Phylogenetic Inference using RevBayes

An introduction to Calibrated Time Trees & Divergence Time Estimation

Sebastian Höhna and Tracy A. Heath

1 Background

Estimating nodes ages (*i.e.,* branch lengths in proportion to time) is confounded by the fact that the rate of evolution and time are intrinsically linked when inferring genetic differences between species.

$$\text{branch length} = \text{rate} \times \text{time} \tag{1}$$

A model of substitution rate and divergence time must be applied to tease apart rate and time. When applied in methods for divergence time estimation, the resulting trees have branch lengths that are proportional to time. External node age estimates from the fossil record or other sources are necessary for inferring the real-time (or absolute) ages of lineage divergences (Figure **??**).

1.1 Modeling lineage-specific substitution rates

Many factors can influence the rate of substitution in a population such as mutation rate, generation time, and selection. As a result, many models have been proposed that describe how substitution rate may vary across the Tree of Life. The simplest model, the molecular clock, assumes that the rate of substitution remains constant over time (Zuckerkandl and Pauling 1962). However, many studies have shown that molecular data (in general) violate the assumption of a molecular clock and that there exists considerable variation in the rates of substitution among lineages.

Several models have been developed and implemented for inferring divergence times without assuming a strict molecular clock and are commonly applied to empirical data sets. Some models assume that rates are heritable and autocorrelated over the tree, others model rate change as a step-wise process, and others assume that the rates on each branch are independently drawn from a single distribution. Many of these models have been applied as priors using Bayesian inference methods. The implementation of dating methods in a Bayesian framework provides a flexible way to model rate variation and obtain reliable estimates of speciation times, provided the assumptions of the models are adequate. When coupled with numerical methods, such as MCMC, for approximating the posterior probability distribution of parameters, Bayesian methods are extremely powerful for estimating the parameters of a statistical model and are widely used in phylogenetics.

1.1.1 Some models of lineage-specific rate variation:

Global molecular clock: a constant rate of substitution over time (Zuckerkandl and Pauling 1962).

Local molecular clocks: Closely related lineages share the same rate and rates are clustered by subclades (Kishino et al. 1990; Rambaut and Bromham 1998; Yang and Yoder 2003; Drummond and Suchard 2010)

Compound Poisson process: Rate changes occur along lineages according to a point process and at rate-change events, the new rate is a product of the old rate and a Γ-distributed multiplier (Huelsenbeck et al. 2000).

Autocorrelated rates: substitution rates evolve gradually over the tree

- *Log-normally distributed rates:* the rate at a node is drawn from a log-normal distribution with a mean equal to the parent rate (Thorne et al. 1998; Kishino et al. 2001; Thorne and Kishino 2002)
- *Cox-Ingersoll-Ross Process:* the rate of the daughter branch is determined by a non-central χ^2 distribution. This process includes a parameter that determines the intensity of the force that drives the process to its stationary distribution (Lepage et al. 2006).

Uncorrelated rates: The rate associated with each branch is drawn from a single underlying parametric distribution such as an exponential or log-normal (Drummond et al. 2006; Rannala and Yang 2007; Lepage et al. 2007).

Mixture model on branch rates: Branches are assigned to distinct rate categories according to a Dirichlet process (Heath et al. 2012).

1.2 Priors on node times

There are many components that make up a Bayesian analysis of divergence time. One that is often overlooked is the prior on node times, often called a tree prior. This model describes how speciation events are distributed over time. When this model is combined with a model for branch rate, Bayesian inference allows you to estimate relative divergence times. Furthermore, because the rate and time are confounded in the branch-length parameter, the prior describing the branching times may have a strong effect on divergence time estimation.

We can separate the priors on node ages into different categories:

Phenomenological: models that make no explicit assumptions about the biological processes that generated the tree. These priors are conditional on the age of the root.

- *Uniform distribution:* This simple model assumes that internal nodes are uniformly distributed between the root and tip nodes (Lepage et al. 2007; Ronquist et al. 2012b).
- *Dirichlet distribution:* A flat Dirichlet distribution describes the placement of internal nodes on every path between the root and tips (Kishino et al. 2001; Thorne and Kishino 2002).

Mechanistic: models that describe the biological processes responsible for generating the pattern of lineage divergences.

- *Population-level processes:* Coalescent processes describe the time, in generations, between coalescent events and allow for the estimation of population-level parameters (Kingman 1982). Furthermore, they describe demographic processes (suitable for describing differences among individuals in the same species/population).
- *Species-level processes:* Birth-death stochastic branching models describe lineage diversification (suitable for describing the timing of divergences between samples from different species) (Kendall 1948; Thompson 1975; Nee et al. 1994; Rannala and Yang 1996; Yang and Rannala 1997; Höhna 2015).

We cover the different process that can be used for tree and divergence time prior distributions in detail in the RB_DiversificationRate_Tutorial because of their interest in estimating diversification rates and patterns.

1.3 Calibration to absolute time

Without external information to calibrate the tree, divergence time estimation methods can only reliably provide estimates of relative divergence times and not absolute node ages; or if one would know the (average) rate of substitutions. Calibration information can come from a variety of sources including "known" substitution rates (often secondary calibrations estimated from a previous study), dated tip sequences from serially sampled data (time-stamped virus data or ancient DNA), or geological date estimates (fossils or biogeographical data). Age estimates from fossil organisms are the most common form of divergence time calibration information. These data are used as age constraints on their putative ancestral nodes. There are numerous difficulties with incorporating node age estimates from fossil data including disparity in fossilization and sampling, uncertainty in dating, and correct phylogenetic placement of the fossil. Thus, it is critical that careful attention is paid to the paleontological data included in phylogenetic divergence time analyses. With an accurately dated and identified fossil in hand, further consideration is required to determine how to apply the node-age constraint. If the fossil is truly a descendant of the node it calibrates, then it provides a reliable minimum age bound on the ancestral node time. However, maximum bounds are far more difficult to come by.

Bayesian methods provide a way to account for uncertainty in fossil calibrations. Prior distributions reflecting our knowledge (or lack thereof) of the amount of elapsed time from the ancestral node to is calibrating fossil are easily incorporated into these methods. A nice review paper by Ho and Phillips (2009) outlines a number of different parametric distributions appropriate for use as priors on calibrated nodes.

Uniform distribution: Typically, you must have both maximum and minimum age bounds when applying a uniform calibration prior (though some methods are available for applying uniform constraints with soft bounds). The minimum bound is provided by the fossil member of the clade. The maximum bound may come from a bracketing method or other external source. This distribution places equal probability across all ages spanning the interval between the lower and upper bounds.

Normal distribution: When applying a biogeographical date (*e.g.,* the Isthmus of Panama) or a secondary calibration (a node age estimate from a previous study), the normal distribution can be a useful calibration prior. This distribution is always symmetrical and places the greatest prior weight on the mean. Its scale is determined by the standard deviation parameter.

Exponential distribution: The exponential distribution is characterized by a single rate parameter and is useful for calibration if the fossil age is very close to the age of its ancestral node. The expected (mean) age difference under this distribution is equal to the 1/rate. Under the exponential distribution, the greatest prior weight is placed on node ages very close to the age of the fossil with diminishing probability to ∞. As the rate parameter is increased, this prior density becomes strongly informative, whereas very low values of the rate result in a fairly non-informative prior (Figure 3a).

Log-normal distribution: An offset, log-normal prior on the calibrated node age places the highest probability on ages somewhat older than the fossil, with non-zero probability to ∞.

1.4 Integrating Fossil Occurrence Times in the Speciation Model

Calibrating Bayesian divergence-time estimates using parametric densities (as described in the previous section: Sec. 1.3) are typically applied in a multiplicative manner such that the prior probability of a calibrated node age is the product o the probability coming from the tree-wide speciation model and the probability under the calibration density (Heled and Drummond 2012; Warnock et al. 2012; 2015). However, when using fossil information, it is also possible to account for the fact that the fossils are part of the same diversification process (*i.e.,* birth-death model) that generated the extant species using the fossilized birth-death (FBD) model described in Stadler (2010) and Heath et al. (2014). This model simply treats the fossil observations as part of the prior on node times. The fossilized birth-death process provides a model for the distribution of speciation times, tree topology, and distribution of lineage samples before the present (*i.e.,* non-contemporaneous samples like fossils or viruses). Thus, it provides a reasonable prior distribution for analyses combining morphological or DNA data for both extant and fossil taxa—*i.e.,* the so-called 'total-evidence' or 'tip-dating' approaches described by Ronquist et al. (2012a) (also see Pyron (2011)). When matrices of discrete morphological characters for both living and fossil species are unavailable, the fossilized birth-death model imposes a time structure on the tree by marginalizing over all possible attachment points for the fossils on the extant tree (Heath et al. 2014), therefore, some prior knowledge of phylogenetic relationships is important, much like for calibration-density approaches.

2 Overview

This tutorial covers how to estimate divergence times and time calibrated phylogenies. The key concepts of this tutorial are node- and fossil-calibrations (assuming a global molecular clock). A good overview about best practices for node- and fossil-calibrations is found in Parham et al. (2012), Warnock et al. (2012), Joyce et al. (2013) and Warnock et al. (2015).

In this tutorial you will perform a Bayesian inference to estimate a time-calibrated phylogeny. In the first part we will demonstrate you how to set up a basic model for time-calibrated phylogeny inference. Throughout we will use the global molecular clock rate. In a later tutorial we will cover other models for the molecular clock, *e.g.,* relaxed clock methods. The first analysis will use an informative prior distribution on the root age (crown age) to calibrate the phylogeny. In the second analysis you will use informative node- and fossil-calibrations instead of the informative prior on the root age. In the third analysis you will use minimum and maximum age constraint (both using hard and soft constraints) to calibrate the phylogeny. All the assumptions will be covered more in detail later in this tutorial.

Requirements

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

- RB_Getting_Started
- RB_Basics_Tutorial
- RB_CTMC_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 on the way. You may only need the RB_Basics_Tutorial for a more in-depth discussion of concepts in `Rev`.

3 Data and files

We provide the data file of DNA sequences required for this tutorial. You may want to use your own data instead.

→ Create a folder called **data** and download the following files:

- `primates_cytb.nex`: Alignment of the *cytochrome b* subunit from 23 primates representing 14 of the 16 families (*Indriidae* and *Callitrichidae* are missing).

Below you will also find two tables with the calibration dates that we will use later in this tutorial (see Table 1 and Table 2). Table 1 will be used for calibrating clades/nodes with informative priors (*e.g.*, normal distributions). Table 2 will be used to calibrate clade/nodes by specifying minimum and maximum ages using hard and soft constraints.

Table 1: Node information used for calibrating divergence times in the primate tree from Perelman et al. (2011).

Clade	Age range (My)	Citation
Simiiformes	43 ± 4.5	Seiffert et al. (2003)
Lorisiformes	40 ± 3	Franzen et al. (2009); Poux and Douzery (2004)
Catarrhini	29 ± 6	Poux and Douzery (2004)
Platyrrhini	23.5 ± 3	Hodgson et al. (2009); Kay et al. (2008)
Primates	90 ± 6	Matsui et al. (2009); Steiper et al. (2004); Tavaré et al. (2002)

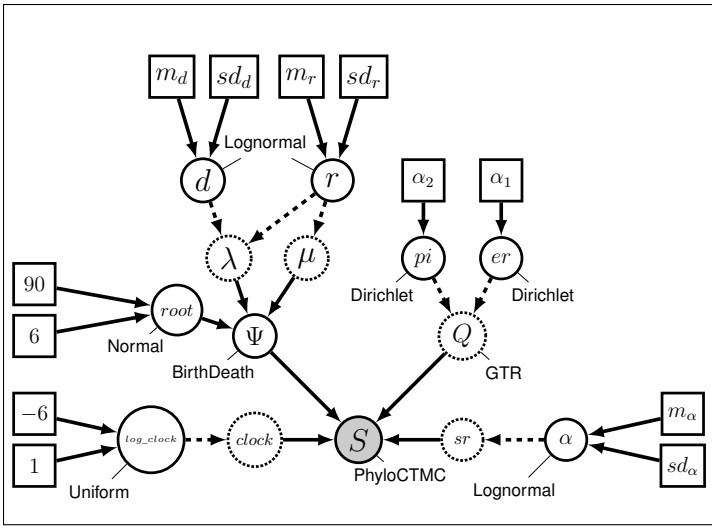
Table 2: Calibration intervals used in Springer et al. (2012) to calibrate nodes in the primate tree.

Clade	Min Age (My)	Max Age (My)	Citation
Simiiformes	28.3	56	Seiffert et al. (2003)
Lorisiformes	37.1	56	Franzen et al. (2009); Poux and Douzery (2004)
Catarrhini	20.55	37.3	Poux and Douzery (2004)
Platyrrhini	11.8	37.3	Hodgson et al. (2009); Kay et al. (2008)

4 Divergence time estimation using an informative prior on the root age

4.1 Getting Started

The first section of this exercise involves: (1) setting up a general time reversible (GTR) substitution model (Tavaré 1986) with gamma distributed rate variation among sites (Yang 1994) for an alignment of the cytochrome b subunit; (2) use an informative prior on the root age to date the phylogeny; (3) approximating the posterior probability of the tree topology and node ages (and all other parameters) using MCMC, and; (4) summarizing the MCMC output by computing the maximum *a posteriori* tree. This analysis is mostly equivalent to the analysis performed in the RB_CTMC_Tutorial.




```
m_d <- ln( ln(450/2) / 90 )
sd_d <- 0.587405*2
d ~ dnLognorma(m_d,sd_d)
m_r <- ln( ln(450/2) / 90 )
sd_r <- 0.587405*2
r ~ dnLognorma(m_r,sd_r)
lambda := d + r
mu := r
root ~ dnNormal(90.0,6.0)
psi ~ dnBirthDeath( lambda, mu, root, taxa)

alpha1 <- v(1,1,1,1,1,1)
alpha2 <- v(1,1,1,1)
er ~ dnDirichlet( alpha1 )
pi ~ dnDirichlet( alpha2 )
Q_mol := fnGTR(er, pi)

m_alpha <- 5.0
sd_alpha <- 0.587405
alpha ~ dnLognormal( m_alpha, sd_alpha )
sr := fnDiscretizeGamma( alpha, alpha, 4, false )
log_clock ~ dnUniform(-6,1)
clock := 10^log_clock

seq ~ dnPhyloCTMC( tree=psi, Q=Q_mol, sr
                   branchRates=0.01, type="DNA" )
seq.clamp( data )
```

Figure 1: An example phylogenetic model depicted in graphical-model notation (left) and the corresponding specification in the `Rev` language (right). This example shows the basic outline of the root calibration that we will use in the first exercise.

The general structure of the model is represented in Figure 1. This figure shows the full model graph.

4.2 Loading the Data

→ You should have already downloaded the data files in Section 3. Links to additional files, including the scripts to run these analyses can be found on the RevBayes tutorials website. Remember that the data file should be in a directory called **`data`** that is in your current working directory.

First load in the sequences using the **`readDiscreteCharacterData()`** function.

```
data <- readDiscreteCharacterData("data/primates_cytb.nex")
```

Executing these lines initializes the data matrix as the respective `Rev` variable.

Next we will specify some useful variables based on our dataset. The variable **data** has *member functions* that we can use to retrieve information about the dataset. These include the number of species (**n_species**) and the tip labels (**taxa**). Each of these variables will be necessary for setting up different parts of our model (*e.g.,* the birth-death process prior).

```
n_species <- data.ntaxa()
taxa <- data.taxa()
```

Additionally, we set up a counter variable for the number of moves that we already added to our analysis. [Recall that moves are algorithms used to propose new parameter values during the MCMC simulation.] This will make it much easier if we extend the model or analysis to include additional moves or to remove some moves.

```
mvi = 0
```

You may have noticed that we used the **=** operator to create the move index. This simply means that the variable is not part of the model. You will later see that we use this operator more often, *e.g.,* when we create moves and monitors.

With the data loaded, we can now proceed to specify our substitution model.

4.3 General Time Reversible (GTR) Substitution Model

The GTR model requires that we define and specify a prior on the six exchangeability rates, which we will describe using a flat Dirichlet distribution. As we did previously for the Dirichlet prior on base frequencies, we first define a constant node specifying the vector of concentration-parameter values using the **v()** function:

```
er_prior <- v(1,1,1,1,1,1)
```

This node defines the concentration-parameter values of the Dirichlet prior distribution on the exchangeability rates. Now, we can create a stochastic node for the exchangeability rates using the **dnDirichlet()** function, which takes the vector of concentration-parameter values as an argument and the **~** operator. Together, these create a stochastic node named **er**, see Figure 1:

```
er ~ dnDirichlet(er_prior)
```

The Dirichlet prior on our parameter **er** creates a *simplex* of values that sum to 1.

For each stochastic node in our model, we must also specify a proposal mechanism if we wish to estimate that parameter.

```
moves[++mvi] = mvSimplexElementScale(er)
```

We can use the same type of distribution as a prior on the 4 stationary frequencies ($\pi_A, \pi_C, \pi_G, \pi_T$) since these parameters also represent proportions. Specify a flat Dirichlet prior density on the base frequencies:

```
pi_prior <- v(1,1,1,1)
pi ~ dnDirichlet(pi_prior)
```

The node **pi** represents the π node in Figure 1. Now add the simplex scale move on the stationary frequencies to the moves vector:

```
moves[++mvi] = mvSimplexElementScale(pi)
```

We can finish setting up this part of the model by creating a deterministic node for the GTR instantaneous-rate matrix **Q**. The **fnGTR()** function takes a set of exchangeability rates and a set of base frequencies to compute the instantaneous-rate matrix used when calculating the likelihood of our model.

```
Q := fnGTR(er,pi)
```

4.4 Setting up the Gamma Model

Create a constant node called **alpha_prior_mean** and a second constant node called **alpha_prior_sd** for the lognormal prior on the gamma-shape parameter (this is represented as the constant rate parameter in Figure 1):

```
alpha_prior_mean <- 5.0
alpha_prior_sd <- 0.587405
```

Then create a stochastic node called **alpha** with an lognormal prior (this represents the stochastic node for the α-shape parameter in Figure 1):

```
alpha ~ dnLognormal( alpha_prior_mean, alpha_prior_sd )
```

The way the ASRV model is implemented involves discretizing the mean-one gamma distribution into a set number of rate categories, k. To specify this, we need a deterministic node that is a vector that will hold the set of k rates drawn from the gamma distribution with k rate categories. The **fnDiscretizeGamma()** function returns this deterministic node and takes three arguments: the shape and rate of the gamma

distribution and the number of categories. Since we want to discretize a mean-one gamma distribution, we can pass in **alpha** for both the shape and rate.

Initialize the **gamma_rates** deterministic node vector using the **fnDiscretizeGamma()** function with **4** bins:

```
gamma_rates := fnDiscretizeGamma( alpha, alpha, 4 )
```

The random variable that controls the rate variation is the stochastic node **alpha**. We will apply a simple scale move to this parameter.

```
moves[++mvi] = mvScale(alpha, weight=2.0)
```

For more information on ASRV please read the RB_CTMC_Tutorial.

4.5 Tree Prior: Tree Topology and Node Ages

The tree (the topology and node ages) is a stochastic node in our phylogenetic model. In Figure 1, the tree is denoted Ψ. We will assume a constant-rate birth-death process as the prior distribution on the tree. The distribution in RevBayes is **dnBDP()**. For more information on tree priors please read the RB_DiversificationRate_Tutorial.

For the birth-death process we need a speciation rate and extinction rate parameter. Instead of prior distributions on these parameters directly, we will specify lognormal prior distributions on the diversification and turnover rates.

```
diversification_mean <- ln( ln(n_species/2.0) / 90 )
diversification_sd <- 0.587405*2
diversification ~ dnLognormal(mean=diversification_mean,sd=diversification_sd)
moves[++mvi] = mvScale(diversification,lambda=1.0,tune=true,weight=3.0)

turnover_mean <- ln( ln(n_species/2.0) / 90 )
turnover_sd <- 0.587405*2
turnover ~ dnLognormal(mean=turnover_mean,sd=turnover_sd)
moves[++mvi] = mvScale(turnover,lambda=1.0,tune=true,weight=3.0)

### Transform the parameters
birth_rate := diversification + turnover
death_rate := turnover
```

In our reference publication, Perelman et al. (2011) used a normal distribution with mean of 90.0 MYA with stdev = 6.0 as the prior distribution on the root age. The normal distribution itself is defined on the complete real line (*i.e.,* from $-\infty$ to $+\infty$), however, we know that the root age of primates is definitely larger than 0 (it happen before the present) and smaller than, say, 5000 MYA. Thus, we truncate the

normal distribution. This also has the advantage that the type of the variable for the root age is a positive real number, instead of a real number.

```
root_time ~ dnNormal(mean=90.0,sd=6.0,min=0.0,max=5000.0)
moves[++mvi] = mvScale(root_time,weight=2.0)
```

Additionally, we know that we do not have all primate species included in this data set. We only have 23 out of the approximately 450 primate species. Thus, we use a sampling fraction to represent this incomplete taxon sampling (Höhna et al. 2011; Höhna 2014).

```
rho <- n_species/450
```

Next, specify the **tree** stochastic node by passing in the tip labels **taxa** to the **dnBDP()** distribution:

```
psi ~ dnBDP(lambda=birth_rate, mu=death_rate, rho=rho, rootAge=root_time,
    samplingStrategy="uniform", condition="survival", taxa=taxa)
```

Some types of stochastic nodes can be updated by a number of alternative moves. Different moves may explore parameter space in different ways, and it is possible to use multiple different moves for a given parameter to improve mixing (the efficiency of the MCMC simulation). In the case of our rooted tree, for example, we can use both a nearest-neighbor interchange move without and with changing the node ages (**mvNarrow** and **mvNNI**) and a fixed-nodeheight subtree-prune and regrafting move (**mvFNPR**). For overviews about moves on tree see Lakner et al. (2008), Höhna et al. (2008) and Höhna and Drummond (2012). We also need moves that change the ages of the internal nodes; which are for example the **mvSubtreeScale** and **mvNodeTimeSlideUniform**. These moves do not have tuning parameters associated with them, thus you only need to pass in the **psi** node and proposal **weight**.

```
moves[++mvi] = mvNarrow(psi, weight=5.0)
moves[++mvi] = mvNNI(psi, weight=1.0)
moves[++mvi] = mvFNPR(psi, weight=5.0)
moves[++mvi] = mvFNPR(psi, weight=2.0)
moves[++mvi] = mvSubtreeScale(psi, weight=3.0)
moves[++mvi] = mvNodeTimeSlideUniform(psi, weight=15.0)
```

The weight specifies how often the move will be applied either on average per iteration or relative to all other moves. Have a look at the MCMC tutorial for more details about moves and MCMC strategies: http://revbayes.github.io/tutorials.html

4.6 Monitoring specific clade ages

The exercise in this tutorial involves looking at specific age estimates. There are two ways in RevBayes how to obtain age estimates. First, you can look into the generated *maximum a posteriori* tree in FigTree.

Second, you can add deterministic variables for the ages that you are interested in and look at the values in `Tracer`. Both approaches are useful and could be used together. For the second approach to work we need to create these deterministic variables.

We start with a deterministic node monitoring the age of the *Catarrhini*. This involves create a clade object. A clade object simply contains a number of species names. Note, the names need to match **exactly**. Then, we use the **`tmrca`** function which will record the time of the most recent common ancestor of this clade.

```
clade_catarrhini = clade("Pan_paniscus", "Macaca_mulatta")
age_catarrhini := tmrca(psi, clade_catarrhini)
```

Next, a deterministic node monitoring the age of the *Platyrrhini*:

```
clade_platyrrhini = clade("Alouatta_palliata", "Callicebus_donacophilus")
age_platyrrhini := tmrca(psi, clade_platyrrhini)
```

Then, a deterministic node monitoring the age of the *Simiiformes*:

```
clade_simiiformes = clade("Cebus_albifrons", "Macaca_mulatta")
age_simiiformes := tmrca(psi, clade_simiiformes)
```

And finally, a deterministic node monitoring the age of the *Lorisiformes*:

```
clade_lorisiformes = clade("Loris_tardigradus", "Galago_senegalensis")
age_lorisiformes := tmrca(psi, clade_lorisiformes)
```

4.7 The Global Molecular Clock Model

The global molecular clock assumes that the rate of substitution is constant over the tree and over time (Fig. 1). Since we calibrated the tree with an informative distribution on the root age we will estimate the clock rate. Here we use a uniform distribution on the logarithm of the clock rate, which signifies our uncertainty of the magnitude of the clock rate. We will say that every magnitude or clock rates between 10^{-6} and 10^{1} are equally probably a priori.

```
logClockRate ~ dnUniform(-6,1)
clockRate := 10^logClockRate

moves[++mvi] = mvSlide(logClockRate)
```

4.8 Putting it All Together

We have fully specified all of the parameters of our phylogenetic model—the tree topology with branch lengths, and the substitution model that describes how the sequence data evolved over the tree with branch lengths. Collectively, these parameters comprise a distribution called the *phylogenetic continuous-time Markov chain*, and we use the **PhyloCTMC** constructor function to create this node. This distribution requires several input arguments: (1) the **tree** with branch lengths; (2) the instantaneous-rate matrix **Q**; (3) the clock rate, and; (4) the **type** of character data.

Build the random variable for the character data (sequence alignment).

```
# the sequence evolution model
seq ~ dnPhyloCTMC(tree=psi, Q=Q, branchRates=clockRate, siteRates=gamma_rates, type="
    DNA")
```

Once the **PhyloCTMC** model has been created, we can attach our sequence data to the tip nodes in the tree.

```
seq.clamp(data)
```

[Note that although we assume that our sequence data are random variables—they are realizations of our phylogenetic model—for the purposes of inference, we assume that the sequence data are "clamped".] When this function is called, RevBayes sets each of the stochastic nodes representing the tips of the tree to the corresponding nucleotide sequence in the alignment. This essentially tells the program that we have observed data for the sequences at the tips.

Finally, we wrap the entire model to provide convenient access to the DAG. To do this, we only need to give the **model()** function a single node. With this node, the **model()** function can find all of the other nodes by following the arrows in the graphical model:

```
mymodel = model(Q)
```

4.9 Performing an MCMC Analysis Under the Global Clock Model

In this section, will describe how to set up the MCMC sampler and summarize the resulting posterior distribution of trees.

4.9.1 Specifying Monitors

For our MCMC analysis, we need to set up a vector of *monitors* to record the states of our Markov chain. The monitor functions are all called **mn***, where * is the wildcard representing the monitor type. 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[++mni] = mnModel(filename="output/primates_cytb_root_calibration.log",
    printgen=10, separator = TAB)
```

The **mnFile** monitor will record the states for only the parameters passed in as arguments. We use this monitor to specify the output for our sampled trees and branch lengths.

```
monitors[++mni] = mnFile(filename="output/primates_cytb_root_calibration.trees",
    printgen=10, separator = TAB, psi)
```

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

```
monitors[++mni] = mnScreen(printgen=1000, clockRate, root_time, age_simiiformes)
```

4.9.2 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)
```

We may wish to run the **.burnin()** member function. Recall that this function **does not** specify the number of states that we wish to discard from the MCMC analysis as burnin (*i.e.,* the samples collected before the chain converges to the stationary distribution). Instead, the **.burnin()** function specifies a *completely separate* preliminary MCMC analysis that is used to tune the scale of the moves to improve mixing of the MCMC analysis.

```
mymcmc.burnin(generations=10000,tuningInterval=250)
```

Now, run the MCMC:

```
mymcmc.run(generations=30000)
```

When the analysis is complete, you will have the monitored files in your output directory.

→ Look at the file called **output/primates_cytb_root_calibration.log** in Tracer.

4.10 Exercise 1

We are interested in the divergence time estimate between *Simiiformes*, *Platyrrhini*, *Catarrhini*, *Lorisiformes* and all primates.

To obtain an estimate of the divergence time we read in the tree trace and build the annotated maximum *a posteriori* tree.

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

Fill in the following table as you go through the tutorial.

→ Look at the file called **`output/primates_cytb_root_calibration.tree`** in `FigTree`.

Table 3: Estimated divergence times in our primates example*.

	Primates		Simiiformes		Platyrrhini		Catarrhini		Lorisiformes	
Clock Model	*Mean Estimate*	*Credible interval*	*Mean Estimate*	*Credible interval*	*Mean Estimate*	*Credible interval*	*Mean Estimate*	*Credible interval*	*Mean Estimate*	*Credible interval*
4 Root Calibration										
5 Node Calibration										
6 Hard Bounds										
7 Soft Bounds										

*you can edit this table

5 Informative node calibration

In the previous section we calibrated the phylogeny using an informative prior on the root age only. This part of the exercise will involve specifying a informative prior on internal nodes in our tree: (1) *Semiiformes*, the split between *New World Monkeys* and *OldWorld Monkeys*; (2) *Platyrrhini*; (3) *Catarrhini*; and (4) *Lorisifomres*, the split between *galagids* and *lorisids*.

In `RevBayes`, calibrated internal nodes are treated differently than in many other programs for estimating species divergence times (*e.g.,* BEAST). This is because the graphical model structure used in `RevBayes` does not allow a stochastic node to be assigned more than one prior distribution. By contrast, the common approach to applying calibration densities as used in other dating softwares leads to incoherence in the calibration prior (for detailed explainations of this see Warnock et al. 2012; Heled and Drummond 2012; Heath et al. 2014). More explicitly, common calibration approaches assume that the age of a calibrated node is modeled by the tree-wide diversification process (*e.g.,* birth-death model) *and* a parametric density parameterized by the occurrence time of a fossil (or other external prior information). This can induce a calibration prior density that is not consistent with the birth-death process or the parametric prior distribution. Thus, approaches that condition the birth-death process on the calibrated nodes are more statistically coherent (Yang and Rannala 2006).

In `RevBayes`, calibration densities are applied in a different way, treating fossil observation times like data. The graphical model in Figure **??** illustrates how calibrated nodes are specified in the directed acyclic graph (DAG). Here, the age of the calibration node (*i.e.,* the internal node specified as the MRCA of the fossil and a set of living species) is a deterministic node—*e.g.,* denoted o_1 for fossil $\mathcal{F}_1$—and acts as an offset on the stochastic node representing the age of the fossil specimen. The fossil age, $\mathcal{F}_i$, is specified as a stochastic node and clamped to its *observed* age in the fossil record. The node $\mathcal{F}_i$ is modeled using a distribution that describes the waiting time from the speciation event to the appearance of the observed fossil. Thus, if the MCMC samples any state of Ψ for which the age of $\mathcal{F}_i$ has a probability of 0, then that state will always be rejected, effectively calibrating the birth-death process without applying multiple prior densities to any calibrated node (Fig. **??**).

5.1 Adding node calibrations

Additional to the model that we used in the previous exercise we will add calibrations to some clades (see Table 1).

First, we specify the calibration for the catarrhini split using a normal distribution with mean 29 and standard deviation of 6.

```
obs_age_catarrhini ~ dnNormal(age_catarrhini,6)
obs_age_catarrhini.clamp(29)
```

Next, we specify the calibration for the Platyrrhini split using a normal distribution with mean 23.5 and standard deviation of 3.

```
obs_age_platyrrhini ~ dnNormal(age_platyrrhini,3)
obs_age_platyrrhini.clamp(23.5)
```

Then, we specify the calibration for the split between *New World Monkeys* and *OldWorld Monkeys* using a normal distribution with mean 43 and standard deviation of 4.5.

```
obs_age_simiiformes ~ dnNormal(age_simiiformes,4.5)
obs_age_simiiformes.clamp(43)
```

Finally, we specify the calibration for the lorisiformes (the split between *galagids* and *lorisids*) using a normal distribution with mean 40 and standard deviation of 3.

```
obs_age_lorisiformes ~ dnNormal(age_lorisiformes,3)
obs_age_lorisiformes.clamp(40)
```

5.2 Exercise 2

- Copy the file **RootCalibration.Rev** and call it **NodeCalibration**.
- Add the node calibrations outlines above.
- Rename the output files (*i.e.*, the filenames for the monitors).
- Run the MCMC analysis.
- Look at the output in **Tracer**.
- Fill in the table.

6 Hard-bounded node calibrations

Additional to the model that we used in the previous exercise we will add calibrations to some clades (see Table 2).

First, we specify the calibration for the catarrhini split using a uniform distribution with minimum of 20.55 and maximum of 37.3.

```
min_age_catarrhini <- 20.55
max_age_catarrhini <- 37.3
width_age_prior_catarrhini <- (max_age_catarrhini-min_age_catarrhini)/2.0
mean_age_prior_catarrhini <- min_age_catarrhini + width_age_prior_catarrhini
obs_age_catarrhini ~ dnUniform(age_catarrhini - width_age_prior_catarrhini,
    age_catarrhini + width_age_prior_catarrhini)
obs_age_catarrhini.clamp( mean_age_prior_catarrhini )
```

Next, we specify the calibration for the Platyrrhini split using a uniform distribution with minimum of 11.8 and maximum of 37.3.

```
min_age_platyrrhini <- 11.8
max_age_platyrrhini <- 37.3
width_age_prior_platyrrhini <- (max_age_platyrrhini-min_age_platyrrhini)/2.0
mean_age_prior_platyrrhini <- min_age_platyrrhini + width_age_prior_platyrrhini
obs_age_platyrrhini ~ dnUniform(age_platyrrhini - width_age_prior_platyrrhini,
    age_platyrrhini + width_age_prior_platyrrhini)
obs_age_platyrrhini.clamp( mean_age_prior_platyrrhini )
```

Then, we specify the calibration for the split between *New World Monkeys* and *OldWorld Monkeys* using a uniform distribution with minimum of 28.3 and maximum of 56.

```
min_age_simiiformes <- 28.3
max_age_simiiformes <- 56
width_age_prior_simiiformes <- (max_age_simiiformes-min_age_simiiformes)/2.0
mean_age_prior_simiiformes <- min_age_simiiformes + width_age_prior_simiiformes
obs_age_simiiformes ~ dnUniform(age_simiiformes - width_age_prior_simiiformes,
    age_simiiformes + width_age_prior_simiiformes)
obs_age_simiiformes.clamp( mean_age_prior_simiiformes )
```

Finally, we specify the calibration for the lorisiformes (the split between *galagids* and *lorisids*) using a uniform distribution with minimum of 37.1 and maximum of 56.

```
min_age_lorisiformes <- 37.1
max_age_lorisiformes <- 56
width_age_prior_lorisiformes <- (max_age_lorisiformes-min_age_lorisiformes)/2.0
mean_age_prior_lorisiformes <- min_age_lorisiformes + width_age_prior_lorisiformes
obs_age_lorisiformes ~ dnUniform(age_lorisiformes - width_age_prior_lorisiformes,
    age_lorisiformes + width_age_prior_lorisiformes)
obs_age_lorisiformes.clamp( mean_age_prior_lorisiformes )
```

6.1 Exercise 3

- Copy the file **RootCalibration.Rev** and call it **HardBoundsNodeCalibration.Rev**.
- Add the node calibrations outlines above.
- Rename the output files (*i.e.*, the filenames for the monitors).
- Run the MCMC analysis.
- Look at the output in **Tracer**.
- Fill in the table.

7 Soft-bounded node calibrations

Additional to the model that we used in the previous exercise we will add calibrations to some clades (see Table 2).

First, we specify the calibration for the catarrhini split using a uniform distribution with minimum of 20.55 and maximum of 37.3.

```
min_age_catarrhini <- 20.55
max_age_catarrhini <- 37.3
width_age_prior_catarrhini <- (max_age_catarrhini-min_age_catarrhini)/2.0
mean_age_prior_catarrhini <- min_age_catarrhini + width_age_prior_catarrhini
obs_age_catarrhini ~ dnSoftBoundUniformNormal(min=age_catarrhini -
    width_age_prior_catarrhini, max=age_catarrhini + width_age_prior_catarrhini, sd
    =2.5, p=0.95)
obs_age_catarrhini.clamp( mean_age_prior_catarrhini )
```

Next, we specify the calibration for the Platyrrhini split using a uniform distribution with minimum of 11.8 and maximum of 37.3.

```
min_age_platyrrhini <- 11.8
max_age_platyrrhini <- 37.3
width_age_prior_platyrrhini <- (max_age_platyrrhini-min_age_platyrrhini)/2.0
mean_age_prior_platyrrhini <- min_age_platyrrhini + width_age_prior_platyrrhini
obs_age_platyrrhini ~ dnSoftBoundUniformNormal(min=age_platyrrhini -
    width_age_prior_platyrrhini, max=age_platyrrhini + width_age_prior_platyrrhini, sd
    =2.5, p=0.95)
obs_age_platyrrhini.clamp( mean_age_prior_platyrrhini )
```

Then, we specify the calibration for the split between *New World Monkeys* and *OldWorld Monkeys* using a uniform distribution with minimum of 28.3 and maximum of 56.

```
min_age_simiiformes <- 28.3
max_age_simiiformes <- 56
width_age_prior_simiiformes <- (max_age_simiiformes-min_age_simiiformes)/2.0
mean_age_prior_simiiformes <- min_age_simiiformes + width_age_prior_simiiformes
obs_age_simiiformes ~ dnSoftBoundUniformNormal(min=age_simiiformes -
    width_age_prior_simiiformes, max=age_simiiformes + width_age_prior_simiiformes, sd
    =2.5, p=0.95)
width_age_prior_simiiformes)
obs_age_simiiformes.clamp( mean_age_prior_simiiformes )
```

Finally, we specify the calibration for the lorisiformes (the split between *galagids* and *lorisids*) using a uniform distribution with minimum of 37.1 and maximum of 56.

```
min_age_lorisiformes <- 37.1
max_age_lorisiformes <- 56
width_age_prior_lorisiformes <- (max_age_lorisiformes-min_age_lorisiformes)/2.0
mean_age_prior_lorisiformes <- min_age_lorisiformes + width_age_prior_lorisiformes
obs_age_lorisiformes ~ dnSoftBoundUniformNormal(min=age_lorisiformes -
    width_age_prior_lorisiformes, max=age_lorisiformes + width_age_prior_lorisiformes,
    sd=2.5, p=0.95)
width_age_prior_lorisiformes)
obs_age_lorisiformes.clamp( mean_age_prior_lorisiformes )
```

7.1 Exercise 4

- Copy the file **RootCalibration.Rev** and call it **SoftBoundsNodeCalibration.Rev**.
- Add the node calibrations outlines above.
- Rename the output files (*i.e.,* the filenames for the monitors).
- Run the MCMC analysis.
- Look at the output in Tracer.
- Fill in the table.

References

Drummond, A., S. Ho, M. Phillips, and A. Rambaut. 2006. Relaxed Phylogenetics and Dating with Confidence. PLoS Biology 4:e88.

Drummond, A. and M. Suchard. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biology 8:114.

Franzen, J. L., P. D. Gingerich, J. Habersetzer, J. H. Hurum, W. von Koenigswald, and B. H. Smith. 2009. Complete primate skeleton from the middle Eocene of Messel in Germany: morphology and paleobiology. PLoS One 4:e5723.

Heath, T., M. Holder, and J. Huelsenbeck. 2012. A dirichlet process prior for estimating lineage-specific substitution rates. Molecular Biology and Evolution 29:939–955.

Heath, T. A., J. P. Huelsenbeck, and T. Stadler. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Heled, J. and A. J. Drummond. 2012. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Systematic Biology 61:138–149.

Ho, S. Y. and M. J. Phillips. 2009. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. Systematic Biology Page syp035.

Hodgson, J. A., K. N. Sterner, L. J. Matthews, A. S. Burrell, R. A. Jani, R. L. Raaum, C.-B. Stewart, and T. R. Disotell. 2009. Successive radiations, not stasis, in the South American primate fauna. Proceedings of the National Academy of Sciences 106:5534–5539.

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. Defoin-Platel, and A. Drummond. 2008. Clock-constrained tree proposal operators in Bayesian phylogenetic inference. Pages 1–7 *in* 8th IEEE International Conference on BioInformatics and BioEngineering, 2008. BIBE 2008.

Höhna, S. and A. J. Drummond. 2012. Guided Tree Topology Proposals for Bayesian Phylogenetic Inference. Systematic Biology 61:1–11.

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.

Huelsenbeck, J. P., B. Larget, and D. L. Swofford. 2000. A compound Poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Joyce, W. G., J. F. Parham, T. R. Lyson, R. C. Warnock, and P. C. Donoghue. 2013. A divergence dating analysis of turtles using fossil calibrations: an example of best practices. Journal of Paleontology 87:612–634.

Kay, R. F., J. Fleagle, T. Mitchell, M. Colbert, T. Bown, and D. W. Powers. 2008. The anatomy of Dolichocebus gaimanensis, a stem platyrrhine monkey from Argentina. Journal of Human Evolution 54:323–382.

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

Kingman, J. F. C. 1982. On the genealogy of large populations. Journal of Applied Probability 19:27–43.

Kishino, H., T. Miyata, and M. Hasegawa. 1990. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. Journal of Molecular Evolution 31:151–160.

Kishino, H., J. L. Thorne, and W. J. Bruno. 2001. Performance of a divergence time estimation method under a probabilistic model of rate evolution. Molecular Biology and Evolution 18:352–361.

Lakner, C., P. van der Mark, J. P. Huelsenbeck, B. Larget, and F. Ronquist. 2008. Efficiency of Markov Chain Monte Carlo Tree Proposals in Bayesian Phylogenetics. Systematic Biology 57:86–103.

Lepage, T., D. Bryant, H. Philippe, and N. Lartillot. 2007. A general comparison of relaxed molecular clock models. Molecular Biology and Evolution 24:2669.

Lepage, T., S. Lawi, P. Tupper, and D. Bryant. 2006. Continuous and tractable models for the variation of evolutionary rates. Mathematical biosciences 199:216–233.

Matsui, A., F. Rakotondraparany, I. Munechika, M. Hasegawa, and S. Horai. 2009. Molecular phylogeny and evolution of prosimians based on complete sequences of mitochondrial DNAs. Gene 441:53–66.

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

Parham, J. F., P. C. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, et al. 2012. Best practices for justifying fossil calibrations. Systematic Biology 61:346–359.

Perelman, P., W. E. Johnson, C. Roos, H. N. Seuánez, J. E. Horvath, M. A. Moreira, B. Kessing, J. Pontius, M. Roelke, Y. Rumpler, et al. 2011. A molecular phylogeny of living primates. PLoS Genetics 7:e1001342.

Poux, C. and E. J. Douzery. 2004. Primate phylogeny, evolutionary rate variations, and divergence times: a contribution from the nuclear gene IRBP. American Journal of Physical Anthropology 124:01–16.

Pyron, R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of Lissamphibia. Systematic Biology Page syr047.

Rambaut, A. and L. Bromham. 1998. Estimating divergence dates from molecular sequences. Molecular Biology and Evolution 15:442–448.

Rannala, B. and Z. Yang. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution 43:304–311.

Rannala, B. and Z. Yang. 2007. Inferring Speciation Times under an Episodic Molecular Clock. Systematic Biology 56:453–466.

Ronquist, F., S. Klopfstein, L. Vilhelmsen, S. Schulmeister, D. L. Murray, and A. P. Rasnitsyn. 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the hymenoptera. Systematic Biology 61:973–999.

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. 2012b. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Systematic Biology 61:539–542.

Seiffert, E. R., E. L. Simons, and Y. Attia. 2003. Fossil evidence for an ancient divergence of lorises and galagos. Nature 422:421–424.

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.

Stadler, T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 267:396–404.

Steiper, M. E., N. M. Young, and T. Y. Sukarna. 2004. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid–cercopithecoid divergence. Proceedings of the National Academy of Sciences 101:17021–17026.

Tavaré, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Some Mathematical Questions in Biologyï£¡DNA Sequence Analysis 17:57–86.

Tavaré, S., C. R. Marshall, O. Will, C. Soligo, and R. D. Martin. 2002. Using the fossil record to estimate the age of the last common ancestor of extant primates. Nature 416:726–729.

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

Thorne, J., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15:1647–1657.

Thorne, J. L. and H. Kishino. 2002. Divergence time and evolutionary rate estimation with multilocus data. Systematic Biology 51:689–702.

Warnock, R. C., J. F. Parham, W. G. Joyce, T. R. Lyson, and P. C. Donoghue. 2015. Calibration uncertainty in molecular dating analyses: there is no substitute for the prior evaluation of time priors. Proceedings of the Royal Society B: Biological Sciences 282:20141013.

Warnock, R. C., Z. Yang, and P. C. Donoghue. 2012. Exploring uncertainty in the calibration of the molecular clock. Biology letters 8:156–159.

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

Yang, Z. and B. Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo Method. Molecular Biology and Evolution 14:717–724.

Yang, Z. and B. Rannala. 2006. Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds. Molecular Biology and Evolution 23:212–226.

Yang, Z. and A. D. Yoder. 2003. Comparison of likelihood and bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. Systematic Biology 52:705–716.

Zuckerkandl, E. and L. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189–225 *in* Horizons in Biochemistry (M. Kasha and B. Pullman, eds.) Academic Press, New York.

Version dated: March 22, 2016