BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 3 2000
Pages 296–297

RRTree: Relative-Rate Tests between groups of sequences on a phylogenetic tree

Marc Robinson-Rechavi[1,3,*] and Dorothée Huchon[2]

[1]Department of Zoology, Tel Aviv University, Ramat Aviv 69978, Israel and [2]Institut des Sciences de l'Evolution, Université Montpellier 2, 34095 Montpellier, France

Received on June 17, 1999; revised on August 27, 1999; accepted on October 20, 1999

Abstract

Summary: *RRTree is a user-friendly program for comparing substitution rates between lineages of protein or DNA sequences, relative to an outgroup, through relative rate tests. Genetic diversity is taken into account through use of several sequences, and phylogenetic relations are integrated by topological weighting.*

Availability: *The ANSI C source code of RRTree, and compiled versions for Macintosh, MS-DOS/Windows, SUN Solaris, and CGI, are freely available at http://pbil.univ-lyon1.fr/software/rrtree.html*

Contact: *marc.robinson@ens-lyon.fr*

Introduction

Substitution rates between sequences are routinely compared using relative-rate tests that originally inferred the evolutionary rate of two sequences relative to a third one (Sarich and Wilson, 1973; Wu and Li, 1985). It is one of the most common ways of testing molecular clock hypotheses. Recent developments of the relative-rate test allow the comparison of more than two sequences (Li and Bousquet, 1992; Takezaki *et al.*, 1995), and taking sampling and phylogeny into account (Robinson *et al.*, 1998).

Available programs computing relative-rate tests usually have little concern for user-friendliness, with the notable exception of PHYLTEST (Kumar, 1995), or for the variety of formats used for sequence data. Lintre (Takezaki *et al.*, 1995), K2WuLi (Jermiin, 1997) and PHYLTEST each require a specific file format, and are available only for PC (and Unix for Lintre). Lintre uses a specific tree file format, which makes exchange with other programs, or even printing the trees, problematic. K2WuLi only allows comparison of one sequence to another with one outgroup, whereas PHYLTEST and Lintre use several sequences, but without taking into account phylogenetic information. Although Lintre can compare amino-acid sequences, and the three codon positions separately, none of these programs implements a model specific to protein-coding DNA sequences.

The program RRTree (Relative-Rate Test with a tree) presented here, (1) generalizes the relative-rate tests to any number of sequences, with phylogenetic weighting and (2) answers the aim of bringing the relative-rate test to biologists in a user-friendly way, and notably those who use large datasets. It also presents a relevant choice of distance methods (coding or non coding DNA, protein sequences).

Algorithm

The methods used in RRTree are mostly described in Robinson *et al.* (1998), although we added here a larger choice of distance methods. See also Wu and Li (1985), Li and Bousquet (1992), and Takezaki *et al.* (1995). Lengths of internal branches, and the associated variances, are estimated by a least squares estimate as in Li and Bousquet (1992). The main originality of the implemented algorithm is weighting by the phylogenetic topology (phylogenetic uncertainties are allowed).

Sequences are grouped in ingroups, to be compared, and an outgroup, each containing any number of sequences (at least one). The distance between two groups of sequences is the mean of all distances between all pairs of sequences of the two groups, weighted by their phylogenetic positions and eventually (depending on method) the number of nucleotides compared. A full description of the weighting scheme may be found in Robinson *et al.* (1998).

To allow use of independent phylogenetic information, a rooted guide tree must be provided, with the two ingroups monophyletic relative to the outgroup. Multifurcations can be specified by internal branches of length 0 in a bifurcating tree, or by 'real' multifurcations in the form (A, B, C). Lowly supported nodes can be discarded by the program, and thus considered as additional multifurcations. The nodes to discard are defined by a user-specified threshold on scores, usually bootstrap percentages, but it

*To whom correspondence should be addressed.
[3]Present address: Laboratoire de Biologie Moléculaire et Cellulaire, Ecole Normale Supérieure de Lyon, 6 allée d'Italie, 69364 Lyon cedex 07, France, marc.robinson@ens-lyon.fr

may be any other measure of support. The default value is set to 50, Berry and Gascuel (1996) having shown that discarding nodes under 50% bootstrap frequency yields trees significantly closer to the true tree than the fully resolved reconstructed tree. But the user is free to provide any other threshold, or keep all nodes. If no tree is provided, RRTree gives equal weight to all sequences within a group.

RRTree analyzes all types of sequences (i.e. coding/non-coding, nuclear/mitochondrial, nucleotides/amino acids). For coding nucleotide sequences different types of substitutions are computed, according to the method of Li (1993) and Pamilo and Bianchi (1993). The program allows evolutionary rate evaluations at different divergence or saturation levels of the sequences, through choice of the relevant distance. Synonymous and non-synonymous transversions may in particular be useful when there is a GC-content difference between compared sequences, since transition computations, but not transversion computations, are then biased (Galtier and Gouy, 1995).

The exact probability associated with the test is computed using integration by parts, assuming a normal distribution of the mean number of substitutions per site.

Input and Output

A single datafile of aligned sequences is compulsory to use RRTree. The following formats are implemented: MASE, CLUSTAL, GDE, FASTA, PHYLIP, NEXUS, MEGA, Lintre, and PHYLTEST. These formats cover the vast majority of programs biologists routinely use. The last two allow easy comparison with clock test results from these programs. RRTree automatically detects the file format and sequence type, while allowing manual specification. An eventual tree file should be in the NEWICK format.

To run RRTree the user needs to respond to queries, default options being systematically proposed, relevant to the type of data and the previous choices of the user, to help the non-specialist user. Names can be specified for the groups of sequences, making the output clearer. To facilitate multiple runs of RRTree on the same or similar datasets, a command file can be automatically or manually created, and used to run the program. This has been found especially useful to avoid re-entering lineage attribution of sequences when running the program several times on a large dataset.

The standard output, on the screen or in a file, includes for each pair of ingroup comparisons: topological weights given to all sequences, mean GC content of the two ingroups and the outgroup (DNA/RNA sequences only), mean substitution rates relative to the outgroup, the rate difference between the two ingroups, its standard deviation, and the probability associated to the test. All means are weighted when a tree is provided. An optional output file can contain the same results in a table, for analysis of large amounts of results under statistical software such as Microsoft Excel or Statview.

Acknowledgements

Many functions used in RRTree were adapted from the code of PHYLO_WIN (Galtier *et al.*, 1996), and we warmly thank Nicolas Galtier for letting us use them. Manolo Gouy, Guy Perrière and Tal Pupko gave helpful advice. Thanks to all beta-users who signaled bugs or suggested improvements. D.H. is supported by a MENESR grant (N 97132), and part of this work was done when M.R.R. was a recipient of a Lavoisier grant from the French Ministry of Foreign Affairs. RRTree is hosted by the ftp site of the Pôle BioInformatique Lyonnais.

References

Berry,V. and Gascuel,O. (1996) On the interpretation of bootstrap trees: appropriate threshold of clade selection and induced gain. *Mol. Biol. Evol.*, **13**, 999–1011.

Galtier,N. and Gouy,M. (1995) Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl Acad. Sci. USA*, **92**, 11317–11321.

Galtier,N., Gouy,M. and Gautier,C. (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.*, **12**, 543–548.

Jermiin,L.S. (1997) K2WuLi: A program to conduct relative-rate tests from DNA sequences. http://jcsmr.anu.edu.au/dmm/humgen.

Kumar,S. (1995) *PHYLTEST: Phylogeny Hypothesis Testing*. Penn State University, University Park, PA 16801.

Li,P. and Bousquet,J. (1992) Relative-rate test for nucleotide substitutions between two lineages. *Mol. Biol. Evol.*, **9**, 1185–1189.

Li,W.H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.*, **36**, 96–99.

Pamilo,P. and Bianchi,N.O. (1993) Evolution of the Zfx and Zfy genes: rates and interdependence between the genes. *Mol. Biol. Evol.*, **10**, 271–281.

Robinson,M., Gouy,M., Gautier,C. and Mouchiroud,D. (1998) Sensitivity of the relative-rate test to taxonomic sampling. *Mol. Biol. Evol.*, **15**, 1091–1098.

Sarich,V.M. and Wilson,A.C. (1973) Generation time and genomic evolution in primates. *Science*, **179**, 1144–1147.

Takezaki,N., Rzhetsky,A. and Nei,M. (1995) Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.*, **12**, 823–833.

Wu,C.I. and Li,W.H. (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl Acad. Sci. USA*, **82**, 1741–1745.