RRF: An R package for evolutionary dates, rates, and priors using relative rate framework

Qiqing Tao[1,2], Sudip Sharma[1,2], and Sudhir Kumar[1,2*]

[1]Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA 19122, USA., [2]Department of Biology, Temple University, Philadelphia, PA 19122, USA.

Abstract

The relative rate framework (RRF) estimates evolutionary rates and divergence times using a phylogeny with branch lengths. RRF is the fundamental basis of RelTime, a relaxed clock method for molecular dating of large phylogenies. RRF has also been used to develop computationally efficient and accurate methods for testing autocorrelation of branch rates in a phylogeny (CorrTest) and guiding the selection of birth-death speciation tree prior (ddBD) for use in Bayesian molecular dating. We have programmed RRF in R to provide open-source software to estimate divergence times, infer evolutionary rates, conduct CorrTest, and build a ddBD tree prior. RRF input is a standard newick file with branch lengths. It outputs in formats directly useable or connected with other visualization software and packages (e.g., MEGA, ggplot, and FigTree). The source code and example datasets are available from GitHub (https://github.com/cathyqqtao/RRF).

1. Introduction

The relative rate framework (RRF) is a system of equations that relate relative rates between sister lineages with their evolutionary (sequence) divergences (Tamura et al. 2018). A lineage emanating from a node consists of the branch connecting to that node (stem) and all the branches in the descendant clade (**Fig. 1A**). In RRF, the principle of minimum rate change among lineages and their descendants is applied, enabling an analytical solution for estimating times and lineage rates from branch lengths (Tamura et al., 2018). Computer simulations and empirical data analyses have shown that RRF can perform well in estimating divergence times and is very fast for large phylogenies (review in Tao *et al.*, 2020). Its use has been cited in hundreds of research articles already. But, RRF has only been available in MEGA (Kumar *et al.*, 2012; Tamura *et al.*, 2021), which is not useable in the R ecosystem. So, we have programmed RRF in R for its broader application.

In addition to functions to estimate dates and rates in the RRF package, we have implemented two other useful functions that have been recently developed by employing RRF. One is for testing the presence of autocorrelation of branch rates in a phylogeny (CorrTest) (Tao *et al.*, 2019). Knowledge of autocorrelation is interesting biologically and useful when selecting a clock model in Bayesian dating analyses. The second function is to generate a data-driven birth-and-death tree prior (ddBD) for Bayesian dating analyses (Tao *et al.*, 2021). Using ddBD speciation prior in MCMCTree produces better time estimates than the default setting, especially when the number of calibrations is small (Tao *et al.*, 2021).

We describe our R implementations, input requirements, outputs produced, and interpreting results in the following. Because of their analytical nature, all the RRF calculations are ultra-fast even for large phylogenies; it took less than one second for phylogenies with 1000 tips (**Fig. 1B**).

2. Functions and implementations

2.1. Relative rate framework (RRF)

The RRF package requires a rooted phylogeny with branch lengths derived from molecular or non-molecular (e.g., morphological characters) sequences. The program accepts a text file containing a rooted newick tree. The users can also provide a file containing a list of names of tips in the outgroup. When an outgroup file is supplied, it will be used even if the newick tree is rooted. Three primary RRF functions provided are *rrf.rates*, *rrf.times*, and *rrf.rates.times*.

The *rrf.rates* function uses a phylogeny with branch lengths as input. It outputs a table of relative lineage rates and a tree with lineage rates in the newick format. Note that RRF estimates lineage rates rather than branch rates. Lineage rates are assigned to the stem branches of the corresponding lineages (**Fig. 1A**) in the output newick tree. Also, only the ingroup tips are included in the output because the equality of rates between ingroup and outgroup clades cannot be tested (Kumar *et al.*, 2016). The output tree can be plotted and branches colored based on the lineage rates in R (**Fig. 1C**).

The *rrf.times* function has the same requirements and characteristics as *rrf.rates*, except that it outputs a table of relative node times. The third function, *rrf.rates.times*, computes both relative rates and node times simultaneously. It outputs a table of rates and times and a timetree with colored lineage rates (**Fig. 1C**).

2.2. Functions based on the RRF framework

The *corrtest* function is included in the RRF package to classify the rate variation among branches and lineages in a phylogeny. It is based on a machine learning approach that uses the relative rates generated by RRF and produces a CorrScore that ranges from 0 to 1 (Tao *et al.*, 2019). A high CorrScore is suggestive of autocorrelated rates. A *P*-value is also produced to assess the statistical significance of the rejection of

the independent rate model. The mean value and standard deviation of rates will be output if the user provides an anchor node time to convert relative node times to absolute node times. The other inputs required by *corrtest* are a tree with branch lengths, and the number of sister rate resampling replicates that is recommended for small phylogenies (<50 tips) (Tao *et al.*, 2019).

The *ddbd* function estimates the Birth-Death (BD) speciation tree parameters prior to molecular dating use in the MCMCTree (Yang, 2007). It uses relative times obtained using RRF (Tao *et al.*, 2021). This function requires a tree with branch lengths along with one anchor time. There is an option to estimate the proportion of species sampled (i.e., sampling fraction). The *ddbd* function produces estimates of birth and death rates and sampling fraction if it is not user-provided. A figure showing the density distribution of node ages and the fitted BD curve will be plotted (**Fig. 1D**). Note that while the inferred BD parameters are useful for molecular dating, these parameters are not direct estimates of birth and death rates because many combinations of parameter values can result in the same kernel density (Tao *et al.*, 2021).

3. Conclusions

In conclusion, RRF will be a convenient tool for efficiently generating rates and times for small and large phylogenies. Its use will also facilitate a better selection of priors in Bayesian dating analysis. One can compare Bayesian estimates obtained using these priors with estimates obtained directly from RRF/RelTime to evaluate the robustness of inferred times. Ultimately, the RRF package application can enable the Tree of Life's dating with greater accuracy and precision, important for inferring organism evolution, diversification dynamics, phylogeography, and biogeography studies.

Acknowledgments

We thank Jose Barba-Montoya and Sara Vahdatshoar for testing the R package and providing many comments.

Funding

Support by a grant from the National Institutes of Health (GM-0126567-01) to S.K.

Conflict of Interest: none declared.

References

dos Reis,M. *et al.* (2012) Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc R Soc B*, **279**, 3491-3500.

Kumar,S. *et al.* (2012) MEGA-CC: Computing core of molecular evolutionary genetics analysis program

for automated and iterative data analysis. *Bioinformatics*, **28**, 2685–2686.

Kumar,S. *et al.* (2016) MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*, **33**, 1870–1874.

Tamura,K. *et al.* (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol Biol Evol*, **38**, 3022–3027.

Tao,Q. *et al.* (2019) A machine learning method for detecting autocorrelation of evolutionary rates in large phylogenies. *Mol Biol Evol*, **36**, 811–824.

Tao,Q. *et al.* (2021) Data-driven speciation tree prior for better species divergence times in calibration-poor molecular phylogenies. *Bioinformatics*, **37**, i102–i110.

Tao,Q. *et al.* (2020) Efficient Methods for Dating Evolutionary Divergences. In, Ho,S.Y.W. (ed), *The Molecular Evolutionary Clock*. Springer US, pp. 197–220.

Yang,Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*, **24**, 1586–1591.

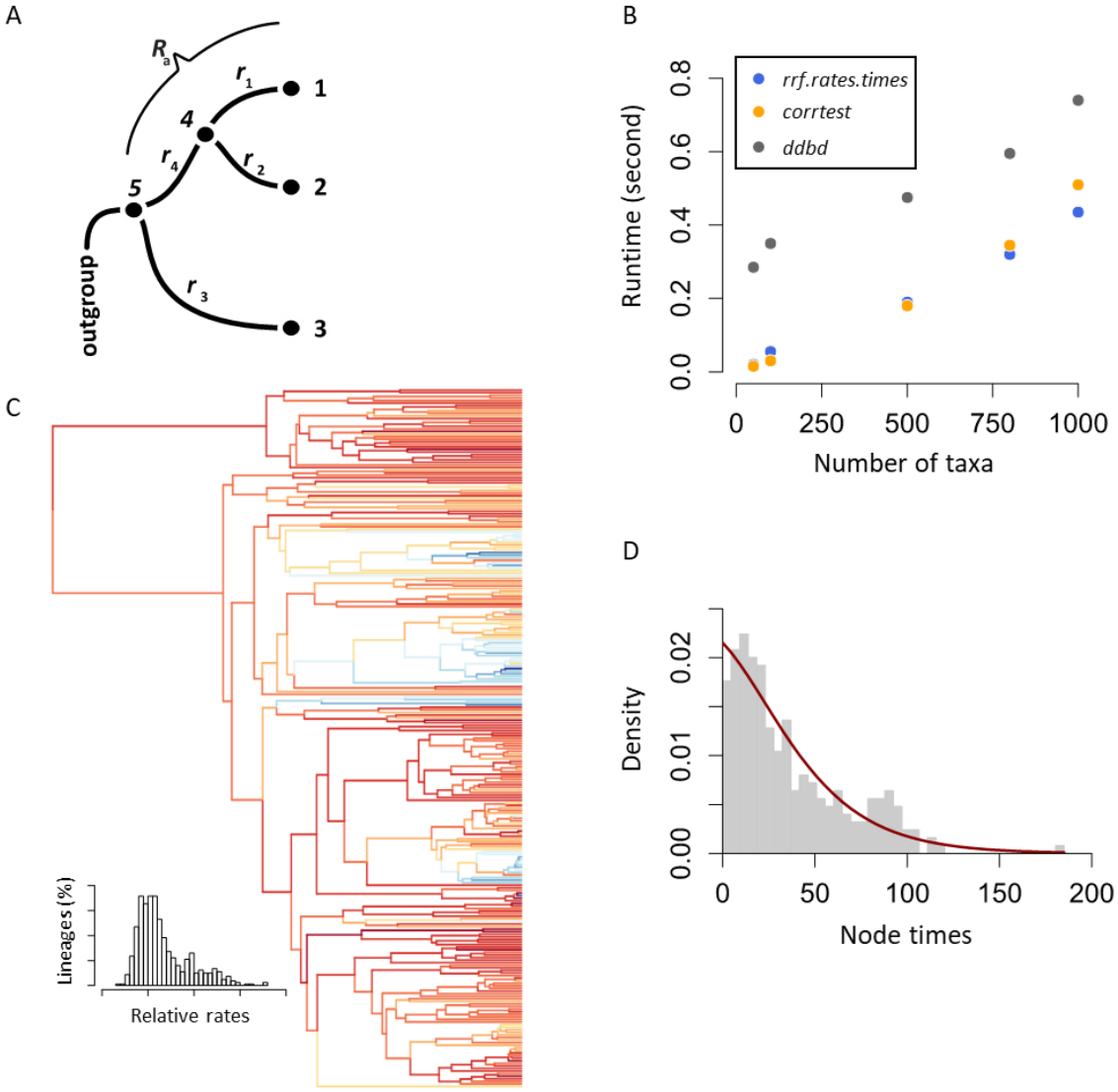

Figure 1. (A) An example three-taxon phylogeny, rooted by an outgroup, with branch rates (*r*) and lineage rates (*R*) shown. The lineage rate R_a includes the evolutionary rate of stem branch (r_4) and evolutionary rates of all the descendent branches (r_1 and r_2). (B) Time taken by *rrf.rates.times*, *corrtest,* and *ddbd* functions for small and large phylogenies. Each point represents the average of runtimes of two phylogenies: one simulated under the BD process with molecular rates drawn from an independent lognormal branch rate model, and the other simulated using the same BD process but with rates drawn from an autocorrelated branch rate model. All parameters used in the simulation were derived empirically. (C) A timetree produced by *rrf.rates.times* in which branches are colored based on the value of relative rates; the inset shows the distribution of relative rates. For this analysis, branch lengths of the phylogeny shown were estimated using a sequence alignment from dos Reis *et al.* (2012). (D) The density distribution of node ages and the fitted curve used to generate BD parameters by *ddbd*.