Phylogenetic Inference using `RevBayes`

Branch-Specific Diversification Rate Estimation

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. 2015)—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_{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_{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, ρ, of survival. Finally, all extinct lineages are pruned and only the reconstructed tree remains (Figure 1).

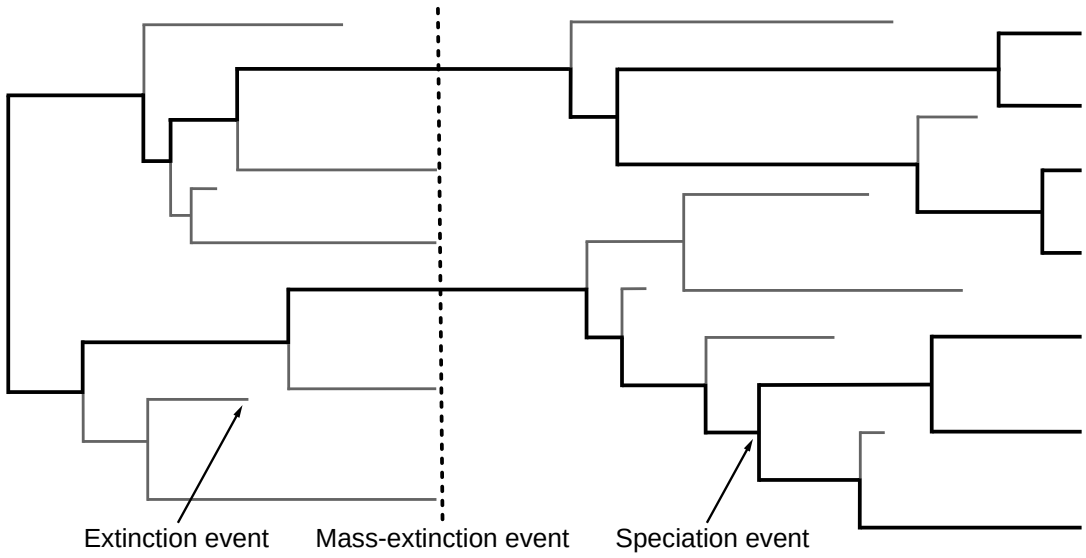

Figure 1: A realization of the birth-death process with mass extinction. Lineages that have no extant or sampled descendant are shown in gray and surviving lineages are shown in a thicker black line.

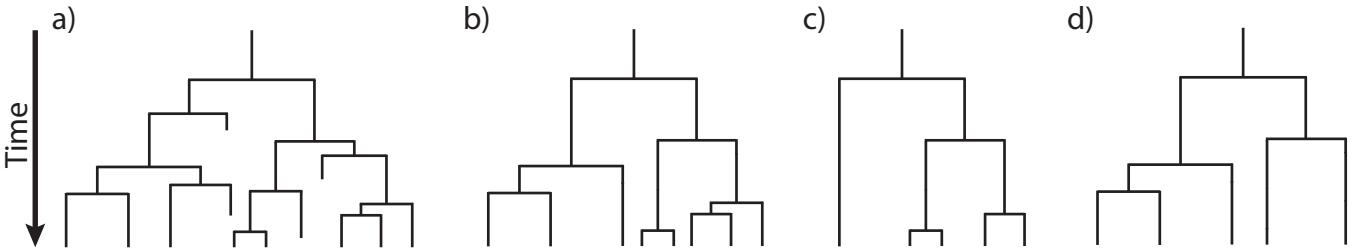

Figure 2: **Examples of trees produced under a birth-death process.** The process is initiated at the first speciation event (the 'crown-age' of the MRCA) when there are two initial lineages. At each speciation event the ancestral lineage is replaced by two descendant lineages. At an extinction event 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, ρ. (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 Branch-Specific Speciation & Extinction Rates

3.1 Outline

This tutorial describes how to specify a branch-specific branching-process models in `RevBayes`; a birth-death process where diversification rates vary among branches, similar to (Rabosky 2014). The probabilistic graphical model is given for each component of this tutorial. The goal is to obtain estimate of branch-specific diversification rates using Markov chain Monte Carlo (MCMC).

3.2 Requirements

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

- RB_Getting_Started
- RB_Basics_Tutorial
- RB_BayesFactor_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`. The RB_BayesFactor_Tutorial introduced Bayesian model selection by means of Bayes factors, which can be skipped by readers familiar with Bayesian model selection. 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 Branch-Specific Birth-Death Model

The basic idea behind the model is that speciation and extinction rates are allowed to vary across branches of the tree (see Figure 3). Unfortunately, it is not possible to model rates drawn from a continuous distribution directly, as done for example in `BAMM`, because in that case one needs to integrate over any number of possible rate shifts, any time of these shifts and most importantly over all possible new rates. This is unfeasible to do and failure to do so has been shown to make parameter estimates unreliable.

Here we adopt an approach using few discrete rate categories instead. This allows us to numerically integrate over all possible rate categories using a system of differential equation originally described by Maddison et al. (2007) (see also FitzJohn et al. (2009) and FitzJohn (2010)). The numerical procedure beaks time into very small time interval and sums over all possible events occurring in that interval (see Figure 4).

→ You don't need to worry about any of the technical details. It is important for you to realize that this model assumes that new rates at a rate-shift event are drawn from a given set (discrete) of rates.

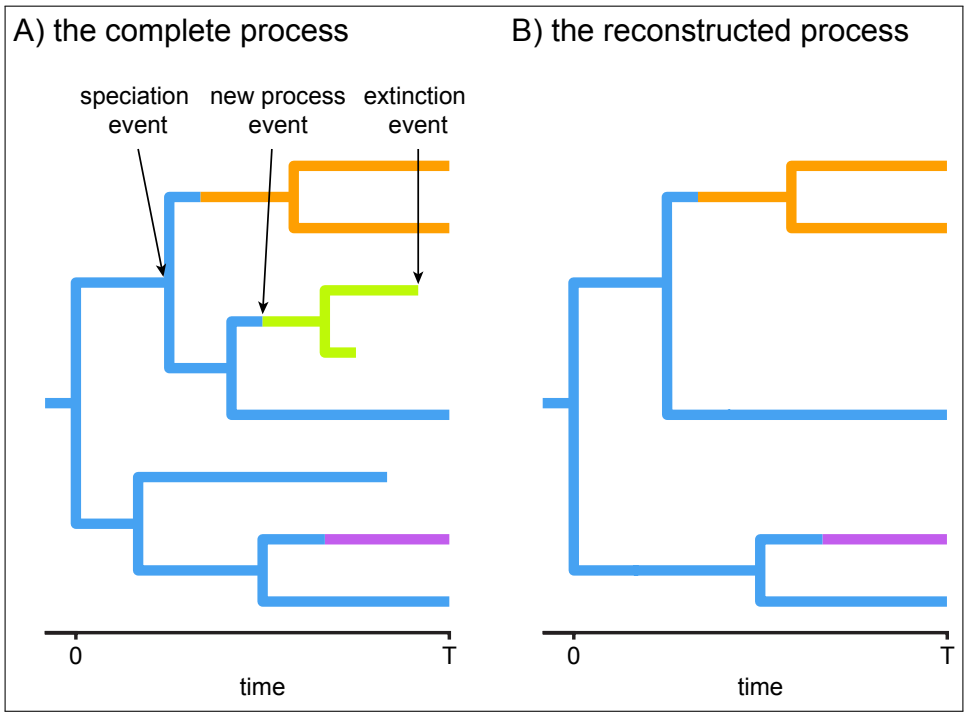

Figure 3: Cartoon of a branch-specific birth-death process. On the left we see the full process. On the right we only see the branches of the reconstructred tree, thus missing one rate-shift event.

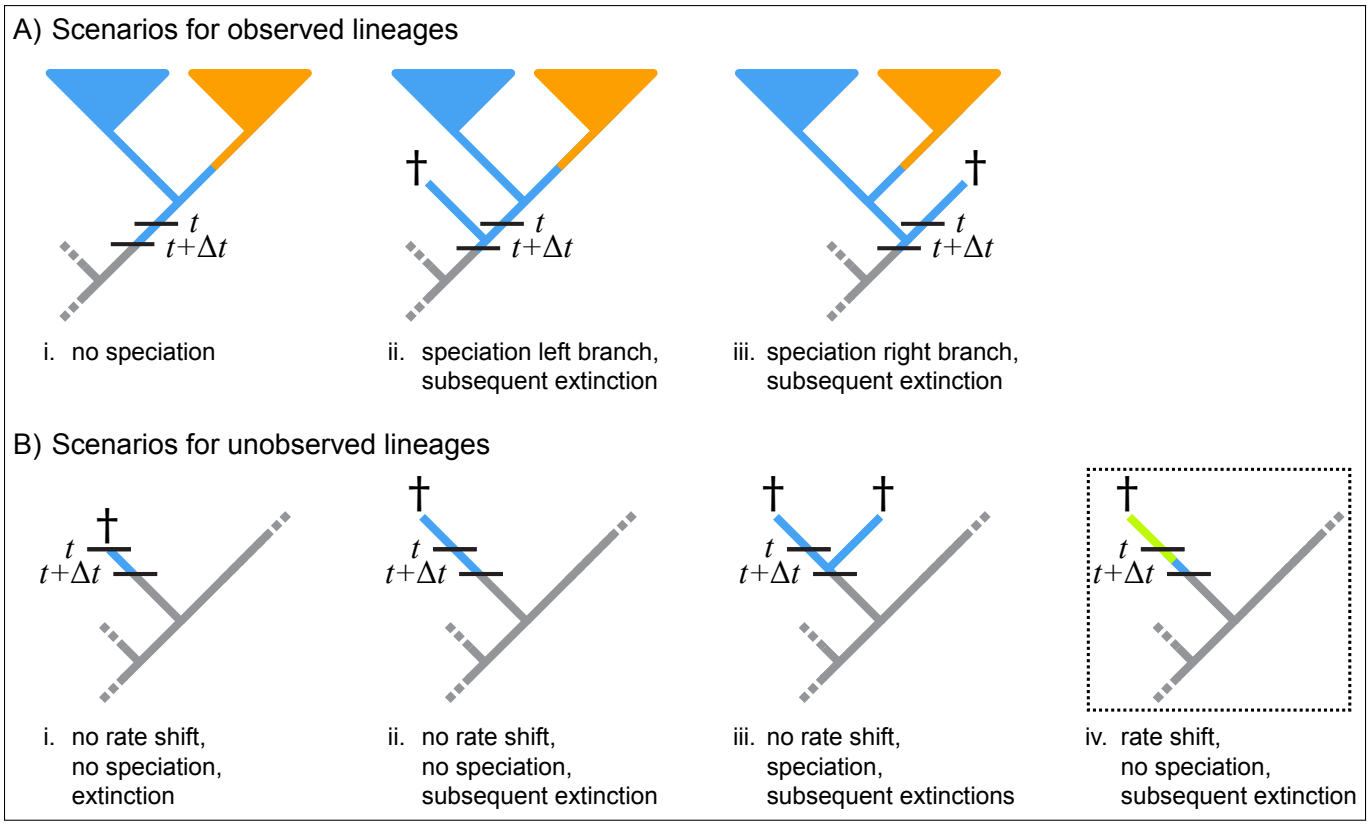

Figure 4: Cartoon of the likelihood computation using numerical integration.

5.1 Read the tree

Begin by reading in the observed tree.

```
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:

```
mi = 0
```

Finally, we create a helper variable that specifies the number of discrete rate categories.

```
NUM_RATE_CATEGORIES = 10
```

Using this variable we can easily change our script to use more or fewer categories and test the impact.

5.2 Specifying the model

5.2.1 Priors on rates

Instead of using a continuous probability distribution we will use a discrete approximation of the distribution, as done for modeling rate variation across sites (Yang 1994) and for modeling relaxed molecular clocks (Drummond et al. 2006). That means, we assume that the speciation rates are drawn from one of the N quantiles of the lognormal distribution. For this we will use the function **fnDiscretizeDistribution** which takes in a distribution as its first argument and the number of quantile as the second argument. The return value is a vector of quantiles. We use it as a deterministic variable and every time the parameters of the base distribution (*i.e.,* the lognormal distribution in our case) change the quantiles will update automatically as well. Thus we only need to specify parameters for our base distribution, the lognormal distribution. We choose a stochastic variable for the mean parameter of the lognormal distribution drawn from a uniform prior distribution between -10 and 10 signifying a large amount of uncertainty. Consequently we get an expected speciation rate between $exp(-10) = 4.54e-05$ and $exp(-10) = 22026.47$, which should definitely cover the range of speciation rates. Additionally, we choose a fixed standard deviation of $0.587405 * 2$ for the speciation rates because it represents two orders of magnitude variance in the rate categories.

```
mean_speciation ~ dnUniform(-10,10)
moves[++mi] = mvSlide(mean_speciation,delta=1,tune=true,weight=5)
sd_speciation <- 0.587405*2
```

```
speciation := fnDiscretizeDistribution( dnLognormal(mean_speciation, sd_speciation),
    NUM_RATE_CATEGORIES )
```

For each of the speciation rate categories we need a extinction rate category. We are completely free to choose how we construct these rate categories. For example, we could chose a similar discretization of a lognormal distribution using its quantiles to provide different extinction rate categories. The only drawback of this approach is that low speciation rates are paired by definition with low extinction rates and high extinction rates are paired with high extinction rates. This is because rates are pared by their index, thus the speciation rate with index j is paired with the extinction rate with index j. We could avoid this by estimating N independent rate categories instead of assuming that the rate categories are computed by the quantile function.

Another option is to simply assume that the relative extinction rate is equal for all rate category. We use this model choice in this exercise. Hence, we only need to create one random variable for the relative extinction rate.

```
relative_extinction ~ dnBeta(1,1)
moves[++mi] = mvSlide(relative_extinction,delta=1,tune=true,weight=5)
```

Then we multiply each speciation rate category by the relative extinction rate to obtain the extinction rates per rate category.

```
for (i in 1:NUM_RATE_CATEGORIES) {
    extinction[i] := relative_extinction * exp( speciation[i] )
}
```

Next, we need a rate parameter for the rate-shifts events. Again, there we do not have much prior information about this rate and choose an exponential distribution with rate 1.0. As usual for rate parameter, we apply a scaling move to the **event_rate** variable.

```
event_rate ~ dnExponential(1.0)
moves[++mi] = mvScale(event_rate,lambda=1,tune=true,weight=5)
```

Additionally, we need a parameter for the category of the process at root. We use a uniform prior distribution on the indices 1 to N since we do not have any prior information in which rate category the process

is at the root. The move for this random variable is a random integer walk because the random variable is defined only on the indices (*cf.* with real number).

```
root_category ~ dnUniformNatural(1,NUM_RATE_CATEGORIES)
moves[++mi] = mvRandomIntegerWalk(root_category,weight=1)
```

5.2.2 Incomplete Taxon Sampling

We know that we have sampled 369 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 <- 369/450
```

5.2.3 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()
```

5.2.4 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 ~ dnHBDP(lambda=speciation, mu=extinction, rootAge=root_time, rho=rho,
   rootState=root_category, delta=event_rate, taxa=taxa )
```

And then we attach data to it.

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

This specific implementation of the branch-specific birth-death process augments the tree with rate-shift events. In order to sample the number, the location, and the types of the rate-shift events, we have to apply special moves to the tree. These moves will not change the tree but only the augmented rate-shift events. We use a **mvBirthDeathEvent** to add and remove events, a **mvEventTimeSlide** and **mvEventTimeBeta** move to change the time and location of the events, and a **mvDiscreteEventCategoryRandomWalk** to change the new rate categories after an event occured.

```
moves[++mi] = mvBirthDeathEvent(timetree,weight=2)
moves[++mi] = mvEventTimeSlide(timetree,weight=2)
moves[++mi] = mvEventTimeBeta(timetree,weight=2)
moves[++mi] = mvDiscreteEventCategoryRandomWalk(timetree,weight=2)
```

In the first place we are interested in the branch-specific rates. So far we do not have any variables that directly give us the number of rate-shift events per branch or the rates per branch. Fortunately, we can construct deterministic variables and query these properties from the tree. These function are made available by the branch-specific birth-death process distribution.

```
num_events := timetree.numberEvents()
avg_lambda := timetree.averageSpeciationRate()
avg_mu    := timetree.averageExtinctionRate()

total_num_events := sum( num_events )
```

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 an extended-newick monitor. The extended-newick monitor writes the tree to a file and adds parameter values to the branches and/or nodes of the tree. We can thus print the tree with the average speciation and extinction rates at the branches into a files. We will need this file later to estimate and visualize the posterior distribution of the rates at the branches.

```
monitors[2] = mnExtNewick(filename="output/primates_BSBD.trees", isNodeParameter=FALSE
   , printgen=10, separator = TAB, tree=timetree, avg_lambda, avg_mu)
```

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

```
monitors[3] = mnScreen(printgen=10, event_rate, mean_speciation, root_category,
    total_num_events)
```

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 branch-specific rates by attaching them to the tree. This is actually done automatically in our **mapTree** function.

```
treetrace = readTreeTrace("output/primates_BSBD.trees", treetype="clock")
map_tree = mapTree(treetrace,"output/primates_BSBD_MAP.tree")
```

Now you can open the tree in `FigTree`.

→ The `Rev` file for performing this analysis: `mcmc_BSBD.Rev`.

5.4 Exercise

- Run an MCMC simulation to estimate the posterior distribution of the speciation rate and extinction rate.
- Visualize the branch-specific rates in `FigTree`.

- Do you see evidence for rate decreases or increases? What is the general trend?
- Run the analysis using a different number of categories, *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., S. Ho, M. Phillips, and A. Rambaut. 2006. Relaxed Phylogenetics and Dating with Confidence. PLoS Biology 4:e88.

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.

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

FitzJohn, R. G. 2010. Quantitative traits and diversification. Systematic Biology 59:619–633.

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, B. Boussau, B. R. Moore, N. Lartillot, T. A. Heath, J. P. Huelsenbeck, and F. Ronquist. 2015. RevBayes: Bayesian Phylogenetic Inference Using Graphical Models and an Interactive Model Specification Language. submitted .

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.

Maddison, W., P. Midford, and S. Otto. 2007. Estimating a binary character's effect on speciation and extinction. Systematic Biology 56:701.

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

Rabosky, D. L. 2014. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. PLoS One 9:e89543.

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.

Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. Journal of Molecular Evolution 39:306–314.

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: February 23, 2016