geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees

Matthew W. Pennell [1*], Jonathan M. Eastman [1 *], Graham J. Slater [2], Joseph W. Brown [1,3], Josef C. Uyeda [1], Richard G. FitzJohn [4], Michael E. Alfaro [5], and Luke J. Harmon [1]

[1]Department of Biological Sciences and Institute for Bioinformatics and Evolutionary Studies, University of Idaho, Moscow, ID 83844, U.S.A., [2]Department of Paleobiology, National Museum of Natural History, Smithsonian Institution, Washington, D.C. 20013, U.S.A., [3]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI 48109, U.S.A., [4]Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia, [5]Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095, U.S.A.

Associate Editor: Prof. David Posada

ABSTRACT

Summary: Phylogenetic comparative methods are essential for addressing evolutionary hypotheses with interspecific data. The scale and scope of such data has increased dramatically in the last few years. Many existing approaches are either computationally infeasible or inappropriate for data of this size. To address both of these problems, we present `geiger v2.0`, a complete overhaul of the popular R package `geiger` (Harmon *et al.*, 2008). We have re-implemented existing methods with more efficient algorithms and have developed several new approaches for accomodating heterogeneous models and data types.

Availability: This R package is available on the CRAN repository http://cran.r-project.org/web/packages/geiger/. All source code is also available on github http://github.com/mwpennell/geiger-v2. `geiger v2.0` depends on the `ape` package (Paradis *et al.*, 2004).

Contact: mwpennell@gmail.com

1 INTRODUCTION

In the past few decades, phylogenetic trees have become an key component of evolutionary research. This development has been fueled by the increased availability of robust time-calibrated phylogenies for many groups, in addition to an expanding number of statistical techniques for inferring patterns and processes from comparative data (reviewed in Pennell and Harmon, 2013). Among the many R packages developed for phylogenetic and comparative data, `geiger` (Harmon *et al.*, 2008) has been a primary utility for making macroevolutionary inferences from phylogenetic trees.

However, in the ∼6 years since the initial release of `geiger`, the data available for comparative biology have changed substantially. For some groups, we now have phylogenies and corresponding trait data with thousands (and even tens of thousands) of species (e.g., Jetz *et al.*, 2012). `geiger v2.0` is a complete overhaul of the previous release (Harmon *et al.*, 2008), designed to scale up comparative methods to large data sets. To do so, we have taken two complementary tacks. The first is to improve algorithms and implementations to increase computational efficiency of existing methods. The second is to expand the suite of statistical approaches to allow for heterogeneity in both models and data types across the phylogeny.

In this applications note, we briefly describe the methods now in `geiger`, with a particular focus on novel implementations and algorithms. Most of these methods have been previously published elsewhere in some form and we refer readers to the relevant publications for full explanations (for an overview of the main features of the package, see Appendix 1 in Supplementary Online Material).

2 METHODS

2.1 Fitting simple models of character evolution using ML

Fitting and comparing models of trait evolution can provide insight into many macroevolutionary questions (Pennell and Harmon, 2013). The two “workhorse” functions in `geiger` for fitting models of trait evolution using maximum likelihood, `fitContinuous` and `fitDiscrete`, have both been completely re-implemented. The previous version of `fitContinuous` calculated the likelihood of a set of continuous characters (e.g., body size) having evolved under a model using a variance-covariance (vcv) matrix approach. This involves inverting the vcv, which is extremely computationally intensive, making the method infeasible for large trees. FitzJohn (2012) demonstrated that using a “pruning”-based algorithm (Felsenstein, 1973) allows for much more efficient likelihood calculations. This algorithm is used the `diversitree` package (FitzJohn, 2012). (For related algorithms, see Freckleton (2012) and the `phylolm` package (Ho and Ané, 2014)). The approach has now been extended to all the models in `fitContinuous`. In addition to improving the efficiency of the algorithm, we have improved numerical optimization procedures and implemented a novel method to simultaneously estimate model parameters and an additional term to account for measurement error.

[*]these authors contributed equally

For the case of modeling discrete characters, the most commonly used models are the Mk models (Pagel, 1994), which can be fit to data with the function `fitDiscrete`. We have re-written this function for `geiger v2.0`, again using an alternative, more efficient algorithm for calculating the likelihood of observing trait values under a model (FitzJohn *et al.*, 2009). In the new version of `fitDiscrete`, the computational time required for the likelihood calculation scales approximately linearly with the number of character states, whereas the previous version scaled superlinearly (see Appendix 2 in SOM for details). This improvement allows for more complex models to be efficiently fit to large phylogenies.

2.2 Bayesian methods for fitting models of character evolution

A major addition to `geiger` is the implementation of several Bayesian methods for fitting models of trait evolution to comparative data. These include the AUTEUR approach of Eastman *et al.* (2011), which uses reversible jump Markov chain Monte Carlo machinery (Green, 1995) to move across multi-rate models of various complexity. The implementation of this method in `geiger v2.0` improves upon the original by allowing model partitions to be constrained a priori and alternate models to be compared (Eastman *et al.*, 2013b). Additionally, `geiger` now includes: a method for fitting models to phylogenies including unresolved clades using Approximate Bayesian Computation (MECCA; Slater *et al.*, 2012b); a method for including fossil information as priors on node states (Slater *et al.*, 2012a); and a posterior predictive simulation approach for assessing the adequacy of common models of trait evolution (Slater and Pennell, 2013). These types of approaches, which allow for greater complexity both in models and data, will be essential to making robust evolutionary inferences from large comparative datasets.

2.3 Inferring shifts in the rate of lineage diversification

Alfaro *et al.* (2009) developed an approach, MEDUSA, to detect shifts in diversification rates from molecular phylogenies using a stepwise-AIC algorithm. A single-rate birth-death model is fit to the entire tree, then the tree is partitioned into two rate classes, breaking the tree at all possible nodes. The partition which improves the fit of the model is then fixed and the process is repeated, breaking the tree into three partitions, and so on, until a stopping criterion is reached. MEDUSA can be applied to both fully bifurcating and unresolved trees.

For this release of `geiger`, the MEDUSA algorithm has been improved in a number of ways. First, it has been re-coded so that it is now orders of magnitude faster and scales well to large trees; this version of MEDUSA has already been applied to a phylogeny of all 9,993 extant bird species (Jetz *et al.*, 2012). Second, we used simulations to develop a threshold-AIC value as a stopping criterion (see Appendix 3 in SOM for details) in order to correct for multiple comparisons–the larger the phylogeny, the greater the threshold for accepting a new partition to the model (a similar approach was developed by Thomas and Freckleton (2012) for studying variation in rates of phenotypic evolution). Last, we have developed tools for summarizing MEDUSA analyses across a distribution of trees, such as from a Bayesian posterior or from non-parametric bootstrapping, so that uncertainty in both topology and branch lengths can be accomodated.

3 CONCLUSION

In this note we provide a broad overview of the methods now available in `geiger`. We have not discussed some methods implemented in `geiger` (e.g., ‘Congruification’ for time-scaling large trees; Eastman *et al.*, 2013a) and many of the nuances of the methods described here have been left out. We refer readers to associated publications and the package documentation for more information.

It is an exciting time for macroevolutionary research. We now have access to data sets of unparalleled size and a wide variety of new statistical approaches with which to analyze them. We hope that the software presented here will help researchers address some fundamental and long-standing questions in macroevolution.

ACKNOWLEDGEMENTS

We are grateful to the large community of researchers–both method developers and users–who have been so generous with their time and expertise. We specifically thank Brian O’Meara, Liam Revell, Dan Rabosky, Carl Boettiger, Jeremy Beaulieu, Dan Wegmann and Simon Uribe-Convers who gave numerous helpful insights in the development of this work. We thank David Posada, Daniele Silvestro and two anonymous reviewers for comments on this note.

Funding: MWP was supported by a Bioinformatics and Computational Biology Graduate Fellowship from the University of Idaho and a NSERC PGS-D Fellowship. This work was also supported by NSF (DEB 0919499, 1208912 and 0918748).

REFERENCES

Alfaro, M. E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D. L., Carnevale, G. and Harmon, L. J. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences*, **106**, 13410–13414.

Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L. and Harmon, L. J. (2011) A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution*, **65**, 3578–3589.

Eastman, J. M., Harmon, L. J. and Tank, D. C. (2013a) Congruification: support for time scaling large phylogenetic trees. *Methods in Ecology and Evolution*, **4**, 688–691.

Eastman, J. M., Wegmann, D., Leuenberger, C. and Harmon, L. J. (2013b) Simpsonian ‘evolution by jumps’ in an adaptive radiation of anolis lizards. ArXiv:1305.4216.

Felsenstein, J. (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. *American Journal of Human Genetics*, **25**, 471–492.

FitzJohn, R. G. (2012) Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, **3**, 1084–1092.

FitzJohn, R. G., Maddison, W. P. and Otto, S. P. (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Systematic Biology*, **58**, 595–611.

Freckleton, R. P. (2012) Fast likelihood calculations for comparative analyses. *Methods in Ecology and Evolution*, **3**, 940–947.

Green, P. J. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, **82**, 711–732.

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. and Challenger, W. (2008) Geiger: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

Ho, L. S. T. and Ané, C. (2014) A linear-time algorithm for gaussian and non-gaussian trait evolution models. *Systematic Biology*. DOI:10.1093/sysbio/syu005.

Jetz, W. G., Thomas, G. H., Joy, J. B., Hartmann, K. and Mooers, A. (2012) The global diversity of birds in space and time. *Nature*,

491, 444–448.

Pagel, M. (1994) Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, **255**, 37–45.

Paradis, E., Claude, J. and Strimmer, K. (2004) Ape: Analyses of phylogenetics and evolution in r language. *Bioinformatics*, **20**, 289–290.

Pennell, M. W. and Harmon, L. J. (2013) An integrative view of phylogenetic comparative methods: connections to population genetics, community ecology, and paleobiology. *Annals of the New York Academy of Sciences*, **1289**, 90–105.

Slater, G. J., Harmon, L. J. and Alfaro, M. E. (2012a) Integrating fossils with molecular phylogenies improves inference of trait evolution. *Evolution*, **66**, 3931–3944.

Slater, G. J., Harmon, L. J., Wegmann, D., Joyce, P., Revell, L. J. and Alfaro, M. E. (2012b) Fitting models of continous trait evolution to incompletely sampled comparative data using approximate bayesian computation. *Evolution*, **66**, 752–762.

Slater, G. J. and Pennell, M. W. (2013) Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Systematic Biology*. DOI:10.1093/sysbio/syt066.

Thomas, G. H. and Freckleton, R. P. (2012) Motmot: models of trait macroevolution on trees. *Methods in Ecology and Evolution*, **3**, 145–151.