display dialogue boxes containing input parameters for tree and reticulogram reconstruction algorithms. This menu also allows the user to redraw the tree or reticulogram graphical representation. A tree or reticulogram can be visualized using Hierarchical, Radial or Axial drawing. The user can select the root of the tree or reticulogram, make branches (edges) all equal or proportional to their real length or display the names of the objects.

Colors

The Colors pop-up menu allows the user to set the color of objects, branches or reticulation branches in the graphical representation of a tree or a reticulogram. To change the color of the selected item(s), select a new color in the pop-up menu.

Help manual and samples

A complete help guide containing a detailed description of each menu and method implemented is supplied with the package. A Sample folder provided with the application includes a number of input and output files for T-REX.

IMPLEMENTATION

The Windows 9x/NT version of T-REX was written by Vladimir Makarenkov in C++ programming language in Visual C++ 6.0 and MFC environment. This version runs under Windows 95 or a later version of Windows. The Macintosh version of T-REX was implemented in C++ and Pascal by Vladimir Makarenkov and Philippe Casgrain. The 32-bit DOS version, as well as the source code for UNIX and MS DOS systems are also available at the T-REX web site at http://www.fas.umontreal.ca/biol/casgrain/en/labo/t-rex.

DISCUSSION

As a practical application T-REX has been used in a number of biological studies. For instance, it was employed in Makarenkov and Leclerc (1999) to analyze Case's data (see for example Case, 1978) of the immunological distances between different species of frogs. Figure 1 shows a phylogenetic tree (on its left part) and a reticulogram (on its right part) inferred from a matrix of evolutionary distance between nine species of frogs. In the latter study, four tree fitting algorithms included in the T-REX package were compared to Felsenstein's Fitch algorithm (1989). As to reticulogram reconstruction, the reticulation branches depicted by dashed lines in the right part of Figure 1 represent conflicting signals between different parts of the phylogenetic tree, whose branches are depicted by full lines. In a recent study, Makarenkov and Legendre (2001) considered two applications of reticulation networks constructed using the T-REX application. The first example produced a spatially-constrained reticulogram representing the postglacial dispersal of freshwater fish in the Quebec Peninsula. For these data, the reticulogram was shown to provide a better model of postglacial dispersal than a classical tree structure. The second example of the latter paper depicted the morphological differentiation of muskrats in a river valley in Belgium. Makarenkov and Legendre (2000) used T-REX to explore how the new reticulation algorithm could be applied to represent homoplasy in the phylogenetic tree of primates. In addition, Levasseur *et al.* (2000) used T-REX for estimating trees from incomplete DNA-hybridization matrices.

ACKNOWLEDGEMENTS

The author is grateful to Philippe Casgrain, Olivier Gascuel, Alain Guénoche, Pierre-Alexandre Landry, François-Joseph Lapointe, Bruno Leclerc and Pierre Legendre for their contributions to the T-REX package. This research was supported by NSERC grant number OGP7738 to P.Legendre.

REFERENCES

Buneman,P. (1971) The recovery of trees from measures of dissimilarity. In Hodson,F.R., Kendall,D.G. and Tautu,P. (eds), *Mathematics in Archaeological and Historical Sciences*. Edinburgh University Press, Edinburgh, pp. 387–395.

Case,S.M. (1978) Biochemical systematics of members of the genus *Rana* native to western North America. *Syst. Zool.*, **27**, 299–311.

De Soete,G. (1984) Additive-tree representations of incomplete dissimilarity data. *Qual. Quant.*, **18**, 387–393.

Felsenstein,J. (1989) PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics*, **5**, 164–166.

Gascuel,O. (1997) Concerning the NJ algorithm and its unweighted version UNJ. In Mirkin,B., McMorris,F.R., Roberts,F. and Rzhetsky,A. (eds), *Mathematical Hierarchies and Biology. DIMACS, Series in Discrete Mathematics and Theoretical Computer Science*. American Mathematical Society, Providence, RI, pp. 149–171.

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. *Mol. Biol. Evol.*, **13**, 818–823.

Lapointe,F.-J., Legendre,P., Rohlf,J., Smouse,P. and Sneath,P. (2000) Special Section dedicated to the reticulate evolution. *J. Classif.*, **17**, 153–195.

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, New York, pp. 149–154.

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

Makarenkov,V. and Leclerc,B. (1999) An algorithm for the fitting of a tree metric according to a weighted least-squares criterion. *J. Classif.*, **16**, 3–26.

Makarenkov,V. and Legendre,P. (2000) Improving the additive tree representation of a dissimilarity matrix using reticulations. In Kiers,H.A.L., Rasson,J.-P., Groenen,P.J.F. and Schader,M. (eds),

Data Analysis, Classification and Related Methods. Springer, New York, pp. 35–40.

Makarenkov,V. and Legendre,P. (2001) The reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.*, submitted.

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

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

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

Swofford,D.L. (1998) PAUP: phylogenetic analysis using parsimony and othermethods (computer program). Version 4.0. Sinauer, Sunderland, MA.

Yushmanov,S.V. (1984) Construction of a tree with p leaves from 2p-3 elements of its distance matrix. *Matematicheskie Zametki*, **35**, 877–887.

Zaretskii,K. (1965) Construction of a tree on the basis of a set of distances between its leaves. *Uspekhi Mat. Nauk*, **20**, 90–92.