

Phylogenetics

GEIGER: investigating evolutionary radiations

Luke J. Harmon^{1,2,*}, Jason T. Weir³, Chad D. Brock⁴, Richard E. Glor⁵
and Wendell Challenger⁶

¹Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA, ²Biodiversity Research Centre, ³Department of Zoology, University of British Columbia, Vancouver, BC, V6T1Z4, Canada ⁴IGERT Program in Evolutionary Modeling, Washington State University, Pullman, WA 99163, ⁵Department of Biology, University of Rochester, Rochester, NY 14627, USA and ⁶Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, BC, V5A1S6, Canada

Received on May 31, 2007; revised on September 18, 2007; accepted on October 21, 2007

Advance Access publication November 15, 2007

Associate Editor: Keith Crandall

ABSTRACT

Summary: GEIGER is a new software package, written in the R language, to describe evolutionary radiations. GEIGER can carry out simulations, parameter estimation and statistical hypothesis testing. Additionally, GEIGER's simulation algorithms can be used to analyze the statistical power of comparative approaches.

Availability: This open source software is written entirely in the R language and is freely available through the Comprehensive R Archive Network (CRAN) at <http://cran.r-project.org/>.

Contact: lukeh@uidaho.edu

1 INTRODUCTION

Phylogenetic trees are now available for a wide range of groups. Each of these trees contains a large amount of information about the history of diversification in the group of interest. Many tools have been developed to extract this information (Freckleton and Harvey, 2006; Nee, 2001; O'Meara *et al.*, 2006). However, the required software is diffuse, working on different platforms and often requiring unique data formats.

Here, we describe a new software package that is dedicated to the analysis of phylogenetic comparative data. We call this package GEIGER, since its main purpose is to detect and describe evolutionary radiations. GEIGER is written in the cross-platform software language R and complements four other packages: APE (Paradis *et al.*, 2004), apTreeshape (Bortolussi *et al.*, 2006), OUCH (Butler and King, 2004) and LASER (Rabosky, 2006). In addition to permitting a range of tests that are not implemented by any of these packages, GEIGER's simulation algorithms can be used to analyze the statistical power of comparative approaches implemented in other packages. Below, we describe the features of the software and illustrate a range of questions that it might be used to address. Full descriptions of all functions are found in GEIGER's online help files (<http://cran.r-project.org/src/contrib/Descriptions/geiger.html>).

*To whom correspondence should be addressed.

2 DESCRIPTION

2.1 Simulation capabilities

GEIGER can simulate both phylogenetic trees and phenotypic characters. Trees are simulated under a general birth–death model in which all lineages share a fixed probability of speciating or going extinct per unit time. When a simulation is complete the resulting phylogenetic tree can be displayed using R's built-in plotting functions (Fig. 1A), subjected to additional analyses in the R framework, exported in variety of formats (e.g. NEXUS, PHYLIP), or saved as a graphics file.

GEIGER also permits several types of tree pruning that are central to phylogenetic comparative analyses. To facilitate comparison between simulated trees (which include both extinct and extant taxa) and actual phylogenies (which typically include only extant taxa), GEIGER permits the pruning of extinct lineages. To generate trees that mimic the incomplete sampling that characterizes most published phylogenies, GEIGER also includes a function to randomly prune taxa from trees. Using tree simulation and random pruning, one can investigate the effects of incomplete sampling on a parameter of interest. This is important because incomplete sampling has a non-random effect on the distribution of branch lengths in a phylogenetic tree and can therefore bias some types of analyses. Consider the gamma statistic, which is used to detect speed-ups or slow-downs in the rate of cladogenesis (Pybus and Harvey, 2000). Because older lineages are less likely to be excluded than younger lineages when sampling is incomplete, the gamma value resulting from an analysis that includes only some of the known taxa in a group is expected to be biased toward the observation of a slow-down in diversification rate over time (Pybus and Harvey, 2000). To compensate for this bias, GEIGER implements a simulation procedure originally suggested by Pybus and Harvey (2000). To conduct this test, a null distribution for the gamma statistic under incomplete sampling is produced by growing a set of pure-birth phylogenies to a size equal to the known species diversity of the group, pruning out taxa randomly from each tree corresponding to the number of missing taxa in the data set, then calculating the gamma statistic (MCCR test, Pybus and Harvey, 2000).

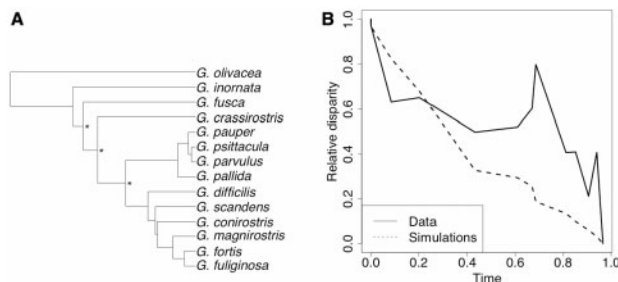


Fig. 1. (A) Relative cladogenesis test applied to *Geospiza* finches; asterisks denote clades that are significantly more diverse than expected. (B) Disparity-through-time plot (Harmon *et al.*, 2003) for morphological measurements of the same set of species. Data courtesy of D. Schluter.

GEIGER also allows simulation of both discrete and continuous character evolution on phylogenetic trees. For a discrete character, one must specify the character state at the root of the tree and a transition matrix that includes instantaneous rates of change among all possible states. The character can have any number of states, and the rate matrix need not be symmetrical. One can also simulate two or more correlated discrete traits by specifying a transition matrix that describes the transitions among all possible combinations of characters (Pagel, 1999a).

To simulate n continuous characters, the program requires an $n \times n$ evolutionary variance-covariance matrix. This matrix describes the expected variances and covariances among characters per unit time; characters evolve under a multivariate Brownian motion model described by this matrix. Although numerous programs are available for the simulation of evolution via Brownian motion, GEIGER permits these types of simulations to be combined with tree simulations. For example, using GEIGER, one can grow birth-death trees and simulate the evolution of a set of characters on those trees.

2.2 Parameter estimation

GEIGER allows estimation of several key parameters associated with species diversification and character evolution. We focus on estimates of average rates of net diversification (speciation-extinction). There are currently three approaches to estimating this rate. One is based on waiting times between successive speciation events and is already implemented via the APE package; this method requires a complete phylogenetic tree with accurate ultrametric branch lengths. The two types of rate estimates provided by GEIGER require less certainty about the phylogenetic tree, but they are also less powerful and may result in estimates with larger sampling error. One estimate relies strictly on a clade's age and species diversity (Magallón and Sanderson, 2001). The second uses the sum of all branch lengths in an ultrametric tree to obtain the Kendall-Moran estimate of diversification rate (Nee, 2001).

To estimate parameters associated with the evolution of multivariate continuous characters, we have implemented a new approach described in Revell *et al.* (2007) that obtains an unbiased estimate of the multivariate variance-covariance

matrix for traits evolving under multivariate Brownian motion. For discrete characters, we provide a function to find the maximum-likelihood transition rate (q) for multistate characters with equal and symmetrical transition rates among character states. This function implements Felsenstein's pruning algorithm (Felsenstein, 1981) in R, and provides a framework for future work using more complex models.

2.3 Hypothesis testing

Although numerous programs are available for the simulation of evolution, GEIGER provides a unified framework within which to implement several new types of tests as well as previously available tests that are also possible in other independent programs. First, GEIGER may be used to test whether particular clades are extraordinarily diverse or undiverse by calculating the probability of obtaining a number of extant species larger or smaller than the realized number given some specified rate and interval of time. For this calculation, we have corrected a copy-proof error in Magallón and Sanderson's (2001) equation 11a; the corrected version is available from the corresponding author.

GEIGER can also calculate the relative cladogenesis statistic, which examines the distribution of numbers of descendants for each branch in a phylogenetic tree existing at some specified time. Under a homogeneous model, the number of descendants should follow a known distribution (Schluter, 2000); lineages with more, or fewer, descendants than expected under this distribution may be hypothesized as lineages that were exceptionally successful or unsuccessful. This test, applied to the Galapagos finches, is shown in Figure 1A.

For character evolution, GEIGER provides a suite of functions designed to investigate the tempo and mode of evolutionary change on a phylogenetic tree. First, GEIGER includes functions to carry out tests of various likelihood models for both continuous and discrete data (Pagel, 1999b). These tests compare the fit of a constant rate Brownian model to various alternatives, including concentrations of character change early or late in the tree or at speciation events, and constrained models of evolution (Butler and King, 2004). Importantly, one can directly compare the likelihood fit of all of these models in an AIC framework. Second, GEIGER implements the approach for comparing levels of phenotypic disparity within and among clades that is described in Harmon *et al.* (2003). This approach provides a running average disparity for clades of a given age range through the history of the tree, and compares that to the expectation under a null model of Brownian motion (illustrated for the Galapagos finches in Fig. 1B). Finally, GEIGER can carry out phylogenetic ANOVA or MANOVA using simulation (Garland *et al.*, 1993).

3 CONCLUSION

As the availability of phylogenetic trees increases, there is an increasing need to develop flexible software packages that permit a variety of comparative analyses. GEIGER expands existing phylogenetic and comparative utilities via the freely distributed, cross-platform R framework. Furthermore,

extensions to our software are facilitated by the flexibility of the R language.

ACKNOWLEDGEMENTS

We thank S. Rogers, A. Mooers, D. Schluter and W. Maddison for helpful comments. L.J.H. was funded by the University of British Columbia Biodiversity Research Centre. C.D.B.'s work was supported by an NSF IGERT Fellowship (NSF BCS-0549425).

Conflict of Interest: none declared.

REFERENCES

- Bortolussi, N. *et al.* (2006) apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics*, **22**, 363–364.
- Butler, M.A. and King, A.A. (2004) Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am. Nat.*, **164**, 683–695.
- Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, **17**, 368–376.
- Freckleton, R.P. and Harvey, P.H. (2006) Detecting non-Brownian trait evolution in adaptive radiations. *PLoS Biol.*, **4**, 2104–2111.
- Garland, T. *et al.* (1993) Phylogenetic analysis of covariance by computer simulation. *Syst. Biol.*, **42**, 265–292.
- Harmon, L.J. *et al.* (2003) Tempo and mode of evolutionary radiation in iguanian lizards. *Science*, **301**, 961–964.
- Magallón, S. and Sanderson, M.J. (2001) Absolute diversification rates in angiosperm clades. *Evolution*, **55**, 1762–1780.
- Nee, S. (2001) Inferring speciation rates from phylogenies. *Evolution*, **55**, 661–668.
- O'Meara, B.C. *et al.* (2006) Testing for different rates of continuous trait evolution. *Evolution*, **60**, 922–933.
- Pagel, M. (1999a) The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.*, **48**, 612–622.
- Pagel, M. (1999b) Inferring the historical patterns of biological evolution. *Nature*, **401**, 877–884.
- Paradis, E. *et al.* (2004) APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20**, 289–290.
- Pybus, O.G. and Harvey, P.H. (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proc. R. Soc. Lond. Ser. B*, **267**, 2267–2272.
- Rabosky, D.L. (2006) LASER: a maximum likelihood toolkit for detecting temporal shifts in diversification rates from molecular phylogenies. *Evol. Bioinform. Online*, **2006**, 257–260.
- Revell, L.J. *et al.* (2007) A phylogenetic approach to determining the importance of constraint on phenotypic evolution in the neotropical lizard *Anolis cristatellus*. *Evol. Ecol. Res.*, **9**, 261–282.
- Schluter, D. (2000) *The Ecology of Adaptive Radiations*. Oxford University Press, Oxford.