Bayesian phylogenetic inference
MrBayes (Practice)

The aim of this tutorial is to give a very short introduction to *MrBayes*. There is a website with most information about *MrBayes*: www.mrbayes.net (which includes several installers and instructions on how to compile your own *MrBayes* executable) and a detailed manual in http://mrbayes.sourceforge.net/mb3.2_manual.pdf.

This document provides a step-by-step tutorial for analyzing a set of DENV-3 nucleotide sequences using MrBayes3.2. The data are 33 sequences from the *env* gene of sampled in Asia.

MrBayes reads sequence alignment files in NEXUS format, which can contain both sequence data and phylogenetic trees in block units. An example of a NEXUS file containing a data block is shown below:

```
#NEXUS
begin DATA;
    Dimensions ntax=4 nchar=25;
    Format datatype=NUCLEOTIDE gap=-;
    Matrix

ID_1
ATCAGACTGTTTGAGTGAGTGAAGT
ID_2
ATCAGGTTGTTTGAGTGAGGGGAGT
ID_3
ATCAGGCTGTTTGAGTGAGGGTACT
ID_4
ATCAGGCTGTTTGAGTGAGGGGACT
;
End;
```

Getting started

MrBayes is a command line software, which means you have to start and run it in a terminal. If you have a Windows, you can start *MrBayes* by double clicking on the executable and a terminal with *MrBayes* will start.

One of the most useful commands to begin in the help command. Just type:

MrBayes > **help**

to view the full list of available commands. You can get the help to a specific command by typing:

MrBayes > **help** *<command>*

All *MrBayes* commands are NOT case sensitive and are converted internally to all lower-case letters. So there is no need to worry to type Execute, execute, EXECUTE or exEcUtE. There is a complete list of commands in the *MrBayes* installation directory, called *commred_mb3.2.pdf*.

Brief Analysis

This initial analyses can be regarded as a first step before specifying any of the more advanced models. You may also use your own dataset here, but make sure it is small enough so that the MCMC runs can finish within this tutorial.

1. Reading the sequence file. Type in:

MrBayes> **execute** <*your_data_file.nex*>

to load your sequence file. Note that you either need to specify the full pathname or otherwise the file must exist in the directory from where *MrBayes* was executed.

2. Selecting model and parameters. In this step, the substitution model and parameters are specified. For this, you need to use the commands lset (l for likelihood) and prset (pr for prior). The lset command is used to specify the characteristics of your phylogenetic model. Nst (number of substitution types) describes the type of DNA substitution matrix (for the JC/F81, lset nst=1, for the K2P/HKY, lset nst=2; for the SYM/GTR, lset nst=6). The command rates describes the proportion of invariable sites (inv) and the heterogeneity of the substitution rates along variable sites (gamma). For example, to select the GTR substitution model with gamma-distributed rate variation across sites and a proportion of invariable sites, type

MrBayes> **lset nst=6 rates=invgamma**

To specify the a priori distributions for the evolutionary parameters of interest, *MrBayes* uses the command prset. The standard priors in *MrBayes* are non-informative and work well for the majority of the analyses. For example, to change the interval of the prior of the parameter I, and of the parameter G, type respectively:

MrBayes> **prset pinvar=uniform(0.00,1.00)**

MrBayes> **prset shape=uniform(0.00,200.00)**

We can have a look at an overview of the phylogenetic model by typing:

MrBayes> **showmodel**

3. Speeding up our analysis via parallelization (BEAGLE). The high-performance computational library BEAGLE comes pre-installed in *MrBayes* versions 3.2 and higher. To see current BEAGLE devices, type:

MrBayes> **showbeagle**

To see current BEAGLE options, type:

MrBayes> **help set**

To turn on the BEAGLE library using vector-parallelization on the CPU, type:

MrBayes**> set usebeagle = yes; set beagleprecision = single; set beaglesse = yes**

4. Running an analysis. To run an MCMC analyses, type

MrBayes> **mcmc ngen=1000000 samplefreq=1000 printfreq=500 diagnfreq=1000**

The command ngen describes the number of generations of the MCMC run, while the command samplefreq determines the frequency at which samples will be taken, which will determine the number of trees in the final sample, in this case 1000 (ngen/samplefreq).

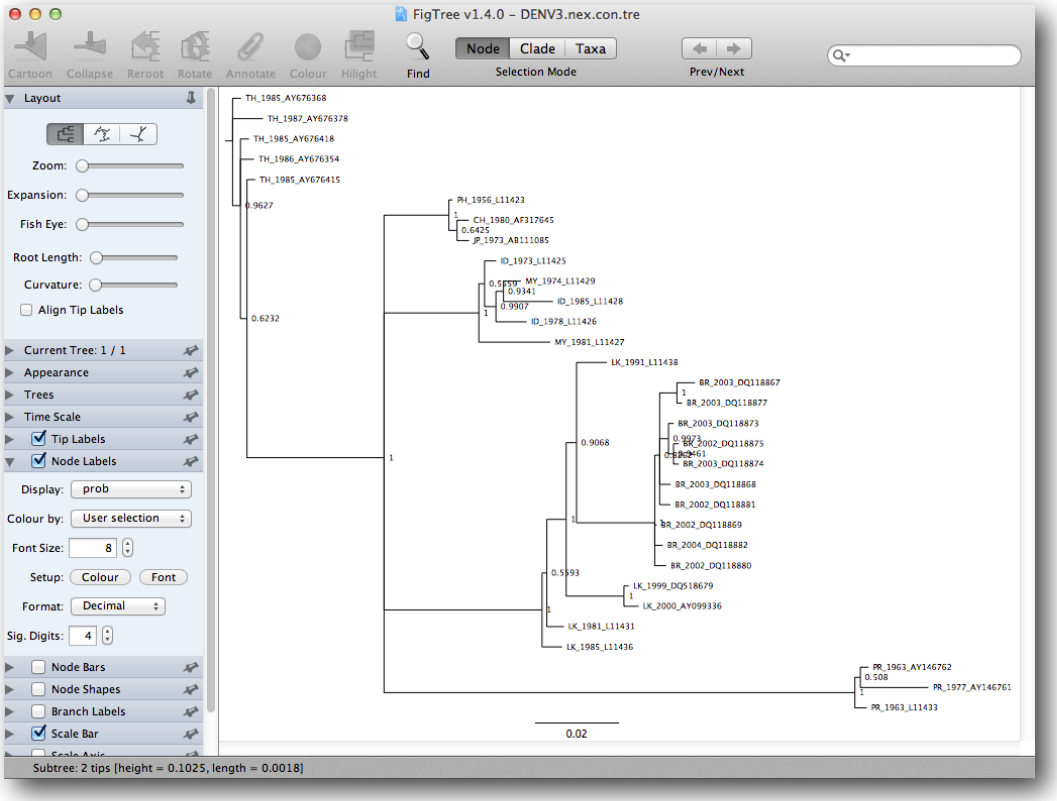
5. Summarizing the results. There are two important commands:

MrBayes> **sump burnin=100**

which summarizes the posterior probabilities for the parameters of your model, and will take 10% out as burn-in, and type:

MrBayes> **sumt burnin=100**

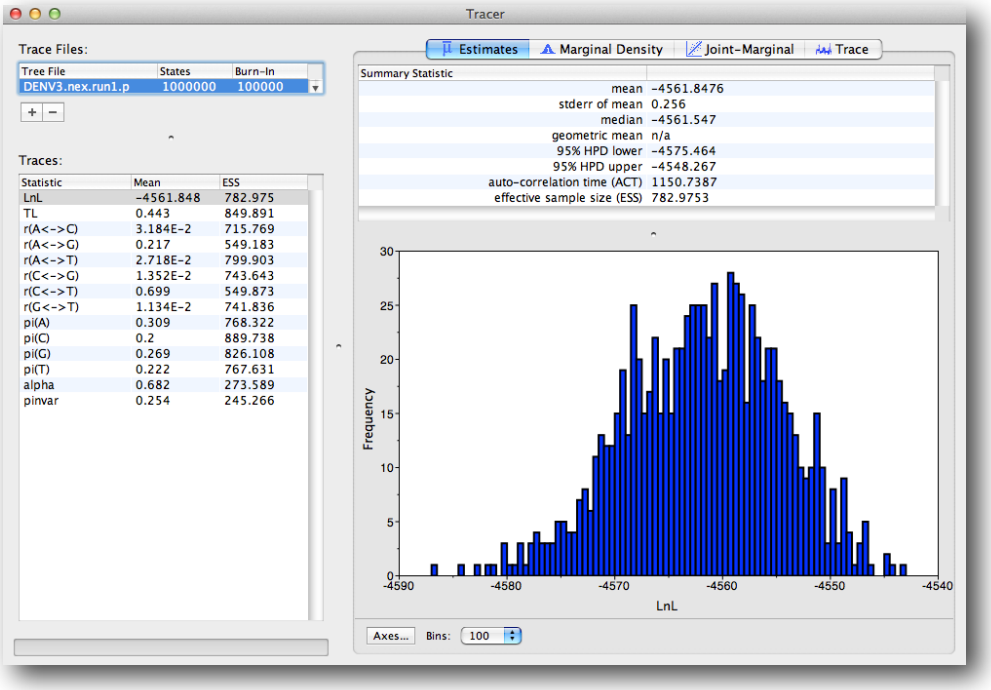
which summarizes the posterior distributions of trees that have been saved in DENV3.nex.run1.t and DENV3.nex.run2.t files in a consensus tree, removing also 10% of the tree files as burn-in. The output of this command included a bipartition table that shows the posterior probabilities for every split found in any tree sampled during the run. After the table is shown a majority-rule consensus tree containing all splits that had posterior probability 0.5 or above. While waiting for the MCMC run to be completed, you can visualize the majority-rule consensus tree in *FigTree* program. There, tick the Node Labels option and select Display prob to show the posterior probabilities above 0.5 in the consensus tree. It should look approximately like this:



Analyzing the MrBayes MCMC output

Before trying to draw any conclusions about the MCMC output, the samples should always be inspected for convergence to the stationary distribution. For this, we are going to use the program *Tracer*. The exact instructions for running *Tracer* differs depending on which computer you are using. Double click on the *Tracer* icon; once running, *Tracer* will look similar irrespective of which computer system it is running on.

Select the "Import Trace File..." option from the 'File' menu. By default, the file names will be DENV3.nex.run1.p and DENV3.nex.runr.p. The file will load and you will be presented with a window similar to the one below. Remember that MCMC is a stochastic algorithm so the actual numbers will not be exactly the same.



On the left hand side is the name of the log file loaded and the traces that it contains. There are traces for the posterior (this is the log of the product of the tree likelihood and the prior probabilities), and the continuous parameters. Selecting a trace on the left brings up analyses for this trace on the right hand side depending on tab that is selected. When first opened, the `posterior' trace is selected and various statistics of this trace are shown under the Estimates tab.

In the top right of the window is a table of calculated statistics for the selected trace. The statistics and their meaning are described below.

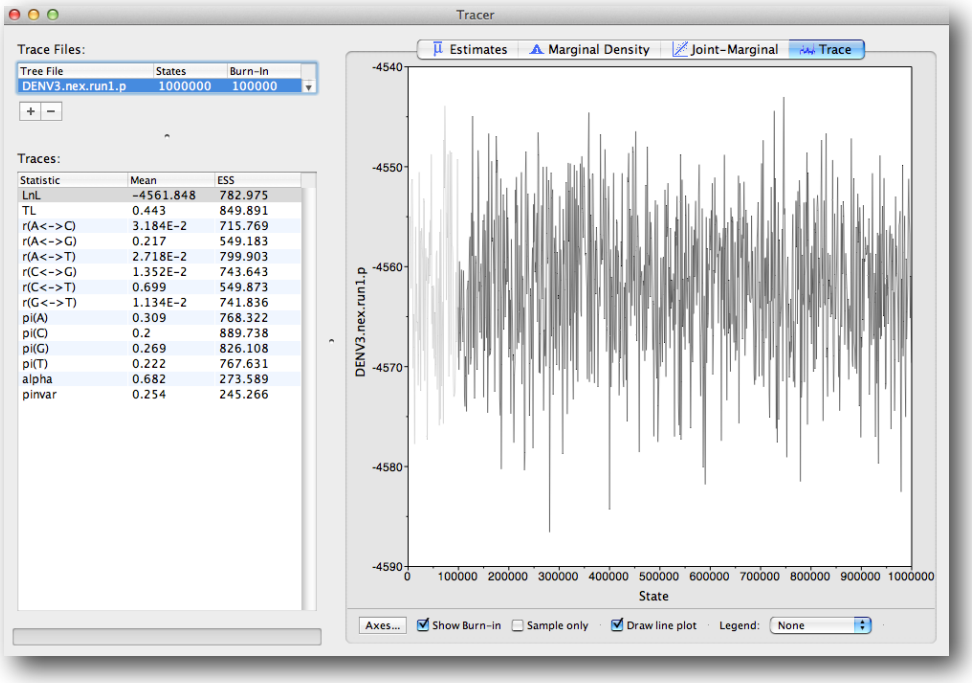
- **Mean** - The mean value of the samples (excluding the burn-in).
- **Stdev** - The standard error of the mean. This takes into account the effective sample size so a small ESS will give a large standard error.
- **Median** - The median value of the samples (excluding the burn-in).
- **95% HPD Lower** - The lower bound of the highest posterior density (HPD) interval. The HPD is the shortest interval that contains 95% of the sampled values.
- **95% HPD Upper** - The upper bound of the highest posterior density (HPD) interval.

- **Auto-Correlation Time (ACT)** - The average number of states in the MCMC chain that two samples have to be separated by for them to be uncorrelated (i.e. independent samples from the posterior). The ACT is estimated from the samples in the trace (excluding the burn-in).

- **Effective Sample Size (ESS)** - The effective sample size (ESS) is the number of independent samples that the trace is equivalent to. This is calculated as the chain length (excluding the burn-in) divided by the ACT.

The effective sample sizes (ESSs) for all the traces are quite OK here (ESSs bellow 100 would be highlighted in red by *Tracer*). A low ESS would mean that the trace contained a lot of correlated samples and thus may not represent the posterior distribution well. In the bottom right of the window is a frequency plot of the samples, which is sufficiently smooth in our case.

If we select the tab on the right-hand-side labelled `Trace' we can view the raw trace, that is, the sampled values against the step in the MCMC chain. There are 1000 samples in the trace (we ran the MCMC for 1,000,000 steps sampling every 1000). Here you can see how the samples are correlated.



Again we have chosen options that produce 1000 samples and with an ESS of about 300 there is still auto-correlation between the samples but 300 effectively independent samples will now provide a reasonable estimate of the posterior distribution. Fortunately, there are no obvious trends in the plot which would suggest that the MCMC has not yet converged, and there are no large-scale fluctuations in the trace which would suggest poor mixing. Now choose the density plot by selecting the tab labeled ‘Density’. Now, some questions:

1. What is the posterior mean estimate and the uncertainty interval for the alpha parameter? Is there rate heterogeneity among sites or are all sites evolving at nearly the same rate? Does the uncertainty interval for the proportion of invariant invariable sites excludes zero?

2. What is the posterior mean tree length (TL)? And what is the mean edge length? (Divide the tree length by the number of edges, which is 2n-3 if n is the number of taxa).

More resources and references

- Tutorials on the *MrBayes* wiki site (http://mrbayes.sourceforge.net/wiki/index.php/Manual)

- Fredrik Ronquist's *MrBayes* page (http://people.sc.fsu.edu/~fronquis/mrbayes/)

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Systematic Biology, 61, 539-542, 2012.

Ayres DL, Darling A, Zwickl DJ, Beerli P, Holder MT, Lewis PO, Huelsenbeck JP, Ronquist F, Swofford DL, Cummings MP, Rambaut A, Suchard MA. BEAGLE: an application programming interface and high- performance computing library for statistical phylogenetics. Systematic Biology, 61, 170-173, 2012.