Published online 6 June 2012 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* ***W573–W579***
doi:10.1093/nar/gks485

T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks

Boc Alix[1], Diallo Alpha Boubacar[2] and Makarenkov Vladimir[2,*]

[1]Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal, QC, H3C 3J7 and [2]Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal, QC, H3C 3P8, Canada

Received February 15, 2012; Revised May 5, 2012; Accepted May 6, 2012

ABSTRACT

T-REX (Tree and reticulogram REConstruction) is a web server dedicated to the reconstruction of phylogenetic trees, reticulation networks and to the inference of horizontal gene transfer (HGT) events. T-REX includes several popular bioinformatics applications such as MUSCLE, MAFFT, Neighbor Joining, NINJA, BioNJ, PhyML, RAxML, random phylogenetic tree generator and some well-known sequence-to-distance transformation models. It also comprises fast and effective methods for inferring phylogenetic trees from complete and incomplete distance matrices as well as for reconstructing reticulograms and HGT networks, including the detection and validation of complete and partial gene transfers, inference of consensus HGT scenarios and interactive HGT identification, developed by the authors. The included methods allows for validating and visualizing phylogenetic trees and networks which can be built from distance or sequence data. The web server is available at: www.trex.uqam.ca.

INTRODUCTION

Phylogenetic trees, i.e. evolutionary trees, additive trees or phylogenies, are the basic structures traditionally used to represent differences between species, and then to analyze those differences statistically (1). Evolutionary relationships among species or other types of taxa can be inferred according to similarities and differences in their genetic or morphological characteristics. Phylogenetic trees can be reconstructed by distance-based methods (2) or by character-based methods (3). The use of distance-based methods is generally, but not necessarily, a two-step process, where distances are first estimated from character data and then a tree is inferred from the estimated distances. The character-based methods, which include Bayesian inference, maximum likelihood and maximum parsimony, rely on the explicit assumption that a set of sequences evolved from a common ancestor by a process of mutation and selection without mixing (e.g. without recombination events).

It is well known, however, that several complex evolutionary mechanisms cannot be adequately described by a phylogenetic tree model (4). Thus, phylogenetic networks should be used when reticulation events, such as hybridization, horizontal gene transfer (HGT), recombination or gene duplication followed by gene loss, have influenced species evolution (5,6). A reticulogram, i.e. reticulated cladogram, is an undirected phylogenetic network capable of portraying reticulate patterns of relationships among organisms (7). Reticulograms have been employed to characterize various phylogenetic and biogeographic mechanisms including hybridization between species, comprising allopolyploidy in plants, microevolution of local populations within a species, historical biogeography events and, finally, homoplasy, which is the portion of phylogenetic similarity resulting from evolutionary convergence (7,8).

An HGT network consists of a traditional phylogenetic, i.e. species, tree with a set of directed branches representing horizontal, i.e. lateral, transfers of the given gene. HGT involves a direct transfer of genetic material from one lineage to another. Bacteria and archaea have developed sophisticated mechanisms of the acquisition of new genes via HGT as an effective way of adaptation to varying environmental conditions (9–12). Two models of HGT have been described in the literature (13). First, and the most popular of them, is the traditional model of complete HGT (14,15). This model assumes that the transferred gene either displaces the orthologous gene of the recipient genome or, when the orthologue is absent in the recipient, is incorporated into the recipient genome as a new gene. The second model is that of partial gene transfer, which involves the formation of mosaic genes (16). A mosaic gene consists of interspersed blocks of sequences having different evolutionary histories but found combined in the resulting allele following HGT and intragenic recombination events (11,17).

*To whom correspondence should be addressed. Tel: +1 514 987 3000 (ext. 3870); Fax: +1 514 987 8477; Email: makarenkov.vladimir@uqam.ca

In this article, we present the T-REX (Tree and reticulogram REConstruction) web server designed to help evolutionary biologists and bioinformatics researchers build, visualize and validate phylogenetic trees, reticulograms and HGT networks. This is a continuation of the T-REX project started in 2001 with the release of the Windows version of the T-REX package (18) whose development stopped in 2006. The most known existing web servers dedicated to the inference and validation of phylogenetic trees are the following: Phylogeny.fr (19), RAxML (20) and PhyML (21) web servers, Phylemon (22) as well as different web server versions of the PHYLIP package (23). While the tree inference methods are provided by many dedicated web servers, the reconstruction of reticulograms and HGT networks is offered only by T-REX. It is worth noting that the most known software intended to detect HGT events, but not available via web services, are the following: LatTrans (24), HorizStory (25), Efficient Evaluation of Edit Paths (26) and PhyloNet package including RIATA-HGT (27).

MATERIALS AND METHODS

The T-REX web server allows users to perform several popular methods of phylogenetic analysis as well as some new phylogenetic applications for inferring, drawing and validating phylogenetic trees and networks, which we have developed. The latest web server version of T-REX includes the following applications:

(1) *Methods for the visualization and interactive manipulation of phylogenetic trees* using hierarchical vertical, hierarchical horizontal, radial and axial types of tree drawing (28). For instance, the Newick Viewer application allows users to visualize a tree coded by its Newick string.
(2) *An application for drawing phylogenetic trees*, allowing for saving them in the Newick format. This program requires the support of the Canvas program by the user's browser. Canvas is supported by many web browsers including Internet Explorer (starting from version 9.0), Firefox (starting from version 2.0) and Safari (staring from version 3.1). A detailed user guide specifying how to create or remove node(s) and branch(es) of a phylogenetic tree is supplied for this application.
(3) *Methods for inferring and validating phylogenetic trees* using distances: traditional Neighbor Joining (NJ, 29), NINJA fast large-scale NJ implementation (30), BioNJ (31), UNJ (32), ADDTREE (33), MW (34), FITCH (23) and Circular order reconstruction (35), maximum parsimony (MS): MS methods from PHYLIP (23) and maximum likelihood (ML): the latest versions of PHYML (36) and RAxML (20) as well as ML methods from PHYLIP (23). For most of the available algorithms, T-REX also carries out bootstrap resampling to assess support of the tree branches. Here, we will give more details on the methods developed by our research group. Circular order reconstruction method (35) builds a phylogeny using a circular order of taxa associated with a given matrix of evolutionary distances. This fitting method was inspired by Yushmanov's (37) article, which introduced the concept of circular orders of taxa corresponding to the clockwise scanning of leaves of a phylogenetic tree. The MW (Method of Weights) procedure searches for the best phylogenetic tree, in the least-squares sense, with respect to the given distance and weight matrices. This method allows for arbitrary weights which may be chosen according to one of the traditional weighting models (34). The tree obtained by any of the seven available distance-based methods is then polished using the procedure of quadratic approximation of its branch lengths [see (28) in the unweighted case and (34) in the weighted case], which is carried out to improve the value of the least-squares criterion while avoiding negative branch lengths.
(4) *Methods for reconstructing phylogenetic trees from a distance matrix containing missing values*, i.e. incomplete matrices. The following four fitting methods are available: Triangles method by Guénoche and Leclerc (38), Ultrametric procedure for the estimation of missing values by Landry *et al.* (39) followed by NJ, Additive procedure for the estimation of missing values by Landry *et al.* (39) followed by NJ, and the Modified Weighted least-squares method (MW*) by Makarenkov and Lapointe (40). The MW* method assigns the weight of 1 to the existing entries, the weight of 0.5 to the estimated entries and the weight of 0 when the entry estimation is impossible. The simulations described in (40) showed that the MW* method clearly outperforms the Triangles, Ultrametric and Additive procedures.
(5) *A method for inferring reticulograms* from distance matrices. The reticulogram reconstruction program first builds a supporting phylogenetic tree using one of the existing tree inferring methods. Following this, a reticulation branch that minimizes the least-squares or the weighted least-squares objective function is added to the tree (or network starting from Step 2) at each step of the algorithm (7). Two statistical criteria, Q_1 and Q_2, have been proposed to measure the gain in fit provided by each reticulation branch:

$$Q_1 = \frac{\sqrt{\sum_{i\in X}\sum_{j\in X}(d(i,j)-\delta(i,j))^2}}{n(n-1)/2-N} \text{ and}$$
$$Q_2 = \frac{\sum_{i\in X}\sum_{j\in X}(d(i,j)-\delta(i,j))^2}{n(n-1)/2-N}. \quad (1)$$

The numerator of these functions is the square root of the sum (or the sum itself) of the quadratic differences between the values of the given evolutionary distance δ and the corresponding reticulogram estimates d, n is the number of taxa in the

considered set X and N is the number of branches in the reticulogram, i.e. total of the phylogenetic tree branches and reticulation branches. The minimum of Q_1 or Q_2 can define a stopping rule for the addition of reticulation branches. A predefined number of reticulation branches, K, can also be added to the supporting tree. The web server version of T-REX also provides the possibility of inferring the supporting tree using one distance matrix and then for adding reticulation branches using another distance matrix. Such an algorithm can be applied for depicting morphological or genetic similarities among given species or for identifying HGT events by using the first distance matrix to infer the species tree and the second matrix (containing the gene-related distances) to infer the reticulation branches representing putative HGTs (7,8).

(6) *Complete and partial HGT detection and validation methods*. The HGT-Detection program aims to determine an optimal, i.e. minimum-cost, scenario of HGTs while proceeding by a gradual reconciliation of the given species and gene trees (12). This algorithm was shown to be faster and generally more effective than the LatTrans (24) and RIATA-HGT (27) techniques. Statistical validation of the obtained gene transfers by bootstrapping can be performed. The HGT bootstrap scores of the predicted gene transfers are obtained by taking into account the uncertainty of the gene tree as well as the number of occurrences of the selected transfers in all minimum-cost HGT scenarios found for the given species tree and the generated gene tree replicates (12,16). The following formula is used to compute the bootstrap score *HGT_BS* of the transfer t obtained while reconciling the species tree T and the gene tree T':

$$HGT\_BS(t) = \left(\sum_{1 \le i \le N_{T'}} \left(\sum_{1 \le k \le N_i} \frac{\sigma_{ki}(t)}{N_i} \times 100\ \% \right) \right) / N_{T'}, \quad (2)$$

where $\sigma_{ki}(t)$ is equal to 1 if the transfer t is a part of the minimum-cost scenario k for the gene tree replicate T_i' and equal to 0, otherwise, $N_{T'}$ is the number of gene tree replicates, i.e. number of HGT bootstrap replicates, generated from re-sampled gene sequences and N_i is the number of minimum-cost scenarios obtained when carrying out the algorithm with the species tree replicate T, which is assumed to be fixed, and the gene tree T_i'. HGT Consensus, Parallel and Interactive versions of the HGT-Detection algorithm are also available. The Consensus version of the algorithm yields a consensus HGT scenario obtained for a given species tree and a set of gene trees. The Parallel version allows the user to speed up the HGT computation by executing the program on a 32-processor Linux cluster, whereas the Interactive version allows the user to pre-define some HGTs and then accept, reject or change the direction of each transfer proposed by the HGT-Detection algorithm during the program execution. A version of the program allowing for identifying Partial HGT scenarios, when only a part of the gene is acquired by the host allele through intragenic recombination, is also provided (16). Note that the results of the HGT detection algorithms depend on the position of the species and gene tree roots. The user can select the tree roots by checking the appropriate check boxes prior to launching the computation.

(7) *MAFFT* (41) *and MUSCLE* (42) *algorithms*, which are among the most widely used multiple sequence alignment tools, are available with slow and fast pairwise alignment options.

(8) *Most common Sequence to Distance transformations*. The following popular substitution models of DNA and amino acids evolution allowing for estimating evolutionary distances from sequence data have been included to T-REX: Uncorrected distance, Jukes–Cantor (43), K80 – two parameters (44), T92 (45), Tajima–Nei (46), Jin–Nei gamma (47), Kimura protein (48), LogDet (49), F84 (50), WAG (51), JTT (52) and LG (53).

(9) *Computation of the Robison and Foulds (RF) topological distance*. This program computes the RF topological distance (54), which is a well-known measure of the tree similarity, between the first tree and all the following trees specified by the user. The trees can be supplied in the Newick or Distance matrix formats. An optimal algorithm described in (55) is carried out to compute the RF metric.

(10) *Newick to Distance matrix and Distance matrix to Newick format conversion*. This application allows the user to convert a phylogenetic tree from the Newick format to the Distance matrix format and vice versa.

(11) *Random phylogenetic tree generation program*. This application generates k random phylogenetic trees with n leaves, i.e. species or taxa, and an average branch length l using the random tree generation procedure described by Kuhner and Felsenstein (56), where the variables k, n and l are defined by the user. The branch lengths of trees follow an exponential distribution. The branch lengths are multiplied by $1 + \mathbf{a}x$, where the variable x is obtained from an exponential distribution ($P(x > k) = \exp(-k)$), and the constant $\mathbf{a}$ is a tuning factor accounting for the deviation intensity (as recommended in (57), the value of $\mathbf{a}$ was set to 0.8). The random trees generated by this procedure have depth of $O(\log(n))$.

RESULTS

The three main types of results provided by the T-REX web server are the following:

(1) *A phylogenetic tree drawing* (Figure 1), *fitting statistics and resulting tree coded in the Newick format*. Fitting statistics include the fitted tree distance

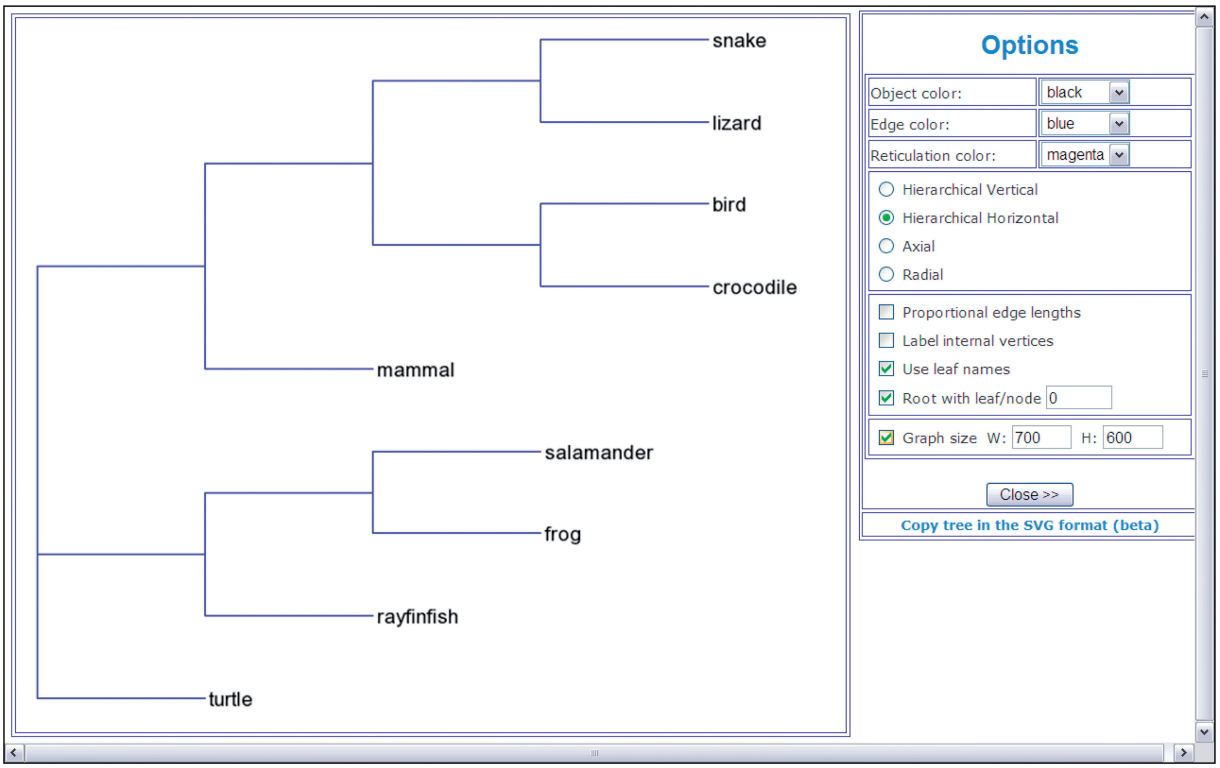

Figure 1. An example of a phylogenetic tree (hierarchical view) showing phylogenetic relationships for a group of nine vertebrate species.

matrix, list of tree branches with their lengths and, if computed, their bootstrap scores. For all distance-based methods, the values of the (weighted) least-squares coefficient, (weighted) average absolute difference, (weighted) maximum absolute difference and the total length of the obtained tree are also provided. The tree presented in Figure 1 was obtained using the NJ inferring method (29). It illustrates phylogenetic relationships for a group of nine vertebrate species.

(2) *A reticulogram drawing* (Figure 2) *and fitting statistics*. Fitting statistics include the fitted reticulogram distance matrix and list of reticulogram branches with their lengths. If the reticulogram reconstruction is performed, T-REX also provides the values of the least-squares criterion as well as the values of the selected stopping criterion Q_1 or Q_2 for the supporting tree topology and for each reticulation branch added to the supporting tree. Figure 2 shows an example of a reticulogram (7,8) depicting phylogenetic relationships for the same group of nine vertebrate species. The reticulation branches linking the pairs of species (bird – mammal) and (lizard – turtle) suggest that the connected species are more closely phylogenetically related to each other than it is depicted by the traditional phylogenetic tree model. The presented network was obtained by adding a predefined number of reticulation branches, $K = 2$, to the supporting phylogenetic tree inferred by NJ (29).

(3) *An HGT network* in which gene transfers are indicated by dashed arrows (Figure 3). Numbers on transfers indicate their order of inference (except for partial HGT detection). HGT bootstrap scores, if computed, are indicated between parentheses and the affected intervals, for Partial HGT detection only, are indicated between brackets. The output file also contains the values of the bipartition dissimilarity (12), Robinson and Foulds topological distance (54) and least-squares coefficient which characterize the proximity between the considered species and gene trees (these statistics are provided at each step of the HGT detection algorithm), list of HGT branches with their bootstrap scores and affected intervals (for Partial HGT detection only). The HGT Consensus algorithm allows the user to infer a consensus HGT scenario. The Interactive version of the HGT-Detection program allows the user to specify some gene transfers as well as to validate all the transfers proposed by the algorithm. Figure 3 shows an example of an HGT network depicting horizontal transfers of the gene *rpl12e* for the group of 14 species originally considered by Matte-Taillez *et. al.* [see Figure 1a in (58)] and subsequently reanalyzed in (12). Five gene transfers were inferred for this example using the standard version of the HGT-Detection algorithm (12). The bootstrap analysis was also carried out for this data set to assess the robustness of the obtained HGTs.

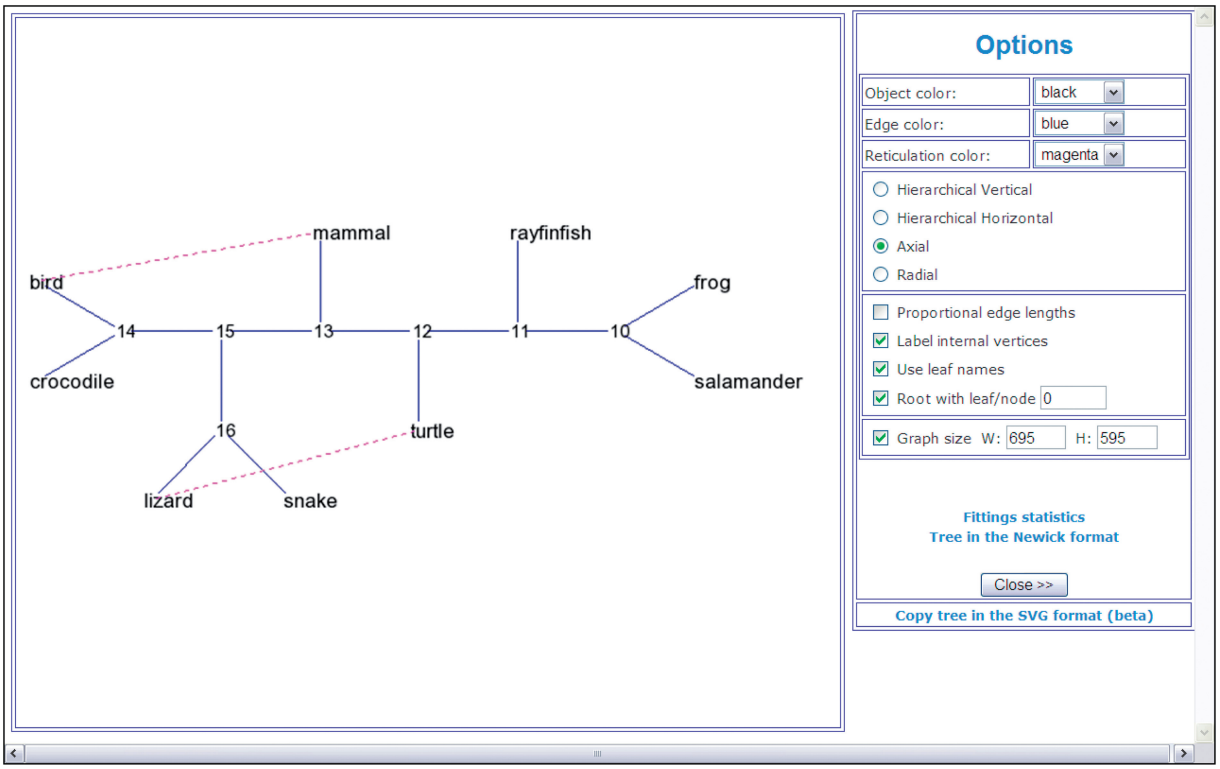

Figure 2. An example of a reticulogram (axial view) showing phylogenetic relationships for a group of nine vertebrate species.

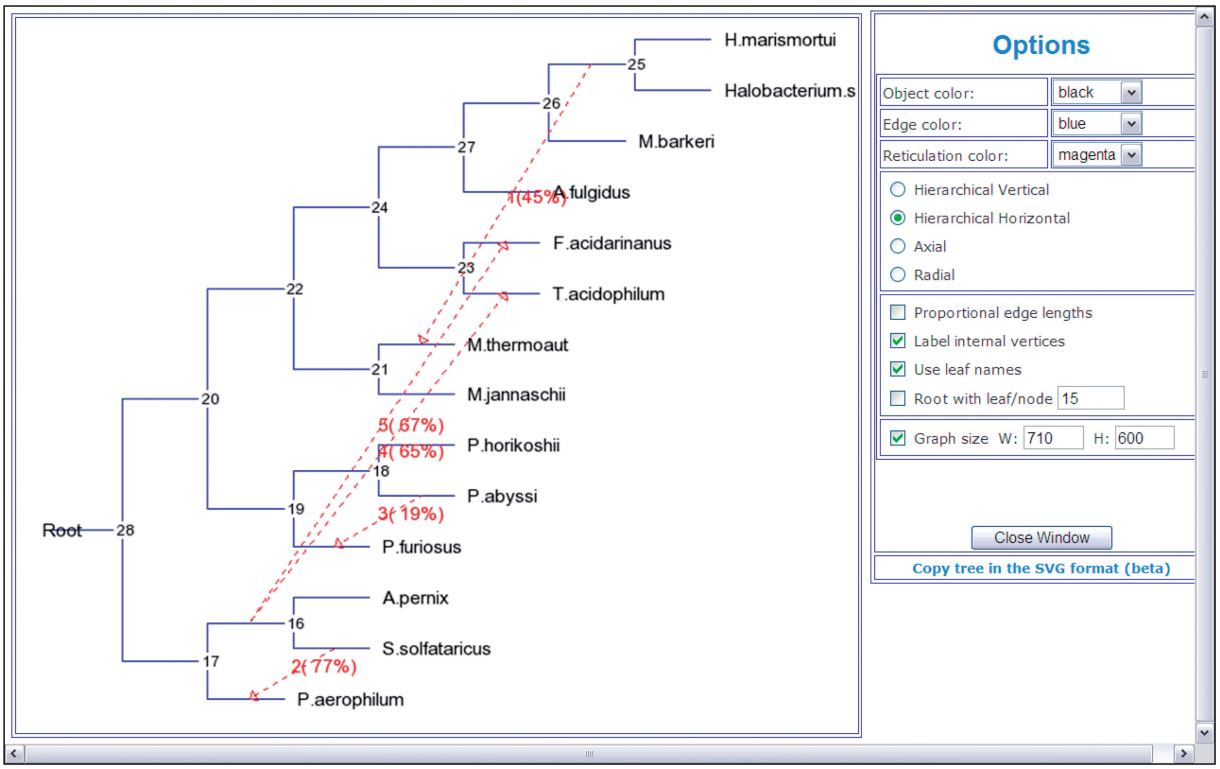

Figure 3. An example of a HGT network showing horizontal transfers of the gene *rpl12e* inferred for the group of 14 species originally considered by Matte-Taillez *et al.* [see Figure 1a in (58)]. Five gene transfers are indicated by dashed arrows. Numbers on transfers designate their order of inference. Bootstrap scores of the obtained transfers are indicated between parentheses.

The T-REX input data can be in the three following formats: Newick, PHYLIP and FASTA. All graphical results provided by the T-REX server can be saved in the SVG (Scalable Vector Graphics) format and then opened and modified (e.g. prepared for a publication or presentation) in the user's preferred graphics editor. All numerical results are given in the text format.

CONCLUSION

In this article, we described the T-REX web server developed to facilitate the inference, validation and visualization of phylogenetic trees and networks. T-REX includes several new and classical methods of phylogenetic analysis. The methods we designed are the following: Reconstruction of phylogenetic trees from distance data according to the least-squares, weighted least-squares and circular order criteria; reconstruction of phylogenetic trees from distance matrices containing missing values using a weighted least-squares approach; computation of the Robinson and Foulds topological distance using an optimal algorithm; reticulogram inference that can be carried out according to several stopping rules; and finally, identification of complete HGT events, including their validation by bootstrap and compilation of the consensus and interactive HGT scenarios as well as partial HGT detection (when any part of the given gene can be incorporated in the recipient allele via intragenic recombination). All these methods have been shown to be effective in the analysis of various traditional, i.e. tree-like, and reticulate, i.e. network-like, evolutionary patterns and could become the methods of choice for the community of evolutionary biologists and bioinformatics researchers.

ACKNOWLEDGEMENTS

The authors thank Philippe Casgrain, Abdoulaye Baniré Diallo, Olivier Gascuel, Alain Guénoche, Joseph Felsenstein, Pierre-Alexandre Landry, François-Joseph Lapointe, Bruno Leclerc, Pierre Legendre, Etienne Lord and Abdellah Mazouzi for providing us with the source code of their methods and/or for their helpful comments. The authors also thank Editor Gary Benson and three anonymous referees for their helpful comments and suggestions.

FUNDING

Natural Sciences and Engineering Research Council of Canada [NSERC-249644-2011]; Nature and Technologies Research Funds of Quebec [FQRNT; individual stipends to A.B. and A.B.D.]. Funding for open access charge: Natural Sciences and Engineering Research Council of Canada [NSERC-249644-2011].

Conflict of interest statement. None declared.

REFERENCES

1. Felsenstein,J. (2004) *Inferring Phylogenies*. Sunderland, MA, Sinauer Associates.
2. Sneath,P.H.A. and Sokal,R.R. (1973) *Numerical taxonomy — The Principles and Practice of Numerical Classification*. In: Freeman,W.H. (ed.). San Francisco, CA.
3. Hennig,W. (1966) In: Dwight Davis,D. and Zangerl,Rainer (eds), *Phylogenetic Systematics*. University of Illinois Press, Urbana, IL.
4. Legendre,P. (2000) Special section on reticulate evolution. *J. Classif*, **17**, 153–195.
5. Huson,D.H. and Bryant,D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–267.
6. Huson,D.H., Rupp,R. and Scornavacca,C. (2011) Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.
7. Legendre,P. and Makarenkov,V. (2002) Reconstruction of biogeographic and evolutionary networks using reticulograms. *Syst. Biol.*, **51**, 199–216.
8. Makarenkov,V. and Legendre,P. (2004) From a phylogenetic tree to a reticulated network. *J. Comput. Biol.*, **11**, 195–212.
9. Doolittle,W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2129.
10. Gogarten,J.P., Doolittle,W.F. and Lawrence,J.G. (2002) Prokaryotic evolution in light of gene transfer. *Mol. Biol. Evol.*, **19**, 2226–2238.
11. Zhaxybayeva,O., Lapierre,P. and Gogarten,J.P. (2004) Genome mosaicism and organismal lineages. *Trends Genet.*, **20**, 254–260.
12. Boc,A., Philippe,H. and Makarenkov,V. (2010) Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.*, **59**, 195–211.
13. Makarenkov,V., Kevorkov,D. and Legendre,P. (2006) Phylogenetic Network Reconstruction Approaches. *Applied Mycology and Biotechnology*, Vol. 6. International Elsevier Series, Bioinformatics, pp. 61–97.
14. Maddison,W.P. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523–536.
15. Page,R.D.M. and Charleston,M.A. (1998) Trees within trees: phylogeny and historical associations. *Trends Ecol. Evol.*, **13**, 356–359.
16. Boc,A. and Makarenkov,V. (2011) Towards an accurate identification of mosaic genes and partial horizontal gene transfers. *Nucleic Acids Res.*, **39**, e144.
17. Hollingshead,S.K., Becker,R. and Briles,D.E. (2000) Diversity of PspA: mosaic genes and evidence for past recombination in *Streptococcus pneumoniae*. *Infect. Immun.*, **68**, 5889–5900.
18. Makarenkov,V. (2001) T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. *Bioinformatics*, **17**, 664–668.
19. Dereeper,A., Guignon,V., Blanc,G., Audic,S., Buffet,S., Chevenet,F., Dufayard,J.-F., Guindon,S., Lefort,V., Lescot,M. *et al*. (2008) Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, **36**, W465–W469.
20. Stamatakis,A., Hoover,P. and Rougemont,J. (2008) A rapid bootstrap algorithm for the RAxML web-servers. *Syst. Biol.*, **75**, 758–771.
21. Guindon,S., Dufayard,J.F., Lefort,V., Anisimova,M., Hordijk,W. and Gascuel,O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.*, **59**, 307–321.
22. Sánchez,R., Serra,F., Tárraga,J., Medina,I., Carbonell,J., Pulido,L., de María,A., Capella-Gutíerrez,S., Huerta-Cepas,J., Gabaldón,T. *et al*. (2011) Phylemon 2.0: a suite of web-tools for molecular evolution, phylogenetics, phylogenomics and hypotheses testing. *Nucleic Acids Res.*, **39**, W470–W474.
23. Felsenstein,J. (1989) PHYLIP—Phylogeny inference package (Version 3.6). *Cladistics*, **5**, 164–166.
24. Hallett,M. and Lagergren,J. (2001) Efficient algorithms for lateral gene transfer problems. In: El-Mabrouk,N., Lengauer,T. and Sankoff,D. (eds), *Proceedings of the Fifth Annual International Conference on Research in Computational Biology*. ACM Press, New-York, pp. 149–156.
25. MacLeod,D., Charlebois,R.L., Doolittle,F. and Bapteste,E. (2005) Deduction of probable events of lateral gene transfer through

comparison of phylogenetic trees by recursive consolidation and rearrangement. *BMC Evol. Biol*, **5**, 27.
26. Beiko,R.G. and Hamilton,N. (2006) Phylogenetic identification of lateral genetic transfer events. *BMC Evol. Biol*, **6**, 15.
27. Than,C., Ruths,D. and Nakhleh,L. (2008) PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinf.*, **9**, 322.
28. Barthélemy,J-P. and Guénoche,A. (1991) *Trees and Proximity Representations*. Wiley, New York.
29. Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, 406–425.
30. Wheeler,T.J. (2009) Large-scale neighbor-joining with NINJA. In: Salzberg,S. and Warnow,T. (eds), *Proceedings of the 9th Workshop on Algorithms in Bioinformatics*. Springer, Verlag, Berlin, pp. 375–389.
31. Gascuel,O. (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.
32. 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.
33. Sattath,S. and Tversky,A. (1977) Additive similarity trees. *Psychometrika*, **42**, 319–345.
34. 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.
35. 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.
36. Guindon,S. and Gascuel,O. (2003) A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.
37. Yushmanov,S.V. (1984) Construction of a tree with p leaves from 2p-3 elements of its distance matrix (in Russian). *Matematicheskie Zametki*, **35**, 877–887.
38. Guénoche,A. and Leclerc,B. (2001) The triangles method to build X-trees from incomplete distance matrices. *RAIRO Operat. Res.*, **35**, 283–300.
39. 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.
40. Makarenkov,V. and Lapointe,F.-J. (2004) A weighted least-squares approach for inferring phylogenies from incomplete distance matrices. *Bioinformatics*, **20**, 2113–2121.
41. Katoh,K., Kuma,K., Toh,H. and Miyata,T. (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.*, **33**, 511–518.
42. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
43. Jukes,T.H. and Cantor,C.R. (1969) Evolution of protein molecules. In: Munro,H.N. (ed.), *Mammalian Protein Metabolism*. Academic Press, New York, pp. 21–123.
44. Kimura,M. (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, **16**, 111–120.
45. Tamura,K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C content biases. *Mol. Biol. Evol.*, **9**, 678–687.
46. Tajima,F. and Nei,M. (1984) Estimation of evolutionary distance between nucleotide sequences. *Mol. Biol. Evol.*, **1**, 269–285.
47. Jin,L. and Nei,M. (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.*, **7**, 82–102.
48. Kimura,M. (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge.
49. Lockhart,P.J., Steel,M.A., Hendy,M.D. and Penny,D. (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, **11**, 605–612.
50. Felsenstein,J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
51. Whelan,S. and Goldman,N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.
52. Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.
53. Le,S.Q. and Gascuel,O. (2008) LG: an improved, general amino-acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.
54. Robinson,D.R. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.
55. Makarenkov,V. and Leclerc,B. (2000) An optimal way to compare additive trees using circular orders. *J. Comp. Biol.*, **7**, 731–744.
56. Kuhner,M. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.
57. Guindon,S. and Gascuel,O. (2002) Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Mol. Biol. Evol.*, **19**, 534–543.
58. Matte-Tailliez,O., Brochier,C., Forterre,P. and Philippe,H. (2002) Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.*, **19**, 631–639.