Updated from:
Relaxed Phylogenetics & Dating with Confidence by Drummond, Rambaut, and Xie
DivTime BEAST Tutorial from Tracy Heath

Lab 09: Divergence Time Estimation with BEAUti and the BEAST

Introduction

Today you will be introduced to the software BEAST, focusing on estimating phylogenies and divergence times when you have calibration information from fossil evidence. This tutorial is modeled after the Divergence Dating (Primates) tutorial available online, but is updated for the current version of BEAST (v1.8). BEAST (Bayesian Evolutionary Analysis by Sampling Trees) uses Bayesian Inference, so expect some familiar terms in this tutorial such as "prior", "MCMC", and "posterior". Refer to the lecture and lab on MrBayes to refresh your memory on what these terms mean.

Today we will explore divergence time estimation using the BEAST and accompanying programs. The goals of the lab are to:

I. Convert a NEXUS to an XML file using BEAUti
II. Run this XML file using BEAST
III. Analyze the output files using TRACER
IV. Generate a tree with posterior probability values in TreeAnnotator
V. View your tree in FigTree

Exercise 1: Using BEAUti to Create an XML file

BEAST uses eXtensible Markup Language (XML) files for initial input. These files allow for the combination of text and additional information. In this way the character matrix data can be stored alongside the analysis specifications that you will make for specific analyses. The XML file specifies sequences, node calibrations, models, priors, output file names, etc. We are starting with a NEXUS file with today's exercise, which we need to convert to an XML. The program BEAUti (Bayesian Evolutionary Analysis Utility) will do this for you and is automatically included when you download BEAST. First open up the BEAUti program, which will be located within your BEAST folder. Now import your NEXUS file. **File → Import Data**. Select the *primates.nex* file that is included with the BEAST example files within the DATA folder.

You will notice the file imports as two partioins "firsthalf" and "secondhalf". This is because the NEXUS file contains this block:

```
begin assumptions;
charset firsthalf = 1-449;
charset secondhalf = 450-898;

end;
```

You might want to do something similar with your own data – if you have several different gene regions represented, each one can be separated as a partition.

Taxa

Select Taxa at the top of the window. In this screen you can create sets of taxa and then later on add calibration for that set's most recent common ancestor (MRCA). Press the "plus" button in the bottom left of the screen. This will create a new set. Let's call this set *ingroup* and include all taxa except Lemur_catta. You can do this by selecting all the taxa and pressing the green arrow to move the taxa into the "Included taxa" right-hand column. Then re-select Lemur_catta and move it back to the left "Excluded taxa" column. Select the Monophyletic? option since we know that the Lemur is the outgroup. This ensures that the ingroup is kept monophyletic during the course of the MCMC analysis. Repeat this to make two additional groups:

1. Human-Chimp: Includes only Homo_sapiens and Pan
2. HomiCerco: Include everything except Lemur, Saimiri, and Tarsius (this contains everything under the hominoid/cercopithoid split)

We may not be too sure about the monophyly of these two groups, so leave the Monophyletic? option unchecked.

Partitions

Now go back to the partitions tab. Select both of the partitions in the box and then select **Unlink Subst. Models**. This will allow the two partitions to have independent substitution models.

We want to link the clock models for these so Select **Link Clock Models**. You can rename the clock model partition to *primatesClock*.

Sites

Click on the "Sites" tab. This shows the evolutionary model setting for BEAST. We will leave both partitions as the HKY substitution model (but check to see what other options are available). Change the Site Heterogeneity Model to **Gamma**. Make sure to do this for both partitions.

Clocks

Click on the "Clocks" tab. The model is currently set to a **Strict clock**, but let's change this to **Lognormal relaxed clock (Uncorrelated)**.

Trees

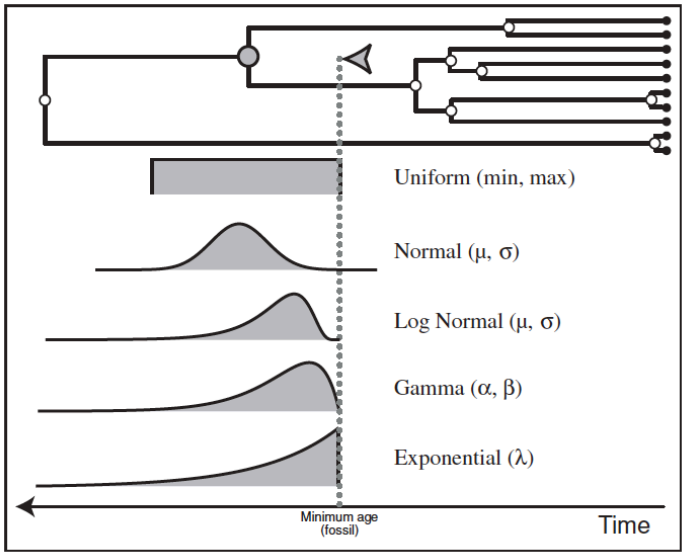
Click on the "Trees" tab. Look at all the different options that you have for the **Tree Prior**. You will want to think carefully about which of these options works best for you if you use this for your own project. Today we'll select the **Yule Process**, which is a simple model of speciation and generally appropriate for the primates data here of sequences from different species. BEAUti

very nicely gives us the citations for these priors, so if you are interested in the differences between these trees, you can delve into those papers.

Priors

Click on the "Priors" tab. These are the priors, which are specified for each parameter in the model. This is where we can specify a prior distribution for some of the divergence times based on our prior fossil knowledge, in other words, calibrate our tree. Click on the button next to **tmrca(human-chimp)**. Select the Normal distribution. We will assume a normal distribution centered around 6 million years with a standard deviation of 0.5 million years. This will give a central 95% range of about 5-7 My, which corresponds to the consensus estimate of the date of the most recent common ancestor of humans and chimps. Following the same procedure set a calibration of 24+/- 0.5 million for the hominoid-cercopithecoid split. Although we created a taxon set for the ingroup (tmrca(ingroup) in the prior table), we are not going to put an informative prior on this.

Question #1: **What do you think about these priors? Good? Bad? I don't know what to think. If you're interested in divergence time estimation, this is a pretty important prior. This figure might help you think about this.**



You will also need to select a prior for **ucld.mean** Set this prior as a diffuse Gamma distribution with shape 0.001 and scale 1000.

MCMC

Click on the "MCMC" tab. The **Length of chain** option specified the number of steps the MCMC will make before finishing. When you run your own analysis you will want to think carefully about what number to set this at. The value of 10,000,000 is the default and completely arbitrary. Change this to 800,000 for today's exercise. The **Echo state to screen** option specifies how often the parameter values in the Markov chain should be displayed on the screen. This output is simply for monitoring the program's progress so can be set to any value (although if set too small, the sheer quantity of information being displayed on the screen will actually slow the program down). For the **Log parameters**, the value should be set relative to the total

length of the chain. Sampling too often will result in very large files with little extra benefit in terms of the precision of the analysis. Sample too infrequently and the log file will not contain much information. You want to aim to store no more than 10,000 sample so this should be set to no less than chain length / 10000. For today, set the screen log to 10,000 and the log parameters to 200. Change the file name stem if you'd like. If you are using Windows, check the box to add .txt suffix so that Windows recognizes these as text files.

Okay, we are now ready to generate the XML file. Simply click on the **Generate BEAST file…** located in the lower right corner. It will ask you if you are sure about your prior settings. Select Continue and tell it where to save this file (I created a separate file for the day to keep track IB200_Lab9) You can check out this file in a text editor to see what it looks like.

Exercise 2: Running BEAST

Now open the BEAST program by double-clicking the icon. Select the XML file that we just created. Keep the options in their default settings and select **Run** in the bottom right corner. If it worked, you should see it running and reporting values to the screen with the increments you specified in BEAUti.

That's it! It ran pretty quickly for me.

Exercise 3: Analyze Your Results with Tracer

The output files from your BEAST run will have saved into the folder with the XML file that you saved. Isn't that wonderfully convenient?

Open the program Tracer. **File → Import Trace File →** and locate the .log file that BEAST generated. For my program, I needed to pull up the Traces panel in the bottom left corner to see all of the statistics. You should see four columns and about 28 rows, some black, some orange, and some red. You will probably have many values in orange and red – this is not good. Examine the Tracer window for your parameters.

Question #2: **Why do you think some of these values are red? What could you do in the analysis to remedy this problem? What do your Tracer windows look like for parameters that are colored black? For parameters that are colored red?**

Initially, the **posterior** value will be highlighted. Select **meanRate** to look at the rate of evolution averaged over the whole tree. Tracer will plot a (marginal posterior) distribution for the selected parameter and also give you statistics such as the mean and median. The 95% HPD stands for highest posterior density interval and represents the most compact interval on the selected parameter that contains 95% of the posterior probability. It can be thought of as a Bayesian analog to a confidence interval.

The **coefficientofVariation** statistic gives a summary of how much the rate of evolution varies from lineage to lineage (expressed as a proportion of the mean rate).

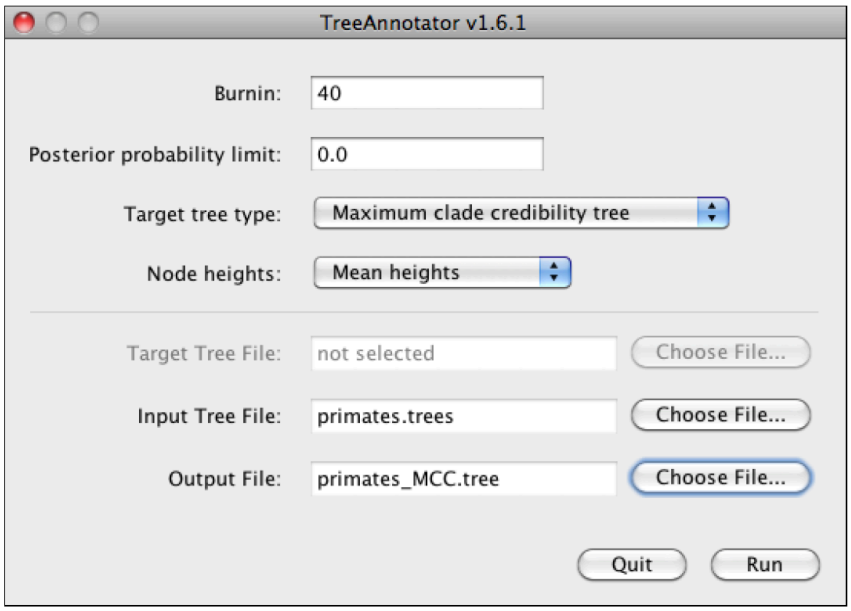
The **treeModel.rootHeight** parameter gives the marginal posterior distribution of the age of the root of the entire tree.

Select the **treeModel.rootHeight** parameter along with the next three (hold shift while selecting). This will show a disiplay of the age of the root and the three MRCAs we specified in BEAUti. The parameter that we used to calibrate the tree (tmrca(human-chimp)) will have posterior distributions very similar to the prior distributions that we specified. Switch the tab at the top of the window to **Marginal Density**. This will display a plot of the marginal posterior densities of each of these date estimates overlayed. What do each of the colors correspond to? Select the legend option at the bottom of the screen to figure this out.

Question #3: **Take a screen shot of the Marginal densities with the legend included.**

Exercise 4: Create a Consensus Tree with TreeAnnotator

BEAST also produces a sample of plausible trees along with its sample of parameter estimates. These need to be summarized using the program TreeAnnotator. This will take the set of trees and find the best supported one. It will then annotate this summary tree with the mean ages of all the nodes and the HPD ranges. It will also calculate the posterior clade probability for each node. Run the TreeAnnotator program and set it up to look like this:



The burnin is the number of trees to remove from the start of the sample. Unlike Tracer which specifes the number of steps as a burnin, in TreeAnnotator you need to specify the actual number of trees. For this run, you specifed a chain length of 800,000 steps sampling every 200 steps. Thus the trees file will contain 4000 trees and so to specify a 1% burnin use the value 40. The Posterior probability limit option speci_es a limit such that if a node is found at less than this frequency in the sample of trees (i.e., has a posterior probability less than this limit), it will not be annotated. The default of 0.5 means that only nodes seen in the majority of trees will be

annotated. Set this to zero to annotate all nodes. For Target tree type you can either choose a specific tree from a file or ask TreeAnnotator to find a tree in your sample. The default option, Maximum clade credibility tree, finds the tree with the highest product of the posterior probability of all its nodes. Choose Mean heights for node heights.This sets the heights (ages) of each node in the tree to the mean height across the entire sample of trees for that clade. For the input file, select the trees file that BEAST created and select a file for the output.
Now press Run.

Exercise 5: View Your Tree in FigTree

We can view this summary tree in FigTree. Open the file that you just created using TreeAnnotator. Try selecting **Node Bars** to get node age error bars. Also turn on **Branch Labels** and select posterior to get it to display the posterior probability for each node (get rid of excessive sig. digits). Under **Appearance** you can also tell FigTree to color the branches by the rate.

Question #4: **Take a screen shot of this tree in FigTree and send it to me.**

Question #5: **My posterior probabilities on the nodes of my tree were pretty high. So should I feel good about this? Think about this in relation to your Tracer analysis.**