Phylogenetic Analysis of Molecular Data (Botany 563)

Computer Lab 09: Bayesian Analysis

Based on a tutorial by Frederik Ronquist (http://www.csit.fsu.edu/~ronquist/mrbayes/LabMLBayes.doc).
For more information refer to the MrBayes manual, which you can find at http://mrbayes.csit.fsu.edu/manual.php

Learning objective:

Get acquainted with MrBayes and the parameters involved in Bayesian Analysis.

A few pointers on MrBayes:

- File format: Nexus
- Kind of data that can be analyzed: nucleotide, amino acid sequences, morphological ("standard") data, restriction site (binary) data, or any mix of these four data TYPEs.
- The input file must be located in the same folder as the MrBayes application
- The name of the input file should not have blank spaces.

Dataset: primates.nex This dataset contains 12 mitochondrial DNA sequences of primates

TASK 1. Introduction. A simple analysis

Getting Started

Double-click the application icon. You should see the information below:

```
                  MrBayes v3.1
            (Bayesian Analysis of Phylogeny)
                        by
        Fredrik Ronquist and John P. Huelsenbeck
            School of Computational Science
                Florida State University
                  ronquist@csit.fsu.edu
      Section of Ecology, Behavior and Evolution
            Division of Biological Sciences
          University of California, San Diego
                  johnh@biomail.ucsd.edu
    Distributed under the GNU General Public License
    Type "help" or "help <command>" for information
          on the commands that are available.
                  MrBayes >
```

The MrBayes > prompt at the bottom, which tells you that MrBayes is ready for your commands.
All of the commands are entered at the MrBayes > prompt.

Executing the dataset and saving your work

TYPE: **execute primates.nex**

You should see evidence that MrBayes has read the data in the DATA block of the Nexus file by outputting a list of taxa and no error message.

Keeping a log file
TYPE: **log start filename=chooseaname.log append**

Specifying a Model
At a minimum two commands, lset and prset, are required to specify the evolutionary model that will be used in the analysis.

TYPE: **lset nst=6 rates=invgamma**

Setting the Priors
There are six TYPEs of parameters in the model: the topology, the branch lengths, the four stationary frequencies of the nucleotides, the six different nucleotide substitution rates, the proportion of invariable sites, and the shape parameter of the gamma distribution of rate variation. The default priors in MrBayes work well for most analyses, and we will not change any of them here.

Checking the Model
TYPE: **showmodel** to check the model before we start the analysis

What model do these parameters correspond to?

Setting up the Analysis
The set-up is done with the mcmcp command. To run the analysis, we use the mcmc command.
TYPE: **mcmcp ngen=10000 temp=0.2 nchains=4 samplefreq=10 savebrlens=yes**
This sets the number of generations to 10,000, the heating parameter to 0.2, the number of chains to 4, and our sampling to 1 every 10 generations, saving branch lengths as well.

By default, MrBayes will run two simultaneous, completely independent analyses starting from different random trees. To change, the command is nruns.

Running the Analysis
TYPE: **mcmc**

MrBayes will first print information about the model and then list the proposal mechanisms that will be used in sampling from the posterior distribution. In our case, the proposals are the following:

```
The MCMC sampler will use the following moves:
   With prob.  Chain will change
      4.17 %   param. 1 (revmat) with Dirichlet proposal
      4.17 %   param. 2 (state frequencies) with Dirichlet proposal
      4.17 %   param. 3 (gamma shape) with multiplier
      4.17 %   param. 4 (prop. invariants) with beta proposal
     20.83 %   param. 5 (topology and branch lengths) with LOCAL
     62.50 %   param. 5 (topology and branch lengths) with extending TBR
```

MrBayes will spend most of its effort changing topology and branch lengths since these are the most

difficult parameters to integrate over.

After the initial log likelihoods, MrBayes will print to screen the state of the chains every 100th generation (regardless of the sample frequency, unless the printfreq paramenter is changed), like this:

```
   Chain results:
       1 -- (-7812.831) (-7523.685) [-7485.569] (-7700.309) * [-7832.045] (-7618.595) (-7776.608) (-7836.826)
     100 -- (-6771.532) (-6857.529) (-6766.678) [-6682.527] * [-6506.277] (-6944.449) (-6784.126) (-6991.307) -- 0:01:39
     200 -- (-6321.464) [-6179.561] (-6338.168) (-6272.242) * (-6339.400) (-6715.840) [-6265.329] (-6599.698) -- 0:01:38
     300 -- (-6201.285) (-6084.899) (-6139.200) [-6049.061] * [-6073.056] (-6359.134) (-6106.834) (-6515.348) -- 0:01:37

…

    1000 -- (-5811.949) [-5737.884] (-5888.234) (-5819.793) * (-5867.377) (-5851.693) (-5784.437) [-5749.264] -- 0:01:21
   Average standard deviation of split frequencies: 0.073946
    1100 -- (-5784.586) [-5730.069] (-5880.476) (-5798.728) * (-5839.268) (-5836.074) (-5779.940) [-5739.170] -- 0:01:20
...
...
    9900 -- (-5743.556) [-5724.246] (-5736.484) (-5731.803) * [-5727.082] (-5732.476) (-5728.482) (-5725.045) -- 0:00:00
   10000 -- (-5738.106) [-5723.380] (-5726.322) (-5727.599) * (-5728.776) (-5732.410) [-5728.515] (-5725.418) -- 0:00:00
Average standard deviation of split frequencies: 0.0001
Continue with analysis? (yes/no):
```

The first column lists the generation number. The following four columns with negative numbers each correspond to one chain in the first run. Each column corresponds to one physical location in computer memory, and the chains actually shift positions in the columns as the run proceeds. The numbers are the log likelihood values of the chains. The chain that is currently the cold chain has its value surrounded by square brackets, whereas the heated chains have their values surrounded by parentheses. When two chains successfully change states, they trade column positions (places in computer memory). If the Metropolis coupling works well, the cold chain should move around among the columns; this means that the cold chain successfully swaps states with the heated chains. If the cold chain gets stuck in one of the columns, then the heated chains are not successfully contributing states to the cold chain, and the Metropolis coupling is inefficient. The analysis may then have to be run longer or the temperature difference between chains may have to be lowered.

The star column separates the two different runs. The last column gives the time left to completion of the specified number of generations. Different moves are used in each generation, so the exact time varies somewhat for each set of 100 generations, and the predicted time to completion will be unstable in the beginning of the run. After a while, the predictions will become more accurate and the time will decrease predictably.

When to Stop the Analysis

Although it is advisable to use convergence diagnostic, such as the standard deviation of split frequencies, to determine run length, there are simpler but less powerful methods of determining when to stop the analysis. Arguably the simplest technique is to examine the log likelihood values (or, more exactly, the log probability of the data given the parameter values) of the cold chain, that is, the values printed to screen within square brackets. In the beginning of the run, the values typically increase rapidly (the absolute values decrease, since these are negative numbers). This phase of the run is referred to as the "burn-in" and the samples from this phase are typically discarded. Once the likelihood of the cold chain stops to increase and starts to randomly fluctuate within a more or less stable range, the run may have reached stationarity, that is, it is producing a good sample from the posterior probability distribution. At stationarity, we also expect different, independent runs to sample similar likelihood values. Trends in likelihood values can be deceiving though; you're more likely to detect problems with

convergence by comparing split frequencies than by looking at likelihood trends.

At the end of the run, MrBayes asks whether or not you want to continue with the analysis. Before answering that question, examine the average standard deviation of split frequencies. As the two runs converge onto the stationary distribution, we expect the average standard deviation of split frequencies to approach zero, reflecting the fact that the two tree samples become increasingly similar. In our case, the average standard deviation is about 0.07 after 1,000 generations and then drops drastically towards the end of the run. Values can differ slightly because of stochastic effects. Given the low value of the average standard deviation at the end of the run (and that this is a sample exersise), there appears to be no need to continue the analysis beyond 10,000 generations.

To stop the analysis, TYPE **no** when MrBayes asks "Continue with analysis? (yes/no):

When you stop the analysis, MrBayes will print several TYPEs of information useful in optimizing the analysis. This is primarily of interest if you have difficulties in obtaining convergence. Since we apparently have a good sample from the posterior distribution already (based on deviation of split frequencies), we will ignore this information for now. We will return to the subject of optimizing the MCMC analysis later.

Summarizing Samples of Substitution Model Parameters

During the run, samples of the substitution model parameters have been written to the .p files every samplefreq generation.

These files are tab-delimited text files that look something like this:

```
[ID: 5848203808]
  Gen  LnL        TL     r(A<->C)  ...  pi(G)     pi(T)     alpha     pinvar
    1  -7559.137  2.044  0.166667  ...  0.250000  0.250000  0.500000  0.000000
   10  -6585.519  2.181  0.090180  ...  0.169513  0.247418  0.569271  0.036162
   ...
 9990  -5728.935  2.527  0.051106  ...  0.076213  0.262107  0.932771  0.194004
10000  -5722.857  2.642  0.051106  ...  0.078383  0.246791  0.721217  0.185486
```

The first number, in square brackets, is a randomly generated ID number that lets you identify the analysis from which the samples come. The next line contains the column headers, and is followed by the sampled values. From left to right, the columns contain: (1) the generation number (Gen); (2) the log likelihood of the cold chain (LnL); (3) the total tree length (the sum of all branch lengths, TL); (4) the six GTR rate parameters (r(A<->C), r(A<->G) etc); (5) the four stationary nucleotide frequencies (pi(A), pi(C) etc); (6) the shape parameter of the gamma distribution of rate variation (alpha); and (7) the proportion of invariable sites (pinvar). If you use a different model for your data set, the .p files will of course be different.

Q: How many samples (lines) should there be in the .p file for this analysis? _____
Explain

TYPE **sump**

The sump command will first generate a plot of the generation versus the log probability of the data (the log likelihood values). If we are at stationarity, this plot should look like 'white noise', that is, there should be no tendency of increase or decrease over time.

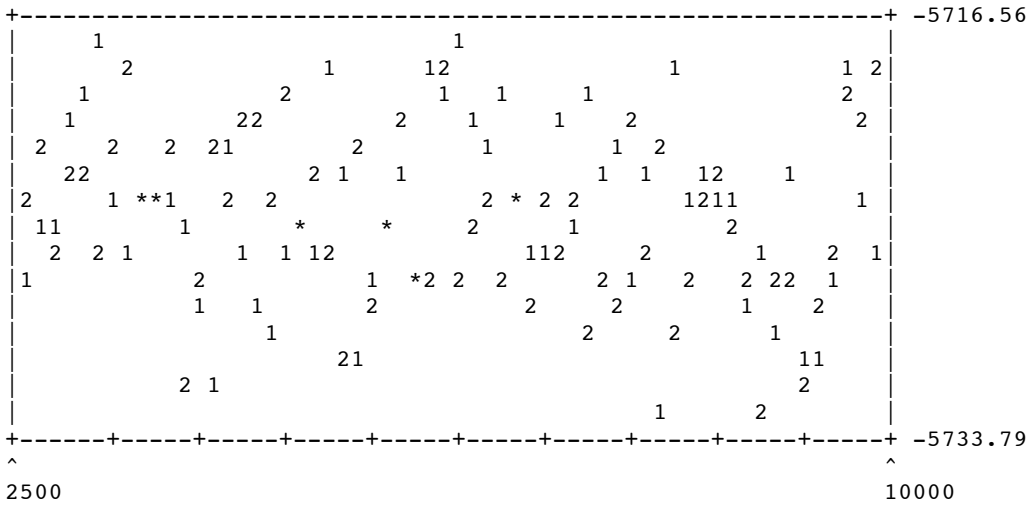
Sketch the plot below.

The burnin is a fixed number of samples from the beginning of the chain to be discarded. Since we sampled every 10th generation, there are 1,000 samples (1,001 to be exact, since the first generation is always sampled).

Based on your plot, how many generations did it take to reach stationarity? _____
So, what would be an appropriate "burnin" (in number of samples)? _____

TYPE **sump burnin=xxx** (here substitute for the value you considered appropriate)

The plot should look something like this (otherwise try higher burnin):



If you see an obvious trend in your plot, either increasing or decreasing, you probably need to run the analysis longer to get an adequate sample from the posterior probability distribution.

At the bottom of the sump output, there is a table summarizing the samples of the parameter values:

```
Model parameter summaries over the runs sampled in files
   "primates.nex.run1.p" and "primates.nex.run2.p":
   (Summaries are based on a total of 1502 samples from 2 runs)
   (Each run produced 1001 samples of which 751 samples were included)
                                          95% Cred. Interval
                                       ----------------------
Parameter       Mean        Variance       Lower         Upper         Median       PSRF *
-------------------------------------------------------------------------------------------
TL            2.885314      0.071446      2.408000      3.464000      2.861000      1.004
r(A<->C)      0.045110      0.000070      0.031406      0.061424      0.044756      0.999
r(A<->G)      0.477554      0.002421      0.387141      0.569168      0.473521      1.040
r(A<->T)      0.038541      0.000060      0.024204      0.053787      0.038402      0.999
r(C<->G)      0.033765      0.000195      0.010836      0.064049      0.032475      1.033
r(C<->T)      0.386144      0.001735      0.311504      0.459011      0.386897      1.012
r(G<->T)      0.018885      0.000133      0.001217      0.045924      0.016801      1.048
pi(A)         0.355301      0.000162      0.329769      0.377153      0.356808      1.092
pi(C)         0.320831      0.000131      0.299953      0.341757      0.321312      1.152
pi(G)         0.081154      0.000048      0.068816      0.098442      0.081723      1.014
pi(T)         0.242714      0.000104      0.226573      0.264684      0.242113      1.000
alpha         0.714405      0.037305      0.419210      1.147286      0.673545      1.007
pinvar        0.185901      0.004887      0.025889      0.307468      0.188027      1.009
-------------------------------------------------------------------------------------------
   * Convergence diagnostic (PSRF = Potential scale reduction factor [Gelman and Rubin, 1992],
uncorrected) should approach 1 as runs converge. The values may be unreliable if you have a small
number of samples. PSRF should only be used as a rough guide to convergence since all the
assumptions that allow one to interpret it as a scale reduction factor are not met in the
phylogenetic context.
```

For each parameter, the table lists the mean and variance of the sampled values, the lower and upper boundaries of the 95 % credibility interval, and the median of the sampled values. The parameters are the same as those listed in the .p files: the total tree length (TL), the six reversible substitution rates (r(A<->C), r(A<->G), etc), the four stationary state frequencies (pi(A), pi(C), etc), the shape of the gamma distribution of rate variation across sites (alpha), and the proportion of invariable sites (pinvar). Note that the six rate parameters of the GTR model are given as proportions of the rate sum (the Dirichlet parameterization). This parameterization has some advantages in the Bayesian context; in particular, it allows convenient formulation of priors. If you want to scale the rates relative to the G-T rate, just divide all rate proportions by the G-T rate proportion.

The last column in the table contains a convergence diagnostic, the Potential Scale Reduction Factor (PSRF). If we have a good sample from the posterior probability distribution, these values should be close to 1.0. If you have a small number of samples, there may be some spread in these values, indicating that you may need to sample the analysis more often or run it longer. In our case, we can probably easily obtain more accurate estimates of some parameters by running the analysis slightly longer. 2.10. Summarizing Samples of Trees and Branch Lengths Trees and branch lengths are printed to the .t files. These files are Nexus-formatted tree files.

Summarizing Tree and branch length information

TYPE **sumt burnin=xxx** (same value as before)

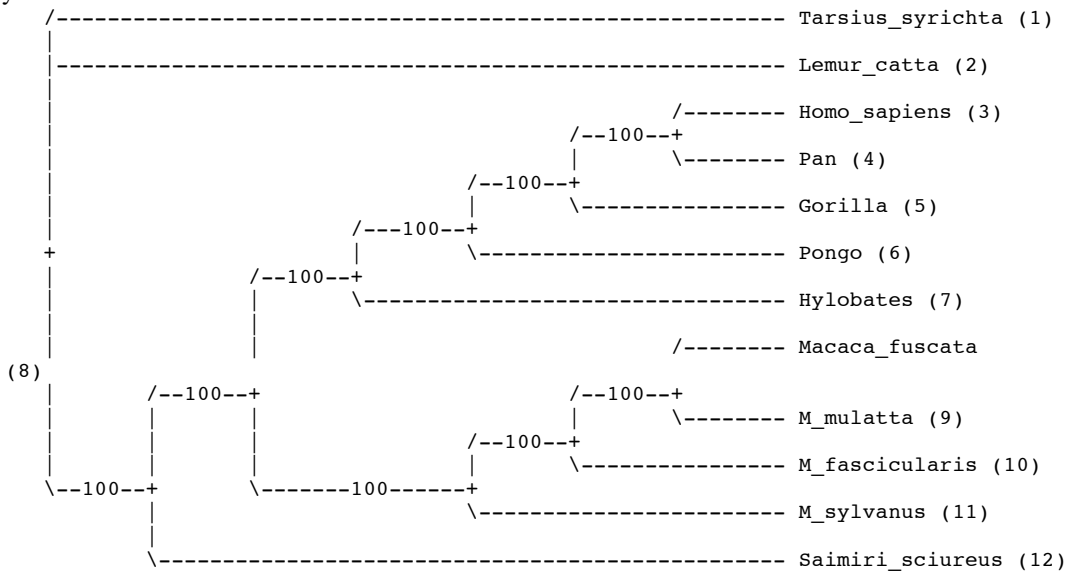
The sumt command will output, among other things, summary statistics for the taxon bipartitions, a tree with clade credibility (posterior probability) values, and a phylogram (if branch lengths have been saved). The summary statistics (see below) describes each partition in the “dot-star” format (dots for the taxa that are on one side of the partition and stars for the taxa on the other side; for instance, the first

partition (ID 1) is the terminal branch leading to taxon 8 since it has a star in the 8th position and a dot in all other positions). Then it gives the number of times the partition was sampled (#obs), the probability of the partition (Probab.), the standard deviation of the partition frequency (Stdev(s)), the mean (Mean(v)) and variance (Var(v)) of the branch length, the Potential Scale Reduction Factor (PSRF), and finally the number of runs in which the partition was sampled (Nruns). In this analysis, there is overwhelming support for a single tree, so all partitions have a posterior probability of 1.0.

```
Summary statistics for taxon bipartitions:
ID -- Partition     #obs    Probab.  Stdev(s)   Mean(v)   Var(v)    PSRF  Nruns
-------------------------------------------------------------------------------
 1 -- ....*.......  1502  1.000000  0.000000  0.058828  0.000235  1.002    2
 2 -- ..***.......  1502  1.000000  0.000000  0.082206  0.000426  1.000    2
 3 -- ..**........  1502  1.000000  0.000000  0.030594  0.000152  0.999    2
 4 -- ..*.........  1502  1.000000  0.000000  0.049234  0.000104  1.002    2
 5 -- ...*........  1502  1.000000  0.000000  0.063670  0.000207  1.001    2
 6 -- ..**********  1502  1.000000  0.000000  0.271433  0.003850  0.999    2
 7 -- .**********  1502  1.000000  0.000000  0.472390  0.007137  1.010    2
 8 -- ..*********.  1502  1.000000  0.000000  0.113412  0.002208  0.999    2
 9 -- ......*.....  1502  1.000000  0.000000  0.165887  0.000861  1.017    2
10 -- ..*****.....  1502  1.000000  0.000000  0.122957  0.001186  1.009    2
11 -- ........*...  1502  1.000000  0.000000  0.022199  0.000039  1.003    2
12 -- .......*....  1502  1.000000  0.000000  0.018330  0.000042  1.002    2
13 -- ..........*.  1502  1.000000  0.000000  0.070973  0.000449  1.043    2
14 -- .*.........  1502  1.000000  0.000000  0.329111  0.003588  1.008    2
15 -- ..........*  1502  1.000000  0.000000  0.429012  0.004224  1.003    2
16 -- .........*..  1502  1.000000  0.000000  0.056534  0.000137  1.009    2
17 -- .......****.  1502  1.000000  0.000000  0.248962  0.002162  1.005    2
18 -- .....*......  1502  1.000000  0.000000  0.139867  0.000708  1.001    2
19 -- .......**...  1502  1.000000  0.000000  0.035640  0.000107  1.020    2
20 -- ..****......  1501  0.999334  0.000942  0.054652  0.000435  1.022    2
21 -- .......***..  1501  0.999334  0.000942  0.049466  0.000355  1.010    2
-------------------------------------------------------------------------------
```

The clade credibility tree gives the probability of each partition or clade in the tree, and the phylogram gives the branch lengths measured in expected substitutions per site.

Clade credibility values:



```
Phylogram:
              /-------------------------------------- Tarsius_syrichta (1)
              |
              |--------------------------- Lemur_catta (2)
              |
              |                                                    /---- Homo_sapiens (3)
              |                                                 /--+
              |                                                 |  \----- Pan (4)
              |                                              /--+
              |                                              |  \----- Gorilla (5)
              |                                           /--+
              +                                           |  \------------ Pongo (6)
              |                                        /--+
              |                                        |  \--------------- Hylobates (7)
              |                                        |
              |                                        |                        /-- Macaca_fuscata (8)
              |                          /-------------+                     /--+
              |                          |             |                     |  \-- M_mulatta (9)
              |                          |             |                  /--+
              |                          |             |                  |  \----- M_fascicularis (10)
              |                          \-------------+------------------+
              |                                        |                  \----- M_sylvanus (11)
              |                                        |
              |                                        \---------------------------------- Saimiri_sciureus (12)
   |--------| 0.100 expected changes per site

Credible sets of trees (6 trees sampled):
      99 % credible set contains 1 trees
```

In the background, the sumt command creates three additional files. The first is a .parts file, which contains the list of taxon bipartitions, their posterior probability (the proportion of sampled trees containing them), and the branch lengths associated with them (if branch lengths have been saved). The branch length values are based only on those trees containing the relevant bipartition. The second generated file has the suffix .con and includes two consensus trees. The first one has both the posterior probability of clades (as interior node labels) and the branch lengths (if they have been saved) in its description. A graphical representation of this tree can be generated in Rod Page's program TreeView. The second tree only contains the branch lengths and it can be imported into a wide range of tree-drawing programs such as MacClade and Mesquite. The third file generated by the sumt command is the .trprobs file, which contains the trees that were found during the MCMC search, sorted by posterior probability.

To quit MrBayes, simply TYPE **quit**.

TASK 2. Exploring the effect of heat on mixing.

Here we will look at the output that is printed at the end of the run (use .log file).
One way to verify that there is good mixing is to look at is the table of acceptance rates of the proposal mechanisms used in your analysis. The Metropolis proposals used by MrBayes work best when their acceptance rate is neither too low nor too high. A rough guide is to try to get them within the range of 10% to 70%; rates outside this range are not necessarily a big problem but they typically mean that the analysis is inefficient.

Look at the Chain swap information table and report the rate values (upper diagonal) here:

1 —> 2 : _____

2 —> 3 : _____

3 —> 4 : _____

The default heating value is temp=0.2. To examine the effect of the heating parameters on the mixing of the chains, repeat the analysis on the primates.nex dataset, but this time use a much lower and much higher temperature (values around 0.01 and 0.8 are fine). You can do this by modifying the temp parameter in the mcmcp command, something like:

```
mcmcp ngen=10000 temp=0.8 nruns=2 nchains=4 samplefreq=10;
```

The file bayesblock.rtf contains a MrBayes block that you can copy and paste into the dataset (after the data semicolon, before the "end;") to run it in batch mode.

Report:

A) temp you used: _____

Did the swap rates of among cold and heated chains go up or down? _____

How many trees were sampled?

How many trees are in the credible set of trees?

B) temp you used: _____

Did the swap rates of among cold and heated chains go up or down? _____

How many trees were sampled?

How many trees are in the credible set of trees?

Based on your results, **answer** the following:

1) Do the heating parameter value affect your search? How?

2) If acceptance rates for the swaps between adjacent chains (the values close to the diagonal in the swap statistics matrix) are low, then would you increase or decrease the temperature?

For more information on how to check and to proceed when it is difficult to get convergence, refer to the MrBayes manual.

TASK 3. (Optional)

As time allows, do the following:

1. Calculate a 95% credible interval for the shape parameter of the gamma distribution (alpha). Does it change with different heating values?

2. For cases where more than a single tree is sampled, what is the posterior probability of the different topologies (hint: .trprobs file)?