Simulating Trees with a Fixed Number of Extant Species

Syst. Biol. 60(5):676–684, 2011

DOI:10.1093/sysbio/syr029
Advance Access publication on April 11, 2011

Simulating Trees with a Fixed Number of Extant Species

TANJA STADLER*

Institut für Integrative Biologie, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland;
**Correspondence to be sent to: Institut für Integrative Biologie, ETH Zürich, Universitätsstr. 16, 8092 Zürich, Switzerland;*
E-mail: tanja.stadler@env.ethz.ch.

Received 5 March 2010; reviews returned 2 July 2010; accepted 6 January 2011
Associate Editor: Tiffani Williams

Abstract.—In this paper, I develop efficient tools to simulate trees with a fixed number of extant species. The tools are provided in my open source R-package TreeSim available on CRAN. The new model presented here is a constant rate birth–death process with mass extinction and/or rate shift events at arbitrarily fixed times 1) *before* the present or 2) *after* the origin. The simulation approach for case (2) can also be used to simulate under more general models with fixed events after the origin. I use the developed simulation tools for showing that a mass extinction event cannot be distinguished from a model with constant speciation and extinction rates interrupted by a phase of stasis based on trees consisting of only extant species. However, once we distinguish between mass extinction and period of stasis based on paleontological data, fast simulations of trees with a fixed number of species allow inference of speciation and extinction rates using approximate Bayesian computation and allow for robustness analysis once maximum likelihood parameter estimations are available. [Birth–death model; mass extinction; phylogenetic tree; rate shift; simulation.]

In order to test hypotheses on speciation and extinction rates in phylogenetic trees, the phylogenetic trees are compared with speciation and extinction models. For simple models and simple questions, this can be done analytically. For example, when assuming that the speciation and extinction rates remained constant through time, the maximum likelihood speciation and extinction rates for a given tree can be calculated (Nee et al. 1994). Also, under this constant rate model, the distribution of the γ statistic (Pybus and Harvey 2000) can be calculated, which is used to test for departure from the constant rate model (Day et al. 2008; Phillimore and Price 2008; Cusimano and Renner 2010). For more complicated models and/or more complex questions toward the speciation and extinction rates, simulating trees under different models is often necessary, as analytic results are not available.

Simulating trees under a given model might seem straightforward. However, simulations become difficult when we want to condition on observing a certain feature in the data, for example, a fixed number of extant species. Until now, only few models can be simulated under such conditions. For instance, there is no tool available, which simulates trees with a fixed number of species and has a mass extinction event at 65 Ma. The widely used program PHYLOGEN (Rambaut 2002) can stop simulations either after a fixed time (e.g., 65 myr after the mass extinction) or once reaching a certain number n of species, but PHYLOGEN cannot condition on both. Therefore, analysis of empirical trees cannot be based on a model, which takes into account both the time of mass extinction and the number of extant species. In common simulation tools, simulations are stopped or mass extinctions happen once a predefined number of species appears, see, for example (Harvey et al. 1994; Crisp and Cook 2009). However, such simulations do not reflect the underlying process, namely events happening at predefined times in the past, for example, 65 Ma. The bias of the usage of such simulations is not clear. In the following, I will outline what simulation tools are *expected* to do and what they *actually* do. In Models and Algorithms section, I develop methods that do simulations as *expected* and investigate the bias of commonly used simulation tools.

Under most common models for speciation and extinction, trees evolve perpetually, and therefore, trees of all ages and number of leaves are possible. In particular, the expected age of the simulated trees is infinite. As empirical trees are finite, the simulated trees have to be conditioned on some aspect in order to compare them with empirical trees. The five natural choices are conditioning

(A) on the number of extant species or

(B) the time since the first species evolved or

(C) the time since the most recent common ancestor (mrca) of the extant species evolved or

(D) the number of extant species and the time since the first species evolved or

(E) the number of extant species and the time since the most recent common ancestor of the extant species evolved.

The choice of condition depends on which aspect we want to emphasize on. I will discuss all five possibilities.

I call the trees consisting of the extant and the extinct lineages *complete trees*. The trees resulting from deleting all lineages that do not have extant descendants are called *reconstructed trees*. Trees inferred from extant species (i.e., not including any fossils) are therefore always reconstructed trees.

Simulating trees under the natural Conditions (A)–(E) turns out to be a nontrivial task. Standard software for simulating trees consider the following neutral model, a *constant rate birth–death process* (BDP): under this model, all species have the same rates at all times—a constant rate of speciation and a constant rate of extinction.

The BDP with extinction rate zero is the pure birth Yule model (Yule 1924).

The BDP can be extended to model mass extinction events or shifts in rates in which case it is called *episodic birth–death process* (EBDP). When considering an EBDP, we distinguish between a forward and a backward EBDP: under a *forward EBDP*, we specify the time of mass extinction events or rate shifts *after* the time of origin of the first species. Under a *backward EBDP*, we specify the time of mass extinction events or rate shifts *prior* to the present. In general, we define a model to be a *forward model* when events are fixed at a certain time *after* origin and a *backward model* when events are fixed at a certain time *before* the present.

When conditioning on (A), the number of extant species, it is not clear what value to take for the time of origin of the process. Commonly, a uniform prior on $(0, \infty)$ for the time of origin is assumed (where time is measured such that today is zero and increasing going into the past)—in other words, each age of the tree is equally likely; a uniform prior is assumed explicitly in Popovic (2004), Aldous and Popovic (2005), and Gernhard (2008) and is assumed implicitly when using the *simple sampling approach* (SSA) for simulating trees under the Yule model (Hartmann et al. 2010).

In the following, I list the standard simulation packages that attempt to simulate under the EBDP:

- PHYLOGEN (Rambaut 2002): This package simulates trees with n species under the forward EBDP by starting with a single species and stopping when n species are first reached. I call this simulation approach the *simple sampling approach* (SSA). Later periods with n species are disregarded, so the tree simulation is stopped too early. I discuss the resulting problems below. For the Yule model, the SSA works correctly, see Hartmann et al. (2010).

  Conditioning on the time since origin is done by PHYLOGEN for the forward and backward EBDP as expected: the process is started with one species and stopped after a fixed time. So Condition (B) is handled correctly with PHYLOGEN. By simulating two trees under (B) with both trees having extant species and then joining the two trees at the origin, the resulting tree is sampled under Condition (C).

- GEIGER (Harmon et al. 2007): This R-package simulates BDP trees by starting with two species and stopping either when (1) first n species are reached or when (2) time t has elapsed. Under both conditions, we do not sample from the intended distribution.

  Under (1), as we stop when first n species existed, we disregard later periods when n species existed. Approach (1) and the SSA differ in one aspect, namely under (1), we start with two species, whereas under the SSA, we start with one species. Under (2), we condition on the time since two species existed. We clearly do not simulate under Condition (B). However, we also do not simulate under Condition (C): one or both the species with which we started might not have extant descendants. So the time since the most recent common ancestor of the extant species might be shorter.

- CASS (Stadler 2006–2009): Using the theory on point processes, I found an efficient way to simulate trees with n species under the BDP (Gernhard 2008). The simulated trees are reconstructed trees. Unfortunately, complete trees cannot be inferred by this approach.

- TreeSample (Hartmann 2007): In a previous paper (Hartmann et al. 2010), we introduced a general simulation approach (GSA). This approach samples trees with n species correctly under forward models. The forward models need to have the property that the species number eventually becomes bigger or smaller than n. Models where species number fluctuate between some value a and b with $a < n < b$ cannot be simulated.

In summary, PHYLOGEN samples under the forward and backward EBDP conditioning on (B) or (C) correctly. Furthermore, if only considering the simulated trees with n final species obtained in the case (B) or (C), PHYLOGEN samples correctly conditioning on (D) or (E). The latter is a very slow approach though. CASS samples reconstructed trees under the BDP conditioning on (A) correctly but cannot sample complete trees. TreeSample samples correctly from a big class of models (including the forward EBDP) conditioning on (A). Note that no method can sample from the backward EBDP under (A). For an overview, see Table 1.

In this paper, I close the obvious gap by providing a method for efficiently sampling complete and reconstructed trees under the backward EBDP conditioning on (A). Furthermore, for the BDP, I provide a very fast method to simulate reconstructed trees under (A), (D), or (E) allowing for incomplete taxon sampling. I also improve the GSA for correctly sampling trees with n species from trees generated by the SSA under any

TABLE 1. Overview over the available simulation tools: specification of which software can simulate under which models

Program	Yule	BDP	Forward EBDP	Backward EBDP
PHYLOGEN	A, B, C	B, C	B, C	B, C
GEIGER	A, C			
CASS	A	A_r		
TreeSample	A	A	A	
TreeSim	A, B, C, D, E	A, B, C, D_r, E_r	A, B, C	A, B, C

Note: An algorithm for simulating under (B) or (C) can be used for simulating under (D) or (E) by only considering realizations with n extant species. The (D) and (E) in the table indicate the availability of faster algorithms than using an algorithm for simulating under (B) or (C). The subscript r indicates that the approach is only available for simulating reconstructed trees but not complete trees

forward model. The algorithms are provided in the R-package TreeSim (Stadler 2010).

Using the developed simulation tools, I investigate lineages-through-time (LTT) plots (Harvey et al. 1994) of birth–death models with mass extinction events and show that they do not differ significantly from LTT plots with rate shifts.

Models and Algorithms

Backward Models: The Backward EBDP

In this section, I define the backward EBDP formally. Time today is set to $t_0 = 0$. Mass extinction and rate shift events in the past are at times $t_1 < t_2 < \cdots < t_m$. Each species at time t with $t_i < t < t_{i+1}$ ($i \in \{0, 1, \ldots, m\}$, where $t_{m+1} := \infty$) has a constant rate λ_i of speciating and a constant rate μ_i of going extinct. At a mass extinction event at time t_i, the fraction ρ_i of species survives, the surviving set is chosen uniformly at random.

Note that $\rho_i = 1$ corresponds to a rate shift from λ_i, μ_i to λ_{i-1}, μ_{i-1}. At time t_0, that is, today, a fraction ρ_0 of all extant species is sampled uniformly at random. The parameters of the EBDP are summarized as $\lambda{=}(\lambda_0, \ldots, \lambda_m)$, $\mu = (\mu_0, \ldots, \mu_m)$, $\rho = (\rho_0, \ldots, \rho_m)$, $t = (t_0, \ldots, t_m)$. In the following, the EBDP is conditioned on obtaining a fixed number n of sampled extant species. Furthermore, assume that the time of origin t_{or} of the process is not known. We assume a uniform prior on $(0, \infty)$ (i.e., each t_{or} is equally likely), as done in Popovic (2004), Aldous and Popovic (2005), and Gernhard (2008). Note that a uniform prior is usually implicitly assumed: for example, when simulating a tree under the BDP without extinction (which is a Yule process; Yule 1924), the common algorithm starts with one species and stops just before the $(n+1)$th species appears. We showed in Hartmann et al. (2010) that this simple sampling algorithm samples n-species Yule trees by implicitly assuming a uniform prior for the time of origin.

A *naive algorithm* of sampling trees with n extant species under the backward EBDP is to explicitly simulate the model in the following way. Start with an initial species at the random time t_{or} in the past. Note that as we cannot sample t_{or} from $(0, \infty)$, we sample t_{or} from $(0, C)$, where C is chosen so large that the probability of a tree being older than C and having n extant species is negligible small. Let the species then evolve according to the parameters λ, μ, ρ, t. Stop after time t_{or} has elapsed, that is, at $t_0 = 0$. If the resulting tree has n sampled extant species, this tree is included into our sample of trees. We proceed until we have the required number of simulated trees. This approach is obviously correct as it simulates the model by explicitly following the model definition. But the approach is very slow as we will often end up with a number different from n. The probability of simulating a tree with n species using this approach is (Gernhard 2008, proof of lemma 3.1),

$$p(n) = \frac{\lambda^{n-1}}{nC} \left(\frac{1 - e^{-(\lambda-\mu)C}}{\lambda - \mu\, e^{-(\lambda-\mu)C}} \right)^n .$$

For example, for $n = 100$ species, $\lambda = 0.02$, $\mu = 0.01$, and $C = 200$, we have $p(100) = 1.3 \times 10^{-6}$, meaning that only each millionth tree has 100 species. In the following, I will provide a much more efficient way of simulating trees with n species under the EBDP process. To obtain a tree with n extant species, we simulate the EBDP backward in time with the following EBDP backward algorithm.

EBDP backward algorithm:

(1) We start at time $t = 0$ with $i = 0$ and the number of species $N = n$. The current tree τ consists of N isolated vertices.

(2) The total number of species is $N \leftarrow \text{round}(N/\rho_i)$. We add $\text{round}(N/\rho_i) - N$ isolated vertices to the tree τ at time t.

(3) We draw from the exponential distribution with parameter $(\lambda_i + \mu_i)N$ the waiting time w.

(4) If $t + w > t_{i+1}$, increment i by 1, add the edge length $t_{i+1} - t$ to each species in the tree alive at t, set $t \leftarrow t_{i+1}$ and go to (2).

(5) Add the length w to each species in the tree alive at t. Set $t \leftarrow t + w$.

(6) The event at time t is with probability $\mu_i/(\lambda_i + \mu_i)$ an extinction event and with probability $\lambda_i/(\lambda_i + \mu_i)$ a speciation. If we sample an extinction event, we set $N \leftarrow N + 1$ and add a new species (isolated vertex at time t) to the tree (forward in time, a species went extinct). If we sample a speciation event and $N > 1$, we pick two species uniformly at random and coalesce their lineages (forward in time, this is a speciation event). If we sample a speciation event and $N = 1$, we set $N \leftarrow 0$.

(7) If $N > 0$ go to (3). Otherwise: return the tree with probability $\frac{1/\lambda_i}{\sum_{j=0}^{m} 1/\lambda_j}$. If no tree is returned, go to (1).

In the Appendix, I establish that the EBDP backward algorithm samples trees correctly from the distribution on n-species trees under the backward EBDP. For the special case of a BDP, we can do the simulation of reconstructed trees much faster as I show in the next section.

BDP with Stochastic Sampling of Extant Species

The BDP with stochastic sampling of extant species is a birth–death process with speciation rate λ, extinction rate μ, and each extant species is included into the final tree with probability ρ. Under the EBDP, we assumed sampling of the fraction ρ of extant species. Note that the approach below only holds for the stochastic sampling scheme as the approach assumes that species are sampled independently from each other. Further note that, on the other hand, it is not clear how to alter the EBDP backward algorithm in order to allow for

stochastic sampling, as we do not know the probability of having m species prior to the mass extinction, given we have $k \leq m$ species after the mass extinction.

In Hartmann et al. (2010), we simulated trees under the BDP with stochastic sampling and with sampling of a constant fraction. The resulting trees cannot be distinguished, so there is no bias induced by using one sampling scheme instead of the other. A mass extinction event is another incomplete sampling event. Using an inductive argument, the tree ancestral to the mass extinction event is not biased by the sampling scheme (stochastic or constant fraction).

In Stadler (2009), it is shown that the birth–death process with stochastic sampling (parameters λ, μ, and ρ) is equivalent to the birth–death process with complete sampling and speciation and extinction rates $\lambda' = \rho\lambda$, $\mu' = \mu - \lambda(1-\rho)$. So the fast birth–death point process approach of Hartmann et al. (2010) for simulating trees can be applied after a transformation of the parameters such that the transformed sampling rate is $\rho' = 1$. The fast approach simulates trees for Conditions (A), (D), and (E) (Hartmann et al. 2010) and is implemented into TreeSim.

Forward Models

Trees with Condition (A) can be simulated under a forward model using the GSA. The idea of the GSA is to sample trees with n species correctly from trees, which were simulated by the SSA (Hartmann et al. 2010). Trees are simulated using the SSA under a forward model until a number $m >> n$ of species is reached or all species are extinct (this might happen as extinction of a tree has a positive probability under birth–death models). The number m is chosen to be sufficiently large, such that the chance of returning back to n species is negligible small (if there is no such m for the considered model, the GSA approach does not work).

Then, the GSA does the following: for each tree τ simulated by the SSA, we determine how many trees with n species we want to sampled from τ—the number is proportional to the total time during which the tree τ has n species. For each n species tree to be sampled from τ, we choose a point in time uniformly at random where τ had n species. The tree τ is cut off at this point in time and becomes our sampled tree with n species. We show in Hartmann et al. (2010) that this approach actually samples trees under Condition (A). This approach is implemented into the package TreeSample.

By sampling several trees with n species from one tree τ, we produce for small sample sizes a bias: we get the pattern subtending the root of τ several times. I therefore changed the method such that either one or no trees are sampled from τ. The probability of sampling is proportional to the time for which τ had n species. This improved GSA method is implemented in my R-package TreeSim—it takes as input a sample of trees generated by the SSA and outputs a correct sample of trees with n species under the considered model.

Implementation

All methods presented in this paper are implemented in my R-package TreeSim (Stadler 2010), which is available on CRAN (http://cran.r-project.org/) and my home page (http://www.tb.ethz.ch/people/tstadler). The trees produced by the algorithms are in the format defined in the standard phylogenetic R-package APE (Paradis et al. 2004). The simulated trees can therefore be used directly for further analysis in R or can be written into a Nexus file using an APE function. The package simulates trees with Conditions (A)–(E) under the backward EBDP, forward EBDP, and the BDP.

For simulating under Condition (B) (respectively (C)), we start with the origin (respectively mrca) and simulate for the required time. We do the same for (D) (respectively (E)) and then only consider realizations that yield the right number of species. For Condition (A) under the backward EBDP, I implemented the method presented in Backward Models: The Backward EBDP section. The TreeSim package samples trees with Condition (A) under a forward model using the GSA explained in Forward Models section, given we have trees simulated under the forward model using the SSA (the SSA for the forward EBDP is implemented in PHYLOGEN). For the BDP, I implemented the much faster approach for (A), (D), and (E) as described in BDP with Stochastic Sampling of Extant Species section.

Application of the Simulation Algorithms

When using molecular data to infer phylogenies, we obtain reconstructed trees. A reconstructed phylogeny not only provides insight on relationships but also enables us to test evolutionary hypotheses on speciation and extinction rates (Harvey et al. 1994; Pybus and Harvey 2000). In the following, I investigate whether we can detect mass extinction events in reconstructed phylogenetic trees. In a complete phylogeny, we clearly see when mass extinction events happen: the LTT plot drops significantly at the time of a mass extinction event (see Fig. 1). It is, however, unclear to what extent we can detect mass extinction events in LTT plots of reconstructed phylogenies.

Previous studies detected, using simulation approaches as implemented in PHYLOGEN, that a mass extinction event produces an antisigmoid LTT curve (Harvey et al. 1994; Crisp and Cook 2009). Crisp and Cook (2009) reveal that this antisigmoid curve is also "sometimes" produced under a model with constant rates interrupted by a phase of stasis (meaning almost no speciation and extinction).

However, in both studies, the rate shift, the mass extinction, or the termination of the simulation happened when a predefined number of species was first reached. For example, once the simulation reached 700 species, then a mass extinction happened and 97% of the species died. In Crisp and Cook (2009), mass extinction events could also happen at a specified time after origin (but *not* at a specified time before present). The study of Crisp and Cook (2009) used the package PHYLOGEN

summarized above. Both studies could not allow for events at fixed times before the present.

The methodology introduced in this study is the first to allow for fixed times of rate changes and mass extinction events, while ensuring that the final tree has the desired number of species. I simulated trees with a mass extinction event and trees with a phase of stasis using the parameter combinations as in Crisp and Cook (2009) but being able to control the time of mass extinction/rate shift. In the next section, I use the same parameters as in Crisp and Cook (2009) to investigate the biases introduced by the different simulation approaches.

Speciation Model with Mass Extinction and Speciation Model with a Stasis Phase Produce the Same LTT Plots

First, I investigated the shape of LTT curves under a model of constant rates ($\lambda = 0.3, \mu = 0.15$, as in Crisp and Cook 2009), interrupted by a mass extinction event at 25 timesteps ago. I simulated 50 phylogenies on 350 extant species using the EBDP backward algorithm. The mass extinction severity was varied: I sampled the fraction $0.5, 0.4, 0.3, 0.2, 0.1, 0.03, 0.01$ as well as all the extant species at 25 timesteps ago. For each parameter combination, I calculated the average LTT plot, which is the LTT plot calculated by computing the average number of lineages over all 50 simulated trees at each timepoint (thus mimicking a mean age chronogram obtained in dating analyses). The average LTT plots of the simulated complete and reconstructed trees are shown in Figure 1. The plots follow an antisigmoid curve. To investigate the impact of incomplete extant species sampling, I redid the simulations with $\rho_0 = 0.5$, that is, the trees have 700 extant species at present, but only 350 species are sampled uniformly at random. The average LTT plot is shown in Figure 2. Again, antisigmoid curves are recovered, which are very similar to those in Figure 1.

Second, I simulated trees under a model with a stasis period. I used the same speciation and extinction rates as in Crisp and Cook (2009) ($\lambda = 0.3, \mu = 0.15$ before and after stasis and $\lambda = 0.05, \mu = 0.025$ during stasis). The length of the stasis period was set to 20. The end of the stasis period was at 25 time units ago. I simulated 50 trees on 700 species and then sampled 350 species uniformly at random. The LTT plots show a plateau between 25 and 45 timesteps ago.

Importantly, in my simulations, the LTT plots from the trees under the stasis model are indistinguishable from the trees under the mass extinction model (where 97% of species went extinct at 25 timesteps ago; see Fig. 3). This makes the conclusions of Crisp and Cook (2009) even stronger: Crisp and Cook (2009) find that the LTT plots of a model with mass extinction mostly have a antisigmoid shape, whereas the LTT plots of a stasis model only sometimes have an antisigmoid shape (fig. 4 in their paper). They therefore concluded that, for their considered legume data having an antisigmoid LTT plot, the process with a stasis period cannot be ruled out. I show here that, as the two processes produce exactly the same LTT plots, each process is equally likely to produce the observed pattern of legume diversification. The reason that Crisp and Cook (2009) only sometimes observed the stasis plateau is that the length of the stasis period was not fixed but was variable, and the variable duration of stasis was very short (the length of the stasis period was defined as the time during which the tree grew from 20 to 25 species).

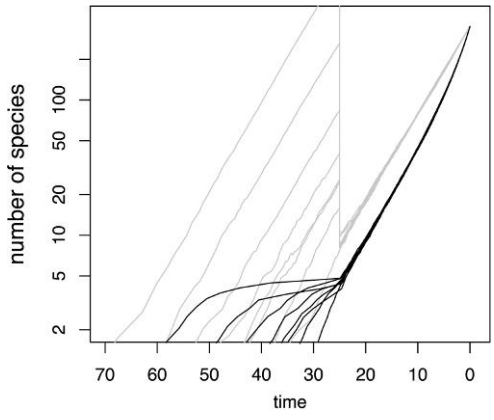

FIGURE 1. LTT plots with a mass extinction event at 25 timesteps ago; complete sampling of extant species: average LTT plot of 50 simulated phylogenies (gray: complete—i.e., phylogenies including extinct species, black: reconstructed—i.e., phylogenies including only the extant species) with 350 extant species under the backward EBDP ($\lambda = (0.3, 0.3), \mu = (0.15, 0.15), t = (0, 25)$). The mass extinction intensity was varied from bottom to top: $\rho = (1, 1)$ (no mass extinction), $\rho = (1, 0.5)$, $\rho = (1, 0.4)$, $\rho = (1, 0.3)$, $\rho = (1, 0.2)$, $\rho = (1, 0.1)$, $\rho = (1, 0.03)$, $\rho = (1, 0.01)$. Recall that $\rho = (1, x)$ means that all extant species are sampled, and the fraction x of species alive prior to the mass extinction survived the mass extinction.

DISCUSSION

Incorrect Simulation Algorithms Produce a Bias

Phylogenetic trees are frequently simulated with PHYLOGEN using SSA. The SSA has the advantage

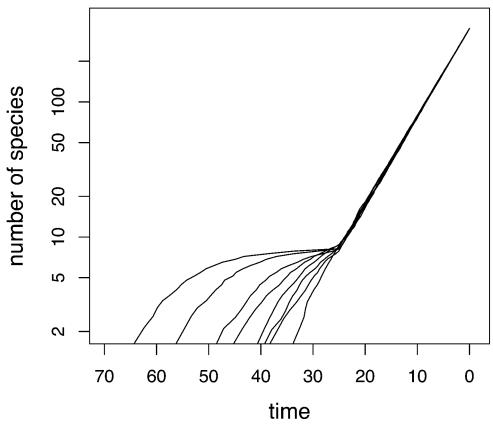

FIGURE 2. Average LTT plot of the reconstructed simulated trees as in Figure 1, but with incomplete extant species sampling ($\rho_0 = 0.5$).

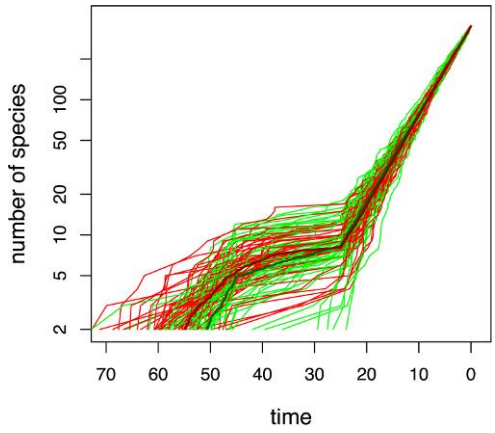

FIGURE 3. Comparison of model with mass extinction and model with stasis period: LTT plots of 50 simulated reconstructed phylogenies with 350 extant species under the backward EBDP ($\lambda = (0.3, 0.3), \mu = (0.15, 0.15), t = (0, 25)$), with mass extinction intensity set to $\rho_1 = 0.03$ and extant species sampling proportion $\rho_0 = 0.5$, are shown in light red. The corresponding average LTT plot is dark red. LTT plots under a model with a stasis period ($\lambda = (0.3, 0.05, 0.3), \mu = (0.15, 0.025, 0.15)$), and no mass extinction but incomplete extant species sampling ($\rho = (0.5, 1, 1)$), are shown in light green. The corresponding average LTT plot is dark green.

of being much faster and easier to implement than the correct approaches presented here. However, the SSA produces a bias on trees under the simple BDP model; therefore, it is necessary to use the more complex approach presented here. In Hartmann et al. (2010), we analyzed how much bias the incorrect SSA approach induces on the simulated trees under a BDP model. We compared the reconstructed trees under the BDP simulated with the SSA to the reconstructed trees simulated with a correct approach using TreeSample. We found that in the SSA trees, speciation times were up to a 3-fold younger. Furthermore, in the SSA trees with high extinction rates, the initial slope of the LTT plots were steeper. Such biases may lead to wrong conclusions. For example, by simulating traits under a gradual trait evolution model on simulated BDP trees, we found no correlation between trait variance and tree age using SSA, whereas we found a correlation using the correct approach. This shows the importance of using correct simulation tools.

Furthermore, my simulations show that reconstructed trees with mass extinction events in the past look like reconstructed trees with a stasis period. This suggests that we may not be able to distinguish, based on the reconstructed tree, between a mass extinction and a model with constant rates interrupted by a stasis period.

Parameter Tuning Such That Simulations and Empirical Trees Coincide is Easy with the New Simulation Approach

An LTT plot of an empirical phylogenetic tree may potentially reveal a lot about the past speciation and extinction rates (Harvey et al. 1994). In a period of constant rates, the slope of the LTT plot is the net speciation rate $\lambda - \mu$. A slope uplift in the LTT plot, with a constant slope before and after the uplift, corresponds to an increase in rates. An antisigmoid curve prior to the uplift is produced by a mass extinction event at the uplift. I showed in this study that a model where constant rates are interrupted by a period of stasis produces the same antisigmoid curve.

As a mass extinction event and a model with a stasis period both yield an antisigmoid curve, we cannot estimate all parameters in an EBDP model simultaneously from only a reconstructed phylogeny—we need to specify a priori if we assume a mass extinction or rate shifts. In order to justify the choice between mass extinction and rate shifts, inferences from paleontological and paleoclimatic studies may need to be introduced, with emphasis on the time of the LTT upturn.

Given we decided on the model with mass extinction (or on the model with stasis period) for an empirical tree with an antisigmoid LTT plot, the goal is to tune the model parameters such that simulations yield the same trees as the empirical tree. Without being able to both change the rates at given timepoints in the simulations and conditioning on the final number of species, it is hard to tune the parameter combinations such that the simulated trees match with empirical trees on their LTT plots. With the simulation approach in this paper, we can fix the time of rate changes and mass extinction events and therefore the adjustment becomes easy: altering λ and μ changes the slopes, whereas altering the mass extinction intensity (respectively the length of the stasis period) changes the length of the plateau.

Note that the ratio μ/λ only influences the LTT plot in the very recent past (Harvey et al. 1994) (the "pull of the present" effect) and cannot be estimated reliable from a single empirical observation due to large stochastic variations (Rabosky 2010). Therefore, this setting plays a minor role toward the simulated trees.

Parameter Estimation

The presented parameter tuning approach is a "visual" approach for fitting the model to the empirical data. The simulation tools developed in this paper provide the first step toward implementing an approximate Bayesian computation (ABC) approach (Marjoram et al. 2003) for estimating rates under the EBDP model. An ABC approach samples from the *approximate* posterior distribution of parameters in the following way. Initial parameters are proposed and a tree is simulated under the parameters. Based on a *summary statistic*, it is decided if the simulated tree matches the empirical data well. If so, the parameters are added to the posterior. A new set of parameters is proposed, for the next simulation. This procedure is iterated until a sufficiently large posterior set of parameters is obtained. Overall, an ABC approach requires a fast algorithm for simulating trees with n species. Furthermore, it requires a robust statistic, which determines if a parameter combination is added to the posterior distribution.

As it is hard to evaluate which statistics in an ABC analysis are "robust" or "good," and an ABC analysis is very time consuming, it would be preferable to have a likelihood function for the backward EBDP model. A likelihood function allows to calculate the maximum likelihood speciation and extinction rates and the mass extinction or rate shift times for a given empirical phylogenetic tree. The likelihood function also allows to do a Bayesian inference to obtain the posterior distribution of speciation and extinction rates.

There are attempts to derive a likelihood function for models with rate shifts (Paradis 1997, 1998; Rabosky 2006). However, the method proposed by Paradis (1997) looks at shifts through a sliding window without allowing for mass extinctions. Therefore, the method detects two shifts in a antisigmoid curve, although there might have been a mass extinction without a shift. Also, the approach detects a rate shift at the time when the "pull of the present" effect appeared. Paradis (1998) can detect different rates in different clades but cannot detect changes in rates at one time in the whole tree. Rabosky (2006) looks at rate shifts but assumes that the tree before a rate shift is independent of the tree after a rate shift. This is an invalid assumption though as extinction after the rate shift influences the reconstructed tree prior to the rate shift. In particular, a mass extinction event influences the LTT plot prior to the event significantly—an antisigmoid curve is produced. A challenge will be to derive a closed-form likelihood expression under the EBDP. The simulation algorithm presented here is crucial for investigating the power of such a maximum likelihood approach.

Assumptions and Limitations of the EBDP Model

I introduce the EBDP as a general model, which accounts for rate shifts and mass extinction events. The main assumptions of the model are as follows: (1) speciation rates and extinction rates are constant in the time intervals between the rate shift events, (2) species are *indistinguishable* (Aldous 2001), meaning that they all have the same rates and sampling probabilities, and (3) mass extinction affects all species in the same way.

Models have been proposed where speciation and extinction rates are more complex than under Assumption (1). For example, extinction rates might be heritable (Rabosky 2009b) or speciation rates might be density dependent (Rabosky and Lovette 2008). Such variations may introduce tree imbalance or a LTT plot with an initial steep slope followed by a flat slope. Antisigmoid LTT plots cannot be explained under these models alone. It is highly debated if and how speciation and extinction rates indeed change, see, for example, Rabosky (2009a). The EBDP model serves as a null model toward testing more complex scenarios.

Assumption (2), namely species are indistinguishable with respect to speciation and extinction rates and sampling probability, is relaxed by models with trait-specific speciation and extinction rates (Maddison 2007; FitzJohn et al. 2009). However, such models do not produce antisigmoid plots. The assumption that all extant species have the same probability of being sampled might be violated due to biased sampling. For example, species are often collected such that they are representative of the morphologic and geographic variation in the taxon sampled. Different sampling schemes can easily be incorporated into the EBDP model: we first simulate the complete trees and then apply the desired sampling procedure to the complete tree. In this paper, I only consider the sampling scheme under which each species has the same probability of being sampled, as this allowed me to compare my results with the Crisp and Cook (2009) study. Furthermore, my goal was to detect mass extinction events, and I did not want to make the LTT plots more complicated by introducing sampling artifacts.

Last, I assumed Assumption (3), that is, that each species is affected by mass extinction in the same random way. At fine scales (species within families, within local communities), there is indication that mass extinctions happen randomly among lineages (Heard and Mooers 2002). On coarser scales, the EBDP model with random mass extinctions can be used as a null model to test for more complex scenarios.

My model allows to fix both the time of rate shifts and the mass extinctions in the past and the number of sampled species. I believe that this assumption meets the a priori knowledge of biologists about their empirical data better than the assumption made implicitly by previous simulation tools, namely that changes occur once a certain number of species is *first* reached.

Funding

This work was supported by ETH Zurich.

Acknowledgments

I would like to thank the editor, the associate editor, and two referees for very valuable comments on the manuscript.

References

Aldous D.J. 2001. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. Stat. Sci. 16:23–34.

Aldous D.J., Popovic L. 2005. A critical branching process model for biodiversity. Adv. Appl. Probab. 37:1094–1115.

Crisp M., Cook L. 2009. Explosive radiation or cryptic mass extinction? Interpreting signatures in molecular phylogenies. Evolution. 63:2257–2265.

Cusimano N., Renner S. 2010. Slowdowns in diversification rates from real phylogenies may not be real. Syst. Biol. 59:458–464.

Day J., Cotton J., Barraclough T. 2008. Tempo and mode of diversification of Lake Tanganyika cichlid fishes. PloS one. 3:1730.

FitzJohn R., Maddison W., Otto S. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. Syst. Biol. 58:595–611.

Gernhard T. 2008. The conditioned reconstructed process. J. Theor. Biol. 253:769–778.

Harmon L., Weir J., Brock C., Glor R., Challenger W., Hunt G. 2007. GEIGER in R—running macroevolutionary simulation, and

estimating parameters related to diversification from comparative phylogenetic data [Internet]. Available from: http://cran.r-project.org/web/packages/geiger/index.html.
Hartmann K. 2007. TreeSample [Internet]. Available from: http://www.tb.ethz.ch/people/tstadler.
Hartmann K., Wong D., Stadler T. 2010. Sampling trees from evolutionary models. Syst. Biol. 59:465–476.
Harvey P.H., May R.M., Nee S. 1994. Phylogenies without fossils. Evolution. 48:523–529.
Heard S., Mooers A. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. Syst. Biol. 51:889.
Maddison W. 2007. Estimating a binary character's effect on speciation and extinction. Syst. Biol. 56:701–710.
Marjoram P., Molitor J., Plagnol V., Tavaré S. 2003. Markov chain Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U.S.A. 100: 15324.
Nee S.C., May R.M., Harvey P. 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. Lond. Ser. B. 344:305–311.
Paradis E. 1997. Assessing temporal variations in diversification rates from phylogenies: estimation and hypothesis testing. Proc. R. Soc. B Biol. Sci. 264:1141.
Paradis E. 1998. Detecting shifts in diversification rates without fossils. Am. Nat. 152:176–187.
Paradis E., Claude J., Strimmer K. 2004. Ape: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.
Phillimore A., Price T. 2008. Density-dependent cladogenesis in birds. PLoS Biol. 6:e71.
Popovic L. 2004. Asymptotic genealogy of a critical branching process. Ann. Appl. Probab. 14:2120–2148.
Pybus O.G., Harvey P.H. 2000. Testing macro-evolutionary models using incomplete molecular phylogenies. Proc. R. Soc. Lond. 267:2267–2272.
Rabosky D. 2006. Likelihood methods for detecting temporal shifts in diversification rates. Evolution. 60:1152–1164.
Rabosky D. 2009a. Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. Ecol. Lett. 12:735–743.
Rabosky D. 2009b. Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. Syst. Biol. 58:629.
Rabosky D. 2010. Extinction rates should not be estimated from molecular phylogenies. Evolution. 64:1816–1824.
Rabosky D., Lovette I. 2008. Density-dependent diversification in North American wood warblers. Proc. R. Soc. B Biol. Sci. 275:2363.
Rambaut A. 2002. PhyloGen: phylogenetic tree simulator package. Department of Zoology, University of Oxford. Available from: http://tree.bio.ed.ac.uk/software/.
Stadler T. 2006–2009. Cass [Internet]. Available from: http://www.tb.ethz.ch/people/tstadler.
Stadler T. 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. J. Theor. Biol. 261:58–66.
Stadler T. 2010. TreeSim in R—simulating trees under the birth-death model [Internet]. Available from: http://cran.r-project.org/web/packages/TreeSim/index.html.
Yule G.U. 1924. A mathematical theory of evolution: based on the conclusions of Dr. J.C. Willis. Philos. Trans. R. Soc. Lond. Ser. B. 213: 21–87.

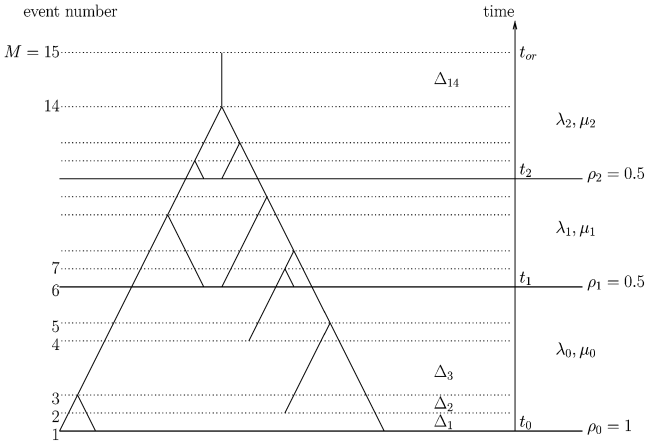

FIGURE A1. EBDP tree: tree that evolved under an EBDP model with $m = 2$ mass extinction events at times t_1 and t_2. The events (speciation or extinction or mass extinction) are numbered increasing going back in time, and the length of the time interval between event j and $j + 1$ is Δ_j.

APPENDIX

I show in the following that obtaining a tree τ with the EBDP backward algorithm has the same probability density as obtaining the tree τ with the naive algorithm. Because the naive algorithm is correct as it simulates the model by exactly following the model definition, I establish that the EBDP backward algorithm is correct.

I will first derive the probability density of obtaining a fixed n-species-tree τ with the EBDP backward algorithm. The events happening in a tree are speciation, extinction, and mass extinction/rate shift. We order all events occurring in tree τ timewise, starting at t_0 with event number 1 and the origin being event number M (see Fig. A1). Let the time interval between event j and $j + 1$ have length Δ_j, $j \in \{1, \ldots, M - 1\}$. Note that M corresponds to the event of the origin appearing. Let n_j be the number of species in interval Δ_j, and let $\tilde{\lambda}_j, \tilde{\mu}_j$ be the speciation and extinction rate in interval Δ_j. Now, to obtain the density of a fixed tree, we have to multiply the probabilities of having the correct events (speciation, extinction, and rate shift) and multiply the probability densities of observing all the interval lengths between events, going back in time.

Let event $j + 1$ be a speciation event. The probability density of observing interval length Δ_j between event j and $j + 1$ and observing event $j + 1$ is (combining (3) and (6) in the algorithm),

$$\frac{\tilde{\lambda}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}(\tilde{\lambda}_j + \tilde{\mu}_j)e^{-(\tilde{\lambda}_j + \tilde{\mu}_j)n_j\Delta_j}.$$

Note that the origin of the tree is considered by the backward algorithm as a speciation event.

Similarly, when event $j + 1$ is an extinction event, the probability density of observing interval length Δ_j between event j and $j + 1$ and observing event $j + 1$ is (combining (3) and (6) in the algorithm),

$$\frac{\tilde{\mu}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}(\tilde{\lambda}_j + \tilde{\mu}_j)e^{-(\tilde{\lambda}_j + \tilde{\mu}_j)n_j\Delta_j}.$$

For $j + 1$ being a mass extinction/rate shift event, the probability density of observing interval length Δ_j between event j and $j + 1$ is the probability that nothing happens between j and $j + 1$,

$$e^{-(\tilde{\lambda}_j + \tilde{\mu}_j)n_j\Delta_j}.$$

Furthermore, as we choose the species undergoing an event uniformly at random, each tree is equally likely, given the sequence of the M events with the interval lengths Δ_j. Therefore, the probability density of the tree τ is,

$$f_{\text{backward}}(\tau) \propto \frac{1/\tilde{\lambda}_M}{\sum_{j=0}^{m} 1/\lambda_j} \prod_{j=1}^{M-1} e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j} \prod_{\substack{j=2 \\ j\text{: spec. event}}}^{M} \tilde{\lambda}_j \prod_{\substack{j=2 \\ j\text{: ext. event}}}^{M-1} \tilde{\mu}_j$$

$$\propto \prod_{j=1}^{M-1} e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j} \prod_{\substack{j=2 \\ j\text{: spec. event}}}^{M-1} \tilde{\lambda}_j \prod_{\substack{j=2 \\ j\text{: ext. event}}}^{M-1} \tilde{\mu}_j.$$

For calculating the probability density of obtaining tree τ under the naive algorithm, we go from the past to the present. Recall that we have a uniform prior for the origin of the tree. For now, let the prior be defined on $(0, C)$. With probability density $1/C$, we sample the right time of origin. Then, we calculate the probability density of observing interval length Δ_j followed by the right event. Let event j be a speciation event. The probability density of observing interval length Δ_j and observing event j is,

$$\frac{\tilde{\lambda}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}(\tilde{\lambda}_j + \tilde{\mu}_j)e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j}.$$

Analogue, when event j is an extinction event, the probability density of observing interval length Δ_j and observing event j is,

$$\frac{\tilde{\mu}_j}{\tilde{\lambda}_j + \tilde{\mu}_j}(\tilde{\lambda}_j + \tilde{\mu}_j)e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j}.$$

For j being a mass extinction/rate shift event, the probability density of observing Δ_j is the probability that nothing happens between j and $j+1$,

$$e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j}.$$

Furthermore, as we choose the species undergoing an event uniformly at random, each tree is equally likely, given the sequence of the M events with the interval lengths Δ_j. Therefore, the probability density of the tree τ is,

$$f_{\text{naive}}(\tau) \propto \frac{1}{C} \prod_{j=1}^{M-1} e^{-(\tilde{\lambda}_j+\tilde{\mu}_j)n_j\Delta_j} \prod_{\substack{j=2 \\ j\text{: spec. event}}}^{M-1} \tilde{\lambda}_j \prod_{\substack{j=2 \\ j\text{: ext. event}}}^{M-1} \tilde{\mu}_j$$

$$\propto f_{\text{backward}}(\tau).$$

This shows that for all C, in particular $C \to \infty$, the tree τ has the same chance of being simulated under the naive and the backward algorithm. This establishes the correctness of the EBDP backward algorithm.