Tutorial: Molecular dating

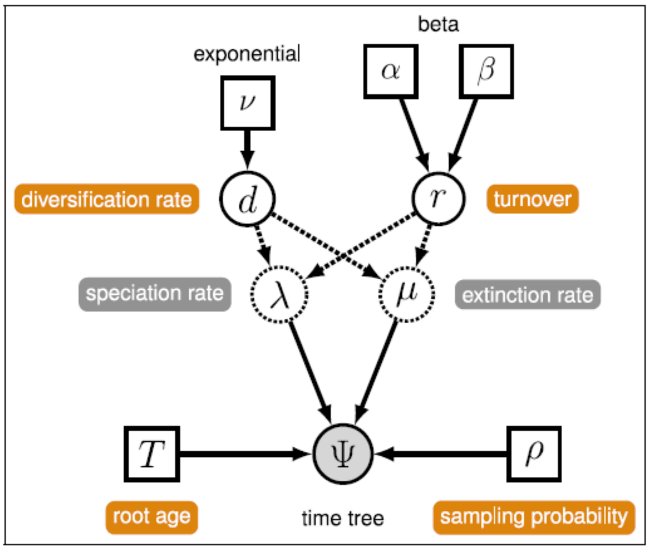
Seraina Klopfstein, Tracy Heath & Fredrik Ronquist
RevBayes workshop, ACEBB, The University of Adelaide. November 17-21, 2014

A skeletal file performing a time-tree analysis on a group of bears is available under the filename "Bears_skeletal.Rev". Add your blocks here. Don't forget to keep checking whether your variables contain what you have set them up for, and if they are of the correct type.

Exercise 1: Birth-Death process & tree calibrations

1.a: Diversification and turnover rate

We will now use a birth-death prior, instead of a uniform tree prior as we used it for the non-clock tree. The birth-death process has two parameters, the birth rate and the death rate. It is however more convenient to parameterize it with the diversification rate (which is equal to birth rate minus death rate) and the turnover (death rate divided by birth rate). The first can take any positive value, while the turnover is a value between zero and one - because the death rate should be smaller than the birth rate given that we observe taxa at in the present.



Set up the model as follows:

- exponential prior with mean 0.1 on the **diversification** rate. Use a scale move.
- beta(1,1) prior on the **turnover** rate. Use a slide move.

Question: why should the turnover-rate be below than one?

To get the birth and death rates, we define deterministic nodes. Use the following formulae:

```
birth_rate := diversification / abs(1.0 - turnover)

death_rate := (turnover * diversification) / abs(1.0 - turnover)
```

The "abs()" function makes sure that our denominator is always above zero, and we get a positive real value for both birth and death rates.

Finally, we need to specify the sampling probability, which is equal to the number of taxa we have sampled divided by the total number of taxa. There are 147 described species of caniforms and we have sampled 10 of them. From this information, calculate the sampling probability and define it as the constant node "`rho`".

1.b: Assemble a birth-death tree prior with constraints

We will use two internal node and one root calibration to calibrate the tree. The root node is a special case as it also corresponds to the tree age, which is a parameter of the Birth-Death prior. In addition, it does not need a topology constraint - because all taxa in a tree are always monophyletic. So we already set this calibration here, by defining a stochastic node with a lognormal distribution, according to the following information:

- use as an offset the age of the oldest fossil canid: 38 Mya.
- as a mean for the exponential distribution, use the oldest carnivor (49 Mya)
- use the distribution `dnLNorm(mu, sigma)`

To calculate the µ parameter of the lognormal distribution, we need to know that the mean of the lognormal distribution is $e^{\mu+\sigma^2/2}$ (with µ and σ being the mean and standard deviation on the non-log scale). The parameters µ and σ are used in RevBayes to parametrize the Lognormal distribution function `dnLnorm`. The whole formula will look like this:

```
mean <- 49-38
sigma <- 0.25
MU <- ln(mean) - ((stdv*stdv) * 0.5)
root_time ~ dnLnorm(MU, sigma, offset=38)
```

For the two internal node calibrations, we have to create topology constraints on which we will condition the birth-death process tree topology. To do this, the "clade" function is used. Check in the file how the two clades are constrained and loaded into a time tree, together with the birth and death rate, sampling fraction, and root age.

Finally, we set the starting value of the time tree to the tree T that we have been reading in at the beginning of the file. We will NOT add any topology moves later on, so that the trees that are sampled during the MCMC effectively have a fixed topology.

Questions:

- Could we just have "clamped" the tree instead?
- What assumption does the "samplingStrategy='uniform'" in the dnBDP distribution refer to? Do you think that this assumption is well met in our case (consider not only the bears but also the outgroups)?

1.c: Setting up node calibrations

The node calibration script is already added to the skeletal Rev script. It uses a trick to set up the node calibration which might be confusing at first. The problem is that the age of a node is determined by its tree object, so except for the root, we cannot directly access it. This becomes clear when reading the line

```
tmrca_Ursidae := tmrca(timetree,clade_Ursidae)
```

- the time of the most recent common ancestor of all Ursidae is here defined as a deterministic node depending on the tree. So we cannot also define this age as a stochastic node and associate a prior distribution to it.

So what we do is the following: Instead of assuming that our fossil age gives us a distribution on the age of the node, we say that the age of the fossil is a stochastic node, and that this age follows a certain distribution with respect to the age of our calibration node. This stochastic node is then clamped to the observed age of the fossil.

This is what is detailed in the script. In order to elucidate this approach, do the following exercise:

- make a graphical drawing of a tree, calibration node and associated fossil
- add a calibration prior to the drawing for the age of the calibration node
- now add the actual prior distribution as we have set it on the age of the fossil. Why are some of the numbers in the script negative?

Exercise 2: Relaxed-clock models

2.a: Set up a relaxed-clock model

The base rate of the clock is here obtaining a lognormal prior (see skeletal file). To this base rate, we now add a **relaxed-clock model**, i.e., a model that accounts for variation in evolutionary rates among the branches of the tree.

For setting up the relaxed-clock model, have a look at the script which specifies an uncorrelated lognormal model.

[Extra exercise for the quick ones: Modify this part of the script in order to obtain branch-specific rates with a gamma instead of a lognormal distribution. Set the rate parameter of the gamma prior to 1. Both the mean and the variance of the gamma distribution are then equal to the shape parameter. As a hyperprior on the shape parameter, set an exponential distribution with a mean of 0.5 (don't forget to add a scale move to this parameter). Copy-paste your new gamma-distributed relaxed-clock model into the corresponding section of the skeletal file.]

For everyone: have a look at the remainder of the file, and answer the following questions:

- What do the final branch-specific rates that are passed to the phyloCTMC distribution consist of?
- Which priors will have an influence on the actual age estimates?

2.b: Run the node-dating analysis

Run the node dating MCMC, and interpret the outcome:

- Compare the resulting consensus tree with the tree that we used as a starting tree. How high is the node support for each node? Why? What has changed during the MCMC? What stayed the same?

- What can you say about convergence? Which parameters were most difficult to get convergence on? Use Tracer.

Exercise 3: Running priors only

In node dating, calibration priors on node ages will interact with each other, which means that the effective age calibrations might be much more narrow than the priors we have put on the individual nodes. To investigate this, run the analysis again, but this time without data, in order to obtain the effective prior distributions on the node ages. To do so, use the following command:

```
mymcmc.run(generations=10000,underPrior=true)
```

Look at the resulting tree and node age priors in Tracer!