Phylogenetic Inference using `RevBayes`

Total-evidence Dating under the FBD Model

Tracy A. Heath, April Wright, and Walker Pett

1 Introduction

Ronquist et al. (2012)

1.1 Models

1.1.1 Sequence Evolution

Point to other tutorials (e.g., GTR stuff)

1.1.2 Morphological Character Change

Mk models and ascertainment bias

1.1.3 Lineage-Specific Substitution Rates

Clocks (Zuckerkandl and Pauling 1962) and relaxing them

1.1.4 Lineage Diversification and Sampling

Birth-death processes and FBD

2 Prerequisites

What do you need to know before doing this?

2.1 Requirements

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

- RB_Getting_Started
- RB_Basics_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(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

- **bears_cytb.nex**: description of file/data (we also need more descriptive file names).

4 Exercise: Title

4.1 Getting Started

4.2 Creating Rev Files

For complex models and analyses, it is best to create Rev script files that will contain all of the model parameters, moves, and functions. In this exercise, you will work primarily in your text editor and create a set of modular files that will be easily managed and interchanged.

4.3 Start the Master Rev File and Import Data

> Open your text editor and create the master Rev file called mcmc_TEFBD.Rev in the **scripts** directory.
>
> Enter the Rev code provided in this section in the new model file.

The first command will load in the sequences using the **readDiscreteCharacterData()** function and assign it to a variable called **cytb**.

```
cytb <- readDiscreteCharacterData("data/bears_cytb.nex")
```

Next, import the morphological character matrix and assign it to the variable **morpho**.

```
morpho <- readDiscreteCharacterData("data/bears_morphology.nex")
```

Now we read in the full list of taxa and create a workspace object with the total number of taxa.

```
taxa <- readTaxonData("data/bears_taxa.tsv", delimiter=TAB)
n_taxa <- taxa.size() # the number of taxa
```

If you open the bears taxa file (**bears_taxa.tsv**), you'll notice that this is a tab-separated file of all of the taxon names, with the age in millions of years ago (mya) in the second column. An age of 0.0 indicates an extant bear species. We will use this information to allow fossils to be incorporated as tips in the analysis.

Notice that the two data matrices have different numbers of taxa. The last bit of data preparation we will do is to add any taxa that are not found in the molecular partition (i.e. are only found in the fossil data) to the molecular partition as missing data, and vice versa. In order for all the taxa to appear on the same tree, they all need to be part of the same dataset, as opposed to present in separate datasets. This ensures that there is a unified taxon set that contains all of our tips.

```
cytb.addMissingTaxa( taxa )
morpho.addMissingTaxa( taxa )
```

Now that our data are loaded and prepared, we can specify the fossilized birth-death model, and the MCMC moves associated with it. As we work, we will be creating moves to govern how values are sampled for our various priors and parameters. We'll first create an iterator to hold all of our moves.

```
mvi = 1
```

4.4 The Fossilized Birth-Death Process

Open your text editor and create the fossilized birth-death model file called `model_FBDP_TEFBD.Rev` in the **scripts** directory.

Enter the `Rev` code provided in this section in the new model file.

Two key parameters of the FBD process are the birth rate (the rate at which lineages are added to the tree) and the death rate (the rate at which lineages are removed from the tree). We'll place exponential priors on both of these values. An exponential prior with a $\lambda = 10$ places a higher probability on values less than 0.4 for these parameters.

```
birth_rate ~ dnExponential(10)
death_rate ~ dnExponential(10)
```

Now that the priors have been specified, we give RevBayes some information on how to sample values for our parameters. We'll use a scaling move, which changes the value sampled multiplicatively with the tuning parameter. We will use three different tuning parameters, which govern the size of the move. Including multiple tuning parameters improves mixing.

```
moves[mvi++] = mvScale(birth_rate, lambda=0.01, weight=3.0)
moves[mvi++] = mvScale(birth_rate, lambda=0.1, weight=3.0)
moves[mvi++] = mvScale(birth_rate, lambda=1.0, weight=3.0)
moves[mvi++] = mvScale(death_rate, delta=0.01, weight=3.0)
moves[mvi++] = mvScale(death_rate, delta=0.1, weight=3.0)
moves[mvi++] = mvScale(death_rate, delta=1, weight=3.0)
```

In order to print nodes in our graph to output (also called *monitoring*), we need to create deterministic nodes for the diversification and turnover. Deterministic nodes are value transformations between existing stochastic nodes. So we will define diversification and turnover as deterministic nodes.

```
diversification := birth_rate - death_rate
turnover := death_rate/birth_rate
```

All extant bears are represented in this dataset. Therefore, we can fix the sampling probability of extant lineages to 1.

```
rho <- 1.0
```

The rate of sampling fossils (ψ), on the other hand is not known. We will use an exponential prior on this parameter as well, and use a slide move to sample values from our distribution.

```
psi ~ dnExponential(10)
moves[mvi++] = mvScale(birth_rate, lambda=0.01, weight=3.0)
moves[mvi++] = mvScale(birth_rate, lambda=0.1, weight=3.0)
moves[mvi++] = mvScale(birth_rate, lambda=1.0, weight=3.0)
```

Under the FBD model, the process is conditioned on the age of the origin, or the start of the process. We will specify a uniform distribution on the age of the origin. If you looked in the bears taxa file, you might notice that the age of the oldest fossil is slightly younger than the upper bound of the uniform distribution on the origin age. For this parameter, we will use a sliding window move. A sliding window move samples within an interval (defined by **`delta`**) . Sliding window moves can be tricky for small values, as the window may overlap zero. However, for parameters such as the origin, there is little risk of this being an issue.

```
origin_time ~ dnUnif(37.0, 55.0)
moves[mvi++] = mvSlide(origin_time, delta=0.01, weight=10.0)
moves[mvi++] = mvSlide(origin_time, delta=0.1, weight=10.0)
moves[mvi++] = mvSlide(origin_time, delta=1, weight=10.0)
```

All the parameters of the FBD process are now defined. The next step is to combine these parameters to define the tree prior as the FBD.

```
tree_prior = dnFBDP(origin=origin_time, lambda=birth_rate, mu=death_rate, psi=psi, rho
    =rho, taxa=taxa)
```

Next, we will define the **fbd_tree** variable as a random variable. It will be used to generate trees under the FBD process that conform to our clade constraints.

```
fbd_tree ~ dnConstrainedTopology(tree_prior, constraints)
```

Finally, we can also create deterministic nodes for other quantities we might be interested in monitoring. Below, we will define a monitor that prints the number of fossils that are inferred to be 'sampled ancestors' - lineages that are present in the phylogeny, and have descendants present on the tree. We will also define a deterministic node for the age of the crown group of bears, using the previously-defined extant bear constraint (Section **??**).

```
sa := fbd_tree.numSampledAncestors();
crown := tmrca(fbd_tree, clade_extant)
```

4.5 The Uncorrelated Exponential Relaxed-Clock Model

Open your text editor and create the lineage-specific branch-rate model file called `model_UCExp_TEFBD.Rev` in the **scripts** directory.

Enter the `Rev` code provided in this section in the new model file.

4.6 The General-Time Reversible Gamma-Rates Model of Sequence Evolution

Open your text editor and create the molecular substitution model file called `model_GTRG_TEFBD.Rev` in the **scripts** directory.

Enter the `Rev` code provided in this section in the new model file.

4.7 Modeling the Evolution of Binary Morphological Characters

Open your text editor and create the morphological character model file called `model_Morph_TEFBD.Rev` in the **scripts** directory.

Enter the `Rev` code provided in this section in the new model file.

4.8 Complete MCMC File

Return to the master `Rev` file you created in Section 4.3 called `mcmc_TEFBD.Rev` in the **scripts** directory.

Enter the `Rev` code provided in this section in this file.

4.8.1 Source Model Scripts

```
source("scripts/model_FBDP_TEFBD.Rev")

source("scripts/model_UCExp_TEFBD.Rev")

source("scripts/model_GTRG_TEFBD.Rev")

source("scripts/model_Morph_TEFBD.Rev")
```

4.8.2 Create Model Object

```
mymodel = model(sf)
```

4.8.3 Specify Monitors and Output Filenames

```
mni = 1
monitors[mni++] = mnModel(filename="output/bears.log", printgen=10)
monitors[mni++] = mnFile(filename="output/bears.trees", printgen=10, fbd_tree)
monitors[mni++] = mnScreen(printgen=10, age_extant, num_samp_anc, origin_time)
```

4.8.4 Set up the MCMC

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

mymcmc.run(generations=10000)
```

Save and close all files.

4.9 Run it

```
./rb
```

Execute the MCMC analysis:

```
source("scripts/mcmc_TEFBD.Rev")
```

```
   Processing file "scripts/mcmc_TEFBD.Rev"
   Successfully read one character matrix from file 'data/bears_cytb.nex'
   Successfully read one character matrix from file 'data/bears_morphology.nex'
   Processing file "scripts/model_FBDP_TEFBD.Rev"
   Processing of file "scripts/model_FBDP_TEFBD.Rev" completed
   Processing file "scripts/model_UCExp_TEFBD.Rev"
   Processing of file "scripts/model_UCExp_TEFBD.Rev" completed
   Processing file "scripts/model_GTRG_TEFBD.Rev"
   Processing of file "scripts/model_GTRG_TEFBD.Rev" completed
   Processing file "scripts/model_Morph_TEFBD.Rev"
   WARNING: There are 19 characters incompatible with the specified coding bias. These characters
   will be excluded.
   Processing of file "scripts/model_Morph_TEFBD.Rev" completed

   Running MCMC simulation
   This simulation runs 1 independent replicate.
   The simulator uses 163 different moves in a random move schedule with 267 moves per iteration

Iter      |     Posterior  |    Likelihood  |        Prior  |    age_extant  |  num_samp_anc  |   origin_time  |   elapsed  |       ETA  |
-------------------------------------------------------------------------------------------------------------------------------------------------
0         |      -8174.01  |       -8053.8  |      -120.209  |       34.8641  |            0   |       44.4332  |  00:00:00  |  --:--:--  |
10        |      -4654.95  |       -4611.2  |      -43.7495  |       4.32618  |            7   |       45.4494  |  00:00:01  |  --:--:--  |
20        |      -4294.05  |      -4266.91  |      -27.1443  |       4.58804  |            7   |       46.5636  |  00:00:01  |  00:08:19  |
30        |      -4267.35  |      -4233.41  |       -33.94  |        6.8467  |            6   |       45.9177  |  00:00:02  |  00:11:04  |
40        |      -4226.63  |      -4188.32  |      -38.3037  |       6.40484  |            8   |       44.3696  |  00:00:02  |  00:08:18  |
```

4.10 Summarize Your Results

4.10.1 Evaluate MCMC

4.10.2 Summarize Tree

Start up RevBayes at the command line. You should do this from within the **RB_TotalEvidenceDating_FBD_Tutorial** directory.

```
./rb
```

Read in the MCMC sample of trees from file.

```
trace = readTreeTrace("output/bears.trees")

for(i in 1:trace.size())
{
    trees[i] = fnPruneTree(trace.getTree(i), pruneTaxa=v(taxa[17],taxa[20]))
}

trace_pruned = treeTrace(trees)
mccTree(trace_pruned, "output/bears.mcc.tre" )
```

See Fig. 1

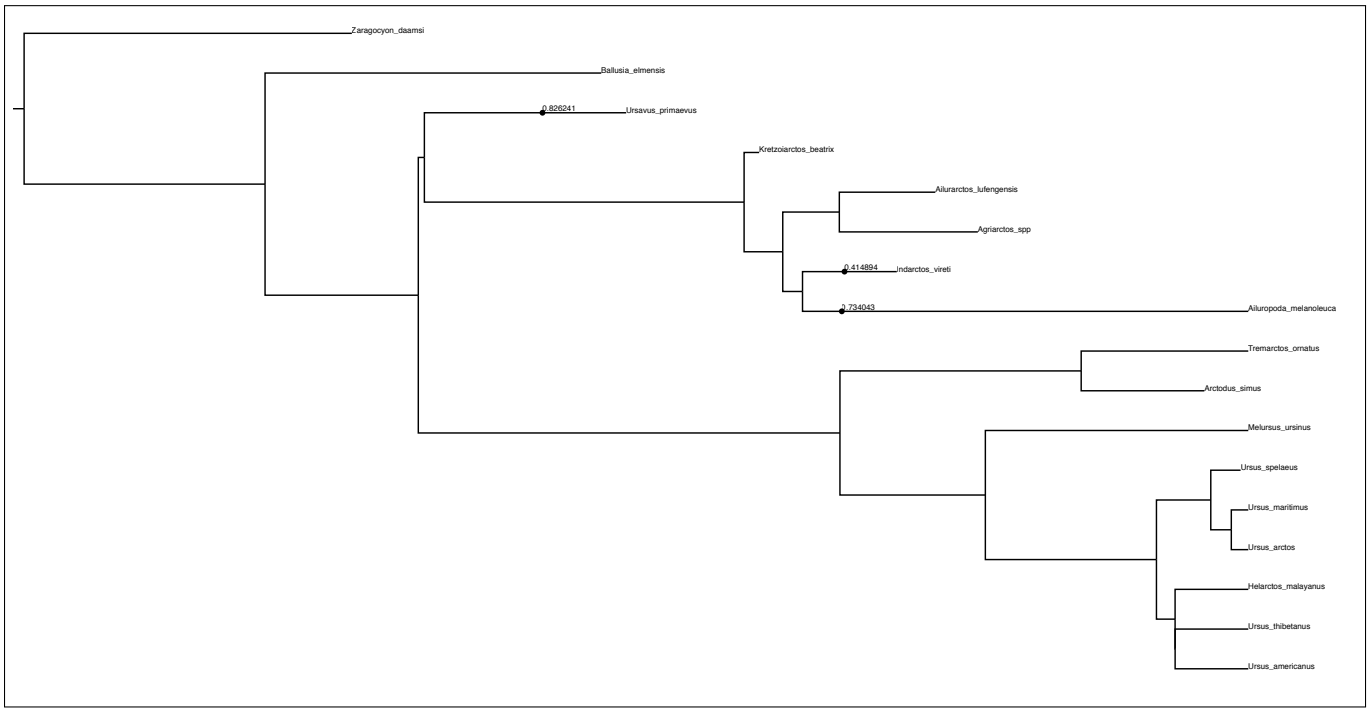

Figure 1: This is a place-holder figure.

References

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

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: December 19, 2016