BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 11 2013, pages 1367–1374
doi:10.1093/bioinformatics/btt153

Phylogenetics

Advance Access publication March 29, 2013

Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes

Sebastian Höhna

Department of Mathematics, Stockholm University, Kräftriket Hus 6, S-10691 Stockholm, Sweden

Associate Editor: David Posada

ABSTRACT

Motivation: Diversification rates and patterns may be inferred from reconstructed phylogenies. Both the time-dependent and the diversity-dependent birth–death process can produce the same observed patterns of diversity over time. To develop and test new models describing the macro-evolutionary process of diversification, generic and fast algorithms to simulate under these models are necessary. Simulations are not only important for testing and developing models but play an influential role in the assessment of model fit.

Results: In the present article, I consider as the model a global time-dependent birth–death process where each species has the same rates but rates may vary over time. For this model, I derive the likelihood of the speciation times from a reconstructed phylogenetic tree and show that each speciation event is independent and identically distributed. This fact can be used to simulate efficiently reconstructed phylogenetic trees when conditioning on the number of species, the time of the process or both. I show the usability of the simulation by approximating the posterior predictive distribution of a birth–death process with decreasing diversification rates applied on a published bird phylogeny (family Cettiidae).

Availability: The methods described in this manuscript are implemented in the R package TESS, available from the repository CRAN (http://cran.r-project.org/web/packages/TESS/).

Contact: hoehna@math.su.se

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 10, 2012; revised on March 24, 2013; accepted on March 27, 2013

1 INTRODUCTION

The birth–death process is arguably the most commonly used process to model species diversity (Nee, 2006; Ricklefs, 2007). Nee *et al.* (1994a) introduced the birth–death process to estimate diversification rates from reconstructed phylogenetic trees. It was then used to estimate speciation and extinction rates that are constant over time and equal for all species (Yang and Rannala, 1997). In recent years, the constant equal-rate birth–death process has been extended to time-dependent speciation and extinction rates (Rabosky, 2006; Rabosky and Lovette, 2008; Alfaro *et al.*, 2009; Morlon *et al.*, 2011); trait-dependent speciation and extinction rates (Maddison *et al.*, 2007; FitzJohn *et al.*, 2009); and diversity-dependent speciation rates (Etienne *et al.*, 2012). The models used for parameter inference have become increasingly complex and different models—such as the time-dependent speciation and the diversity-dependent speciation models—can produce almost indistinguishably close fits to observed phylogenetic trees. Thus, studies investigating the different patterns generated by the different models need to be compared. An evident approach is to use simulations under one model and inference under any of the available models. Furthermore, simulations play an important role in investigations of which parameters can be estimated (Quental and Marshall, 2009; Liow *et al.*, 2010).

However, simulating trees has gained less attention and remains a challenge in several situations (Hartmann *et al.*, 2010). One might be interested in simulating trees under three different conditions when the process starts with a single species at time $t_0 = 0$:

- Simulating a tree for a given time t.
- Simulating a tree for a given number of species n.
- Simulating a tree for a given number of species n that have evolved over time t.

Additionally, the conditions can be modified if the process starts with the most recent common ancestor (MRCA), and thus, there are two species at time $t_0 = 0$ (Stadler, 2011). I consider the time-dependent speciation and extinction rates birth–death process as the model of choice for simulations in this article. This process contains the constant-rate birth–death process as a special case.

Simulations under the first condition, a given time t of the process, are straightforward from a theoretical point of view. One only needs to simulate the exponentially distributed waiting times until the next event, either a speciation event or an extinction event, and pick a random species. Nevertheless, implementing this forward simulation might be challenging and inefficient in situation when the rates are not constant and both rates are high. Additionally, there is no available software for simulating reconstructed trees for a given time t when the rates vary continuously. An overview of available software packages for simulating trees under a time-dependent birth–death process is given in Stadler (2011).

Simulating phylogenies under models with a non-zero extinction rate is challenging when one is interested in simulating a tree with a fixed number n of extant species (Stadler, 2011). The general sampling approach (GSA) introduced by Stadler (2011) in the program *TreeSim*, is the currently only available approach to simulate trees under the first and second condition. However, *TreeSim* only implements changes of speciation and extinction rates at specific predefined times. Furthermore, when one chooses a model with high species turnover (the extinction rate being close to the speciation rate), then the GSA takes a long time to terminate because many speciation and extinction events

are simulated and the process converges only slowly to the terminating species number.

In this article, I will derive the joint probability density of the reconstructed phylogeny using a new approach that shows that each speciation time in the reconstructed tree is independent and identically distributed (*iid*). I will then develop an algorithm that uses the cumulative distribution function for fast simulation of the reconstructed tree under any—continuous or discrete—time-dependent speciation and extinction rate function. I apply this new simulation technique to sample from the posterior predictive distribution of an empirical dataset, the bird family Cettidae.

2 METHODS

The strategy of the simulation algorithm, developed in the following section, is to derive the probability density of a single speciation event independent of all other speciation events in the reconstructed tree. Once the probability density is known, it is (at least numerically) possible to compute the inverse cumulative distribution function and thus samples from the desired distribution. The simulation procedure only needs to draw $n-1$ speciation events for a tree of size n and does not need to draw all speciation events that are pruned because they do not have extant descendants. I will start by defining the birth–death process.

The process starts with a single species at time t_0. Every species has the same speciation and extinction rate at any time t, denoted by $\lambda(t)$ and $\mu(t)$, respectively. The time until a speciation event of a given species is exponentially distribution with rate $\lambda(t)$. At a speciation event, a species gives birth to exactly one new species. The time until an extinction event of a given species is exponentially distributed with rate $\mu(t)$. Furthermore, let $N(t)$ denote the number of species at time t. Following standard notation, I will use t_0 as the time of the origin or start of the process and T as the present time when the process was stopped.

The observation of a tree generated by the birth–death process containing extant as well as extinct species is called a complete tree (Fig. 1a). Removing all extinct branches gives the reconstructed tree (Nee *et al.*, 1994b) (Fig. 1b). From the reconstructed tree with n extant species, I obtain the set of speciation times $\{t\} = \{t_0, \ldots, t_{n-1}\}$. In this article, I consider the reconstructed sampled tree as the data.

Let me begin with stating some known probability densities of the birth–death process, the probability of survival $P(N(T)>0 \mid N(t)=1)$, the probability of exactly one descendant $P(N(T)=1 \mid N(t)=1)$ and the probability of n descendants $P(N(T)=n \mid N(t)=1)$ (Kendall, 1948; Thompson, 1975; Nee *et al.*, 1994b; Morlon *et al.*, 2011). Although I will follow closely Nee *et al.* (1994b), I modified the process to start at time t and stop at time T instead of starting at time 0 and stopping at $T-t$ and thus is a generalization of the original equations in Nee *et al.* (1994b). Finally, I derive the probability density of a single speciation event, following the idea of Thompson (1975), by using the above-mentioned probability densities.

First, I define the rate function r as

$$r(t,s) = \int_t^s (\mu(z) - \lambda(z))dz. \tag{1}$$

The probability of a single species alive at time t surviving until time T $[P(N(T)>0 \mid N(t)=1)]$ is given (Nee *et al.*, 1994b, Equation 24) by

$$P(N(T)>0 \mid N(t)=1) = \left(1 + \int_t^T \left(\mu(s)\exp(r(t,s))\right)ds\right)^{-1}. \tag{2}$$

The probability of n species at time T when starting with a single species at time t is (Nee *et al.*, 1994b, Equation 3)

$$\begin{aligned} P(N(T) &= n \mid N(t) = 1) \\ &= (1 - P(N(T) > 0 \mid N(t) = 1)\exp(r(t,T)))^{n-1} \\ &\quad \times P(N(T) > 0 \mid N(t) = 1)^2 \exp(r(t,T)) \end{aligned} \tag{3}$$

which leads to the probability of obtaining a single lineage at time T

$$P(N(T) = 1 \mid N(t) = 1) = P(N(T) > 0 \mid N(t) = 1)^2 \exp(r(t,T)). \tag{4}$$

Equation (3) can be derived from a geometric distribution with parameter $P(N(T)>0 \mid N(t)=1)\exp(r(t,T))$ multiplied by the probability of survival of the process $P(N(T)>0 \mid N(t)=1)$. Thus, the probability of n species at time T given that we had 1 species at time t $[N(t)=1]$ and the species survives, denoted by $S(1,t,T)$, is

$$\begin{aligned} P(N(T) &= n \mid N(t) = 1, S(1,t,T)) \\ &= (1 - P(N(T)>0 \mid N(t)=1)\exp(r(t,T)))^{n-1} \\ &\quad \times P(N(T)>0 \mid N(t)=1)\exp(r(t,T)). \end{aligned} \tag{5}$$

Now, following the reasoning of Thompson (1975, p. 54–58), I give the joint probability distribution of all speciation times $\{t\}$ of the reconstructed tree given that we sample (or stop) the process at time T

$$\begin{aligned} &P_{BD}(\{t\} \mid N(t_0) = 1, T) \\ &= P(N(T) = 1 \mid N(t_0) = 1) \\ &\quad \times \prod_i^{n-1} \left(i \times \lambda(t_i) \times P(N(T) = 1 \mid N(t_i) = 1)\right) \end{aligned} \tag{6}$$

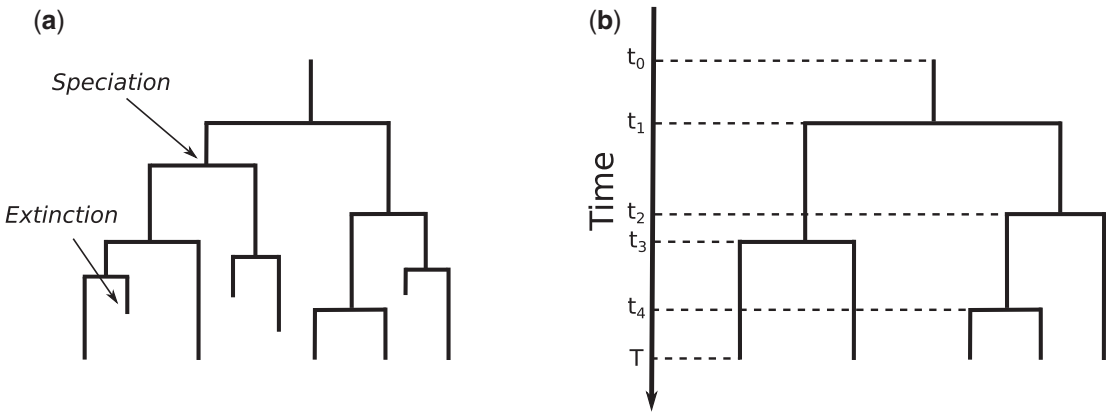

Fig. 1. A simulated birth–death tree starting with a single species. (**a**) The complete tree containing both extant and extinct species. (**b**) The reconstructed tree containing only extant species

The intuitive interpretation of the probability density is that the process starts with one species that survives until the present time $[P(N(T) = 1 \mid N(t_0) = 1)]$. Then, for each speciation event, we multiply with the density to obtain a speciation event at time t_i $[i\lambda(t_i)]$ and the probability that this new species has one sampled descendant at the present $[P(N(T) = 1 \mid N(t_i) = 1)]$. A short proof is given in the Appendix.

Using Equations (3) and (6), we can compute the probability of the conditioned process on n lineages at the present time T results

$$P_{BD}(\{t\} \mid N(t_0) = 1, N(T) = n, T) = \prod_{i}^{N-1} \frac{i \times \lambda(t_i) \times P(N(T) = 1 \mid N(t_i) = 1)}{1 - P(N(T) > 0 \mid N(t_0) = 1)\exp(r(t_0, T))} \quad (7)$$

Note, the factor i in front of $\lambda(t_i)$ is simply a combinatorial constant representing the $(n-1)!$ possible orderings of the speciation events (Aldous and Popovic, 2005). Hence, the probability density of each single speciation event is *iid* by the definition of independence $P_{BD}(t_j \mid \{t\}\backslash\{t_j\}, N(t_0) = 1, N(T) = n, T) = P_{BD}(t_j \mid N(t_0) = 1, N(T) = n, T)$, and have the probability density

$$P_{BD}(t) \propto \lambda(t) P(N(T) = 1 \mid N(t) = 1) \quad (8)$$

This has previously been shown by Yang and Rannala (1997) for constant rates and been exploited by Höhna *et al.* (2011) for estimations of model parameters.

I will now show how to use these probability distributions to simulate reconstructed trees under the three different conditions.

2.1 Simulating reconstructed trees for a fixed time *t* and number of species *n*

I will start developing the algorithm for simulating reconstructed trees when conditioning on both—the time of the process t and the number of sampled extant species n—although this might not be the common or natural condition for simulating reconstructed trees. It is, however, used by the two other simulation conditions. Random speciation events are drawn by solving the inverse cumulative distribution function of the speciation times.

From Equation (7) we have that the joint probability of all speciation times in the reconstructed tree and established that all speciation times are *iid* with the probability density given in Equation (8). Thus, to obtain $n-1$ speciation times (or $n-2$ in the case when starting at the MRCA), one needs to draw $n-1$ random draws from the above distribution. Such random draws can be obtained by first drawing a random number u form the Uniform (0,1) distribution. Then, the speciation time t^* for which the cumulative distribution function of Equation (8) equals u, gives the randomly drawn speciation time:

$$u = \frac{\int_0^{t^*} P_{BD}(t)dt}{\int_0^{T} P_{BD}(t)dt} = \frac{\int_0^{t^*} \lambda(t) P(N(T) = 1 \mid N(t) = 1)dt}{\int_0^{T} \lambda(t) P(N(T) = 1 \mid N(t) = 1)dt} \quad (9)$$

Figure 2 shows an example of how to draw random speciation times in the sampled reconstructed tree.

The algorithm how to construct a random from the set of speciation time is given below.

2.2 Simulating reconstructed trees for a fixed time *t*

I start by simulating the number of species after the time t. Equation (3) gives the probability of n species at time t. Therefore, one only needs

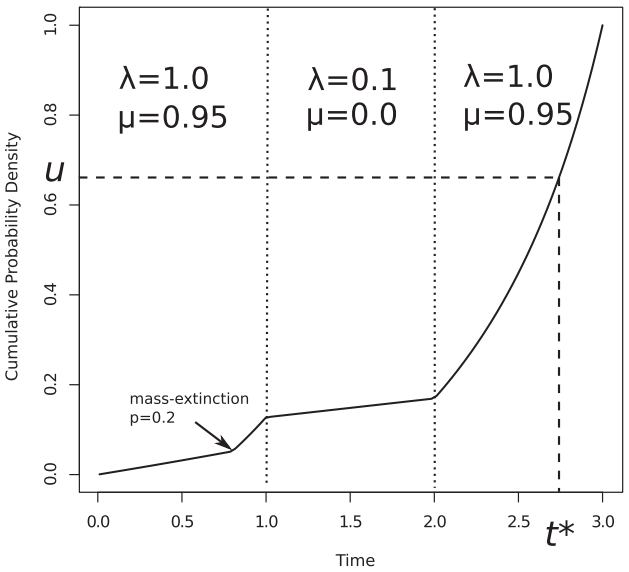

Fig. 2. An example birth–death process with rates of $\lambda = 1.0$ if $t < 1.0$ or $t > 2.0$ and $\lambda = 0.1$ otherwise; $\mu = 0.95$ if $t < 1.0$ or $t > 2.0$ and $\mu = 0.0$ otherwise. Additionally, there is a mass-extinction event at time $t = 0.8$ with a survival probability of $P = 0.2$ per species. The algorithm for drawing new speciation times t^* draws a random $u\backslash\text{simunif}(0, 1)$, and searches the t^* for which $\text{CDF}(t^*) = u$

to make a random draw $u \sim \text{unif}(0, 1)$ and find the maximal n for which the cumulative distribution function of Equation (3) is smaller or equal to u:

$$\max_n \left(u \geq \sum_{i=1}^{n} P(N(t) = i \mid N(t_0) = 1) \right) \quad (10)$$

Once the number of species is obtained by a random draw, the speciation times can be simulated using the above algorithm conditioning on both, the number of species and the time of the process.

2.3 Simulating reconstructed trees for a fixed number of species *n*

A tree with a fixed number of extant species can be simulated by first simulating the time of the process T given that the number of species was $N(T) = n$. Let us assume that one draws a random time T with some probability $P(T)$ and maximum value *max*, then simulates a tree under the birth–death process. A new time T and a new tree are drawn until the tree contains n species at time T. Thus, using Bayes Theorem, the probability of the time of the process (or the time of the origin/MRCA) is

$$P(T \mid n) = \frac{P(N(T) = n \mid N(t_0) = 1, T) \times P(T)}{\int_0^{\max} P(N(t) = n \mid N(t_0) = 1, t) \times P(t)dt}. \quad (11)$$

If a uniform distribution as the prior distribution on the time of the process is assumed, then the probability $P(T)$ in the numerator cancels out with the probability $P(t)$ in the denominator. However, in the general form, when one wants to specify some informative prior distribution of the time of the process, then the integral needs to be computed numerically.

Only a random draw from the distribution given in Equation (11) needs to be obtained and the same algorithm as for conditioning on the number of species and the time of the process can be used.

2.4 Starting at the MRCA

In the previous sections, I only considered the situations when the process starts with $N(t_0) = 1$. Now I will give the necessary modifications if one wants to condition on the MRCA. Let the process start with two species and both survive until the present $S(2, t_0, t)$, then Equation (5) needs to be modified to condition on two species to start with

$$\begin{aligned} P(N(t) = n \mid N(t_0) = 2, S(2, t_0, t)) \\ = \sum_{i=1}^{n-1} (P(N(t) = i \mid N(t_0) = 1, S(1, t_0, t)) \\ \times P(N(t) = n - i \mid N(t_0) = 1, S(1, t_0, t))) \\ = (n-1)(P(N(T) > 0 \mid N(t) = 1)^2 \exp(r(t, T)))^2 \\ \times (1 - P(N(T) > 0 \mid N(t) = 1) \exp(r(t, T)))^{n-2} \end{aligned} \tag{12}$$

The rationale behind the derivation is that one could consider two independent processes, both starting with one species at time $t_0 = 0$, and the sum of the species of both processes equals n.

The algorithms for simulating reconstructed phylogenies remain unchanged, and only Equation (3) needs to be replaced by Equation (12).

2.5 Reconstruction of a tree from speciation times

In the previous sections, I only elaborated the methods how to draw the set of speciation times. The reconstructed phylogeny can be created from the set of speciation times by first creating a set of n species. Then, repeatedly join two random species together. The time of the i-th speciation event is the i-th largest speciation time of the previously created set of speciation times. This procedure is closely related to the UPGMA method (Sokal and Michener, 1958; Felsenstein, 2004). The resulting tree is from the correct distribution because the speciation times are *iid* and the birth–death process implies a uniform distribution on labeled histories (Aldous and Popovic, 2005).

2.6 Modeling mass-extinction events and random taxon sampling

The rate functions have been provided as general functions of time. Following Nee *et al.* (1994b), mass-extinction events are modeled by

$$r(t, s) = \int_t^s (\mu(z) - \lambda(z))dz - \sum_{t_m \in (t,s]} \log(\rho_m) \tag{13}$$

where t_m is the time and ρ_m the survival probability of the m-th mass-extinction event.

Then, the probability of a single species alive at time t surviving until time T $[P(N(T) > 0 \mid N(t) = 1)]$ from Equation (2) changes to

$$\begin{aligned} P(N(T) > 0 \mid N(t) = 1) \\ = \left(1 - \int_t^T \left(\mu(s) \exp(r(t, s))\right) ds - \sum_{t_m \in (t,s]} (\rho_m - 1) \exp(r(t, t_m))\right)^{-1} \end{aligned} \tag{14}$$

Note that here $1 - \rho_m$ represents the extinction probability at the m-th mass-extinction event. Furthermore, random taxon sampling is the same as a mass-extinction event at the present time (Nee *et al.*, 1994b).

2.7 Implementation and numerical integration

The methods described to simulate trees are implemented in the R (R Core Team, 2012) package *TESS* (Tree Evolution Simulation Software). TESS can handle any type of rate functions and is designed for fast simulation of reconstructed trees. Furthermore, TESS computes the likelihood function [Equation (6)] conditioned on the time of the process, survival of the process or the time and number of taxa at present. The likelihood function can be used for Bayesian and Maximum Likelihood inference, posterior prediction and model testing (see the Supplementary Material).

There are no closed form solutions known for the equations presented above. Only for the special cases when the rates are constant, one can compute the probability densities analytically. Nevertheless, R provides several methods for numerical integration. TESS uses the package *deSolve* (Soetaert *et al.*, 2010) to perform the numerical integration. If only a single draw or few draws from a distribution are needed, then TESS performs a simple Monte Carlo sampling instead.

3 RESULTS

In this section, I compare the performance of TESS to TreeSim and show a potential application: posterior prediction of reconstructed phylogenies.

3.1 Performance

One of the major shortcomings of the GSA as implemented in TreeSim is the long computation time, especially in situation when $\lambda(t) \approx \mu(t)$. Additionally, TreeSim cannot handle continuous increasing or decreasing rate functions but only constant rates in predefined intervals. Thus, I designed a challenging but hopefully interesting function of the rates. The speciation rate $\lambda(t) = 1.0$ for $t \in [0.0, 0.8]$, then is increased for a time of rapid radiation as $\lambda(t) = 2.0$ for $t \in (0.8, 1.0]$ and switches back to the original rate when $t > 1.0$. The extinction rate changes at the same time points. First, the rate is comparably high $\mu(t) = 0.95$ for $t \in [0.0, 0.8]$ and then slows down to $\mu(t) = 0.5$ for $t \in (0.8, 1.0]$ before switching back to the original value as well. Additionally a mass-extinction event occurs at time $t = 0.8$ and each species survives with probability $P = 0.2$. The process resembles the situation when the diversity increases only slowly but a high species turnover happens until a mass-extinction event removes most species but also opens new niches, which results into a time of rapid radiation. Furthermore, each extant species was sampled with probability $P = 0.5$.

I ran both R packages on a MacBook Pro with a 2.3 GHz Intel Core i7 using only one core at a time to produce as fair as possible results. Simulations with the given rates where performed for $n \in \{10, 20, \ldots, 200\}$ species sampled at the present time. For each n, 100 trees were simulated and the median running time was reported (Fig. 3). Note that the distribution of running times is skewed for TreeSim. Therefore, the median is a more stable estimator than the mean but the median is also much lower. TESS outperforms TreeSim up to four orders of magnitude. Furthermore, TESS computation time is mostly independent of the number of species, and the main challenge is to compute the integral of the probability of survival. Additionally, TESS can simulate a batch of trees faster once the integral of the survival probability is precomputed, which is shown in the simulations of 10 000 trees (TESS 10K).

3.2 Posterior prediction

For each simulated tree, I computed the posterior predictive distribution can be used to assess the fit of the model to the observed data (Gelman *et al.*, 2003). Posterior predictive tests have the advantage to take the uncertainty in the model parameters into account, if applied an a Bayesian analysis (Bollback,

2002). If only the point estimate is used, e.g. the maximum likelihood estimate, then it is possible to compute the distribution of number of species at any time t using Equation (3) or the probability of n species in the reconstructed tree at time t given in Nee *et al.* (1994a). Instead, when the uncertainty of the estimated parameters is of interest, an analytical solution is infeasible in most situations because commonly there does not exist a closed form solution for the posterior distribution, as for the model considered in this article. Here I will present an example how simulating reconstructed trees can be used to approximate the posterior predictive distribution and evaluate the fit of the model. If the observed statistic fall outside the 95% confidence interval, then we can reject the model.

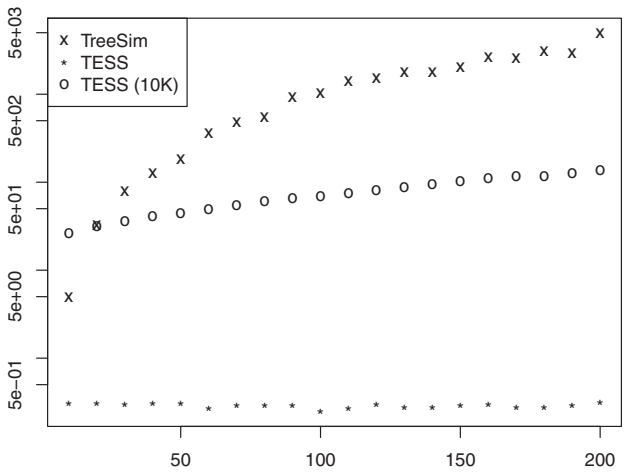

Fig. 3. The birth–death process with the rates of $\lambda = 1.0$ if $t < 0.8$ or $t > 1.0$ and $\lambda = 2.0$ otherwise; $\mu = 0.95$ if $t < 0.8$ or $t > 1.0$ and $\mu = 0.5$ otherwise. Additionally, there is a mass-extinction event at time $t = 0.8$ with a survival probability of $P = 0.2$ per species. Trees with $n \in \{10, 20, \ldots, 200\}$ species were simulated and the computation time was measured. The plot shows the median running time of 100 repetitions of TreeSim for one simulated tree and TESS for 1 and 10 000 trees, respectively

I used the species tree phylogeny from (Aström *et al.*, 2011, Additional file 7) containing the bird family Cettiidae. Phillimore and Price (2008) showed that most bird phylogenesis show patterns of a diversification slowdown. The rate functions with a slowdown in diversification can be modeled by the functions $\lambda(t) = \lambda_0 + \lambda_1 \exp(-\alpha t)$ and $\mu(t) = \mu$. I choose an exponential prior distribution for each parameter with rate 0.1, which is comparably flat but still puts most prior probability on small parameter values and hence on more simple models, i.e. the parameter combination $\mu = 0$ represents the pure birth model. The likelihood is obtained from Equation (6), starting at the MRCA and conditioned on the survival of both species.

The posterior distribution was approximated using a Markov chain Monte Carlo algorithm (Metropolis *et al.*, 1953; Hastings, 1970). The parameter values were changed using a scaling proposal and each parameter was updated once per iteration. The MCMC was run for 55 000 iterations and the first 5000 iterations were considered as the burnin period. Convergence was assessed using the Geweke test of convergence to same distribution in two parts of the run (Geweke, 1991). Every 10th sample from the posterior distribution was used to simulated one tree conditioned on the time of the process (Fig. 4), number of taxa at the present time and the γ-statistic (Pybus and Harvey, 2000). The histograms of the predicted number of taxa and γ-statistic values indicate that the observed values fall well within the 95% confidence intervals. The observed lineage-through-time (LTT) plot and the posterior predictive LTT plots both show significant decrease in the rate of diversification. More generally, the observed LLT plot falls well within the ranges of posterior predicted LTT curves. This indicates that the model fits the data. However, the large variance in the prediction shows the uncertainty and possibly a different model could fit the data more closely.

The MCMC algorithm as well as a basic posterior predictive test is available in TESS. The MCMC algorithm is kept as

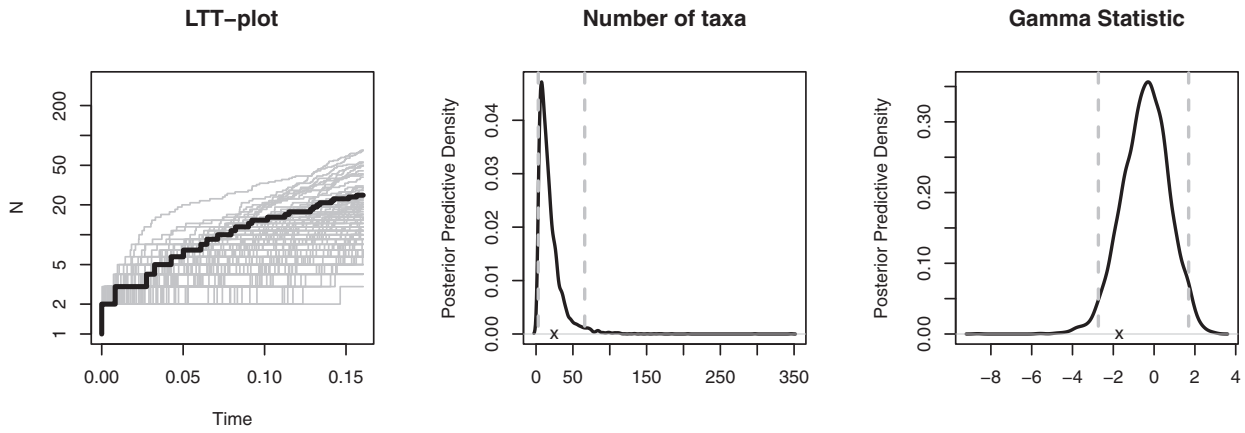

Fig. 4. Posterior predictive draws were obtained by approximating the posterior distribution with an MCMC algorithm and then simulating trees conditioned on the time of the process. For the first 100 trees the LTT curves (left plot), for each tree the number of taxa at present (middle plot), the γ-statistic (right plot) are plotted. The thin gray lines represent the random draws, and the thicker black line the observed data. The histograms contain the 95% confidence interval (dashed lines) and the observed value marked with 'x'. See the additional file *PosteriorPredictionCettiidae.R*

general as possible so that any possible model might be used. See the additional file *PosteriorPredictionCettiidae.R* that was used for this study.

4 DISCUSSION

In the present article, I only considered simulations of reconstructed phylogenies. Most estimated phylogenies are reconstructed phylogenies because no information about extinct species is available, such as ancient molecular sequences, and only the extant species are included. Nevertheless, phylogenetic trees including some extinct species might provide more information in estimating the time-dependent diversification rates. In these cases, either the GSA or simulating the missing speciation and extinction events following the idea of Bokma (2008) are useful.

Another drawback of the GSA is that events—e.g. mass-extinction events—cannot be specified to happen at a given time before the present when one conditions on the number of species being sampled. In this manuscript, I did not consider these circumstances, although the simulation procedure can be modified to accommodate events at specific times before the present without major difficulties. In this case, the distribution of the time of the origin (or the MRCA) needs to be considered backward in time instead of forward. Finally, once the time of the process is obtained, the simulation of the single speciation events remains unchanged.

The simulation algorithm presented here in this article relies on the factorization of the probability density function. Previously, only for the constant-rate birth–death process has such a factorization been known (Yang and Rannala, 1997; Höhna *et al.*, 2011). I derived the factorization for the time-dependent birth–death process. Unfortunately, other types of birth–death process, e.g. the diversity-dependent birth–death process (Etienne *et al.*, 2012) or the protracted speciation process (Etienne and Rosindell, 2012), do not have a know factorization. Nevertheless, the time-dependent birth–death process might be used to approximate other birth–death processes, such as the diversity-dependent birth–death process (Rabosky and Lovette, 2008).

It is possible to simulate tree using an MCMC algorithm, once the likelihood equations are known. However, if no direct simulation technique from the desired distribution is known, then the MCMC algorithm must start from some arbitrary starting values. Furthermore, samples are random draws from the specified distribution only after the chain has reached its stationary distribution. When the chain has reached its stationarity distribution and how long it takes to do so is a challenge in itself (Gelman *et al.*, 2003). Thus, direct sampling, as I propose here, is safer, faster and can even be used for starting values of MCMC runs.

Several of the more complicated birth–death processes, e.g. containing different rates for different species (FitzJohn *et al.*, 2009) or heritable extinction rates (Rabosky, 2009), do not have a known likelihood function. Instead, approximate Bayesian computation (ABC) can be applied to estimate the parameters of interest (Beaumont *et al*, 2002) although ABC has been criticized and should be used cautiously (Robert *et al*, 2011). The birth–death simulations implemented in TESS rely on the likelihood function, and therefore, neither ABC is necessary nor are these simulations needed for parameter estimation. Nonetheless, the simulation techniques can be used to test the summary statistics used in ABC. Currently, only the number of species at the present time, the time of the origin and the γ-statistic have been used as summary statistics for ABC under a birth–death model (Rabosky, 2009). Using TESS, it might be possible to test the performance of these summary statistics and develop new ones.

5 CONCLUSION

Simulating reconstructed phylogenies is important for developing and testing macro-evolutionary processes but has been a challenge for non-constant birth–death models (Hartmann *et al.*, 2010). In the present manuscript, I present an algorithm that can be used to simulate reconstructed phylogenies when conditioning on the number of species being sampled or the time of the process or both. The presented algorithms can accommodate any time-dependent rate function—continuous and discrete—and can model mass-extinction events as well as periods of rapid radiation. Compared with the currently only available software for these types of simulations, TreeSim, my implementation is four orders of magnitude faster for 200 taxa and does not need significantly more computational time to simulate larger trees. Thus, simulations under even more complex models and many taxa at present are now feasible.

I derived the probability density of the sampled reconstructed phylogenetic tree. The derivation follows the ideas of Thompson (1975) and Nee *et al.* (1994b). The newly derived probability density is provided in a general and new form. It shows that the speciation times are *iid*, which has previously only been shown for constant rates (Yang and Rannala, 1997). This enables for instance the *Diversified Sampling* under time-dependent speciation and extinction rates (Höhna *et al.*, 2011).

R is widely used for inferring diversification patterns (Harmon *et al.*, 2008; O'Meara, 2012). It provides a good framework to implement new methods that can be shared and used freely. The methods of this article are implemented in the R package TESS. TESS constructs trees in the *phylo* format of APE (Paradis *et al.*, 2008). Thus, the trees simulated in TESS can be directly analyzed in APE and other tools using the *phylo* format. The implementation is kept general so that any time-dependent rate function can be used. Additionally, an implementation of the likelihood computation is provided that can be used either in a maximum likelihood or in a Bayesian inference, as shown in the posterior prediction example.

ACKNOWLEDGEMENTS

I am very grateful to Fredrik Olsson for countless discussions on several aspects of this and related projects, to Tom Britton and two anonymous reviewers for comments on the manuscripts, Rich FitzJohn for his comments on the manuscript and the R package and the associate editor.

Conflict of Interest: none declared

REFERENCES

Aldous,D. and Popovic,L. (2005) A critical branching process model for biodiversity. *Adv. Appl. Probab.*, **37**, 1094–1115.

Alfaro,M.E. *et al.* (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl Acad. Sci.*, **106**, 13410–13414.

Aström,P. *et al.* (2011) Non-monophyly and intricate morphological evolution within the avian family Cettiidae revealed by multilocus analysis of a taxonomically densely sampled dataset. *BMC Evol. Biol.*, **11**, 352.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2045–2035.

Bokma,F. (2008) Bayesian estimation of speciation and extinction probabilities from (in) complete phylogenies. *Evolution*, **62**, 2441–2445.

Bollback,J.P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, **19**, 1171–1180.

Etienne,R. *et al.* (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proc. Natl Acad. Sci.*, **279**, 1300–1309.

Etienne,R.S. and Rosindell,J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Syst. Biol.*, **61**, 204–213.

Felsenstein,J. (2004) *Distance Matrix Methods. Inferring Phylogenies*. Sinauer Associates, Sunderland, MA, USA.

FitzJohn,R. *et al.* (2009) Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies., *Syst. Biol.*, **58**, 595–611.

Gelman,A. *et al.* (2003) Model checking and improvement. In: *Bayesian Data analysis*, 2nd edn. Chapman & Hall/CRC, New York.

Geweke,J. (1991) *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. Federal Reserve Bank of Minneapolis, Research Department.

Harmon,L.J. *et al.* (2008) GEIGER: investigating evolutionary radiations. *Bioinformatics*, **24**, 129–131.

Hartmann,K. *et al.* (2010) Sampling trees from evolutionary models., *Syst. Biol.*, **59**, 465–476.

Hastings,W.K. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

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

Kendall,D.G. (1948) On the generalized "Birth-and-Death" process. *Ann. Math. Statist.*, **19**, 1–15.

Liow,L.H. *et al.* (2010) When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? *Syst. Biol.*, **59**, 646–659.

Maddison,W. *et al.* (2007) Estimating a binary character's effect on speciation and extinction. *Syst. Biol.*, **56**, 701.

Metropolis,N. *et al.* (1953) Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**, 1087–1092.

Morlon,H. *et al.* (2011) Reconciling molecular phylogenies with the fossil record. *Proc. Natl Acad. Sci.*, **108**, 16327–16332.

Nee,S. (2006) Birth-death models in macroevolution. *Ann. Rev. Ecol. Evol. Syst.*, **37**, 1–17.

Nee,S. *et al.* (1994a) Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **344**, 77–82.

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

O'Meara,B.C. (2012) Evolutionary Inferences from phylogenies: a review of methods. *Annu. Rev. Ecol. Syst.*, **43**, 267–285.

Paradis,E. *et al.* (2004) APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**, 289–290.

Phillimore,A.B. and Price,T.D. (2008) Density-dependent cladogenesis in birds. *Plos Biol.*, **6**, e71.

Pybus,O.G. and Harvey,P.H. (2000) Testing macro–evolutionary models using incomplete molecular phylogenies. *Proc. R. Soc. Lond. B Biol. Sci.*, **267**, 2267–2272.

Quental,T.B. and Marshall,C.R. (2009) Extinction during evolutionary radiations: reconciling the fossil record with molecular phylogenies. *Evolution*, **63**, 3158–3167.

Rabosky,D. (2006) Likelihood methods for detecting temporal shifts in diversification rates. *Evolution*, **60**, 1152–1164.

Rabosky,D.L. (2009) Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Syst. Biol.*, **58**, 629–640.

Rabosky,D. and Lovette,I. (2008) Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, **62**, 1866–1875.

R Core Team. (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. htpp://www.R-project.org (18 April 2013, date last accessed).

Ricklefs,R. (2007) Estimating diversification rates from phylogenetic information. *Trends Ecol. Evol.*, **22**, 601–610.

Robert,C.P. *et al.* (2011) Lack of confidence in approximate Bayesian computation model choice. *Proc. Natl Acad. Sci.*, **108**, 15112–15117.

Soetaert,K. *et al.* (2010) Solving differential equations in R: Package deSolve. *J. Stat. Softw.*, **33**, 1–25.

Sokal,R. and Michener,C. (1958) A statistical method for evaluating systems relationships. *Univ. of Kans. Sci. Bull.*, **38**, 1409–1438.

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

Thompson,E. (1975) *The Likelihood Approach. Human Evolutionary Trees*. Cambridge University Press, Cambridge.

Yang,Z. and Rannala,B. (1997) Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. *Mol. Biol. Evol.*, **14**, 717–724.

APPENDIX A

I give here a short version of the proof of the joint probability density of all speciation times of a reconstructed tree. First, let me restate the probability of obtaining exactly one species at time T with $T > t$. At any given time s, $t < s < T$ the number of species could be between one and infinity; however, all but one species must go extinct until time T (Thompson, 1975):

$$\begin{aligned} &P(N(T) = 1 \mid N(t) = 1) \\ &\quad = \sum_{k=1}^{\infty} k \times P(N(s) = k \mid N(t) = 1) \times P(N(T) = 1 \mid N(s) = 1) \\ &\qquad \times (P(N(T) = 0 \mid N(s) = 1))^{k-1} \end{aligned} \tag{A.1}$$

The factor k arises because any of the k currently alive species could survive until the present and we need to consider each possibility.

Let us transform the reconstructed tree from Figure 1 into a set of branches $\mathbf{b} = \{\{t_0, t_1\}, \{t_1, t_2\}, \{t_1, t_3\}, \{t_2, t_4\}, \{t_2, T\}, \{t_3, T\}, \{t_3, T\}, \{t_4, T\}, \{t_4, T\}\}$. There are two types of branches, those that end at a speciation event and those that end at the sampling time T. The probability density of a branch that ends at the present time is simply the probability of starting with one species and observing exactly one species

$$P(\{t_i, T\}) = P(N(T) = 1 \mid N(t_i) = 1) \tag{A.2}$$

The probability of any other branch starting at time t_i and ending at a speciation event at time t_j is given by the probability that there were k species at time t_j, one of these speciated with probability $k\lambda(t_j)$, this species survived but all other $k-1$ species went extinct before the present time:

$$\begin{aligned} &P(\{t_i, t_j\}) \\ &\quad = \sum_{k=1}^{\infty} k \times \lambda t_j \times P(N(t_j) = k \mid N(t_i) = 1) \\ &\qquad \times (P(N(T) = 0 \mid N(t_j) = 1))^{k-1} \end{aligned} \tag{A.3}$$

If we multiply the equation by $P(N(T) = 1 \mid N(t_j) = 1)/P(N(T) = 1 \mid N(t_j) = 1)$, we can see that we get exactly Equation (A.1)

$$\begin{aligned} P(\{t_i, t_j\}) &= 1/P(N(T) = 1 \mid N(t_j) = 1) \sum_{k=1}^{\infty} k \times \lambda(t_j) \times P(N(t_j) \\ &= k \mid N(t_i) = 1) \times (P(N(T) = 0 \mid N(t_j) = 1))^{k-1} \times P(N(T) \\ &= 1 \mid N(t_j) = 1) = \lambda(t_i) \times P(N(T) \\ &= 1 \mid N(t_i) = 1)/P(N(T) = 1 \mid N(t_j) = 1) \end{aligned} \tag{A.4}$$

Finally, by multiplying the probability densities of all branches together, we obtain the probability density of the reconstructed tree

$$\begin{aligned} P(\mathbf{b}) &= \prod_{\forall b,\, t_j \neq T} \lambda(t_i) \times P(N(T) = 1 \mid N(t_i) = 1)/P(N(T) \\ &= 1 \mid N(t_j) = 1) \times \prod_{\forall b,\, t_j = T} P(N(T) = 1 \mid N(t_i) = 1) \\ &= P(N(T) = 1 \mid N(t_0) = 1) \prod_{i=1}^{n-1} \lambda(t_i) P(N(T) = 1 \mid N(t_i) = 1) \end{aligned} \tag{A.5}$$

The probability density of all branches $P(\mathbf{b})$ is equivalent to the probability density of all branching times [Equation (6)].