*Bioinformatics*, 32(5), 2016, 789–791
doi: 10.1093/bioinformatics/btv651
Advance Access Publication Date: 4 November 2015
Applications Note

Phylogenetics

# TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates

**Sebastian Höhna[1,2,3,4,*], Michael R. May[3] and Brian R. Moore[3]**

[1]Department of Integrative Biology, [2]Department of Statistics, University of California, Berkeley, CA 94720, USA, [3]Department of Evolution and Ecology, University of California, Davis, CA 95616, USA and [4]Department of Mathematics, Stockholm University, Stockholm, SE-106 91, Sweden

*To whom correspondence should be addressed.

Associate Editor: David Posada

Received on June 19, 2015; revised on October 29, 2015; accepted on October 30, 2015

## Abstract

**Summary:** Many fundamental questions in evolutionary biology entail estimating rates of lineage diversification (speciation–extinction) that are modeled using birth–death branching processes. We leverage recent advances in branching-process theory to develop a flexible Bayesian framework for specifying diversification models—where rates are constant, vary continuously, or change episodically through time—and implement numerical methods to estimate parameters of these models from molecular phylogenies, even when species sampling is incomplete. We enable both statistical inference and efficient simulation under these models. We also provide robust methods for comparing the relative and absolute fit of competing branching-process models to a given tree, thereby providing rigorous tests of biological hypotheses regarding patterns and processes of lineage diversification.

**Availability and implementation:** The source code for TESS is freely available at http://cran.r-project.org/web/packages/TESS/.

**Contact:** Sebastian.Hoehna@gmail.com

## 1 Introduction

Stochastic-branching process models (e.g. birth–death models) describe the process of diversification that gave rise to a given study tree, and include parameters such as the rate of speciation and extinction. Statistical analysis of lineage diversification rates involves: (i) simulating trees under null and alternative hypotheses; (ii) estimating parameters in, e.g. a Bayesian framework, and; (iii) evaluating the relative and absolute fit of candidate models. These considerations motivated our development of TESS, an R package for statistical inference of lineage diversification rates that allows researchers to address three fundamental questions: (i) *What are the rates of the process that gave rise to my study tree?* (ii) *Have diversification rates changed through time in my study tree?* (iii) *Is there evidence that my study tree experienced mass extinction?* We compare TESS to other software in the accompanying vignette (http://cran.r-project.org/web/packages/TESS/).

## 2 Methods and algorithms

### 2.1 Branching-process models

Inferring rates of lineage diversification is based on the *reconstructed evolutionary process* described by Nee *et al.* (1994); a birth–death process in which only sampled, extant lineages are observed. Our implementation exploits recent theoretical work (Höhna, 2015) that allows the rate of diversification to be specified as an arbitrary function of time (but is constant across lineages at any instant). By adopting this generic approach, we can specify an effectively infinite number of branching-process models in TESS. These possibilities correspond to four main types of diversification models: (i) constant-rate birth–death models; (ii) continuously variable-rate birth–death models; (iii) episodically variable-rate birth–death models, and; (iv) explicit mass-extinction birth–death models.

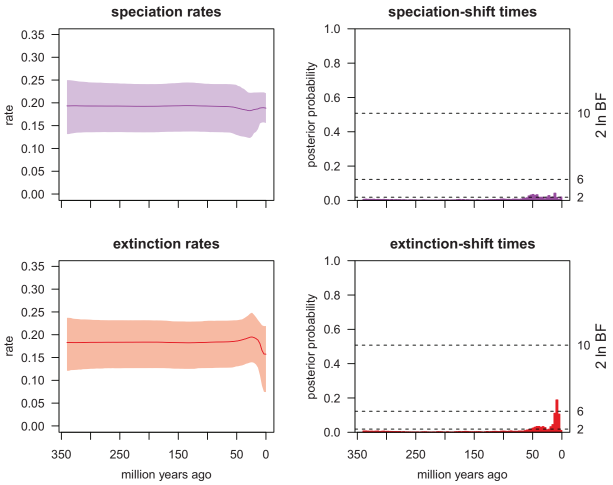

**Fig. 1.** Estimating rates of (and identifying shifts in) lineage diversification through time. Left: Plots of the posterior mean and 95% credible interval for the speciation and extinction rate (upper and lower panels, respectively). Right: Identifying temporal shifts in the speciation and extinction rate (upper and lower panels, respectively) using Bayes factors estimated by rjMCMC. Each bar indicates the posterior probability of at least one rate shift within that interval. Bars that exceed the specified significance threshold (here, 2 ln BF > 6) indicate significant rate shifts. This analysis of the conifer tree from Leslie *et al.* (2012) is provided as an example in the R package and described extensively in the vignette. It reveals a significant shift in the extinction rate ~5 million years ago

## 2.2 Phylogeny simulation

TESS can be used to efficiently simulate phylogenies under any time-dependent birth–death process (Höhna, 2013). Simulation is crucial both for evaluating the performance and statistical behavior of an inference method, and also for evaluating model adequacy (see below).

## 2.3 Phylogenetic data

TESS can be used to estimate parameters of any time-dependent birth–death process from a given tree. Specifically, TESS requires rooted *ultrametric* trees, where all of the tips are sampled at the same time horizon (the present). Additionally, TESS implements various approaches for accommodating incompletely sampled trees, including uniform and diversified species-sampling schemes (Höhna *et al.*, 2011; Höhna, 2014).

## 2.4 Parameter estimation

In TESS, parameters of the branching-process models are inferred in a Bayesian statistical framework. Specifically, we estimate the joint posterior probability density of the model parameters from the study tree using numerical methods; Markov chain Monte Carlo (MCMC) algorithms (Fig. 1). The numerical methods implemented in TESS include adaptive-MCMC algorithms (Haario *et al.*, 1999)—where the scale of the proposal mechanisms is automatically tuned to ensure optimal efficiency (mixing) of the MCMC simulation—and also feature real-time diagnostics to assess convergence of the MCMC simulation to the stationary distribution (the joint posterior probability density of the model parameters). The real-time diagnostics, such as the minimum effective sample size and the Geweke statistic, are used to automatically terminate the MCMC simulation when it has drawn an adequate sample from the posterior.

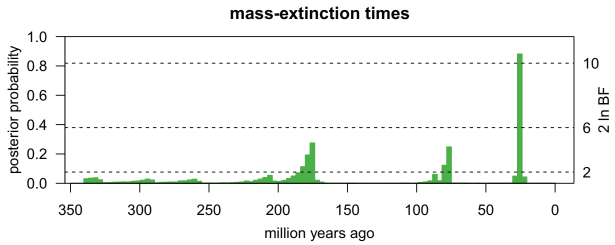

**Fig. 2.** Identifying significant mass-extinction events using rjMCMC. Each bar indicates the posterior probability of at least one mass extinction within that interval. Bars that exceed the specified significance threshold (here, 2 ln BF > 6) are inferred to be significant mass-extinction events. Here we show results for the conifer analysis (Fig. 1), which identifies one significant mass-extinction event that occurred 22 million years ago

## 2.5 Model comparison

Each branching-process model specifies a possible scenario for the diversification process that gave rise to a given study tree. For most studies, several (possibly many) competing branching-process models of varying complexity will be plausible *a priori*. We therefore need a way to objectively identify the best candidate diversification model. Bayesian model selection is based on *Bayes factors* (e.g. Kass and Raftery, 1995). This procedure requires that we first estimate the marginal likelihood of each candidate model, and then compare the ratio of the marginal likelihoods for each pair of candidate models. We have implemented both *stepping-stone sampling* (Xie *et al.*, 2011) and *path-sampling* (Lartillot and Philippe, 2006) algorithms for estimating the marginal likelihoods.

## 2.6 Model adequacy

Bayes factors allow us to assess the *relative* fit of two or more competing branching-process models to a given study tree. However, even the very best candidate model may nevertheless be inadequate in an *absolute* sense. Accordingly, TESS implements methods to assess the absolute fit of a candidate diversification model to a given study tree using *posterior-predictive simulation* (Gelman *et al.*, 1996). The basic premise of this approach is as follows: if the diversification model under consideration provides an adequate description of the process that gave rise to our study tree, then we should be able to use that model to generate new phylogenies that are in some sense ‘similar’ to our study tree. TESS permits use of any test statistic— e.g. the $\gamma$-statistic or any tree-shape statistic to identify rate variation among lineages (Moore *et al.*, 2004)—to measure the similarity between predicted and observed data.

## 2.7 Model averaging

The vast space of possible branching-process models precludes their exhaustive pairwise comparison using Bayes factors. This issue may be addressed by means of model-averaging approaches that treat the model as a random variable (Green, 1995). TESS implements such an approach; the CoMET (CPP on Mass-Extinction Times) model (May *et al.*, 2015), which uses a compound Poisson process (CPP) model and reversible-jump MCMC to average over all possible models (Fig. 2).

# 3 Conclusions

TESS allows users to specify an effectively countless number of diversification models, where each model describes an alternative scenario for the diversification of the tree. TESS can be used to efficiently simulate under and/or infer parameters of these models. Additionally, TESS provides robust methods for assessing the relative fit of competing models to a given tree, providing users with an extremely flexible yet intuitive framework for testing hypotheses regarding the patterns and processes of lineage diversification. TESS is accompanied by a comprehensive vignette that provides detailed explanations of the methods and examples (http://cran.r-project.org/web/packages/TESS/).

*Conflict of Interest*: none declared.

## References

Gelman,A. *et al*. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.*, **6**, 733–807.

Green,P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711.

Haario,H. *et al*. (1999) Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Stat.*, **14**, 375–396.

Höhna,S. (2013) Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes. *Bioinformatics*, **29**, 1367–1374.

Höhna,S. (2014) Likelihood inference of non-constant diversification rates with incomplete taxon sampling. *PLoS One*, **9**, e84184.

Höhna,S. (2015) The time-dependent reconstructed evolutionary process with a key-role for mass-extinction events. *J. Theor. Biol.*, **380**, 321–331.

Höhna,S. *et al*. (2011) Inferring speciation and extinction rates under different species sampling schemes. *Mol. Biol. Evol.*, **28**, 2577–2589.

Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Lartillot,N. and Philippe,H. (2006) Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, **55**, 195.

Leslie,A.B. *et al*. (2012) Hemisphere-scale differences in conifer evolutionary dynamics. *Proc. Natl. Acad. Sci. USA*, **109**, 16217–16221.

May,M.R. *et al*. (2015) A Bayesian approach for detecting mass-extinction events when rates of lineage diversification vary. *Systematic Biology, in press.*

Moore,B.R. *et al*. (2004) Detecting diversification rate variation in supertrees. In: Bininda-Emonds,O.R.P. (ed.) *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pp. 487–533. Kluwer Academic Press, Dordrecht, The Netherlands.

Nee,S. *et al*. (1994) The reconstructed evolutionary process. *Philos. Trans. Biol. Sci.*, **344**, 305–311.

Xie,W. *et al*. (2011) Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, **60**, 150–160.