BIOINFORMATICS APPLICATIONS NOTE Vol. 23 no. 21 2007, pages 2954–2956
doi:10.1093/bioinformatics/btm466

Phylogenetics

The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees

Víctor Soria-Carrasco, Gerard Talavera, Javier Igea and Jose Castresana*

Department of Physiology and Molecular Biodiversity, Institute of Molecular Biology of Barcelona, CSIC,
Jordi Girona 18, 08034 Barcelona, Spain

Received on June 26, 2007; revised on August 13, 2007; accepted on September 6, 2007

Advance Access publication September 22, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: We introduce a new phylogenetic comparison method that measures overall differences in the relative branch length and topology of two phylogenetic trees. To do this, the algorithm first scales one of the trees to have a global divergence as similar as possible to the other tree. Then, the branch length distance, which takes differences in topology and branch lengths into account, is applied to the two trees. We thus obtain the minimum branch length distance or K tree score. Two trees with very different relative branch lengths get a high K score whereas two trees that follow a similar among-lineage rate variation get a low score, regardless of the overall rates in both trees. There are several applications of the K tree score, two of which are explained here in more detail. First, this score allows the evaluation of the performance of phylogenetic algorithms, not only with respect to their topological accuracy, but also with respect to the reproduction of a given branch length variation. In a second example, we show how the K score allows the selection of orthologous genes by choosing those that better follow the overall shape of a given reference tree.

Availability: http://molevol.ibmb.csic.es/Ktreedist.html

Contact: jcvagr@ibmb.csic.es

1 INTRODUCTION

In phylogenetic reconstruction, the application of different methods or the use of different genes may lead to the estimation of different phylogenetic trees (Castresana, 2007; Hillis *et al.*, 2005; Huerta-Cepas *et al.*, 2007). In order to analyze if the resulting trees are congruent, it is fundamental to be able to quantify differences between such trees. Normally, only topology is taken into account for such task, for example, by means of the symmetric difference (Robinson and Foulds, 1981). Few methods have been developed that also take branch length information into account (Hall, 2005; Kuhner and Felsenstein, 1994). These methods have been successfully applied to quantify the performance of different phylogenetic methods in simulated alignments, but they have the drawback that they are not directly applicable to trees with different evolutionary rates. Here, we introduce a new phylogenetic comparison measure that takes branch length information into account after scaling the trees so that they have comparable global evolutionary rates.

2 METHOD

The basis of our method to compare two phylogenetic trees, T and T', is the branch length distance (BLD) introduced by Kuhner and Felsenstein (Felsenstein, 2004; Kuhner and Felsenstein, 1994). This distance is sensitive to the similarity in branch lengths of both trees. Consider the set of partitions present in both trees, that is, the whole set of partitions present in T plus the set of partitions present in T' but not in T. Partitions for external branches are also included. For tree T, we can define an array B of branch lengths associated to each partition ($b_1, b_2, \ldots, b_N$). Branches that do not appear in T (corresponding to partitions that are only present in T') are assigned to 0 in such array. We can similarly define the array B' associated to tree T'. The BLD between trees T and T' is the squared root of the sum of $(b_i - b'_i)^2$ for all partitions. However, BLD depends on the absolute size of the trees being compared, so that two trees with the same shape (topology and relative branch length) but different global rates will give rise to a very high BLD (Kuhner and Felsenstein, 1994), which may be unwanted.

In our method, we introduce a factor, K, to scale tree T' so that both trees, T and T', have a similar global divergence. Thus, we are interested in calculating BLD after scaling T' with a factor K:

$$BLD(K) = \sqrt{\sum_{i=1}^{N} (b_i - Kb'_i)^2} \tag{1}$$

To obtain the value of K that minimizes BLD we differentiate Equation (1). It can be shown that the value of K that makes this derivative zero is:

$$K = \frac{\sum_{i=1}^{N} \left(b_i b'_i\right)}{\sum_{i=1}^{N} b'^2_i} \tag{2}$$

*To whom correspondence should be addressed.

We then substitute this value of K in Equation (1) and obtain the minimum branch length distance or K tree score. It should be taken into account that the K tree score is not symmetric, that is, the result from T to T' may not be the same than from T' to T, and, in consequence, the K score does not have the mathematical properties of a distance. Thus, this score is generally not useful to compare only two trees (although the K factor of Equation (2) can be very valuable for scaling purposes; see below). The K tree score is most useful when there is a tree that serves as reference (T) and several other trees (T') that will be scaled and compared to T. In such cases, trees T' that are similar in shape to T will receive a low K tree score whereas those that are very different will get a relatively higher K score, regardless of their overall rates.

The method that calculates the K tree score (as well as other tree comparison measures) is implemented in a Perl program called Ktreedist.

3 APPLICATIONS

There are several applications of the K tree score. First, it can be used to evaluate the quality of phylogenetic reconstructions in simulated alignments by comparing the true tree to the trees obtained with different phylogenetic methods. For example, the reference tree shown in Figure 1A was used to simulate with SeqGen (Rambaut and Grassly, 1997) 100 alignments of 1000 positions with a GTR model and gamma rate heterogeneity ($\alpha = 1.5$). We then constructed maximum-likelihood (ML) trees from such simulations using Phyml (Guindon and Gascuel, 2003) with two different conditions: without and with rate heterogeneity. To facilitate the comparison between both phylogenetic methods we imposed the topology of the reference tree during the ML reconstructions. After averaging the branch lengths of the 100 reconstructed trees, we obtained one tree for each phylogenetic method. Both trees differed in their overall rates (with the nonrate heterogeneity tree not capturing all substitutions, leading to a K scale factor $\gg 1$) but, importantly, they also differed in their shapes: see, for example, the relative lengths of sp3, sp4 and sp5. The differences in shape were reflected in the K scores: 0.197 for the average tree without rate heterogeneity and 0.030 for the average tree calculated with rate heterogeneity, indicating the better performance of the latter method. (Differences also appeared after averaging the K score from the 100 trees obtained with each method although, in this case, the magnitude of the difference was smaller.) Thus, the K tree score can be used to quantify the different quality in branch length reconstruction of different phylogenetic methods. The K score can also be used with trees that have different topologies. In such cases, nonshared branches that are relatively long will contribute to the K score much more than small conflicting branches. This is different from the symmetric difference (Robinson and Foulds, 1981), in which all topological differences count the same.

In a second example, we show how the K tree score can be used to make an accurate selection of orthologous genes. Orthologs should reflect the same topology of the species tree but they should also give rise, in principle, to a similar tree shape. We extracted from the ENSEMBL database (Hubbard *et al.*, 2007) the tables of pairs of orthologous genes of seven

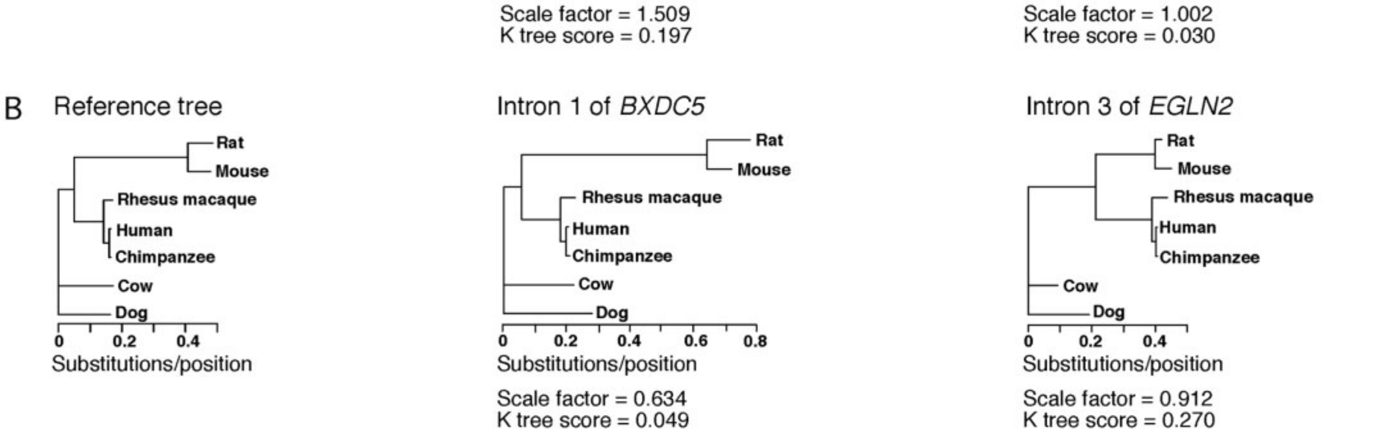

Fig. 1. (**A**) Reference tree used to simulate 100 alignments and the average reconstructions obtained by ML without and with rate heterogeneity. (**B**) Trees obtained with 472 concatenated introns (reference tree) and with two individual introns (intron 1 of *BXDC5* and intron 3 of *EGLN2*).

mammalian species. By matching the pairwise orthology tables, we constructed a set of one-to-one orthologs, and we downloaded the corresponding genes. We then extracted the introns and, after applying several filters (elimination of very long introns, those with problematic alignments, etc.), we obtained a set of 472 putative orthologous introns. Some of these introns produced ML phylogenetic trees that were of unusual shape, which could be due to different rates of evolution in different lineages (heterotachy) or could indicate that they do not come from orthologous genes (hidden paralogy). We then constructed a reference tree (Fig. 1B) with the concatenated alignment of the 472 introns using the RAxML program (Stamatakis, 2006), which can handle very long alignments, with a GTR model of evolution and four rate categories. This tree should reflect the average divergence of the seven genomes and, as expected, rodents showed a higher acceleration in their branches. We then calculated the K score of the trees of all individual introns with respect to the reference tree. We show in Figure 1B the trees of two putative orthologous introns. Intron 1 of *BXDC5*, despite having a high global rate, produced a phylogeny with the same topology and a very similar tree shape to the reference tree. This was reflected in a low K score: 0.049, smaller than the mean of the distribution of K scores of all individual introns (0.104), which is indicative of a very likely ortholog. (The K score would also be low in a similar tree but with a topological conflict affecting a small branch, which would not affect the high probability of orthology.) Intron 3 of *EGLN2* also reproduced the reference topology. However, this tree showed a relatively long basal branch in primates as well as a long branch connecting Euarchontoglires and Laurasiatherians. In consequence, the K score for this tree with respect to the reference is much higher: 0.270. In fact, this value is a clear outlier in the distribution of K scores. Although heterotachy cannot be discarded, the chances that the latter gene contains hidden paralogs in some species are higher than in the first gene. Thus, the K score can be used to establish a certain threshold and make a more accurate selection of orthologous genes.

If orthology is ensured for a set of genes, then a high K tree score with respect to a given reference will be indicative of trees with very fast-evolving species or with a significant amount of other types of heterotachy. These trees are of more difficult reconstruction, and thus the K tree score can be used to select (in a similar way as above) a set of the most reliable genes for estimating species phylogenies.

On a more practical side, the K scale factor [Equation (2)] can be used in instances where it is necessary to scale trees to have equivalent divergences. For example, the linearization of trees by means of a method like nonparametric rate smoothing produces trees with an arbitrary scale when no dates are known for the tree nodes (Sanderson, 1997). In such cases, one can make use of the K scale factor obtained from the comparison between the linearized tree and the original (reference) tree: the scaling of the linearized tree with this K factor will re-establish a genetic distance scale equivalent to that of the original tree.

ACKNOWLEDGEMENTS

J.C. is supported by grant number CGL2005-01341/BOS from the Plan Nacional I+D+I of the MEC (Spain), cofinanced with FEDER funds.

Conflict of Interest: none declared.

REFERENCES

Castresana,J. (2007) Topological variation in single-gene phylogenetic trees. *Genome Biol.*, **8**, 216.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts, pp. 531–533.

Guindon,S. and Gascuel,O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

Hall,B.G. (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Mol. Biol. Evol.*, **22**, 792–802.

Hillis,D.M. *et al.* (2005) Analysis and visualization of tree space. *Syst. Biol.*, **54**, 471–482.

Hubbard,T.J.P. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.

Huerta-Cepas,J. *et al.* (2007) The human phylome. *Genome Biol.*, **8**, R109.

Kuhner,M.K. and Felsenstein,J. (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, **11**, 459–468.

Rambaut,A. and Grassly,N.C. (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

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

Sanderson,M.J. (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Mol. Biol. Evol.*, **14**, 1218–1231.

Stamatakis,A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.