Relaxed molecular clocks and dating
A hands-on practical

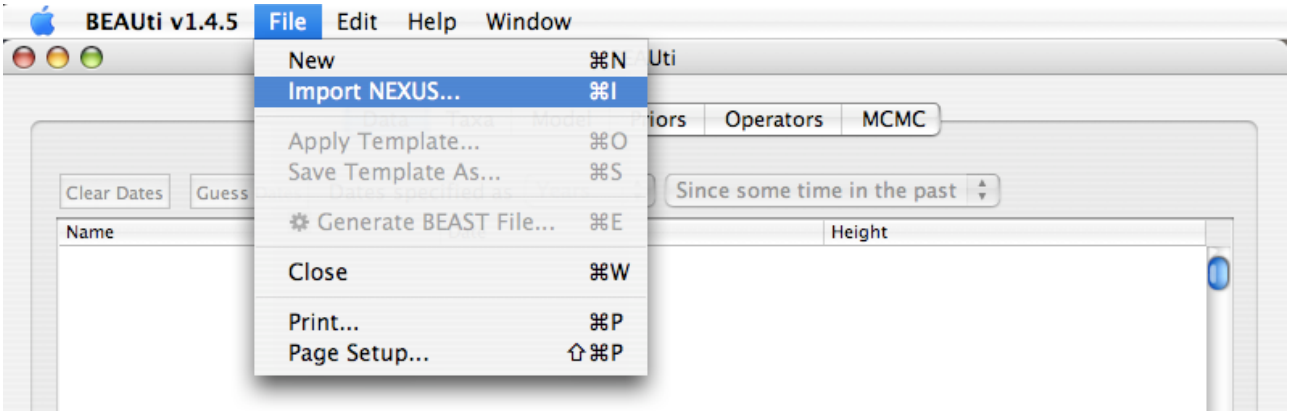
This practical will guide you through the use of BEAUti and BEAST to analyze an alignment of primate sequences and estimate divergence times based on two independent fossil calibrations. BEAST is unique in its ability to estimate the phylogenetic tree and the divergence times simultaneously.

BEAUti

The program **BEAUti** is a user-friendly program for setting the model parameters for BEAST. Run BEAUti by double clicking on its icon.

Loading the NEXUS file

To load a NEXUS format alignment, simply select the **Import NEXUS..**. option from the **File** menu:

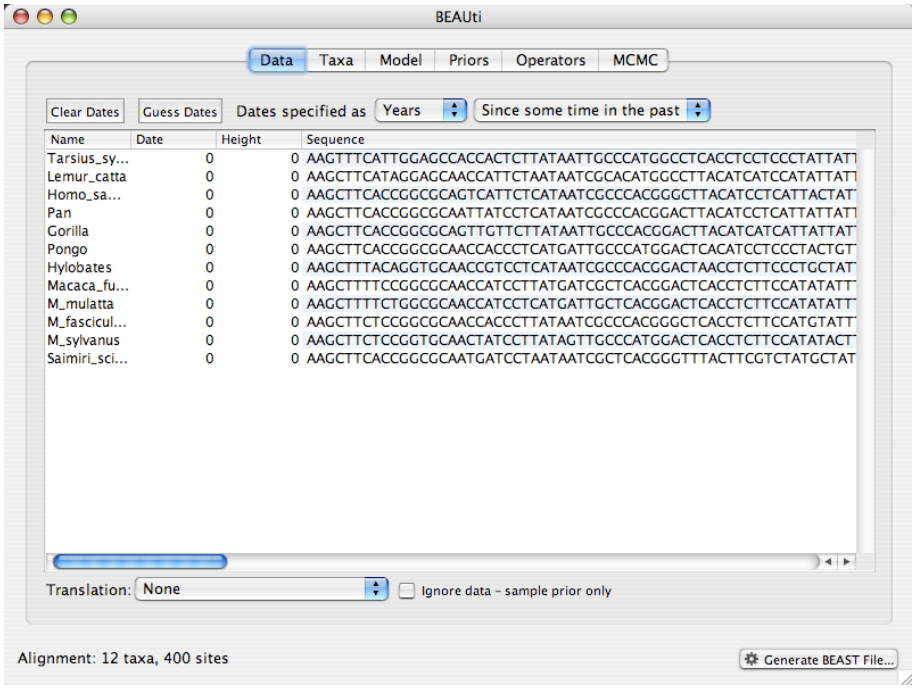


The NEXUS alignment

Select the file called `primates.nex`. This file contains an alignment of mitochondrial sequences from 12 primate species. It looks like this (the lines have been truncated):

```
#NEXUS
begin data;
dimensions ntax=12 nchar=400;
format datatype=dna interleave=no gap=-;
matrix
Tarsius_syrichta  AAGTTTCATTGGAGCCACCACTCTTATAATTGCCCATGGCCTCACCTCCTCCCTATTATTTT…
Lemur_catta       AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCCATATTATTCT…
Homo_sapiens      AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCATTACTATTCT…
Pan               AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCATTATTATTCT…
Gorilla           AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCATTATTATTCT…
Pongo             AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCCCTACTGTTCT…
Hylobates         AAGCTTTACAGGTGCAACCGTCCTCATAATCGCCCACGGACTAACCTCTTCCCTGCTATTCT…
Macaca_fuscata    AAGCTTTTCCGGCGCAACCATCCTTATGATCGCTCACGGACTCACCTCTTCCATATATTTCT…
M_mulatta         AAGCTTTTCTGGCGCAACCATCCTCATGATTGCTCACGGACTCACCTCTTCCATATATTTCT…
M_fascicularis    AAGCTTCTCCGGCGCAACCACCCTTATAATCGCCCACGGGCTCACCTCTTCCATGTATTTCT…
M_sylvanus        AAGCTTCTCCGGTGCAACTATCCTTATAGTTGCCCATGGACTCACCTCTTCCATATACTTCT…
Saimiri_sciureus  AAGCTTCACCGGCGCAATGATCCTAATAATCGCTCACGGGTTTACTTCGTCTATGCTATTCT…
;
end;
```

Once loaded, the list of taxa and the actual alignment will be displayed in the main panel:

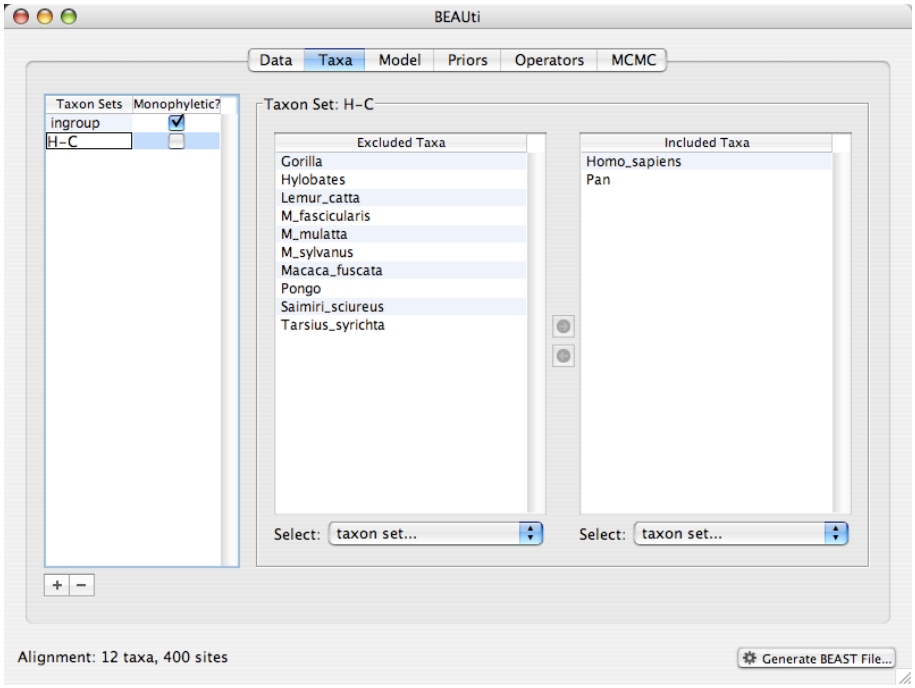


Defining the calibration nodes

Select the `Taxa` tab at the top of the main window. You will see the panel that allows you to create sets of taxa that will enable you to put calibration information for each of their most recent common ancestors (MRCAs). Press the small “plus” button at the bottom left of the panel:

This will create a new taxon set. Rename it by double-clicking on the entry that appears (it will initially be called `untitled1`). Call it `ingroup` (it will contain all taxa except the lemur, which will form the outgroup). In the next table along you will see the available taxa. Select all taxa and press the green arrow button. Move the lemur back into the excluded taxa set. Since we know that lemur is the outgroup, we will set select the checkbox in the `Monophyletic?` column. This will ensure that the ingroup is kept monophyletic during the course of the MCMC analysis.

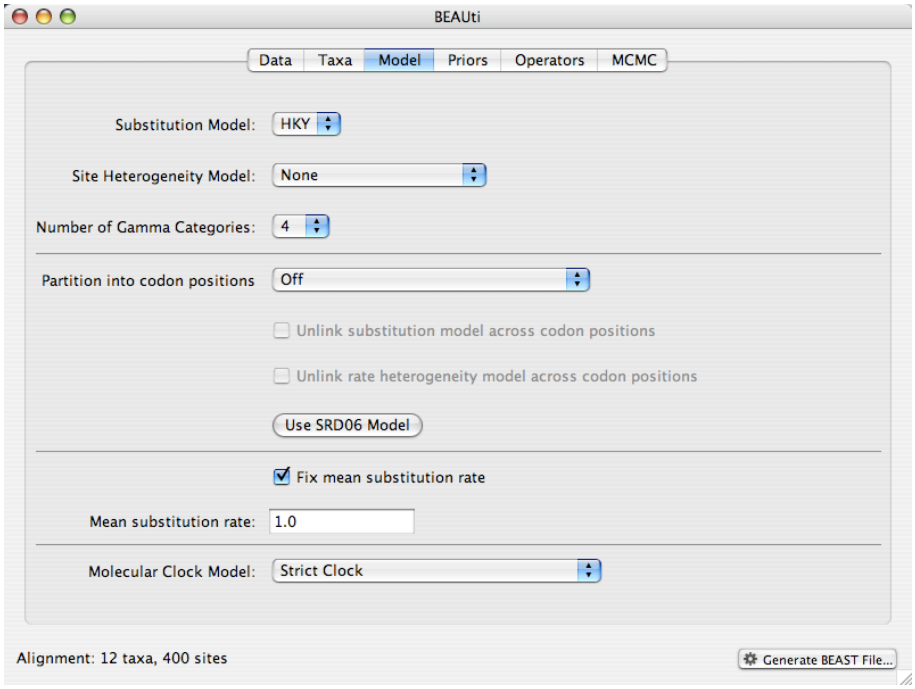
Now repeat the whole procedure creating a set called `H-C` that contains on the human and chimp. The screen should look like this:



Finally, create a taxon group that contains everything under the hominoid/cercopithecoid split (i.e. everything except Lemur, Saimiri and Tarsius). Call this taxon set something like `HomiCerco`.

Setting the evolutionary model

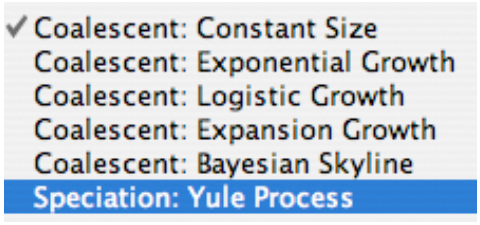
The next thing to do is to click on the Model tab at the top of the main window. This will reveal the evolutionary model settings for BEAST. Exactly which options appear depend on whether the data are nucleotides or amino acids (or nucleotides translated into amino acids). The settings that will appear after loading the Primate data set will be as follows:



Most of the models should be familiar to you. For this analysis, we will make two changes. First you need to turn off the `Fix mean substitution rate` option. This is because we wish to estimate the mean substitution rate (and in doing so the divergence times). Ignore the warning that appears. The second thing we will do is to change the molecular clock model to `Relaxed Clock: Uncorrelated Log-normal` so as to account for lineage-specific rate heterogeneity.

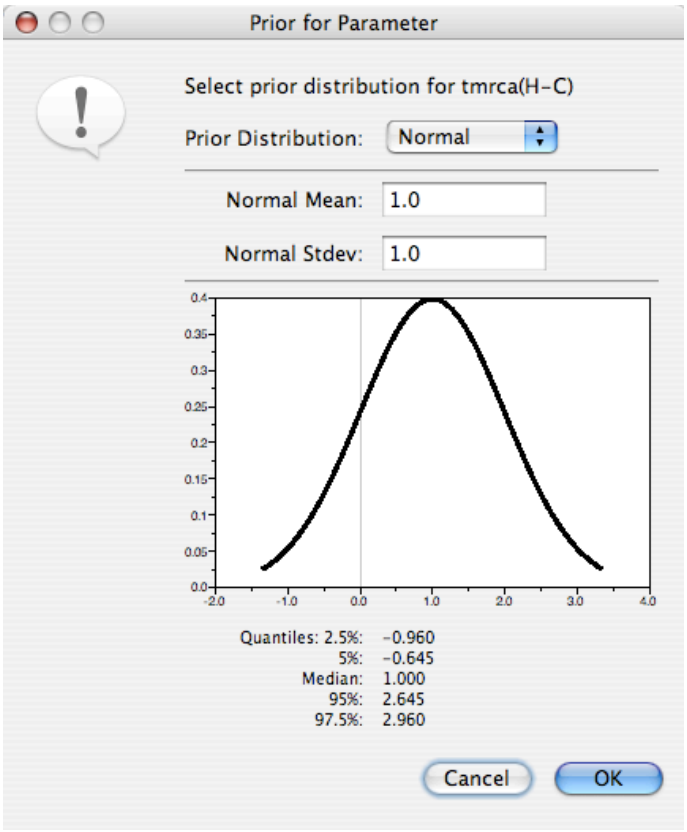
Priors

The next tab allows priors to be specified for each parameter in the model. The first thing to do is to specify that we wish to use a Yule process as the tree prior. This is a simple model of speciation that is more appropriate when considering sequences from different species. Select this from the menu:



We now need to specify a distribution for the divergence of humans and chimpanzees based on our prior fossil knowledge. This is known as calibrating our tree. We will actually use multiple calibrations in this analysis; one on the human-chimp split and one on the hominoid-cercopithecoid split. Click on the button in the table next to `tmrca(H-C)`:

A dialog box will appear allowing you to specify a prior for this MRCA. Select the `Normal` distribution:

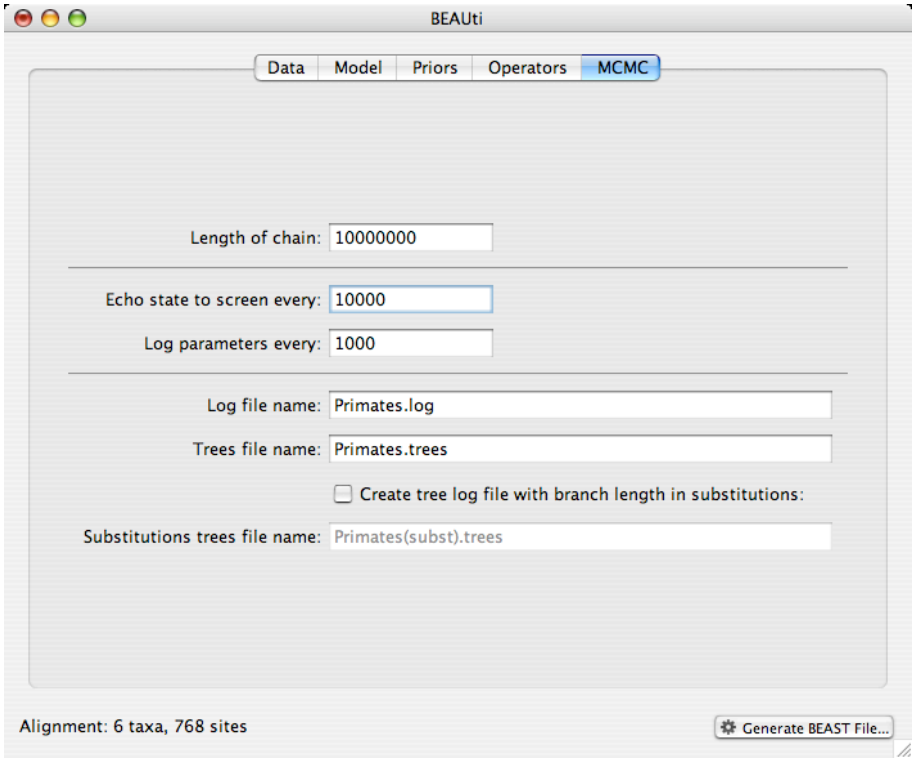


We are going to assume a normal distribution centered at 6 million years with a standard deviation of 0.5 million years. This will give a central 95% range of about 5-7.

Following the same procedure set a calibration of 24 million years +/- 0.5 million (stdev) for the hominoid-cercopithecoid split.

Setting the MCMC options

Ignore the `Operators` tab as this just contains technical settings for the MCMC program. The next tab, `MCMC`, provides settings to control the MCMC:



Firstly we have the `Length of chain`. This is the number of steps the MCMC will make in the chain before finishing. How long this should be depends on the size of the data set, the complexity of the model and the quality of answer required. The default value of 10,000,000 is entirely arbitrary and should be adjusted according to the size of your data set.

For this data set let's initially set the chain length to 2,000,000 as this will run reasonably quickly on most modern computers (a few minutes).

The next options specify how often the current parameter values should be displayed on the screen and recorded in the log file. The screen output is simply for monitoring the programs 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 file, 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 about the distributions of the parameters.

Set the screen log to 10000 and the file log to 200.

The final two options give the file names of the log files for the parameters and the trees. These will be set to a default based on the name of the imported NEXUS file.

- If you are using Windows, we suggest you add the suffix `.txt` to both of these (so, `Primates.log.txt` and `Primates.trees.txt`) so that Windows recognizes these as text files.

Generating the BEAST XML file

We are now ready to create the BEAST XML file. Select **Generate BEAST File...** from the **File** menu and save the file with an appropriate name (we usually end the filename with '.xml'). We are now ready to run the file through BEAST.

Running BEAST

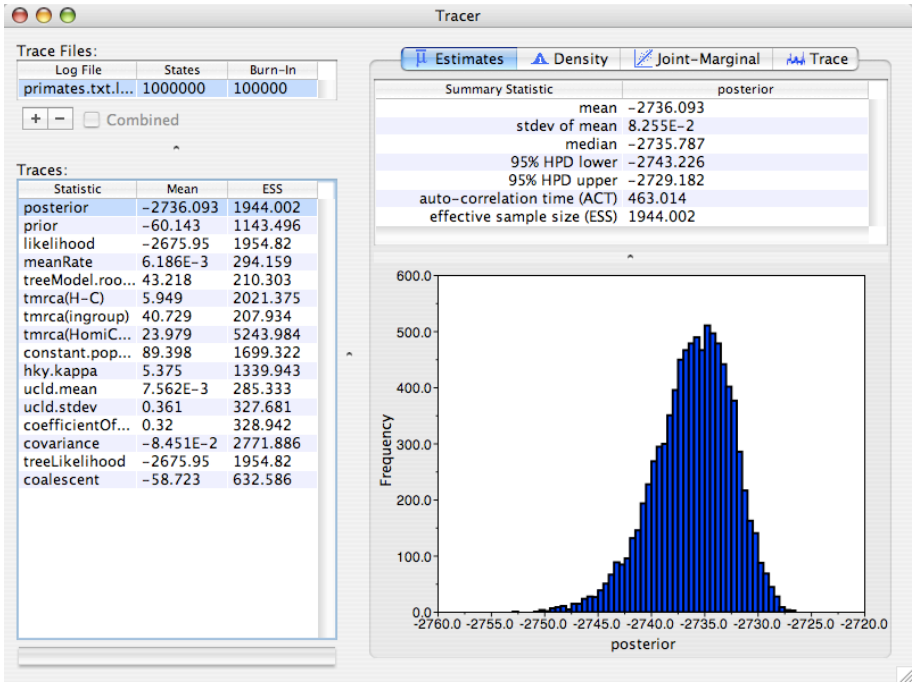
Now run BEAST and when it asks for an input file, provide your newly created XML file as input. BEAST will then run until it has finished reporting information to the screen. The actual results files are save to the disk in the same location as your input file and will look something like this:

```
                  BEAST v1.4.7, 2002-2008
       Bayesian Evolutionary Analysis Sampling Trees
                           by
           Alexei J. Drummond and Andrew Rambaut

               Department of Computer Science
                   University of Auckland
                  alexei@cs.auckland.ac.nz

             Institute of Evolutionary Biology
                  University of Edinburgh
                     a.rambaut@ed.ac.uk

Downloads, Help & Resources:
     http://beast.bio.ed.ac.uk/

Source code distributed under the GNU Lesser General Public License:
     http://code.google.com/p/beast-mcmc/

Additional programming & components created by:
     Roald Forsberg
     Gerton Lunter
     Sidney Markowitz
     Oliver Pybus

Thanks to (for use of their code):
     Korbinian Strimmer

Random number seed: 1185907250052

MacRoman
Parsing XML file: primates.xml
Read alignment, 'alignment':
  Sequences = 12
      Sites = 400
   Datatype = nucleotide
Site patterns 'patterns' created from positions 1-400 of alignment 'alignment'
  pattern count = 199
Creating the tree model, 'treeModel'
  initial tree topology =
(((((((Gorilla,M_mulatta),M_fascicularis),Macaca_fuscata),
((Hylobates,M_sylvanus),Pongo)),(Homo_sapiens,Pan)),
(Saimiri_sciureus,Tarsius_syrichta)),Lemur_catta)
Using discretized relaxed clock model.
  parametric model = logNormalDistributionModel
   rate categories = 22
Creating state frequencies model: Using emprical frequencies from data = {0.3060,
0.3294, 0.1079, 0.2567}
Creating HKY substitution model. Initial kappa = 1.0
Creating site model.
```

```
TreeLikelihood using native nucleotide likelihood core
  Ignoring ambiguities in tree likelihood.
  Partial likelihood scaling off.
Branch rate model used: discretizedBranchRates
Creating swap operator for parameter branchRates.categories (weight=30)
Creating the MCMC chain:
  chainLength=1000000
  autoOptimize=true
  fullEvaluation=2000

Pre-burnin (10000 states)
0              25             50             75            100
|--------------|--------------|--------------|--------------|
************************************************************

state Posterior       Prior          Likelihood     Root Height     Rate
0     -2,735.7205     -59.1451       -2,676.5754    40.1722         6.55132E-3
10000 -2,733.7858     -59.7735       -2,674.0123    42.6459         6.16998E-3
.
.
990000     -2,729.4067     -58.5818       -2,670.8249     42.3833         6.04659E-3
1000000    -2,732.4889     -58.8447       -2,673.6442     38.1462         6.02438E-3

Operator analysis
Operator                                 Pr(accept)  Performance suggestion
hky.kappa                       0.579    0.2900
ucld.mean                       0.722    0.2203
ucld.stdev                      0.456    0.2825
up:ucld.mean down:treeModel.allInternalNodeHeights0.866   0.2153      Try setting
scaleFactor to about 0.8760
swapOperator(branchRates.categories)        0.4417      No suggestions
constant.popSize                0.280    0.2804
treeModel.rootHeight            0.793    0.2136
treeModel.internalNodeHeights            0.2739
subtreeSlide                    3.875    0.3005      Try increasing size to about
4.976102428519987
Narrow Exchange                          0.0031
Wide Exchange                            0.0004
wilsonBalding                            0.0002
```

Analysing the results

Run the program called **Tracer** that you will find in the BEAST package. When the main window has opened, choose **Import Trace File** from the **File** menu and select the file that BEAST has created called `primates.log`. You should now see the following:



On the left hand side is a list of the different parameters and statistics that BEAST has logged. Select `meanRate` to look at the rate of evolution and `treeModel.rootHeight` to look at the marginal posterior distribution of the age of the root of the whole tree. Tracer will plot a distribution for the selected parameter and also give you statistics about each such as the mean. The `95% HPD` stands for *highest posterior density* interval and is the equivalent of confidence intervals. In particular it is the shortest interval that contains 95% of the probability for the selected quantity.

How old is the root of the tree (give the mean and the HPD range)?

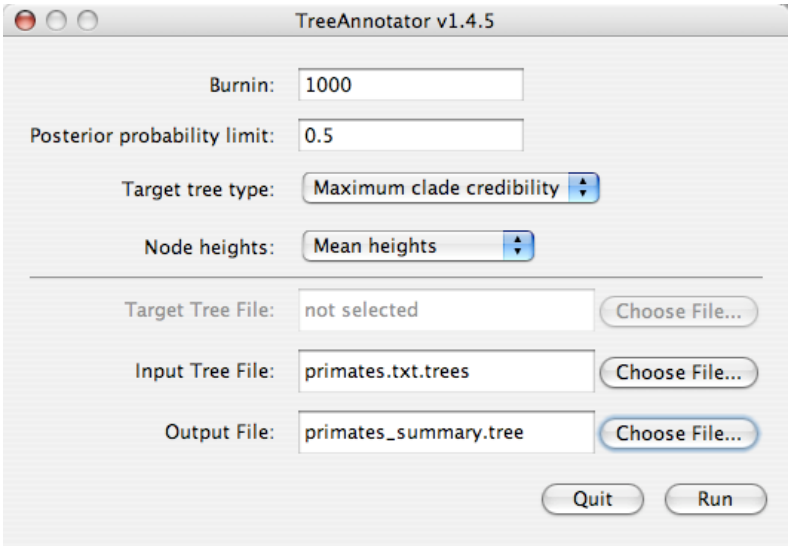
How fast does this gene fragment evolve in apes?

What sources of error does this estimate include?

Is the rate of evolution significantly different on different lineages?

Obtaining a tree

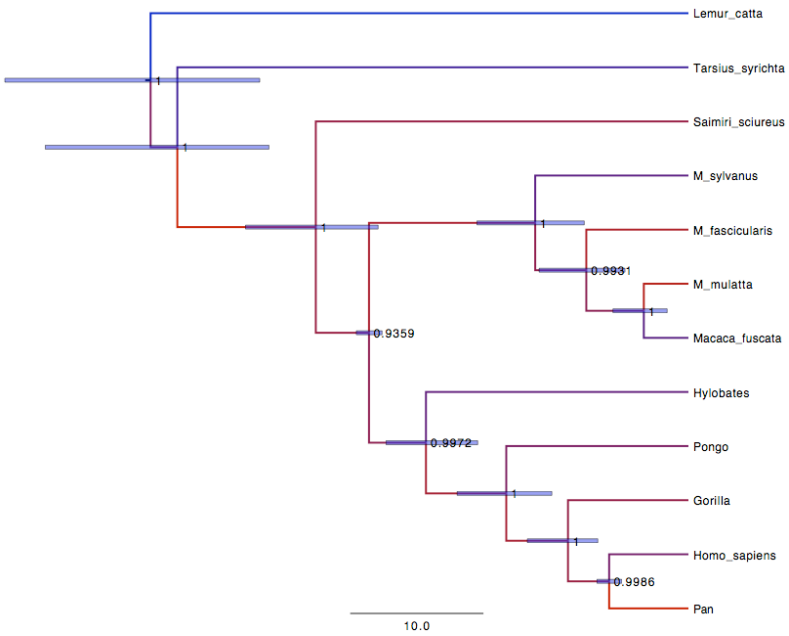
Just like BEAST produces a sample of parameter estimates that need to be summarized, it also produces a sample of plausible trees. These need to be summarized using the program **TreeAnnotator**. This will take the set of trees and find the best supported one. It will then annotate this summary tree 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:



For the input file, select the trees file that BEAST created (by default this will be called `primates.trees`) and select a file for the output (here I called it `primates_summary.tree`. The `burnin` will mean it ignores the first 1000 trees (out of a total of 10000). Choose `Mean heights` for node heights. Now press Run and wait for the program to finish.

Finally, we can look at the tree in another program called **FigTree**. Run this program, and open the `primates_summary.tree` file by using the `Open` command in the `File` menu. The tree should appear. You can now try selecting some of the options in the control panel on the left. 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. Under `Appearance` you can also tell FigTree to colour the branches by the rate.

You should end up with something like this:



Questions

What is the posterior probability of hominoid-cercopithecoid monophyly?

What is the marginal posterior estimate and HPD for the Human-Pongo split?

Advanced Exercises (optional)

Open the BEAST XML file in a text editor and find the <patterns> element in the XML file. It should look like this:

```
<patterns id="patterns" from="1">
     <alignment idref="alignment"/>
</patterns>
```

Add an attribute called “to” with value “200” like so:

```
<patterns id="patterns" from="1" to="200">
     <alignment idref="alignment"/>
</patterns>
```

Re-running the analysis will now only consider the first 200 sites. How do the posterior clade probabilities change? How do the divergence time estimates change?

Comparing your results to the prior

Using BEAUti set up the same analysis but under the **MCMC** options, select the **Sample from prior only** option:

This will allow you to visualize the full prior distribution in the absence of your data. Summarize the trees from the full prior distribution and compare the summary to the posterior summary tree.

What are the main ways in which the prior distribution on trees differs from the posterior distribution?

Are there any surprises?

Check out http://beast.bio.ed.ac.uk/ for more tutorials and an introduction to XML and the BEAST input file. The Manual also has some useful material.