# Phylogenetic Inference using `RevBayes`

*Diversification Rate Estimation with Missing Taxa*

Sebastian Höhna

## 1 Overview: Diversification Rate Estimation

Models of speciation and extinction are fundamental to any phylogenetic analysis of macroevolutionary processes. A prior describing the distribution of speciation events over time is critical to estimating phylogenies with branch lengths proportional to time. Moreover, stochastic branching models allow for inference of speciation and extinction rates. These inferences allow us to investigate key questions in evolutionary biology.

Similarly, diversification-rate parameters are also included as nuisance parameters of other phylogenetic models—*i.e.,* where these diversification-rate parameters are not of direct interest. For example, many methods for estimating species divergence times—such as `BEAST` (Drummond et al. 2012), `MrBayes` (Ronquist et al. 2012), and `RevBayes` (Höhna et al. 2016)—implement 'relaxed-clock models' that include a constant-rate birth-death branching process as a prior model on the distribution of tree topologies and node ages. Although the parameters of these 'tree priors' are not typically of direct interest, they are nevertheless estimated as part of the joint posterior probability distribution of the relaxed-clock model, and so can be estimated simply by querying the corresponding marginal posterior probability densities. In fact, this may provide more robust estimates of the diversification-rate parameters, as they accommodate uncertainty in the other phylogenetic-model parameters (including the tree topology, divergence-time estimates, and the other relaxed-clock model parameters).

### 1.1 Types of Hypotheses for Estimating Diversification Rates

Many evolutionary phenomena entail differential rates of diversification (speciation – extinction); *e.g.,* adaptive radiation, diversity-dependent diversification, key innovations, and mass extinction. The specific study questions regarding lineage diversification may be classified within three fundamental categories of inference problems. Admittedly, this classification scheme is somewhat arbitrary, but it is nevertheless useful, as it allows users to navigate the ever-increasing number of available phylogenetic methods. Below, we describe each of the fundamental questions regarding diversification rates.

**(1) Diversification-rate through time estimation** *What is the (constant) rate of diversification in my study group?* The most basic models estimate parameters of the stochastic-branching process (*i.e.,* rates of speciation and extinction, or composite parameters such as net-diversification and relative-extinction rates) under the assumption that rates have remained constant across lineages and through time; *i.e.,* under a constant-rate birth-death stochastic-branching process model. Extensions to the (basic) constant-rate models include diversification-rate variation through time. First, we might ask whether there is evidence of an episodic, tree-wide increase in diversification rates (associated with a sudden increase in speciation rate and/or decrease in extinction rate), as might occur during an episode of adaptive radiation. A second question asks whether there is evidence of a continuous/gradual decrease in diversification rates

through time (associated with decreasing speciation rates and/or increasing extinction rates), as might occur because of diversity-dependent diversification (*i.e.,* where competitive ecological interactions among the species of a growing tree decrease the opportunities for speciation and/or increase the probability of extinction). A final question in this category asks whether our study tree was impacted by a mass-extinction event (where a large fraction of the standing species diversity is suddenly lost).

**(2) Diversification-rate variation across branches estimation** *Is there evidence that diversification rates have varied significantly across the branches of my study group?* Models have been developed to detect departures from rate constancy across lineages; these tests are analogous to methods that test for departures from a molecular clock—*i.e.,* to assess whether substitution rates vary significantly across lineages. These models are important for assessing whether a given tree violates the assumptions of other inference methods. Furthermore, these models are important to answer questions such as: *What are the branch-specific diversification rates?*; and *Have there been significant diversification-rate shifts along branches in my study group, and if so, how many shifts and along which branches?*

**(3) Character-dependent diversification-rate estimation** *Are diversification rates correlated with some variable in my study group?* Character-dependent diversification-rate models aim to identify overall correlations between diversification rates and organismal features (binary and multi-state discrete morphological traits, continuous morphological traits, geographic range, etc.). For example, one can hypothesize that a binary character, say if an organism is herbivorous/carnivorous or self-compatible/self-incompatible, impact the diversification rates. Then, if the organism is in state 0 (*e.g.,* is herbivorous) it has a lower (or higher) diversification rate than if the organism is in state 1 (*e.g.,* carnivorous).

# 2 Models

We begin this section with a general introduction to the stochastic birth-death branching process that underlies inference of diversification rates in `RevBayes`. This primer will provide some details on the relevant theory of stochastic-branching process models. We appreciate that some readers may want to skip this somewhat technical primer; however, we believe that a better understanding of the relevant theory provides a foundation for performing better inferences. We then discuss a variety of specific birth-death models, but emphasize that these examples represent only a tiny fraction of the possible diversification-rate models that can be specified in `RevBayes`.

## 2.1 The birth-death branching process

Our approach 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. Let $N(t)$ denote the number of species at time $t$. Assume the process starts at time $t_1$ (the 'crown' age of the most recent common ancestor of the study group, $t_{\text{MRCA}}$) when there are two species. Thus, the process is initiated with two species, $N(t_1) = 2$. We condition the process on sampling at least one descendant from each of these initial two lineages; otherwise $t_1$ would not correspond to the $t_{\text{MRCA}}$ of our study group. Each lineage evolves independently of all other lineages, giving rise to exactly one new lineage with rate $b(t)$ and losing one existing lineage with rate $d(t)$ (Figure 1 and Figure 2). Note that although each lineage evolves independently, all lineages share both a common (tree-wide) speciation rate $b(t)$ and a common extinction rate $d(t)$ (Nee et al. 1994; Höhna 2015). Additionally, at certain times, $t_{\mathbb{M}}$, a mass-extinction event occurs and each species existing at that time has the same probability, $\rho$, of survival. Finally, all extinct lineages are pruned and only the reconstructed tree remains (Figure 1).

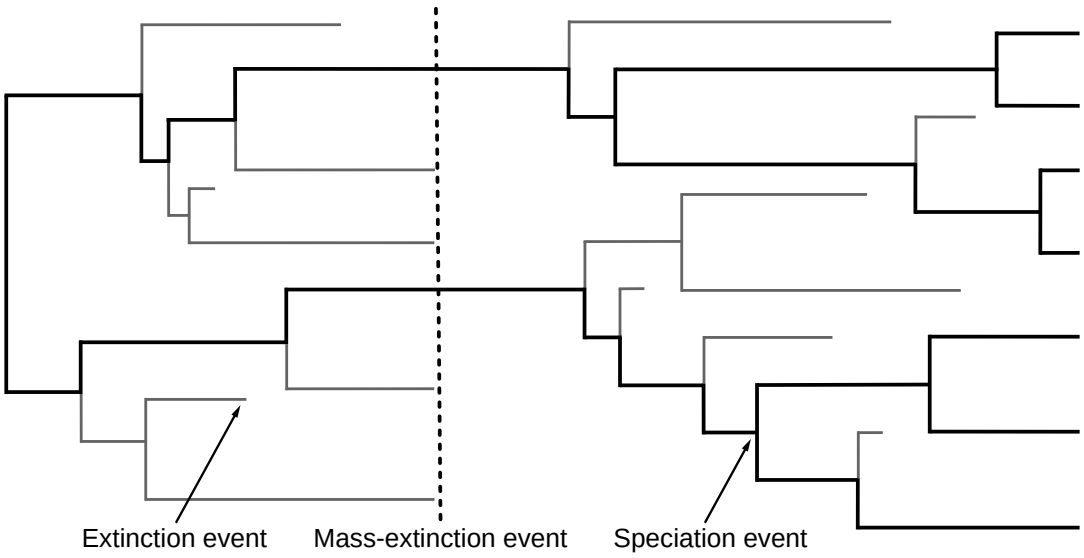


Figure 1: A realization of the birth-death process with mass extinction. Lineages that have no extant or sampled descendant are shown in gray and surviving lineages are shown in a thicker black line.

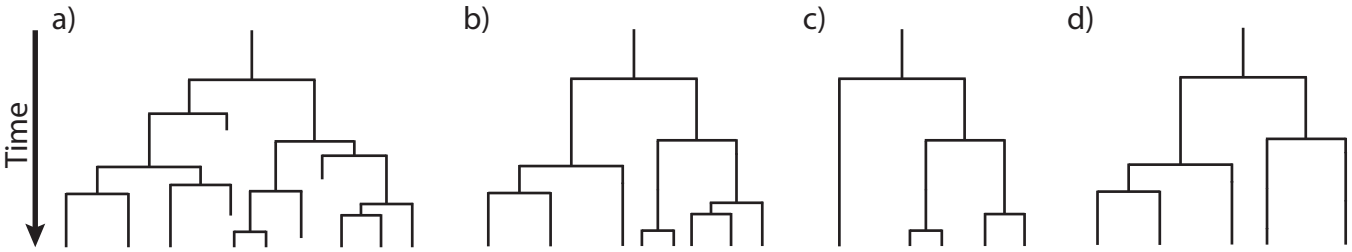


Figure 2: **Examples of trees produced under a birth-death process.** The process is initiated at the first speciation event (the 'crown-age' of the MRCA) when there are two initial lineages. At each speciation event the ancestral lineage is replaced by two descendant lineages. At an extinction event one lineage simply terminates. (A) A complete tree including extinct lineages. (B) The reconstructed tree of tree from A with extinct lineages pruned away. (C) A *uniform* subsample of the tree from B, where each species was sampled with equal probability, $\rho$. (D) A *diversified* subsample of the tree from B, where the species were selected so as to maximize diversity.

To condition the probability of observing the branching times on the survival of both lineages that descend from the root, we divide by $P(N(T) > 0|N(0) = 1)^2$. Then, the probability density of the branching times, $\mathbb{T}$, becomes

$$P(\mathbb{T}) = \frac{\overbrace{P(N(T) = 1 \mid N(0) = 1)^2}^{\text{both initial lineages have one descendant}}}{\underbrace{P(N(T) > 0 \mid N(0) = 1)^2}_{\text{both initial lineages survive}}} \times \prod_{i=2}^{n-1} \overbrace{i \times b(t_i)}^{\text{speciation rate}} \times \overbrace{P(N(T) = 1 \mid N(t_i) = 1)}^{\text{lineage has one descendant}},$$

and the probability density of the reconstructed tree (topology and branching times) is then

$$P(\Psi) = \frac{2^{n-1}}{n!(n-1)!} \times \left( \frac{P(N(T)=1 \mid N(0)=1)}{P(N(T)>0 \mid N(0)=1)} \right)^2 \times \prod_{i=2}^{n-1} i \times b(t_i) \times P(N(T)=1 \mid N(t_i)=1) \quad (1)$$

We can expand Equation (1) by substituting $P(N(T)>0 \mid N(t)=1)^2 \exp(r(t,T))$ for $P(N(T)=1 \mid N(t)=1)$, where $r(u,v) = \int_u^v d(t) - b(t)dt$; the above equation becomes

$$\begin{aligned} P(\Psi) &= \frac{2^{n-1}}{n!(n-1)!} \times \left( \frac{P(N(T)>0 \mid N(0)=1)^2 \exp(r(0,T))}{P(N(T)>0 \mid N(0)=1)} \right)^2 \\ &\quad \times \prod_{i=2}^{n-1} i \times b(t_i) \times P(N(T)>0 \mid N(t_i)=1)^2 \exp(r(t_i,T)) \\ &= \frac{2^{n-1}}{n!} \times \Big( P(N(T)>0 \mid N(0)=1) \exp(r(0,T)) \Big)^2 \\ &\quad \times \prod_{i=2}^{n-1} b(t_i) \times P(N(T)>0 \mid N(t_i)=1)^2 \exp(r(t_i,T)). \end{aligned} \quad (2)$$

For a detailed description of this substitution, see Höhna (2015). Additional information regarding the underlying birth-death process can be found in (Thompson 1975; Equation 3.4.6) and Nee et al. (1994) for constant rates and Lambert (2010); Lambert and Stadler (2013); Höhna (2013; 2014; 2015) for arbitrary rate functions.

To compute the equation above we need to know the rate function, $r(t,s) = \int_t^s d(x) - b(x)dx$, and the probability of survival, $P(N(T)>0|N(t)=1)$. Yule (1925) and later Kendall (1948) derived the probability that a process survives ($N(T)>0$) and the probability of obtaining exactly $n$ species at time $T$ ($N(T)=n$) when the process started at time $t$ with one species. Kendall's results were summarized in Equation (3) and Equation (24) in Nee et al. (1994)

$$P(N(T)>0|N(t)=1) = \left( 1 + \int_t^T \Big( \mu(s) \exp(r(t,s)) \Big) ds \right)^{-1} \quad (3)$$

$$\begin{aligned} P(N(T)=n|N(t)=1) &= (1 - P(N(T)>0|N(t)=1) \exp(r(t,T)))^{n-1} \\ &\quad \times P(N(T)>0|N(t)=1)^2 \exp(r(t,T)) \end{aligned} \quad (4)$$

An overview for different diversification models is given in Höhna (2015).

# 3 Estimating Speciation & Extinction Rates Through Time

## 3.1 Outline

This tutorial describes how to specify different models of incomplete taxon sampling for estimating diversification rates in `RevBayes` (Höhna et al. 2011; Höhna 2014; Höhna et al. 2016). Specifically, we will discuss *uniform*, *diversified*, and *empirical* taxon sampling. All analyses in this tutorial will focus on diversification rate estimation through-time and thus use a birth-death process where diversification rates

vary episodically through time model by piecewise constant rates `RevBayes` (Höhna 2015; May et al. 2016). The probabilistic graphical model is given only once for this tutorial. Finally, you will estimate speciation and extinction rates through-time using Markov chain Monte Carlo (MCMC) and assess the impact of incomplete taxon sampling as well as the sampling scheme.

### 3.2 Requirements

We assume that you have read and hopefully completed the following tutorials:

- RB_Getting_Started
- RB_Basics_Tutorial
- RB_BasicDiversificationRate_Tutorial

Note that the RB_Basics_Tutorial introduces the basic syntax of `Rev` but does not cover any phylogenetic models. You may skip the RB_Basics_Tutorial if you have some familiarity with `R`. We tried to keep this tutorial very basic and introduce all the language concepts and theory on the way. You may only need the RB_Basics_Tutorial for a more in-depth discussion of concepts in `Rev`.

## 4 Data and files

We provide the data file(s) which we will use in this tutorial. You may want to use your own data instead. In the **`data`** folder, you will find the following files

- **`primates_springer.tre`**: Dated primates phylogeny including 369 out of 450 species from Springer et al. (2012).

→ Open the tree **`data/primates_springer.tre`** in `FigTree`.

## 5 Episodic Birth-Death Model

The basic idea behind the model is that speciation and extinction rates are piecewise constant but can be different for different time intervals. Thus, we will divide time into equal time interval with the only exception that the first 20% of the time do not have any rate changes. Our only reason to do so is because there are two few lineages in the reconstructed tree at that time to obtain reliable parameter estimates. An overview of the underlying theory of the specific model and implementation is given in (Höhna 2015).

We assume that the log-transformed rates are drawn from a normal distribution. Furthermore, we will assume that rates are autocorrelated, that is, rates in the next time interval will be centered around the rates in the current time interval. The assumption of autocorrelated rates does not only makes sense biologically but also improves our ability to estimate parameters.

### 5.1 Read the tree

Begin by reading in the observed tree.

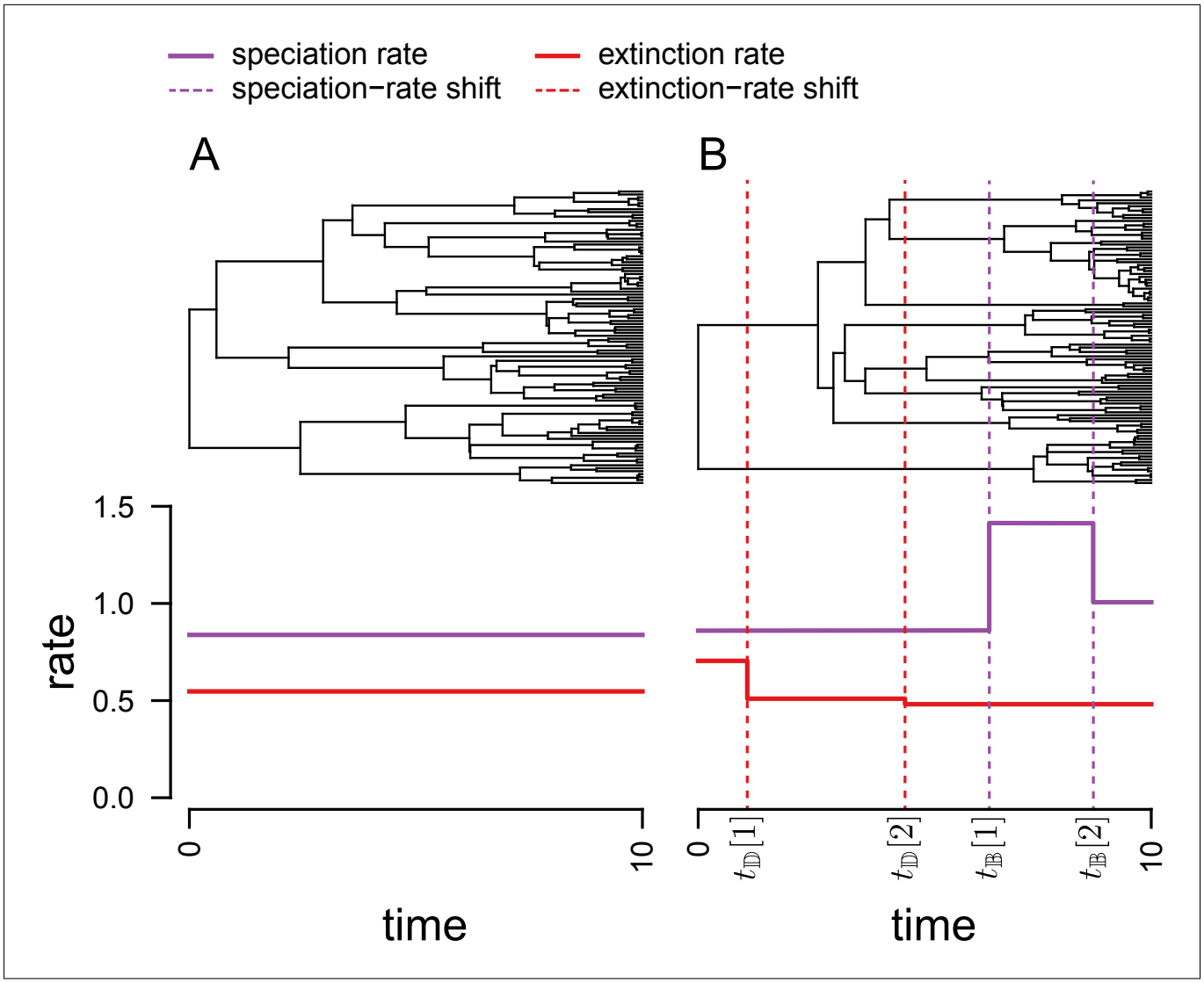


Figure 3: Four scenarios of birth-death models.

```
T <- readTrees("data/primates_springer.tre")[1]
```

From this tree, we can get some helpful variables:

```
taxa <- T.taxa()
```

Additionally, we can initialize an iterator variable for our vector of moves:

```
mvi = 0
```

Finally, we create a helper variable that specifies the number of intervals.

```
NUM_INTERVALS = 10
```

Using this variable we can easily change our script to break-up time into many or few intervals.

## 5.2 Specifying the model

### 5.2.1 Priors on rates

We start by specifying prior distributions on the rates. Each interval specific speciation and extinction rate will be drawn from a normal distribution. Thus, we a parameter for the standard deviation of those normal distributions. We use an exponential hyperprior with rate 1.0 to estimate the standard deviation, but assume that all speciation rates and all extinction rates share the same standard deviation. The motivation for an exponential hyperprior is that it has the highest probability density at 0 which would make he variance between rates 0 too and thus represent a constant rate process. The data will tell us if there should be much variation in rates through time.

```
speciation_sd ~ dnExponential(1.0)
extinction_sd ~ dnExponential(1.0)
```

We apply a simple scaling move on each prior parameter.

```
moves[++mvi] = mvScale(speciation_sd,weight=5.0)
moves[++mvi] = mvScale(extinction_sd,weight=5.0)
```

The second prior parameter that we need to specify is mean of the speciation and extinction rate at present. This is because we are actually modeling rate-changes backwards in time and there is no previous rate for the rate at the present. Thus we use a uniform distribution between -5 and 5 because of lack of prior knowledge.

```
# draw the mean from a uniform distribution
speciation_prior_mean ~ dnUniform(-5.0,5.0)
extinction_prior_mean ~ dnUniform(-5.0,5.0)
```

This time we will apply a simple sliding window move because both parameters are location parameters instead of scale or variance parameters.

```
moves[++mvi] = mvSlide(speciation_prior_mean,weight=5.0)
moves[++mvi] = mvSlide(extinction_prior_mean,weight=5.0)
```

### 5.2.2 Specifying episodic rates

As we mentioned before, we will apply normal distributions as priors for each rate. We begin with the rate at the present. The rates at the present will be specified slightly differently because they are not correlated to any previous rates. Note that we store the variables in vectors.

```
log_speciation[1] ~ dnNormal( mean=speciation_prior_mean, sd=speciation_sd )
log_extinction[1] ~ dnNormal( mean=extinction_prior_mean, sd=extinction_sd )
```

Again, we apply simple sliding window moves for the rates. Normally we would use scaling moves but in this case we work on the log-transformed parameters and thus sliding moves perform better. If you are keen you can test the differences.

```
moves[++mvi] = mvSlide(log_speciation[1], weight=2)
moves[++mvi] = mvSlide(log_extinction[1], weight=2)
```

Now we transform the parameters.

```
speciation[1] := exp( log_speciation[1] )
extinction[1] := exp( log_extinction[1] )
```

Then we repeat the specification for the speciation and extinction rates for each time interval. This can be done efficiently using a **for**-loop. We will use a specific index variable so that we can easier refer to the rate at the previous interval.

```
for (i in 1:NUM_INTERVALS) {
    index = i+1

    log_speciation[index] ~ dnNormal( mean=log_speciation[i], sd=speciation_sd )
    log_extinction[index] ~ dnNormal( mean=log_extinction[i], sd=extinction_sd )

    moves[++mvi] = mvSlide(log_speciation[index], weight=2)
    moves[++mvi] = mvSlide(log_extinction[index], weight=2)

    speciation[index] := exp( log_speciation[index] )
    extinction[index] := exp( log_extinction[index] )

}
```

Finally, we apply moves that slide all values in the vectors, *i.e.,* all speciation or extinction rates, by the same amount. This again considerably improves the efficiency of our MCMC analysis.

```
moves[++mvi] = mvVectorSlide(log_speciation, weight=10)
moves[++mvi] = mvVectorSlide(log_extinction, weight=10)
```

### 5.2.3 Setting up the time inverals

In RevBayes you actually have the possibility unequal time intervals or even different intervals for the speciation and extinction rate. This is achieved by providing a vector of times when each interval ends. Here we simply break-up the most recent 80% of time since the root in equal intervals.

```
interval_times = T.rootAge() * (1:NUM_INTERVALS) / (NUM_INTERVALS) * 0.8
```

### 5.2.4 Incomplete Taxon Sampling

We know that we have sampled 367 out of 450 living primate species. To account for this we can set the sampling parameter as a constant node with a value of 369/450

```
rho <- T.ntips()/450
```

### 5.2.5 Root age

The birth-death process requires a parameter for the root age. In this exercise we use a fix tree and thus we know the age of the tree. Hence, we can get the value for the root from the Springer et al. (2012) tree.

```
root_time <- T.rootAge()
```

### 5.2.6 The time tree

Now we have all of the parameters we need to specify the full episodic birth-death model. We initialize the stochastic node representing the time tree.

```
timetree ~ dnEpisodicBirthDeath(rootAge=T.rootAge(), lambdaRates=speciation,
    lambdaTimes=interval_times, muRates=extinction, muTimes=interval_times, rho=rho,
    taxa=taxa)
```

And then we attach data to it.

```
timetree.clamp(T)
```

Finally, we create a workspace object of our whole model using the **model()** function.

```
mymodel = model(speciation)
```

The **model()** function traversed all of the connections and found all of the nodes we specified.

## 5.3 Running an MCMC analysis

### 5.3.1 Specifying Monitors

For our MCMC analysis, we need to set up a vector of *monitors* to record the states of our Markov chain. First, we will initialize the model monitor using the **mnModel** function. This creates a new monitor variable that will output the states for all model parameters when passed into a MCMC function.

```
monitors[1] = mnModel(filename="output/primates_BSBD.log",printgen=10, separator = TAB
    )
```

Additionally, we create two separate file monitors, one for each vector of speciation and extinction rates. We want to have the speciation and extinction rates stored separately so that we can plot them nicely afterwards.

```
monitors[2] = mnFile(filename="output/primates_EBD_speciation.log",printgen=10,
    separator = TAB, speciation)
monitors[3] = mnFile(filename="output/primates_EBD_extinction.log",printgen=10,
    separator = TAB, extinction)
```

Finally, create a screen monitor that will report the states of specified variables to the screen with **mnScreen**:

```
monitors[4] = mnScreen(printgen=1000, speciation)
```

### 5.3.2 Initializing and Running the MCMC Simulation

With a fully specified model, a set of monitors, and a set of moves, we can now set up the MCMC algorithm that will sample parameter values in proportion to their posterior probability. The **mcmc()** function will create our MCMC object:

```
mymcmc = mcmc(mymodel, monitors, moves)
```

First, we will run a pre-burnin to tune the moves and to obtain starting values from the posterior distribution.

```
mymcmc.burnin(generations=10000,tuningInterval=200)
```

Now, run the MCMC:

```
mymcmc.run(generations=50000)
```

When the analysis is complete, you will have the monitored files in your output directory. You can then visualize the rates through time using R. We have provided a very simple example script for you. Just start R in the main directory for this analysis and then type the following commands:

```
source("RevBayes_scripts/Plot_EBD_RevBayes.R")
plot.EBD("output/primates_EBD","Episodic_Rates")
```

→ The Rev file for performing this analysis: mcmc_EBD.Rev. An R file for plotting the output: Plot_EBD_RevBayes.R.

### 5.4 Exercise

- Run an MCMC simulation to estimate the posterior distribution of the speciation rate and extinction rate.
- Visualize the rate through time using R.
- Do you see evidence for rate decreases or increases? What is the general trend?
- Run the analysis using a different number of intervals, *e.g.,* 5 or 50. How do the rates change?
- Modify the model by specifying a prior on the log-diversification and log-turnover rate and then estimate the diversification rates through time. Do you see any differences in the estimates?

## References

Drummond, A., M. Suchard, D. Xie, and A. Rambaut. 2012. Bayesian phylogenetics with beauti and the beast 1.7. Molecular Biology and Evolution 29:1969–1973.

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. Journal of Theoretical Biology 380:321–331.

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Höhna, S., T. Stadler, F. Ronquist, and T. Britton. 2011. Inferring speciation and extinction rates under different species sampling schemes. Molecular Biology and Evolution 28:2577–2589.

Kendall, D. G. 1948. On the generalized "birth-and-death" process. The Annals of Mathematical Statistics 19:1–15.

Lambert, A. 2010. The contour of splitting trees is a lévy process. The Annals of Probability 38:348–395.

Lambert, A. and T. Stadler. 2013. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. Theoretical Population Biology 90:113–128.

May, M. R., S. Höhna, and B. R. Moore. 2016. A Bayesian Approach for Detecting the Impact of Mass-Extinction Events on Molecular Phylogenies When Rates of Lineage Diversification May Vary. Methods in Ecology and Evolution 7:947–959.

Nee, S., R. M. May, and P. H. Harvey. 1994. The Reconstructed Evolutionary Process. Philosophical Transactions: Biological Sciences 344:305–311.

Ronquist, F., M. Teslenko, P. van der Mark, D. L. Ayres, A. Darling, S. Höhna, B. Larget, L. Liu, M. A. Suchard, and J. P. Huelsenbeck. 2012. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Springer, M. S., R. W. Meredith, J. Gatesy, C. A. Emerling, J. Park, D. L. Rabosky, T. Stadler, C. Steiner, O. A. Ryder, J. E. Janečka, et al. 2012. Macroevolutionary dynamics and historical biogeography of primate diversification inferred from a species supermatrix. PLoS One 7:e49521.

Thompson, E. 1975. Human evolutionary trees. Cambridge University Press Cambridge.

Yule, G. 1925. A mathematical theory of evolution, based on the conclusions of dr. jc willis, frs. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 213:21–87.

Version dated: August 12, 2016