Lab 06:
Bayesian Phylogenetics,
MCMC Convergence Diagnostics,
and Divergence Time Estimation

By Will Freyman, edited by Carrie Tribble,
re-edited by Ixchel González-Ramírez

1 Before you begin

Please download:

1. primates.nex:
   `ib.berkeley.edu/courses/ib200/labs/06/primates.nex`
2. MrBayes:
   `http://mrbayes.sourceforge.net/`
3. Tracer:
   `https://beast.community/tracer`

   Open Tracer after you install it to make sure the install was successful. You may need to install legacy Java to run Tracer:
   `https://support.apple.com/kb/DL1572?locale=en_US`
4. BEAST:
   `http://beast2.org/`

2 Introduction

Bayesian phylogenetic analyses rely on Markov chain Monte Carlo (MCMC) algorithms to approximate the posterior distribution. We say an MCMC analysis has reached *convergence* when it is sampling the parameter values in a proportion that approximates the posterior probability. That is to say, if the tree τ has a 20% posterior probability, then a converged MCMC analysis will produce tree τ approximately 20% of the time. This is true not only for trees, but for all the parameters. The algorithm can produce the posterior distributions of parameters without holding any of the other parameters constant, integrating over all values of those parameters. You can, of course, set parameters to be constant if you want – it is a highly flexible inference framework.

Today we'll run some MCMC analyses and learn how to assess whether or not they have converged.

3 Bayesian inference using MrBayes

With well over 30 thousand citations **MrBayes** [Ronquist et al., 2012] is possibly the most widely used Bayesian phylogenetic software. Though its interface is highly similar to PAUP*, MrBayes has been constantly updated to include many recent advances in phylogenetic modeling.

3.1 MrBayes

1. In terminal, navigate to the folder where you have saved the primate file.

2. Start MrBayes with the command `mb` Just like in PAUP* type:

   ```
   execute primates.nex
   ```

3. Type the command `showmodel`. This shows you the default model MrBayes is set to. Note `nst=1` means there is only 1 substitution rate and `# States = 4` means that there are 4 stationary state frequencies. This corresponds to the F81 model, that introduces the Π parameter, allowing each base to have a different frequency.

4. Instead of the F81 model, let's use the GTR+Γ+I model. Remember this is the generalised time-reversible model with gamma-distributed rate variation across sites and a proportion of invariable sites.

   ```
   lset nst=6 rates=invgamma
   ```

   We have now set up a model in which the tree topology, branch length, the stationary state frequencies, and substitution rates are all jointly estimated from the data. If you wanted to set some of those parameters to be fixed (for example using parameter values from jModelTest) you could type

   ```
   prset statefreqpr=fixed(0.21,0.29,0.13, 0.37)
   ```

   but for now don't fix the stationary state frequencies – let's leave the default flat Dirichlet prior in place.

5. Now let's run an MCMC analysis:

   ```
   mcmc ngen=20000 samplefreq=100 printfreq=100 diagnfreq=1000
   ```

   This means we'll run the MCMC for 20000 generations (or MCMC iterations), and sample parameter values every 100 generations. At the end of the run we'll have 20000/100 = 200 samples from the posterior. This should run relatively quickly.

   By default, we are running 2 parallel analyses, each consisting of Metropolic-coupled MCMC with 4 chains (1 cold and 3 heated).

6. After 20000 generations MrBayes will stop and ask us whether or not to continue. Type `no`. To assess convergence, MrBayes calculates the *average standard deviation of split frequencies* (ASDSF). ASDSF is calculated by comparing clade (or "split") frequencies across independent MCMC runs. As the independent MCMC runs converge, they should each contain the same frequency of each clade. The standard deviation of those frequences across all runs will approach 0.

7. A lot of information about the MCMC analysis will be printed out.

Question 1:
What was your final average standard deviation of split frequencies?

The MCMC *moves* and their acceptance rates are shown. These are the moves (or new proposed states) the algorithm uses to explore parameter space. Look at `https://en.wikipedia.org/wiki/Tree_rearrangement`. All three basic tree rearrangements were used by MrBayes

Which of the moves were tree rearrangements? What were the acceptance rates for these moves?

8. Type the command `sumt` to summarize the trees sampled from the posterior distribution. MrBayes will output a lot of information, and generate a file `primates.nex.con.tre` which you can open in FigTree.

9. We have only seen a tiny bit of what MrBayes can do. For example, if we wanted to use reversible-jump MCMC to sample across the space of all possible substitution models and avoid a priori model testing (like jModelTest) we could have set

   ```
   lset nst=mixed rates=gamma
   ```

   If you have time, give this a try. After the MCMC analysis has run use the command `sump` to get a list of the posterior probabilities of different substitution models.

4 Diagnosing MCMC convergence with Tracer

Tracer is programmed by the same group that developed FigTree and BEAST. It is a handy program to quickly diagnose convergence of MCMC analyses, and get estimates of the posterior distribution of parameter values. However there are many other convergence diagnostics that are not implemented in Tracer, if you are interested in these check out the R package **coda**.

1. Open Tracer. Under the `File` menu select `Import Trace File`.

2. Navigate to the directory in which you ran your MrBayes analysis. For `Format` select `All Files`. Select the file `primates.nex.run1.p`.

3. Click on the `Trace` tab above the graph. This graph (see Figure 1) shows the MCMC samples on the x-axis and the log-likelihood (LnL) values of the trace on the y-axis.

4. View the trace plots of all the parameters by selecting the parameter names in the `Traces:` panel (lower left hand side of window). For example, transition rate for A-C are called `r(A<->C)`.

5. If the MCMC has converged, the trace plots should look like nice “furry caterpillars” such as Figure 2. Your MCMC analysis was much too short to produce furry caterpillars.

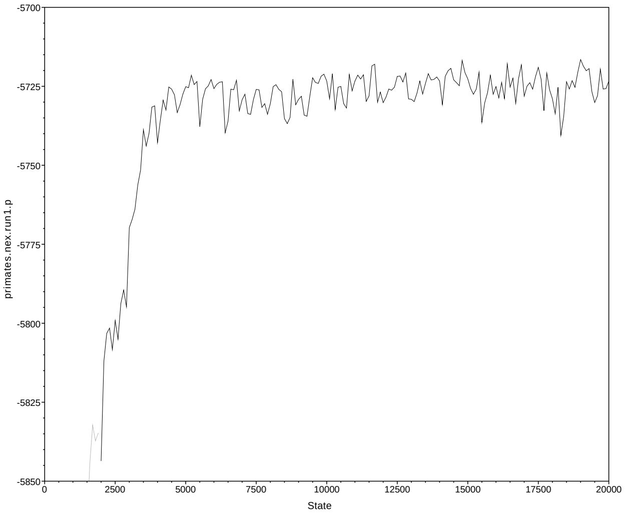

Figure 1: Example trace plot. The plot suggests the burnin should probably be about 5000.

6. Another way to convergence diagnostic is the effective sample size (ESS) value of the trace. These are listed in the `Traces:` panel. ESS values over 200 are considered adequate.

7. By default, Tracer discards the first 10% of samples as *burn-in*. This is shown in the `Trace Files:` panel in the upper left side of the window. The burn-in of an MCMC run is the initial part of a run where the MCMC may have been stuck in a low probability region of parameter space. For the trace in Figure 1, the burn-in should probably be adjusted to at least 5000.

Question 2:
By double-clicking on the *Burn-In* value in the `Trace Files:` panel modify the burn-in to a reasonable amount. How does this affect the ESS values?

By repeating step 1 and 2 above, load the file `primates.nex.run2.p` from the second MrBayes run (remember MrBayes ran two parallel runs). Set an appropriate burn-in for the second run, and then click on `Combined` to view the samples from both MCMC runs combined. How does this affect ESS values?

You now have both ESS and ASDSF values (ASDSF values from Question 1). Do they agree on whether or not convergence has been reached?

Send me a screen shot of your entire Tracer window with combined runs showing.

5 Bayesian Divergence Time Estimation using BEAST

BEAST [Bouckaert et al., 2014] is probably the second most widely used Bayesian phylogenetic software, and it became popular primarily because of the relaxed clock models that were first implemented in BEAST. Though all these models are now implemented in both

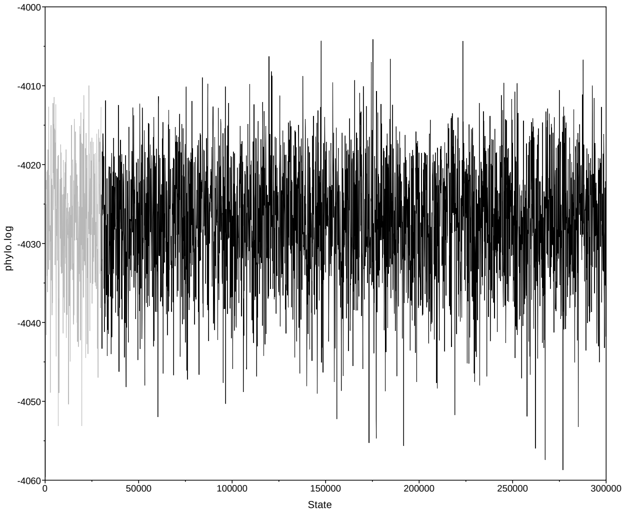

Figure 2: Example "furry caterpillar" trace plot for a nicely converged MCMC analysis. This trace has an effective sample size (ESS) value of 2653.

MrBayes and BEAST, we are going to do this exercise in BEAST since it has a very different user interface.

5.1 Non-Bayesian Divergence Time Estimation

There are a pair of non-Bayesian methods for estimating divergence times implemented in the program **r8s** [Sanderson, 2003]. They are (1) a nonparametric rate smoothing approach, and (2) a semiparametric penalized likelihood method that combines likelihood and the nonparametric rate smoothing penalty function. These have been widely used in the past, but their use is (arguably) decreasing with the advent of Bayesian approaches to fossil dating.

5.2 Tip-Dating Vs. Node-Dating

In this lab we'll be covering the commonly used approach called *node-dating*, in which a fossil is assigned to an internal node of the phylogeny, thus constraining the minimum age of the internal node. A probability density is used as a prior for the calibrated node (see Figure 3). This is problematic because the fossil may be misplaced in the phylogeny, and the choice of a prior is usually arbitrary.

The *tip-dating* (or total evidence) approaches use morphological data from the fossil and extant taxa to place the fossil just as if it was another tip. The difficulty with this approach is that morphological data is needed as well as molecular data. Furthermore, models of morphological character evolution are under developed.

Another approach, that can be used in conjunction with either tip-dating or node-dating is the fossilized birth-death (FBD) process [Heath et al., 2014]. This models speciation, extinction, and the fossil recovery rate together. It can integrate over all possible placements for the fossil taxon (including being a direct descendant) and so does not require specifying an arbitrary prior for the node age.

Unfortunately we only have time to cover node-dating, but you should look into these other approaches for your projects.

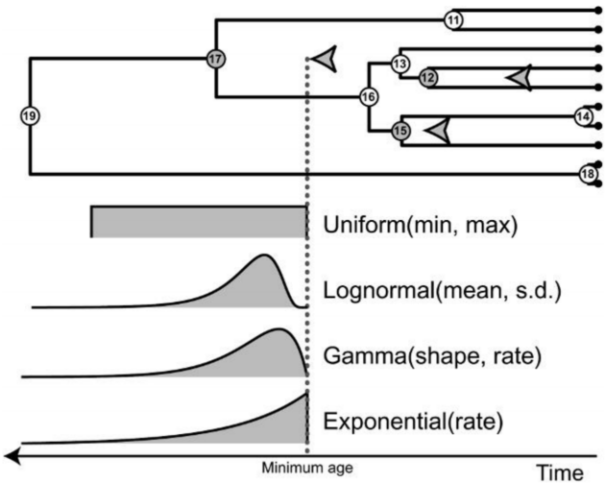

Figure 3: Four probability densities commonly used as priors on node ages, shown here calibrating node 17. The grey dotted line represents the minimum age constraint placed on the age of node 17 by a fossil descendant (grey arrow). Image from Heath [2012].

5.3 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.

1. First open up the BEAUti program, which will be located within your BEAST folder. Now import your NEXUS file. `File -- Import Alignment`. Select the `primates.nex` file.

2. We'll now set up a node age prior on the human-chimpanzee common ancestor. Select the `Priors` tab. Click on the `+` button to add a new prior.

3. In the `Taxon set label:` box label this prior `human-chimp`.

4. Select *Home_sapiens* and *Pan* and add them to the group by clicking on the `>>` box. Click `OK`.

5. Select the Normal distribution. Click the little arrow to the left of the *human-chimp.prior* button.

6. Set the mean to 6 and sigma to 0.5. 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.

7. Set up a second calibration point for the group `HomiCerco` which includes everything except *Lemur*, *Saimiri*, and *Tarsius*. Give this a normal distribution with 24+/- 0.5 million years.

8. Let's also set up a topological constraint to make sure our tree is rooted correctly. Add another prior called `ingroup`, selecting all taxa except the *Lemur*. Constrain this to be monophyletic by clicking the `monophyletic` checkbox.

9. Note the default tree prior is a *Yule Model*. This is a stochastic process that models only speciation. Let's change it to a slightly more realistic *Birth Death Model* that includes extinction.

10. Click on the `Site Model` tab. By default the substitution model is set to Jukes Cantor. Change it to GTR.

11. Click on the `MCMC` tab. Change the `Chain Length` value to 800,000.

12. Under the `File` menu, click `Save As` and save the XML file.

13. Start BEAST and select the XML file you just generated. Click `Run`.

Question 3:
It should run pretty fast. Once it is complete, open the `.log` file in Tracer. How do the ESS values look? Should the MCMC be run longer?

What is the estimated mean age of the ingroup's most recent common ancestor?

14. Now let's summarize the posterior distribution of trees and get the 95% highest posterior density (HPD) ranges for node ages. Open the program **TreeAnnotator** that came with BEAST.

15. Enter an appropriate `Burnin percentage`.

16. 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.

17. For the `Input Tree File` select the `.trees` file that BEAST generated.

18. Enter a name for your output summary tree file and click `Run`.

Question 4:
Open the summary tree in FigTree. Figure out how to add the 95% HPD node age ranges as node bars to the tree. Also add the posterior probabilities on the nodes of the tree. Send me a screen shot.

Please email me the following:

1. The answers to questions 1-4.
2. Screenshots for questions 2 and 4.

References

Remco Bouckaert, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A Suchard, Andrew Rambaut, and Alexei J Drummond. Beast 2: a software platform for bayesian evolutionary analysis. *PLoS Comput Biol*, 10(4):e1003537, 2014.

Tracy A Heath. A hierarchical bayesian model for calibrating estimates of species divergence times. *Systematic biology*, page sys032, 2012.

Tracy A Heath, John P Huelsenbeck, and Tanja Stadler. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Sciences*, 111(29):E2957–E2966, 2014.

Fredrik Ronquist, Maxim Teslenko, Paul van der Mark, Daniel L Ayres, Aaron Darling, Sebastian Höhna, Bret Larget, Liang Liu, Marc A Suchard, and John P Huelsenbeck. Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61(3):539–542, 2012.

Michael J Sanderson. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, 19(2):301–302, 2003.