

## APPLICATION

# PHYLOMETRICS: an R package for detecting macroevolutionary patterns, using phylogenetic metrics and backward tree simulation

Xia Hua\* and Lindell Bromham

Division of Evolution, Ecology and Genetics, Research School of Biology, Australian National University, Canberra ACT 0200, Australia

## Summary

1. Patterns of trait distribution on phylogenies bear the imprint of the macroevolutionary mechanisms that have shaped them. Phylogenetic metrics can be used to summarize these patterns and potentially could be used to uncover the macroevolutionary mechanisms. However, their utility has been limited by the difficulty in constructing proper null distributions of phylogenetic metrics.
2. Here, we present an efficient algorithm to construct the null distributions by generating phylogenies under a trait state-dependent speciation and extinction model, with a fixed number of extant taxa with a given trait state and a fixed sampling fraction of taxa with each trait state.
3. We also provide a pipeline for estimating the false-positive rate and the statistical power of tests on phylogenetic metrics. These are implemented into an R package called PHYLOMETRICS.
4. PHYLOMETRICS also provides a tractable means of assessing model inadequacy in model-fitting approaches that have been widely used to test hypotheses about trait state-dependent diversification processes.

**Key-words:** binary state speciation–extinction, diversification, phylogenetic metrics, R language, trait evolution, tree simulation

## Introduction

Patterns of trait distribution on phylogenies have the potential to reveal macroevolutionary mechanisms (Jablonski 2008; Rabosky & McCune 2010). For example, a trait that is widely dispersed on a phylogeny and mapped to many tips but few internal edges has been interpreted as a sign of ‘negative lineage selection’, where a trait reduces the likelihood that a lineage with that trait will persist or speciate (e.g. Agnarsson *et al.* 2006; Tripp & Manos 2008; Schwander & Crespi 2009; Anacker *et al.* 2011; Bromham 2014). There are two general approaches to inferring macroevolutionary mechanisms from trait distribution: model-fitting and phylogenetic metrics.

Model-fitting approaches allow researchers to test a range of models of trait-dependent diversification and ask which provides the best fit to the observed phylogenetic distribution of traits, evaluated within a likelihood framework. (e.g. Maddison, Midford & Otto 2007; Freckleton, Phillimore & Pagel 2008; Paradis 2008). Much of the recent focus has been directed at derivatives of the binary state speciation–extinction model (BiSSE: Maddison, Midford & Otto 2007), which promises a way of untangling the trait influence on diversification. However, in common with all model-fitting approaches, a comparison of likelihoods does not in itself evaluate the overall

goodness-of-fit of the model to the data. The best-fit model is not necessarily a good description of the underlying processes that produced the observed data. Therefore, model-fitting approaches should ideally be used with tests of model adequacy that have the potential to reject some or all models as inadequate explanations of the data (Bromham, Hua & Cardillo 2016).

Metric-based approaches use particular measures of the phylogenetic distribution of traits to distinguish a significant macroevolutionary pattern by comparing the measures to proper null distributions (Schwander & Crespi 2009; Fritz & Purvis 2010). Compared to model-fitting approaches, metric-based methods have been limited in the extent to which they can reveal macroevolutionary mechanisms for two reasons. First, it is hard to construct a null distribution of a metric, particularly for metrics sensitive to trait prevalence, as no efficient algorithm is currently available to generate a large number of phylogenies of a fixed number of tips under trait-dependent speciation and extinction (Stadler 2011). Secondly, different macroevolutionary mechanisms could lead to similar patterns of trait distribution on phylogenies, so we need to estimate the false-positive rate and the statistical power of a test on a metric to detect macroevolutionary processes under alternative macroevolutionary models. Despite these limitations, metric-based approaches have some advantages over model-fitting approaches. Metric-based approaches reject models by

\*Correspondence author. E-mail: xia.hua@anu.edu.au

absolute criteria, so they have the potential to test for model adequacy. The approaches also require relatively short calculation time and have the ability to compare metric values across many different phylogenies, allowing the evaluation of the generality of some macroevolutionary mechanisms.

In this application note, we describe the R package **PHYLOMETRICS**. **PHYLOMETRICS** aims to provide (i) an efficient algorithm to simulate phylogenies of a fixed number of extant taxa under trait-dependent speciation and extinction; (ii) a pipeline for estimating the false-positive rate and the statistical power of tests on phylogenetic metrics; (iii) a diagnostic test to assess model inadequacy of model-fitting approaches.

### The **PHYLOMETRICS** package

The R package **PHYLOMETRICS** includes two main functions: ‘treesim’ and ‘treestat’. Function `treesim` generates phylogenetic trees with a binary trait given a fixed number of extant taxa and a fixed sampling fraction of taxa with each state. The simulation uses binary state birth–death model (Maddison, Midford & Otto 2007), defined by six parameters (trait state is coded as 0 and 1): speciation rate for state 0 ( $\lambda_0$ ), speciation rate for state 1 ( $\lambda_1$ ), extinction rate for state 0 ( $\mu_0$ ), extinction rate for state 1 ( $\mu_1$ ), transition rate from state 0 to state 1 ( $q_{01}$ ) and transition rate from state 1 to state 0 ( $q_{10}$ ). Four additional input values are the number of sampled extant taxa of each state and the sampling fraction of each state. The simulation algorithm is described in the Appendix S1 (Supporting information).

Function `treestat` conducts a significance test on a phylogenetic metric given two options for null models of trait evolution: a trait distributed randomly across tips or a trait evolving under the threshold Brownian motion developed by Felsenstein (2005). The `treestat` function reads trees generated by the `treesim` function. To apply the function on an empirical phylogeny, users can use the package **APE** (Paradis, Claude & Strimmer 2004) for loading phylogenetic trees. For trees loaded by **ape**, trait state of each tip on a tree is read as a list of tip labels with state 1, or as a vector of state 0 and 1 in the same order as the `tip.labels` of the `tree` class. The phylogenetic metric is loaded from a function. The function can be any of the metric functions included in the package or any metric that users can write as a function, with the state vector and the `tree` class as the first two inputs. We illustrate how to write a metric function and use `treestat` with the function in the Appendix S1, using the example of the gamma statistic that is used to detect changes in the diversification rate over time (Fig. S1 in Appendix S1).

**PHYLOMETRICS** currently includes four metrics that were tested for their ability to detect traits that arise often but then cause their own disappearance, due to either a raised lineage extinction rate or a high trait loss rate (see Bromham, Hua & Cardillo 2016):

1. Tip age rank sum (TARS) tests whether the tips with the trait of interest (state 1) tend to be shorter or longer than those without (state 0), using the Wilcoxon rank-sum test.

2. Number of tips per origin (NoTO) tests whether the minimum number of inferred origins required to explain the pattern of trait distribution is significantly different from that expected under a null model of trait evolution.
3. Sum of sister clade differences (SSCD) tests whether the trait of interest is more or less clustered on a phylogeny than expected under a null model of trait evolution (Fritz & Purvis 2010).
4. Fritz and Purvis D statistic (FPD) calculates the difference between observed SSCD and expected SSCD under Brownian motion, scaled by the difference between SSCD under random distributions of the trait across the tips of the phylogeny and SSCD under Brownian motion (Fritz & Purvis 2010).

### Estimating false-positive rate and statistical power

To use **PHYLOMETRICS** to estimate the ability of a test on a phylogenetic metric to detect a macroevolutionary process of interest, the first step is to design the test and construct the null and alternative macroevolutionary models. For example, if a trait is gained often but then lost often by extinction or trait reversal, its distribution on the phylogeny should be characterized by many independent evolutionary origins, each giving rise to relatively few descendants with the trait, and thus, the number of tips per reconstructed origins (NoTO) of the trait should be lower than expected under the null model where the trait does not have a raised extinction or loss rate. So we can design a test based on NoTO by simulating 1000 sets of trait distributions on the phylogeny under a threshold Brownian motion model and applying a one-tailed test, asking whether the value of NoTO for the trait is lower than over 95% the simulated trait distributions. We define the null model as a ‘Baseline’ model where the trait does not affect speciation and extinction rates, and trait gain and loss rates are equal. We considered three alternatives models: ‘Labile’, where the trait is frequently gained and lost; ‘Dead-End’, where the trait increases the extinction rate; and ‘Suicide’, where trait is gained frequently and causes a raised extinction rate.

The next step is to set the parameter values of the null and alternative macroevolutionary models. Here, we set parameter values arbitrarily for each model.

```

pars_baseline <- c(0.1,0.1,0.01,0.01,0.01,0.01)
pars_labile <- c(0.1,0.1,0.01,0.01,0.1,0.1)
pars_deadend <- c(0.1,0.1,0.01,0.1,0.01,0.01)
pars_suicide <- c(0.1,0.1,0.01,0.1,0.3,0.01)

```

The third step is to simulate phylogenetic trees under the null and alternative macroevolutionary models using function `treesim`. Consider an example with 100 extant taxa with trait and 200 extant taxa without trait, where 60% of the taxa with trait were included in the phylogeny and 50% of the taxa without trait were included. We want to simulate 100 trees under each model.

```

tree_baseline <- replicate(100, treesim (pars_baseline, 200*
0.5, 100*0.6, c(0.5, 0.6)), simplify = F)

```

```

tree_labile <- replicate(100, treesim (pars_labile, 200*0.5,
100*0.6, c(0.5, 0.6)), simplify = F)
tree_deadend <- replicate(100, treesim (pars_deadend, 200*
0.5, 100*0.6, c(0.5, 0.6)), simplify = F)
tree_suicide <- replicate(100, treesim (pars_suicide, 200*
0.5, 100*0.6, c(0.5, 0.6)), simplify = F)

```

The final step is to run the test on the metric for each simulated tree under each macroevolutionary model using function treestat. The function allows users to specify the null model of trait evolution, the number of traits simulated under the null model and the direction of the alternative hypothesis. In this example, we simulate, for each tree, 1000 sets of trait distributions using the threshold Brownian motion model ('TBM') and apply a one-tailed test on NoTO, asking whether the number of tips per reconstructed origin is lower than expected.

```

stat_baseline <- lapply(tree_baseline, treestat, func = noto,
traitevol = "TBM", a = 1000, alternative = "less")
stat_labile <- lapply(tree_labile, treestat, func = noto,
traitevol = "TBM", a = 1000, alternative = "less")
stat_deadend <- lapply(tree_deadend, treestat, func =
noto, traitevol = "TBM", a = 1000, alternative = "less")
stat_suicide <- lapply(tree_suicide, treestat, func = noto,
traitevol = "TBM", a = 1000, alternative = "less")

```

The output is a list of *P*-values for each macroevolutionary model. The test on NoTO is significant when the *P*-value is less than the user-defined significance level, for example 0.05. For trees simulated under the Baseline (null) model, the proportion of significant results is the false-positive rate (the rate at which the test on NoTO rejects the null model when it is true). For trees simulated under each alternative model, the proportion of significant results is the power of test on NoTO to reject the null model when the alternative model is true. This statistical power is different from the power of NoTO to distinguish one alternative model from another. But it can suggest the ability to distinguish different alternative models if the power is high under one model and low under the others.

## Assessing model inadequacy

Inferring evolutionary models using model-fitting approaches such as the BiSSE method (Maddison, Midford & Otto 2007) can have enhanced false-positive rate due to phylogenetic pseudoreplication and violations of model assumptions (Maddison & FitzJohn 2015; Rabosky & Goldberg 2015). Since likelihood ratios do not provide absolute criteria to reject a model, we need an additional diagnostics to assess model inadequacy (Rabosky & Goldberg 2015). Because phylogenetic metrics summarize aspects of tree topology and trait distribution, they can be used to test whether the observed tree could have been generated by a given model, because the trait distribution on the observed tree should fall within the range generated by simulations under that model. Therefore, a significant departure of the observed metric value from the distribution of simulated values leads to the rejection of the macroevolutionary model as being adequate to explain patterns in trait distribution in the observed data.

The assessment of model adequacy starts with tree simulation using treesim under the macroevolutionary model suggested by the model-fitting approach. The value of the metric of trait distribution on each simulated tree and the observed tree is then calculated by treestat. Here, we use treestat to test whether the observed trait distribution (state<sub>obs</sub>) on the observed phylogeny (tree<sub>obs</sub>) has significantly higher or lower NoTO than trees (tree<sub>sim</sub>) simulated under the model selected by the model-fitting approach.

```

stat_obs <- treestat(tree_obs, state = stateobs, func = noto)
stat_sim <- lapply(tree_sim, treestat, func = noto)
p.value <- min(sum(stat_sim <= stat_obs), sum(stat_sim
>= stat_obs))/length(stat_sim)

```

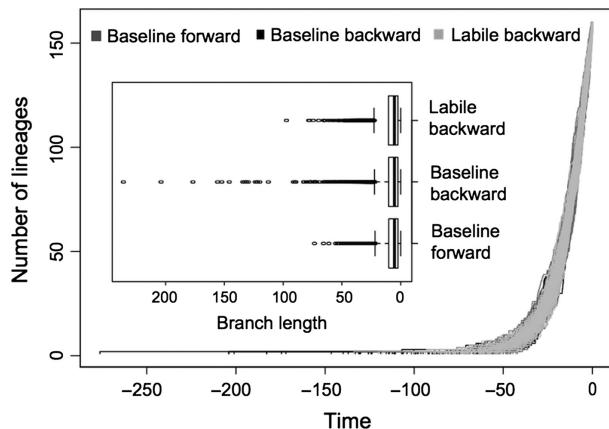
If the *p*.value is less than 0.05, we can conclude that the observed NoTO value falls outside 95% the distribution of NoTO values generated under that model, so the model is unlikely to have produced the observed metric value. We therefore reject that model as being adequate at significant level 0.05.

While we can use this test to reject a model as being adequate, it is important to know that failure to reject a model does not imply that the model is a fair description of the data. Our tree simulation tool is conditioned on the trait prevalence in the observed data set, so it is restricted to exploring the outcome of models given a fixed trait prevalence, and cannot be used to ask how likely that particular pattern of trait prevalence is under the macroevolutionary model being tested. If the test fails to reject the model as being adequate, then other approaches, such as forward tree simulation (FitzJohn 2012) or BiSSE likelihood functions (Maddison, Midford & Otto 2007), are needed to assess how often we expect the macroevolutionary model to generate trees with the observed trait prevalence.

Alternatively, our backward tree simulation allows efficient implementation of an approximated Bayesian computation approach (Marjoram *et al.* 2003). As suggested by Stadler (2011), trees can be simulated under a proposed set of parameters, and then different metrics can be used to decide whether the simulated trees match the observed data in various ways. If so, those parameter values are added to the posterior distribution. A new set of parameters is then proposed and the procedure is iterated until a desirable sample size is reached. To save computation time, the prior distribution of parameters can be defined by the posterior distribution generated by model-fitting approaches, such as the BiSSE method (Maddison, Midford & Otto 2007).

## Strengths and alternatives

The package PHYLOMETRICS provides the first method for simulating phylogenies under trait-dependent speciation and extinction in a backward fashion. This backward simulation assures that all the simulated phylogenies have the same number of extant taxa with the same trait prevalence and tip sampling fractions as the observed data. These properties are essential to construct an informative null distribution for metrics whose values are sensitive to trait prevalence. Metrics describing tip length (e.g. TARS) or distribution of inferred

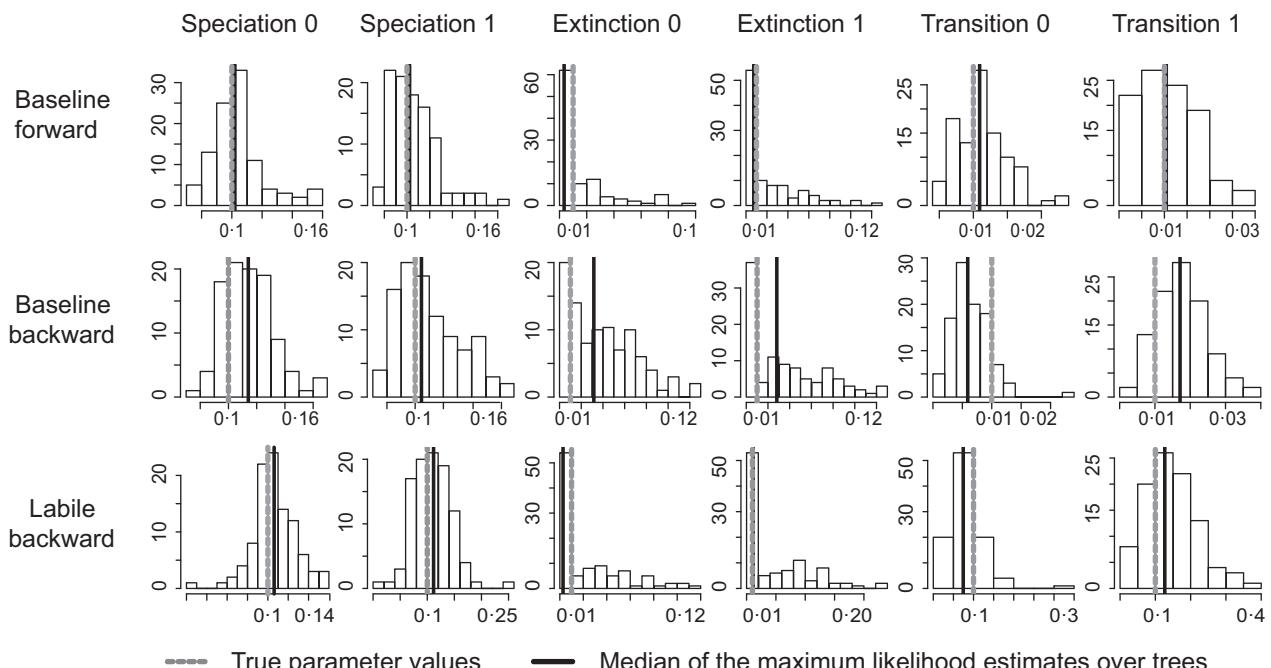


**Fig. 1.** Comparisons of trees generated under backward simulation and forward simulation. Lineage-through-time plots of trees simulated under the Baseline model using forward simulation (dark grey) and using backward simulation (Black), and under the Labile model using backward simulation (light grey), together with boxplots of branch lengths in all the trees simulated under each scenario, show differences in basal branch lengths between the two simulation tools and between the two models.

trait origins (e.g. NoTO) belong to this kind of metrics. Metrics such as these have been used to describe macroevolutionary patterns, but their utility has been limited by the difficulty in generating an appropriate null distribution. An alternative way to simulate phylogenies with the same number of taxa and trait prevalence is to simulate a large number of phylogenies and discard phylogenies that do not meet the

criteria. But the method is time-consuming and becomes impractical when there are many different patterns of trait prevalence that are likely outcomes of the macroevolutionary model being tested. For example, our backward simulation takes about 0.06 s to generate a tree under the Baseline model and 0.10 s under the Labile model, while the forward tree simulation (using ‘simulate’ function in ‘diversitree’ R package; FitzJohn 2012) takes 12.81 s to generate a tree under the Baseline model and fails to generate a tree under the Labile model after several days.

A shortcoming of our backward simulation is that it generates more rare trees than the forward simulation. For example, if backward simulation leads to two basal branches with different trait states, the algorithm continues until coalescence, no matter how rare it could be if trait transition rate is much lower than speciation rate (e.g. the Baseline model). In contrast, forward simulation starts with the root, so it does not constrain the state of basal branches. As a result, the backward simulation produces longer basal branches waiting for rare events to occur than the forward simulation under the Baseline model (Fig. 1), which causes overestimations in extinction rates and biased estimations in trait transition rates (Fig. 2). These biases reduce under the Labile model (Fig. 2), since trait lability makes different states of basal branches equally likely and so the backward simulation does not produce long basal branches (Fig. 1). One way to prevent backward simulation from generating rare trees is to discard trees that do not coalesce for a certain amount of time. This can be done by setting a value to ‘max.t’ in the treesim function. It would be preferable to use a likelihood value as the criterion for acceptance probability of a



**Fig. 2.** Histograms of the maximum likelihood estimates for the six parameters in the Baseline model and the Labile model from trees simulated using backward simulation and forward simulation. The true parameter values are marked by grey dashed lines. The medians of the maximum likelihood estimates are marked by black solid lines. Maximum likelihood estimates are derived from the BiSSE likelihood function (Maddison, Midford & Otto 2007) using ‘diversitree’ (FitzJohn 2012).

tree, but there is no closed-form expression for the state-dependent birth–death processes, making the simulation inefficient.

## Acknowledgements

The authors would like to thank Marcel Cardillo, Emma Day and Camile Moray for providing valuable suggestions by using PHYLOMETRICS on their studies.

## Data accessibility

This manuscript does not use any data. Documentation and source code are freely available on CRAN (<http://cran.r-project.org/web/packages/phylometrics>).

## References

Agnarsson, I., Avilés, L., Coddington, J.A. & Maddison, W.P. (2006) Sociality in Theridiid spiders: repeated origins of an evolutionary dead end. *Evolution*, **60**, 2342–2351.

Anacker, B.L., Whittall, J.B., Goldberg, E.E. & Harrison, S.P. (2011) Origins and consequences of serpentine endemism in the California flora. *Evolution*, **65**, 365–376.

Bromham, L. (2014) Macroevolutionary patterns of salt tolerance in angiosperms. *Annals of Botany*, **115**, 333–341.

Bromham, L., Hua, X. & Cardillo, M. (2016) Detecting macroevolutionary self-destruction from phylogenies. *Systematic Biology*, **65**, 109–127.

Felsenstein, J. (2005) Using the quantitative genetic threshold model for inferences between and within species. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **360**, 1427–1434.

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

Freckleton, R.P., Phillimore, A.B. & Pagel, M. (2008) Relating traits to diversification: a simple test. *The American Naturalist*, **172**, 102–115.

Fritz, S.A. & Purvis, A. (2010) Selectivity in mammalian extinction risk and threat types: a new measure of phylogenetic signal strength in binary traits. *Conservation Biology*, **24**, 1042–1051.

Jablonski, D. (2008) Species selection: theory and data. *Annual Review of Ecology, Evolution, and Systematics*, **39**, 501–524.

Maddison, W.P. & FitzJohn, R.G. (2015) The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology*, **64**, 127–136.

Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003) Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 15324.

Paradis, E. (2008) Asymmetries in phylogenetic diversification and character change can be untangled. *Evolution*, **62**, 241–247.

Paradis, E., Claude, J. & Strimmer, K. (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Rabosky, D.L. & Goldberg, E.E. (2015) Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, **64**, 340–355.

Rabosky, D.L. & McCune, A.R. (2010) Reinventing species selection with molecular phylogenies. *Trends in Ecology & Evolution*, **25**, 68–74.

Schwander, T. & Crespi, B.J. (2009) Twigs on the tree of life? Neutral and selective models for integrating macroevolutionary patterns with microevolutionary processes in the analysis of asexuality. *Molecular Ecology*, **18**, 28–42.

Stadler, T. (2011) Simulating trees with a fixed number of extant species. *Systematic Biology*, **60**, 676–684.

Tripp, E.A. & Manos, P.S. (2008) Is floral specialization an evolutionary dead-end? Pollination system transitions in *Ruellia* (Acanthaceae). *Evolution*, **62**, 1712–1737.

Received 17 August 2015; accepted 12 December 2015

Handling Editor: Robert Freckleton

## Supporting Information

Additional Supporting Information may be found in the online version of this article.

**Appendix S1.** Tree simulation algorithm and an example of working with user-defined metrics.