BIOINFORMATICS APPLICATIONS NOTE Vol. 22 no. 1 2006, pages 117–119
doi:10.1093/bioinformatics/bti720

Phylogenetics

A novel algorithm and web-based tool for comparing two alternative phylogenetic trees

Tom M.W. Nye[1,*], Pietro Liò[2] and Walter R. Gilks[1]

[1]Medical Research Council Biostatistics Unit, Cambridge, UK and [2]University of Cambridge Computer Laboratory, Cambridge, UK

Received on August 15, 2005; revised on September 28, 2005; accepted on October 14, 2005
Advance Access publication October 18, 2005
Associate Editor: Keith A. Crandall

ABSTRACT

Summary: We describe an algorithm and software tool for comparing alternative phylogenetic trees. The main application of the software is to compare phylogenies obtained using different phylogenetic methods for some fixed set of species or obtained using different gene sequences from those species. The algorithm pairs up each branch in one phylogeny with a matching branch in the second phylogeny and finds the optimum 1-to-1 map between branches in the two trees in terms of a topological score. The software enables the user to explore the corresponding mapping between the phylogenies interactively, and clearly highlights those parts of the trees that differ, both in terms of topology and branch length.

Availability: The software is implemented as a Java applet at http://www.mrc-bsu.cam.ac.uk/personal/thomas/phylo-comparison/comparison_page.html. It is also available on request from the authors.

Contact: thomas.nye@mrc-bsu.cam.ac.uk

1 INTRODUCTION

Many biological analyses involve the construction of a phylogenetic tree for some set of sequence data, and a variety of methods for inferring phylogenies are available. However, the choice of phylogenetic method can have a strong influence on the tree obtained for a given set of sequence data, both in terms of its topology and branch lengths. In addition, different gene trees can be obtained for some fixed set of species, where each gene tree is based on a different set of orthologous sequences chosen for the analysis. Comparison between gene trees and species trees can reveal concensus patterns of evolution as well as genes that diverge from this pattern. Methods for comparing phylogenies that are capable of revealing where two trees agree or differ are therefore desirable, in order to assess the quality of phylogenetic trees and analyse different phylogenetic methods.

While a number of algorithms for tree comparison have been developed previously, it seems as though there is no standard tool widely used by the phylogenetics community for visualizing similarities and differences. Several previous approaches to comparing trees (Robinson and Foulds, 1981; Cole *et al*., 2000; Hon *et al*., 2001) have concentrated on finding the maximal common subtree, or have involved computing a series of transformations that map one tree into the other in order to express the dissimilarity between the trees as an edit distance or metric. Our approach is rather different: given two trees to compare, we match up branches that have a similar topological characteristic. The topological characteristic we consider is the partition of leaf nodes determined by each branch in a tree. This process of matching up branches within the two trees under comparison leads to a form of alignment between the trees as opposed to a chain of operations relating them or a metric specifying their dissimilarity. In fact, our approach to tree comparison can be thought of as being analogous to sequence alignment, where instead of conserved letters in a sequence we have branches that share topological features. The alignment obtained is best explored interactively, as implemented in the web-based tool we describe below. Other approaches to tree comparison (Munzner *et al*., 2003; Page, 1995) have matched nodes in two alternative rooted trees according to the ancestors the nodes share. We discuss these methods in a later section.

2 ALGORITHM

Suppose we are given two phylogenetic trees T_1 and T_2 that share the same set of leaves L. The trees may not necessarily be bifurcating, and can be either rooted or unrooted. For simplicity we assume the trees have the same number of branches. Our comparison algorithm has two stages. First every pair of edges (i,j) with $i \in T_1$ and $j \in T_2$ is assigned a score $s(i,j)$, that reflects the topological similarity of the branches i and j. Secondly, branches in the two trees are paired up to optimize the overall score. More formally, this is equivalent to finding a bijection (i.e. a 1-to-l correspondence) $f : T_1 \rightarrow T_2$ between the branches of the trees, that maximizes the quantity

$$\sum_{i \in T_1} s(i, f(i)). \qquad (1)$$

These steps are described in more detail below. The outcome of the algorithm is the correspondence f between branches in the two trees, which we refer to as an alignment.

*To whom correspondence should be addressed.

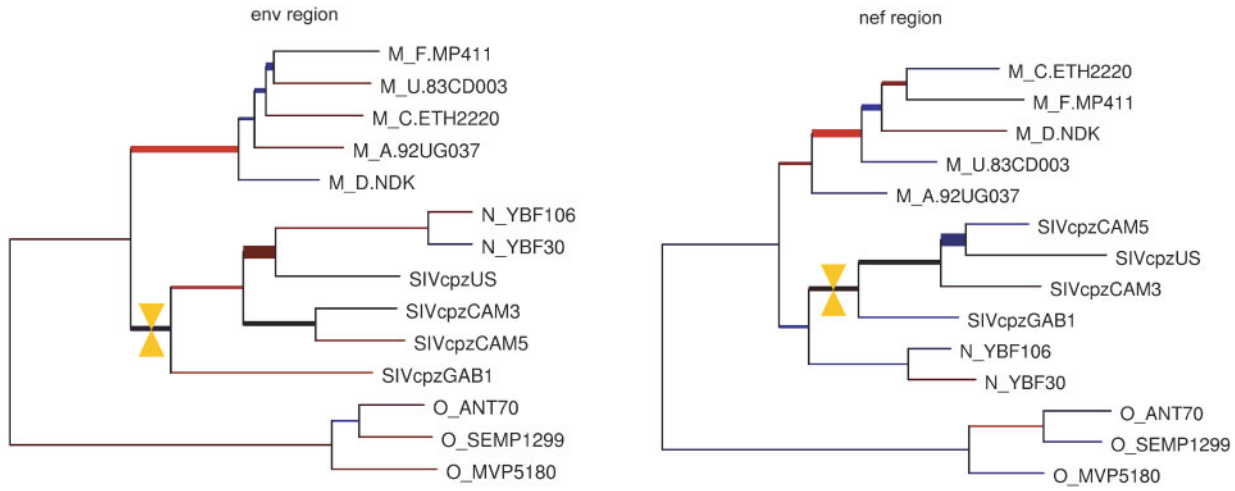

Fig. 1. Comparison of alternative phylogenies for HIV strains. The phylogenies were constructed for strains with fully sequenced genomes, by using sequences from two different genomic regions. Three groups of human strains (M, N and O) are shown together with four simian strains (the SIV group). The phylogenies exhibit two different positions for the N-group, one closer to the simian group (left), and the other closer to the human groups (right). This was probably caused by an ancient recombination event in the N-group ancestor. The thicker branches are those receiving a low topological score: in particular the thickest branches arise from the two alternative positions for the N-group. The topology of the O-group matches exactly between the phylogenies, as indicated by the thin branches, while the internal topology of the M-group differs quite widely. Branches drawn in red are longer than the corresponding branches in the other tree, with the intensity of the colour indicating the level of mismatch. Blue branches conversely denote shorter branches. The yellow markers indicate a match between branches (clicking on any branch results in its match being highlighted and the score displayed). The particular match illustrated here has a topological score of 67%. Although both trees contain a branch separating the N and simian clades from the rest of the tree, those branches are not identified under the constraint of finding the optimal 1-to-1 map between branches.

2.1 Scoring branch pairs

Each edge e in a tree T defines a partition of the leaf nodes into two subsets: cutting the branch divides the tree into two subtrees, and this determines the partition of leaf nodes. The score $s(i,j)$ for any pair of edges (i,j), $i \in T_1$ and $j \in T_2$, is obtained by comparing the two corresponding partitions of the leaf nodes L. Suppose the pair (i, j) gives rise to the two partitions

$$P_{i0} \cup P_{i1} = P_{j0} \cup P_{j1} = L,$$

where P_{i0}, P_{i1} are the two disjoint subsets forming the partition of L corresponding to branch i, and similarly for P_{j0}, P_{j1}. We can then count the number of elements of L shared by the partitions: for $r, s = 0, 1$ define

$$a_{rs} = \frac{|P_{ir} \cap P_{js}|}{|P_{ir} \cup P_{js}|}.$$

For a fixed branch pair (i, j), a_{rs} represents the proportion of elements shared by the sets P_{ir} and P_{js}.

The score $s(i, j)$ is then defined by

$$s(i,\ j) = \max\{\min\{a_{00}, a_{11}\}, \min\{a_{01}, a_{10}\}\}.$$

To illustrate how the score is calculated, consider the following example. Suppose the set of leaf nodes is $\{a, b, c, d, e, f, g\}$ and that for some pair of branches (i, j) we have the partitions

$$P_{i0} = \{a, b\}, \qquad P_{i1} = \{c, d, e, f, g\},$$
$$P_{j0} = \{d, e, f, g\}, \qquad P_{j1} = \{a, b, c\}.$$

We then have

$$s(i, j) = \max\{\min\{0, \frac{1}{7}\}, \min\{\frac{2}{3}, \frac{4}{5}\}\} = \frac{2}{3}.$$

The form of the score $s(i,j)$ defined by the above equation warrants a brief discussion. Given the two partitions corresponding to branches i, j, their respective subsets can be compared in two ways: either P_{i0} is compared with P_{j0} and P_{i1} to P_{j1}, or P_{i0} is compared with P_{j1} and P_{i1} to P_{j0}. The score $s(i,j)$ corresponds to assigning to each of these two modes of comparison the score a_{rs} of the most dissimilar sets, and then picking the mode of comparison that maximizes this score. While other scoring systems could be used, ours has the advantage that when two similar partitions both consist of one large set and one much smaller set, the effect of dissimilarities between the two smaller sets is not dominated by the effect of the two larger sets. As an illustration of this, consider padding out the sets P_{i1} and P_{j0} in the example above with extra elements k, l, m, etc.: the score $s(i, j)$ does not change.

2.2 Finding the optimal tree alignment

As described above, we seek a bijection $f : T_1 \to T_2$ between the branches of the trees that maximizes the quantity given in Equation (1). This problem is solved in $O(n^3)$ steps, where n is the number of leaves, by the Munkres algorithm (Munkres, 1957; Bourgeois and Lassalle, 1971).

3 IMPLEMENTATION

The software is available as a Java applet from the website given in the abstract. The two trees are displayed side-by-side. The mean topological score between internal branches matched under f is displayed, as a global measure of similarity. Clicking with the mouse on any branch in one tree highlights its matching branch in the other tree under the bijection f. Branch thickness is used to indicate the topological score $s(i, f(i))$ for each branch: thicker

branches represent a lower score (although this can be modified by the user). In this way the user's attention is drawn to regions where the two tree topologies differ. Hovering the mouse-pointer over a branch causes its topological score to be displayed. A colour scheme is used to indicate how the well the lengths of each branch i and its match $f(i)$ agree. Shift-clicking on a branch i highlights the branch j in the other tree with the highest score $s(i, j)$.

An example of output from the software is shown in Figure 1. Two alternative phylogenies for strains of HIV viruses, taken from Roques *et al*. (2004), are shown. Further details are given in the figure caption.

4 OTHER APPROACHES

Many approaches to tree comparison involve calculating a single metric to express the dissimilarity between two trees. Of note is the Robinson–Foulds metric (Robinson and Foulds, 1981) that counts the number of partitions of leaf nodes that arise in one tree but not the other. However, there are other methods, similar to the approach presented here, that construct maps between the nodes of rooted trees. The TreeJuxtaposer software of Munzner *et al*. (2003) scores pairs of nodes according to the similarity of their sets of descendants. Given trees T_1 and T_2, it constructs a map $f_1 : T_1 \rightarrow T_2$ between nodes of the two trees, such that each node in T_1 is mapped to the node in T_2 with the most similar set of descendants. Another map $f_2 : T_2 \rightarrow T_1$ is constructed in a similar way. The two maps are not necessarily 1-to-l, and although this approach is similar to ours, the lack of symmetry can make visualization less intuitive. However, as described in the previous section, shift-clicking on branches in our application provides access to equivalents of the maps f_1 and f_2. The TreeMap software package implements a similar approach to TreeJuxtaposer (the method is described in Page, 1995), involving two maps f_1, f_2 between the nodes of rooted trees. TreeMap is particularly adapted to the analysis of trees from host species and their cospeciating parasites.

ACKNOWLEDGEMENT

Funding to pay the Open Access publication charges for this article was provided by the Medical Research Council.

Conflict of Interest: none declared.

REFERENCES

Bourgeois,F. and Lassalle,J.C. (1971) An extension of the Munkres Algorithm for the assignment problem to rectangular matrices. *Commun. ACM*, **14**, 802–804.

Cole,R. *et al*. (2000) An O(n log n) algorithm for the maximum agreement subtree problem for binary trees. *SIAM J. Comput*., **30**, 1385–1404.

Hon,W.K. *et al*. (2001) Improved phylogeny comparisons: non-shared edges, nearest neighbor interchanges, and subtree transfers. *LNCS*, **1969**, 527–538.

Munkres,J. (1957) Algorithms for the assignment and transportation problems. *J. SIAM*, **5**, 32–38.

Munzner,T. *et al*. (2003) TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility. *ACM Trans. graphics*, **22**, 453–462.

Page,R.D.M. (1995) Parallel phylogenies: reconstructing the history of host-parasite assemblages. *Cladistics*, **10**, 155–173.

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

Roques,P. *et al*. (2004) Phylogenetic characteristics of three new HIV-1 N strains and implications for the origin of group N. *AIDS*, **18**, 1371–381.