PS8: Phylogenetics and Forensic Microbiology Practical (Fall 2015)

Download instructions for BEAST:

1. BEAST website: http://beast2.org
2. Install the appropriate version (Mac/Linux/Windows) of BEAST v2.3.1.
3. Install the appropriate version of Tracer v1.6: http://tree.bio.ed.ac.uk/software/tracer/
4. Install the appropriate version of BEAGLE (recommended, but not essential): http://beast.bio.ed.ac.uk/beagle
5. Install the appropriate version of FigTree v1.4.2: http://tree.bio.ed.ac.uk/software/figtree
6. Create a suitably named work directory for your analyses, e.g., Practical.

BEAST package contents:

1. BEAUti - used to create .xml files for phylogenetic analyses; requires a nexus input file with the sequence data.
2. BEAST - performs Bayesian phylogenetic analyses.
3. LogCombiner - used to combine the output file from multiple independent BEAST analyses of the same data set.
4. TreeAnnotator - used to identify the maximum clade credibility (MCC) tree once a BEAST analysis has been completed.

Data sets: HIV-1 RT sequences from Metzker et al. (2002): fasta file, nexus file

Program instructions:

1. Use BEAUti to create an .xml file for a phylogenetic analysis of the HIV-1 RT data collected in the Louisiana vs. Richard J. Schmidt case.
   (a) Download the input file Metzker_pol.nex to your computer and then use the "File → Import Alignment" option in the *Partitions* window to upload it to BEAUti. There should be 42 sequences containing 689 nucleotides.
   (b) Go to the *Site Model* window and select the HKY substitution model; here we will estimate 'kappa' (check the estimate button and set the initial value equal to 2.0) and we will use the empirical frequencies of the bases to estimate the stationary distribution. In addition, we will use a Gamma distribution with 5 categories ('Gamma Category Count') and an estimated shape parameter (initial value = 1.0) to allow for rate variation across sites. We will also allow for a class of invariant sites with an estimated proportion (initial value = 0.5).

(c) Go to the *Clock Model* window and check that the Model is set to 'Strict clock' with Clock.rate equal to 0.005. This is very approximately the substitution rate for the RT gene in units of substitutions per site per year. In this case, leave the estimate button unchecked.

(d) Go to the *Priors* window and select the following priors:

- Tree.t: Coalescent Constant Population;
- gammaShape.s: Uniform with Lower = 0.0, Upper = 10.0, Initial value = 1.0;
- kappa.s: Uniform with Lower = 0.0, Upper = 40.0, Initial value = 2.0;
- popSize.t: $1/X$ (use default values for range and initial value);
- proportionInvariant.s: Uniform with Lower = 0.0, Upper = 1.0, Initial value = 0.5.

(e) Go to the *MCMC* window and set the length of the chain equal to two million steps with Store Every = 10,000 and Pre Burnin = 0. Set Log Every = 10,000 under tracelog, screenlog and treelog.t (this variable has to be set under each of these headings).

(f) When these steps are complete, choose the "File → Save As" command and save the file as Metzker_pol.xml in your work directory. Check that the file has indeed been written to this directory.

2. Create a subdirectory run1 within your work directory and move a copy of the .xml file into it.

3. Click on the program icon to launch BEAST and choose Metzker_pol.xml as the BEAST XML File. If you have installed BEAGLE, then select the 'Use BEAGLE library' option as well. Click on the Run button when you are ready to begin the analysis.

4. BEAST will record the current values of the chain in two files saved to the work directory. One of these will have the file extension .log and will contain the substitution model parameters as well as several tree statistics. The second file will have the file extension .trees and will contain the trees sampled by the Markov chain. In the current example, only one in every 10,000 states visited by the chain is written to the output files. (A third file ending in the extension .state will also appear and can be used to re-start a chain that terminates prematurely.)

5. The progress of your analysis can be monitored in two ways. The BEAST window will show the number of steps completed by the Markov chain, as well as the current values of the posterior probability, the ESS (see below), the likelihood, the prior probability, and the time required per million steps.

6. More detailed information can be obtained by launching the application TRACER. Click on the + button shown in the upper left-hand side of the TRACER window and then select the Metzker_pol.log file from the directory containing the output from the BEAST calculation. This will show the parameter values and tree statistics sampled by the Markov chain and stored in the .log file.

7. Click on the button Estimates to see a histogram of these values along with several summary statistics of the posterior distribution (mean, median, etc.). The 95% HPD is the **highest posterior density interval** which contains 95% of the probability of the posterior distribution; this shows the values of the unknown parameter that are most probable according to the posterior distribution. The **effective sample size** (ESS) is an estimate of the effective number of independent samples from the posterior distribution; because

the successive steps of the Markov chain are correlated, the ESS is usually smaller than the actual number of samples, sometimes substantially so if the chain is slowly mixing. As a rule of thumb, the chain should be run at least until the ESS of each parameter is greater than 200.

8. By clicking on the Trace button in the TRACER window, you can see a plot of the successive values sampled by the Markov chain for each parameter. If the chain is stationary by the end of the burn-in period, then the trace will typically fluctuate in a uniform way across this window. Systematic changes (trends) or very large jumps in the trace usually indicate that the chain has not converged and needs to be run for a larger number of steps.

9. Launch the application TreeAnnotator and select pol_hky.trees for the Input Tree File and pol_hky_MCC.tre for the Output File. Set the Burnin pecentage equal to 50 and select the Mean heights option for Node heights. Click on the Run button to have the program construct the **maximum clade credibility** (MCC) tree, which provides a convenient summary of the different trees visited by the Markov chain. (If the program terminates with an error, try running it again.)

10. The MCC tree will have been saved in the directory containing the other output files generated by the BEAST analysis. Launch the application FigTree and use it to open the file containing the MCC tree. The tree will be displayed in the FigTree window, along with a scale bar that shows the mean number of substitutions per site. The length of each branch is proportional to the time elapsed along that lineage. The leaves (sampled sequences) will be on the right-hand side of the window and the root will be on the left-hand side. Open the Branch Labels menu on the left-hand side of the FigTree window and set the Display option to posterior and then click on the Branch Labels button. This will display the posterior probability that the group of sequences descended from a particular branch are **monophyletic**, i.e., more closely related to one another than to any of the other sequences contained in the sample. Monophyletic groups are also called **clades**. Posterior probabilities close to one indicate that the data provides strong support for the monophyly of those groups. The Export PDF option under the File window can be used to save a picture of the MCC tree.

Instructions for PS 8:

1. Run two independent BEAST analyses of the HIV-1 RT data, using the models and priors described in the instructions given above. Each chain should be run for 20 million steps and separate directories should be used for each run. However, you can create a single .xml file and copy it to each working directory.

   (a) Create a table showing the mean, the 95% HPD interval, and the ESS for the likelihood, the TreeHeight (TMRCA), kappa, gammaShape and proportionInvariant for each of the two runs. How similar are the mean values between the two runs?

   (b) Examine the trace of the log-likelihood for each of the two runs. Do these figures suggest that the runs have converged?

   (c) Construct the MCC tree for each run with the posterior clade probabilities as branch labels. Save each to a PDF and print these out. How strong is the support for the monophyly of the patient and donor HIV-1 RT sequences in each run? Are the donor and patient sequences monophyletic relative to one another?

2. Redo the analyses from (1) using the GTR substitution model with Gamma-distributed rate variation and a class of invariant sites. The GTR model can be selected in the *Site Model* window and you should check the estimate buttons for the AC, AG, AT, CG and GT rates (initial values = 1.0), while leaving the estimate button for the CT rate unchecked. You will also need to choose a prior distribution for each of the five estimated rate parameters, which can be taken to be uniform with Lower = 0.0, Upper = 5.0, and Initial value = 1.0. How do your results compare with those obtained in (1)? Does the substitution model affect the conclusions concerning the relatedness of the patient and victim sequences?

3. Now redo the analyses in (2), but in the *Priors* window of BEAUti select the Yule Model for the Tree.t prior and select a uniform prior for the parameter birthRate.t with Lower = 0.0, Upper = 10, and Initial Value = 1.0. The Yule process is a branching process which assumes that each lineage splits in two at some constant rate (the birth rate). Compare these results with those obtained in (1) and (2)? Which aspects of the analysis are affected by the choice of tree prior?

4. In Bayesian statistics, model choice is often based on quantities called **Bayes' factors**. If D denotes some data and M_1 and M_2 are two models that are proposed to explain the data, the Bayes' factor $K_{1,2}$ of M_1 relative to M_2 is the ratio of the two marginal likelihoods
$$K_{1,2} = \frac{\mathbb{P}(D|M_1)}{\mathbb{P}(D|M_2)}.$$
The quantity $\mathbb{P}(D|M_i)$ is called the **marginal likelihood** because it is obtained by averaging the likelihood function of the data under model M_i over the prior distribution of the parameters on which M_i depends:
$$\mathbb{P}(D|M_i) \equiv \int \mathbb{P}(D|M_i, \Theta) p(\Theta|M_i) d\Theta_i.$$
The Bayes' factor indicates how much more strongly the data supports one model rather than the other: e.g., when $K_{1,2} > 3$, we usually conclude that M_1 is substantially more strongly supported than M_2. Let M_1, M_2 and M_3 denote the three models used in questions (1) - (3). The marginal likelihood of the data under each model can be estimated by exponentiating the mean of the log-likelihood of that estimate. Use this to calculate the Bayes' factor $K_{1,2}$ for model M_1 vs. M_2 and the Bayes' factor $K_{2,3}$ for model M_2 vs. M_3. What do your results say about the HKY model vs. the GTR model? What about the coalescent vs. the Yule process?