Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Evolutionary Bioinformatics

iteRates: An R Package for Implementing a Parametric Rate Comparison on Phylogenetic Trees

James A. Fordyce[1,2], Premal Shah[1–3] and Benjamin M. Fitzpatrick[1,2]

[1]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, TN, USA. [2]National Institute for Mathematical and Biological Synthesis, Knoxville, TN, USA. [3]Department of Biology, University of Pennsylvania, Philadelphia, PA, USA.

ABSTRACT: Patterns of diversification rate variation detected in phylogenetic hypotheses are frequently used to infer historical, ecological, and evolutionary processes. The parametric rate comparison (PRC) is a method for detecting rate variation in trees that models branch lengths as random variables drawn from familiar statistical distributions. iteRates is a library of functions for the R statistical computing environment for implementing PRC on phylogenetic trees. Here, we describe some of the functions in iteRates for subtree identification, tree manipulation, applying the PRC and *K*-clades PRC analyses, and conducting a whole-tree randomization test.

KEYWORDS: chronogram, diversification-rate, phylogeny, rate-variation

CITATION: Fordyce et al. iteRates: An R Package for Implementing a Parametric Rate Comparison on Phylogenetic Trees. *Evolutionary Bioinformatics* 2014:10 127–130 doi: 10.4137/EBO.S16487.

RECEIVED: April 25, 2014. **RESUBMITTED:** June 17, 2014. **ACCEPTED FOR PUBLICATION:** June 17, 2014.

ACADEMIC EDITOR: Jike Cui, Associate Editor

TYPE: Rapid Communication

FUNDING: This work was supported in part by the University of Tennessee, a graduate assistantship to PS from the National Institute for Mathematical and Biological Synthesis, and the US National Science Foundation (DEB 0614223 and 1050947). PS acknowledges support from a David and Lucille Packard Foundation Fellowship awarded to Joshua B. Plotkin, and from a Burroughs Wellcome Fund Career Award.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: jfordyce@utk.edu

This paper was subject to independent, expert peer review by a minimum of two blind peer reviewers. All editorial decisions were made by the independent academic editor. All authors have provided signed confirmation of their compliance with ethical and legal obligations including (but not limited to) use of any copyrighted material, compliance with ICMJE authorship and competing interests disclosure guidelines and, where applicable, compliance with legal and ethical guidelines on human and animal research participants.

Introduction

Macro-evolutionary studies frequently use phylogenetic trees to examine diversification rate variation within and among groups. Variation in diversification rate in a phylogenetic tree can inform evolutionary hypotheses regarding the role of past geological or climatic events, evolutionary novelty, and adaptive and non-adaptive radiations.[1–3] Various methods have been developed to examine diversification rate variation, including, but not limited to, those that examine the accumulation of lineages through time,[4–7] tree balance,[8–10] distribution of tree branch lengths,[11] and the shape of ordered cladogenic events.[12]

Here, we present the R library iteRates, which implements the parametric rate comparison (PRC) test,[13] a new method and approach for identifying rate variation in phylogenetic trees. An in-depth description of the PRC, as well as an examination of its statistical power and false positive rates, can be found in the study by Shah et al.[13] Briefly, the PRC examines the fit of a distribution of branch lengths extracted from a phylogenetic tree to standard statistical distributions.[14] The lengths of terminal branches are treated as censored at the time of sampling. Internal branch lengths and terminal branch lengths are jointly modeled using the censored form of a given distribution. This approach differs from other approaches aimed at identifying rate heterogeneity in a phylogenetic tree in that it does not attempt to estimate the parameters of a particular model of diversification, rather it simply examines the statistical properties of the distribution of branch lengths. By iterating through subtrees in a tree, the PRC can be used to identify subclades where diversification rate differs from the remainder of the tree (see Fig. 1 in Shah et al.[13]). The PRC can also be used to compare a priori defined groups, and does not require that these groups be monophyletic. The PRC can be used both as a hypothesis testing tool

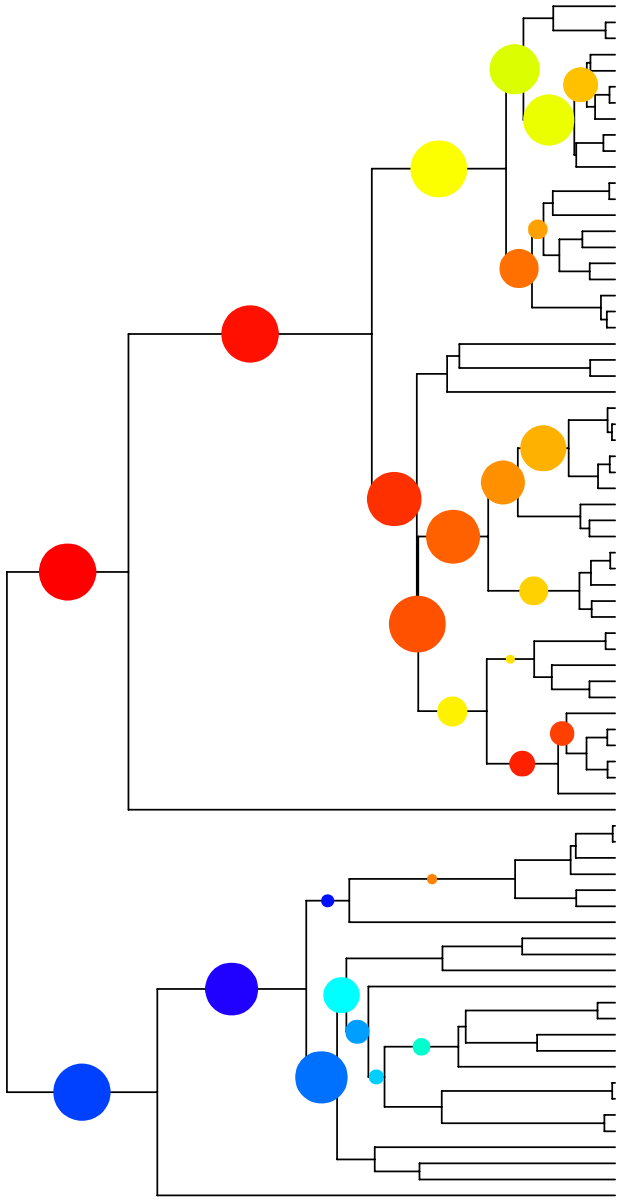

Figure 1. A phylogenetic tree showing regions of rate variation identified by the function comp.subs using the function `color.tree.plot`.
Notes: The colored circles range from hot (red) indicating a relative rate increase to cold (blue) indicating a relative rate decrease. The size of each of the colored circles is scaled to the statistical support for a rate shift at that branch.

and for exploratory data analysis using functions in iteRates library.

Description

PRC – the parametric rate comparison test. The PRC test is implemented using the function `comp.subs`. The function iterates through subtrees of a phylogenetic tree and compares the distribution of branch lengths in that subtree to the remainder of the tree. The function has arguments that allow flexibility for the user to govern how `comp.subs` implements the PRC. An example usage of `comp.subs` showing some important arguments at their default values is as follows:

```
prc.test <-
comp.subs(tree,thr = 6,srt = "drop",mod.
id = c(1,0,0,0))
```

Here, the argument `tree` is an ultrametric phylogenetic tree of object class `phylo`. The argument `thr` indicates the threshold for the minimum number of branches required for a subtree to be considered in the analysis. The argument `srt` determines how the branch that links a subtree to the remainder of the tree is treated in the analysis. The default is to drop the linking branch from the analysis because there is no way of knowing exactly at what point an inferred rate change might have occurred along the branch; however, the user has the option to include this linking branch as part of the subtree or part of the remainder of the tree. There are four different statistical distributions that `iteRates` uses to model the distribution of branch lengths: exponential, Weibull, log-normal, and variable rates.[13,14] The argument `mod.id` is used to indicate which of these distributions are used in the PRC. The default is to consider only an exponential model. When multiple distributions are included in the analysis, `comp.subs` will use the Akaike information criterion (AIC)[15] scores to pick the best-fit model for each subtree vs. remainder of the tree comparison.

There is an expectation that false-positive rates will increase as the tree under scrutiny deviates from a pure-birth model. A whole-tree randomization test can be used to avoid spurious inferences based on observed statistical significance for each node. Here, the topology of the tree is randomized while retaining the observed branching times. For each randomization, the PRC test is employed and the number of statistically significant clades is recorded. Following many randomizations, the observed number of significant clades is compared against the distribution of the number of significant clades from the randomized trees. The whole-tree randomization test is implemented in the function `tree.rand.test`. This function requires an ultrametric tree of object class `phylo` and has arguments that allow the user to determine the number of randomizations and the statistical distributions used.

The results of the PRC can be visualized on a phylogenetic tree using the function `color.tree.plot`. Various options are available to the user to illustrate the relative direction, magnitude, and statistical support for a diversification rate change. The function requires a tree object class `phylo` and the result object from `comp.subs`. Figure 1 shows an example of a phylogenetic tree showing regions with rate variation using default settings.

The PRC test assumes that taxon sampling is complete, although it is robust to incomplete taxon sampling if taxon sampling is random.[13] For situations where there is ambiguity as to what complete taxon sampling might be, for example if there is a recent radiation, or when a tree is based on a higher taxonomic rank (eg, a genus-level phylogeny), the user might

choose to trim a particular amount of time from the tree. For example, a researcher might decide that taxon sampling is complete up until five million years before present in the phylogeny. The function `trimTree` can be used to trim a specified amount of time (or branch length) from the tips of a tree. This function returns a list that contains the trimmed tree and a key indicating the taxa from the original tree that have been collapsed to each terminal node of the trimmed tree.

***K*-clades PRC.** The *K*-clades PRC allows for statistical hypothesis testing of rate variation among a priori defined clades in a tree. All the subtrees in a tree are identified using the function `id.subtrees`. This function returns a list containing all the subtrees that might be compared. The function `comp.fit.subs` implements the *K*-clades PRC test. An example usage of `comp.fit.subs` showing some important arguments at their default values is as follows:

```
cfs<-
comp.fit.subs(tree.Kclades$subtree,focal =
c(55,27,3),k = 3)
```

Here, `tree.Kclades$subtree` is the list of all subtrees provided by `id.subtrees`. The argument `focal` indicates the identifier of the subtrees of interest for comparison. Figure 2 shows a hypothetical phylogenetic tree with the subtrees of interest indicated. The argument k indicates the maximum number of different rates to explore, in this case three. The function will compare all possible models of rate variation, ranging from all subtrees having the same rate, to k number of subtrees having different rates. As in the function `comp.subs`, the user can choose any combination of the four available statistical distributions using the argument `mod.id`. The default is to fit only an exponential distribution (`mod.id = c(1,0,0,0)`). Parameter values, log likelihood, and AIC scores and ΔAIC are returned. The results can be summarized using the function `tab.summary`, which will restrict the returned output to a ΔAIC limit determined by the user and the best-fit model for each k. An example of `tab.summary` and output is as follows:

```
> tab.summary(cfs,daic = 10)
```

k	Groups	total. param	LL	AIC	AICc	dAICc
1	123	1	−249.2265	500.4530	500.5145	13.5576637
2	12vs3	2	−241.3847	486.7694	486.9569	0.0000000
2	23vs1	2	−245.1312	494.2624	494.4499	7.4930709
3	1vs2vs3	3	−240.6848	487.3696	487.7506	0.7937253

In this example, the best model is one that groups the first and second subtrees (subtrees 55 and 27, respectively) to share the same rate, and the third subtree (subtree 3) to be modeled as having a separate rate. Note, based on ΔAIC, modeling each subtree as having a separate rate also falls within the group of "best" models. The `tab.summary` function can be useful when exploring rate variation across numerous subtrees because the number of possible groupings can get quite large; for example, when k = 5, there are 52 different groupings explored.

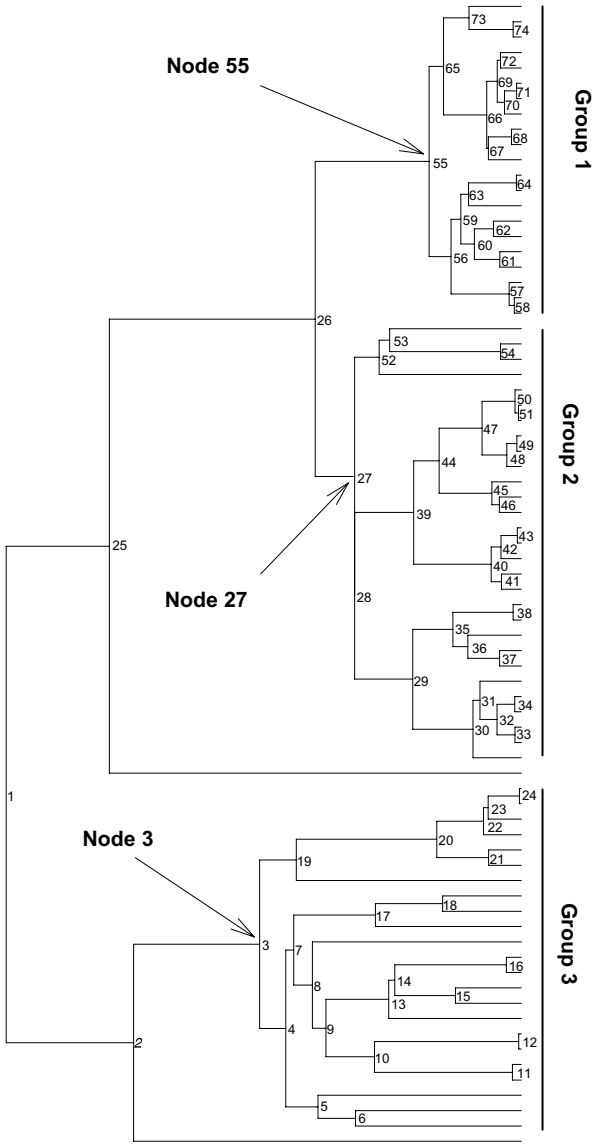

Figure 2. A phylogenetic tree showing the subtrees identified by the function `id.subtrees`. For the example in the text, the subtrees defined by nodes 55, 27, and 3 make up the three groups examined using the *K*-clades PRC, identified by the argument focal = c(55,27,3) in the function `comp.fit.subs`.

Conclusion

The package `iteRates` provides the library functions required to employ the PRC test and the *K*-clades PRC as described by Shah et al.[13] in the R statistical computing environment (R Development Core Team 2011).[16] The package and dependences are available at The Comprehensive R Archive Network (http://cran.r-project.org/). The functions therein will be useful to investigators exploring diversification rate variation across a phylogeny and conducting explicit hypothesis testing among subtrees in a phylogeny.

Acknowledgments

We thank B. O'Meara, C. Hulsey, and T. Near for helpful comments and discussion. Improvement of the PRC was facilitated by helpful comments from C. Boettiger and L. Harmon.

Author Contributions

JAF, PS, and BMF conceived and designed the experiments, and analyzed the data. JAF wrote the first draft of the manuscript. JAF, PS, and BMF contributed to the writing of the manuscript, agreed with the manuscript results and conclusions, jointly developed the structure and arguments for the paper, and made critical revisions and approved the final version. All authors reviewed and approved the final manuscript.

REFERENCES

1. Fordyce JA. Host shifts and evolutionary radiations of butterflies. *Proc R Soc B Biol Sci*. 2010;277(1701):3735–43.
2. Harmon LJ, Schulte JA, Larson A, Losos JB. Tempo and mode of evolutionary radiation in iguanian lizards. *Science*. 2003;301(5635):961–4.
3. Rabosky DL, Lovette IJ. Density-dependent diversification in North American wood warblers. *Proc R Soc B Biol Sci*. 2008;275(1649):2363–71.
4. Morlon H, Potts MD, Plotkin JB. Inferring the dynamics of diversification: a coalescent approach. *PLoS Biol*. 2010;8(9):e1000493.
5. Nee S, Holmes EC, May RM, Harvey PH. Extinction rates can be estimated from molecular phylogenies. *Philos Trans R Soc Lond B Biol Sci*. 1994;344(1307):77–82.
6. Nee S, May RM, Harvey PH. The reconstructed evolutionary process. *Philos Trans R Soc Lond B Biol Sci*. 1994;344(1309):305–11.
7. Rabosky DL. Likelihood methods for detecting temporal shifts in diversification rates. *Evolution*. 2006;60(6):1152–64.
8. Mooers AO, Heard SB. Inferring evolutionary process from phylogenetic tree shape. *Q Rev Biol*. 1997;72(1):31–54.
9. Rohlf FJ, Chang WS, Sokal RR, Kim JY. Accuracy of estimated phylogenies-effects of tree topology and evolutionary model. *Evolution*. 1990;44(6):1671–84.
10. Shao KT, Sokal RR. Tree balance. *Syst Zool*. 1990;39(3):266–76.
11. Alfaro ME, Santini F, Brock C, et al. Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc Natl Acad Sci U S A*. 2009;106(32):13410–14.
12. Pybus OG, Harvey PH. Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc R Soc Lond B Biol Sci*. 2000;267(1459):2267–72.
13. Shah P, Fitzpatrick BM, Fordyce JA. A parametric method for assessing diversification-rate variation in phylogenetic trees. *Evolution*. 2013;67(2):368–77.
14. Venditti C, Meade A, Pagel M. Phylogenies reveal new interpretation of speciation and the Red Queen. *Nature*. 2010;463(7279):349–52.
15. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19(6):716–23.
16. R Development Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (2011); Available at http://www.R-project.org. ISBN 3-900051-07-0.