BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 12 2007, pages 1556–1558
doi:10.1093/bioinformatics/btm135

Phylogenetics

TOPD/FMTS: a new software to compare phylogenetic trees

Pere Puigbò[1,*], Santiago Garcia-Vallvé[1] and James O. McInerney[2]

[1]Evolutionary Genomics Group, Biochemistry and Biotechnology Department, Rovira i Virgili University, Tarragona, Spain and [2]Bioinformatics Laboratory, National University of Ireland, Maynooth, Maynooth, Ireland

Received on January 17, 2007; revised on March 16, 2007; accepted on April 2, 2007

Advance Access publication April 25, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: TOPD/FMTS has been developed to evaluate similarities and differences between phylogenetic trees. The software implements several new algorithms (including the Disagree method that returns the taxa, that disagree between two trees and the Nodal method that compares two trees using nodal information) and several previously described methods (such as the Partition method, Triplets or Quartets) to compare phylogenetic trees. One of the novelties of this software is that the FMTS (From Multiple to Single) program allows the comparison of trees that contain both orthologs and paralogs. Each option is also complemented with a randomization analysis to test the null hypothesis that the similarity between two trees is not better than chance expectation.

Availability: The Perl source code of TOPD/FMTS is available at http://genomes.urv.es/topd.

Contact: ppuigbo@urv.cat

Supplementary information: A complete tutorial and several examples of how to use the software have been included on the home page of the application.

1 INTRODUCTION

Phylogenetic trees have often been compared in molecular evolution studies, because different sets of putatively orthologous genes often yield strongly supported but incompatible tree topologies (Beiko and Hamilton, 2006). Incongruence in tree topologies can be explained by such processes as horizontal gene transfer events (Creevey *et al*., 2004; Garcia-Vallve *et al*., 2003), hidden paralogy (Creevey *et al*., 2004) and model misspecification (Rokas *et al*., 2003). Most archaeal and bacterial genomes contain genes from multiple sources (Doolittle, 1999) and each phylogenetic tree constructed from a protein family reflects the evolutionary history of its sequences. There are also many methods of constructing phylogenetic trees (e.g. Distance, Parsimony or Likelihood), which can produce different trees. Given this situation, it is desirable to compare phylogenetic trees from a set of sequences constructed by different methods and/or to compare phylogenetic trees from different sets of homologs.

Although many methods for comparing phylogenetic trees have been described, for example, nearest-neighbor interchanging (Waterman and Smith, 1978), subtree transfer distance (Allen and Steel, 2001), quartets (Estabrook *et al*., 1985), partition or symmetric difference metrics (Robinson and Foulds, 1981) and path length metrics (Steel and Penny, 1993), very few have been implemented for their use in a program and there is no program with a comprehensive set of implemented methods. For this reason, we have developed the TOPD/FMTS software. TOPD/FMTS compares phylogenetic trees using some of the above methods, but also implements new algorithms. This means a sensitivity analysis can be carried out on any set of results to evaluate methodological properties and biases. TOPD/FMTS combines two programs: (1) the TOPD (TOPological Distance) program, which compares two trees with the same taxa or two pruned trees and (2) the FMTS (From Multiple To Single) program, which converts multi-gene family trees to single-gene family trees. The FMTS program is activated automatically only if one or both trees to be compared are multi-gene family trees, so both programs can work together depending on input data structure. Additionally, each option of this software is complemented with a randomization analysis to test the null hypothesis that the similarity between two trees is not better than chance.

2 PROGRAM OVERVIEW

2.1 Inputs and outputs

The software minimally requires a file containing two trees in PHYLIP format to calculate a distance between them. Alternatively, a file containing a list of trees can be provided in order to calculate the differences between all of them or to compare them with a reference tree. The parameter '−f' followed by the name of the input file is the only mandatory parameter required to run the program. Other parameters can be modified according to the user's requirements (use '−h' to see the complete list of parameters). This software can compare trees with leaf-sets that either completely or partially overlap. If trees only partially overlap, they are pruned to their common leaf-set in order to compare their topologies. The input trees can be rooted or unrooted. If a rooted tree is input, it will be automatically unrooted. Some results are printed in the standard output, by default, but can be easily redirected into an output file using terminal commands. The final results (i.e. the values of the comparison and the percentage of overlapping taxa) are printed in an output file.

*To whom correspondence should be addressed.

2.2 TOPD

The TOPD program compares trees using several methods, which are called 'Nodal,' 'Split,' 'Quartets,' 'Triplets,' and 'Disagree.' The split or partition metrics (Robinson and Foulds, 1981) and quartets and triplets (Estabrook *et al*., 1985) have been described and implemented previously, but this software offers additional possibilities such as the comparison of multigene family trees, the comparison of partially overlapping trees and randomization tests. The nodal method uses the path length metrics described by Steel and Penny (1993). The disagree method uses a novel algorithm described and implemented in this software and is the opposite of the methods that find the most agreement subtree. The agreement method described by Kubicka *et al*. (1995) finds the single greatest agreement subtree when two trees are compared, while our disagree method finds the taxa that produce disagreement between two trees.

The 'Nodal' method constructs pairwise distance matrices from the two input trees using only the leaves that are common to both trees. This is done by comparing the number of nodes that separate each taxon from the other taxa in the tree. If the two trees do not have the same taxa, but have overlapping leaf-sets, the trees are appropriately pruned so they can be compared. Then the differences between the two matrices are calculated to obtain the distance between the two trees. The nodal distance score is calculated using the root-mean-squared distance (RMSD) of these two matrices. The RMSD is 0 for identical trees, and increases as the two trees become more dissimilar. In those cases where two leaf-sets are overlapping but not identical, we have added another score that considers the percentage of taxa that the two trees have in common. This second score is equal to the RMSD if both taxon-sets are the same and becomes proportionally greater when this percentage is reduced, i.e. this score is 0, if two trees are equal and increase depending upon two factors: the dissimilarity between the trees and the number of overlapping leaves (see the equation in http://genomes.urv.es/topd/nodal_e.html).

The 'Disagree' method compares two trees and returns the taxa whose phylogenetic position disagrees in these trees. Penny and Hendy (1985) used the term 'gain' to describe the reduction in the difference when two trees are compared after any taxon is removed. Our disagree method uses an iterative algorithm and can work at four levels of comparison. The computational time needed at each level increases. The method works at level 1 by removing one taxon every time and calculating the gain (reduction in the split distance) between the two trees. The taxon that produces most gain is removed for the following iterations. This procedure is repeated until the split distance is zero (see the algorithm in http://genomes.urv.cat/topd/disagree.html). We have used this algorithm in a thousand comparisons of trees obtained from known protein families. At level 1, ~80% of the comparisons can be solved (i.e. the split distance becomes 0 after removing one taxon or set of taxa). The second, third and fourth levels remove 2, 3 and 4 taxa every time, respectively, and then calculate the gain. When a solution exists, every level solves the comparisons between trees that cannot be solved in the previous level.

2.3 FMTS

The FMTS program can be used to compare two trees, one or both of which are multigene family trees. Until now, trees that contain more than one gene copy per genome could not be compared automatically using any software. The TOPD/FMTS program makes it possible by evaluating each gene copy independently. The FMTS program systematically prunes each gene copy from the multifamily tree to obtain all possible single-gene trees. The result is a set of single-gene family trees. Each tree can be then compared with TOPD, using any of the previously described methods and the result is the mean and SD of all comparisons. In its standard output, the program provides the result of all comparisons and a text file of all of the pruned single-gene family trees. The use of the FMTS program may be computationally expensive when the number of single-gene family trees obtained from a multi-gene family tree is enormous. To overcome this limitation, the FMTS program allows the option of randomly pruning the multi-gene tree by default 100 times. Users, however, can modify this number.

The set of single gene trees obtained with FMTS would contain a mixture of orthologs and paralogs. Those trees can be checked individually, using the TOPD program and a reference species tree, to help to define orthologs and paralogs, or identify horizontally transferred genes. The identification of true orthologs is essential for studying the speciation process. On the other hand, the analysis of paralogs helps to understand the evolution by gene duplication, which is a major force in creating new functionalities (Jordan *et al*., 2001; Lynch and Conery, 2000). Another method capable of dealing with paralogy is the reconciled trees method (Cotton and Page, 2002). But this method tries to infer gene duplication events and estimate species phylogenies, while the FMTS algorithm is helpful to study phylogenies of protein families that contain orthologs and paralogs through the tree comparison with the program TOPD.

2.4 Randomization analysis

This software implements two randomization methods that evaluate whether the similarity between two trees is better than random. In the first method (Guided), all taxa are removed from the tree and randomly reassigned while maintaining the topology of the original tree. This means that the positions of the taxa have been randomly changed. The second method (Random), generates random trees, by a Markovian method, with the same taxa as the original tree but randomly changes the topology of the tree and consequently, the relationships of the taxa. A similar method is used in the Clann program (Creevey and McInerney, 2005). Then a comparison between random trees is calculated using any of the methods allowed by the software. This is repeated as many times as the user requires. By default, the program carries out this random analysis 100 times and the result is the mean and SD of the different repetitions. A critical point can be used to evaluate whether the similarity between two trees is better than random.

ACKNOWLEDGEMENTS

We thank John Bates of the Language Service of the Rovira i Virgili University for his help with writing the manuscript and members of the Bioinformatics Laboratory for discussions. This work has been financed by the project BIO2003-07672 from the 'Ministerio de Ciencia y Tecnología' of the Spanish Government.

Conflict of Interest: none declared.

REFERENCES

Allen,L. and Steel,M. (2001) Subtree transfer operations and their induced metrics on evolutionary trees. *Annals of Combinatorics*, **5**, 1–15.

Beiko,R.G. and Hamilton,N. (2006) Phylogenetic identification of lateral gene transfer events. *BMC Evol. Biol.*, **6**, 15.

Cotton,J.A. and Page,R.D (2002) Going nuclear: gene family evolution and vertebrate phylogeny reconciled. *Proc. R. Soc. Lond. B*, **269**, 1555–1561.

Creevey,C.J. and McInerney,J.O. (2005) Clann: Investigating phylogenetic information through supertree analyses. *Bioinformatics*, **21**, 390–392.

Creevey,C.J. *et al.* (2004) Does a tree-like phylogeny only exist at the tips in the prokaryotes? *Proc. R. Soc. Lond. B. Biol. Sci.*, **271**, 2551–2558.

Doolittle,W.F. (1999) Phylogenetic classification and the universal tree. *Science*, **284**, 2124–2128.

Estabrook,G.F. *et al.* (1985) Comparison of undirected phylogenetic trees based on subtree of four evolutionary units. *Syst. Zool.*, **34**, 193–200.

Garcia-Vallve,S. *et al.* (2003) HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.*, **31**, 187–189.

Jordan,I.K. *et al.* (2001) Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.*, **11**, 555–565.

Kubicka,E. *et al.* (1995) An algorithm to find agreement subtrees. *J. Classification*, **12**, 91–99.

Lynch,M. and Conery,J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science*, **290**, 1151–1155.

Penny,D. and Hendy,M.D. (1985) The use of tree metrics. *Syst. Zool.*, **34**, 75–82.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Rokas,A. *et al.* (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**, 798–803.

Steel,M.A. and Penny,D. (1993) Distribution of tree comparison metrics – some new results. *Systematic Biol.*, **42**, 126–141.

Waterman,M.S. and Steel,M. (1978) On the similarity of dendrograms. *J. Theor. Biol.*, **73**, 789–800.