

Understanding BEAST2 Results:

Simon J. Greenhill (simon@simon.net.nz).

ARC Centre of Excellence for the Dynamics of Language, Australian National University.

Department of Linguistic and Cultural Evolution, Max Planck Institute for the Science of Human History.

Process the Analysis:

1. Open `Tracer` .
2. Drag the log file (cpacific-ctmc-strict.log) into the “Traces” section (or `File-> Import Trace File`).

Step 1. Look for unstable traces.

Because MCMC is a sequential sampling process, we need to worry that our samples are independent enough for our inferences to be statistically robust.

Look at the ESS column – this is the `Effective Sample Size` , which measures how independent each sample from the MCMC chain is. We want each one to be large.

If the ESS is low then it probably means that the parameter is poorly estimated and we shouldn't rely on that parameter: Perhaps the analysis hasn't run long enough, or perhaps we need to put more weight on that parameter in the analysis so it gets modified more often see: <http://beast.bio.ed.ac.uk/increasing-esss>

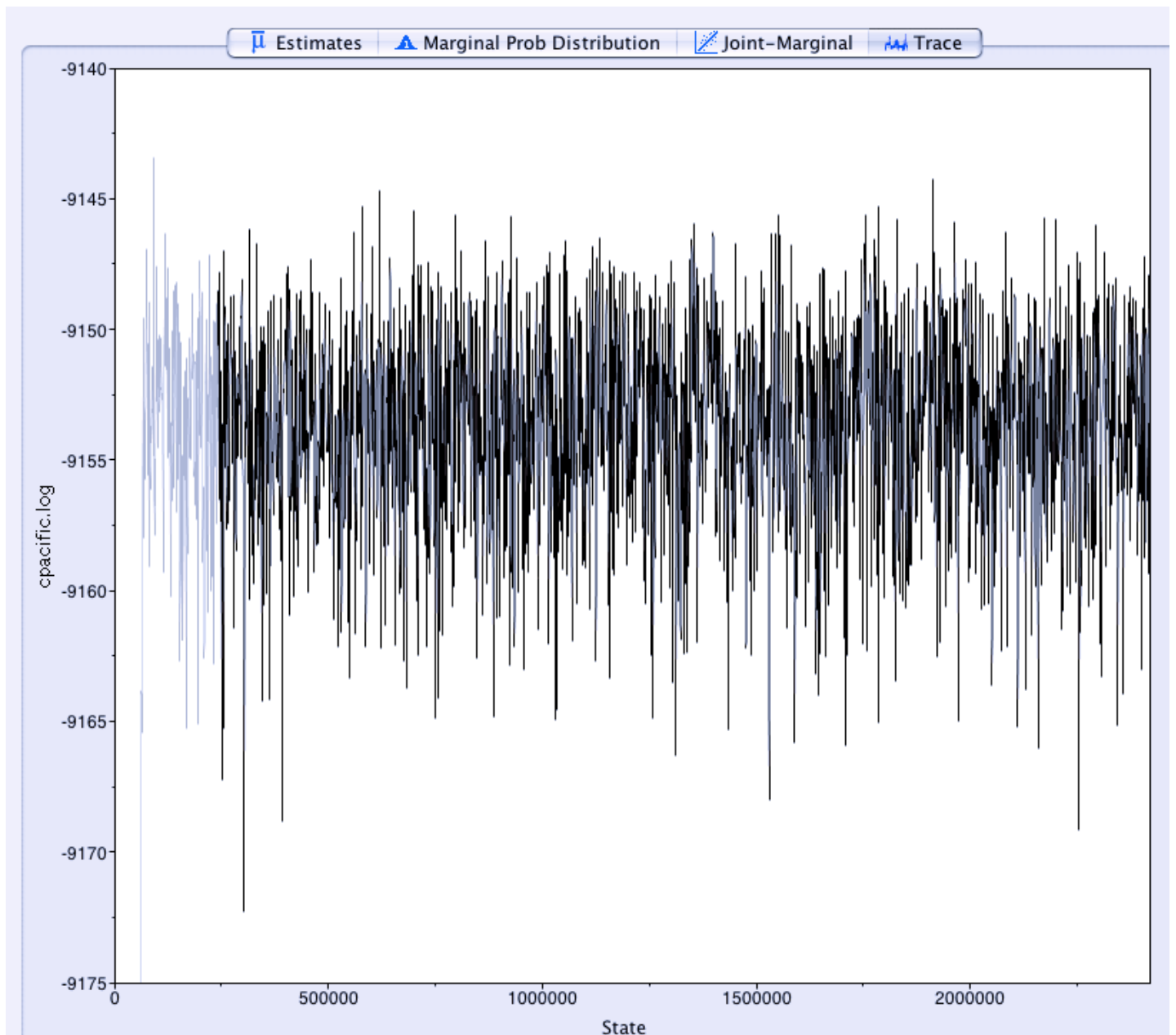
`Tracer` will highlight anything that's below 200, but for publication you probably want > 1000. In the figure below the ESS for gamma_shape_s is slightly low.

[] Look at your data, is there anything with a low ESS?

Traces:			
Statistic	Mean	ESS	T...
posterior	-9153.71	1567	R
likelihood	-9163....	1669	R
prior	10.027	832	R
treeLikelihood.cpacific	-9163....	1669	R
TreeHeight.t:cpacific	0.524	444	R
mutationRate.s:cpacific	1	-	R
gamma_shape_s	2.452	193	R
YuleModel.t:cpacific	12.479	510	R
birthRate.t:cpacific	5.64	2043	R
freqParameter.s:cpac...	0.814	315	R
freqParameter.s:cpac...	0.186	315	R

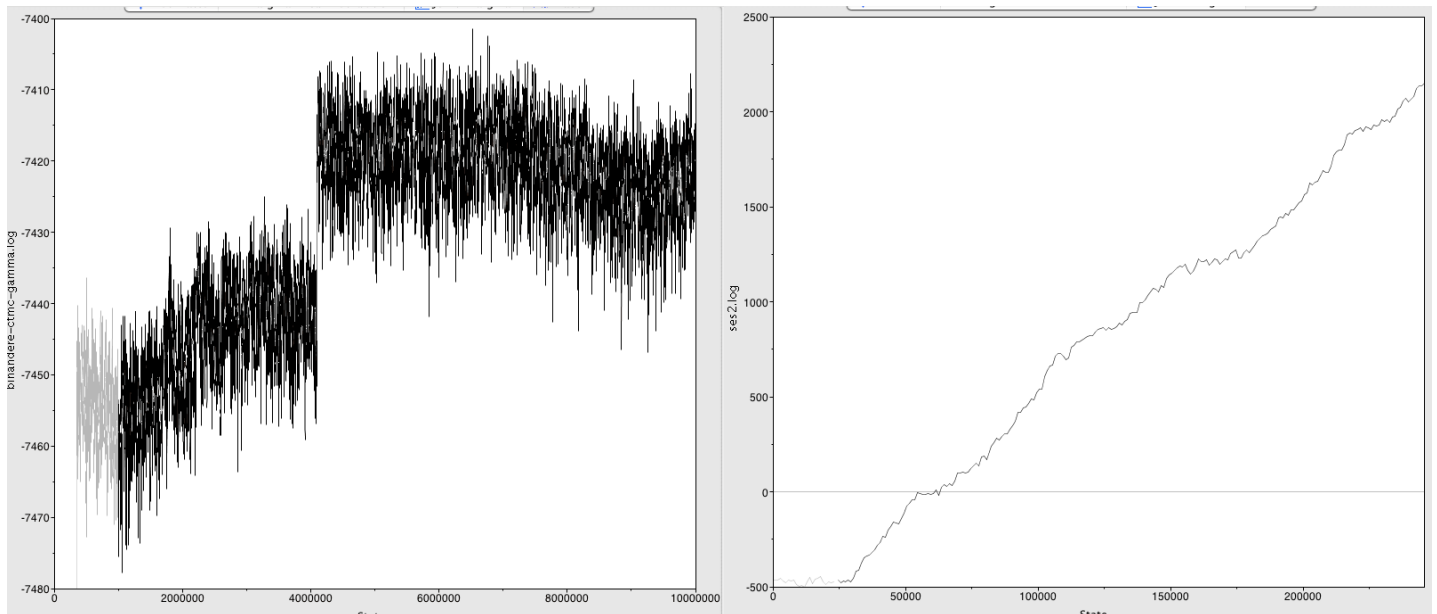
Step 2. Inspect the traces.

Click the **Trace** tab and go through all the traces of each parameter/statistic from the analysis. These are plots of that parameter over time in the analysis. You want to make sure they look like “fuzzy caterpillars”:



... that is, they're consistently estimating a parameter value in the same ballpark (here something around -9153), and not trending up or down.

Here for example, are a couple of bad traces:



... The left one shows a jump in the parameter value. To solve this, increase the burn-in to after the jump (>60,000,000 generations here) and hope you've run the analysis long enough to still get good ESSes in the remainder.

... The right one shows a parameter that's still stabilising. You need to run this analysis a lot longer. Thankfully **BEAST2** can resume analyses so you don't need to start from scratch.

☐ Are all your traces ok?

☐ If not, what are some of the problems you can see?

Step 3. Check **Estimates** :

You should also look at the **Estimates** tab for each parameter. This will show a histogram of the parameter which should be unimodal (i.e. not have multiple peaks) and relatively peaked (i.e. not a flat distribution).

Another key thing to check here is whether the distribution looks truncated at one end. If it is truncated, then the analysis is not allowing that parameter to explore the probability space that it "wants to". There is probably a hard boundary in the analysis from a prior. You should think about relaxing this constraint or removing it to allow the analysis to explore that space properly.

You can also click on the **Estimates** tab to get various statistics about each parameter.

☐ Are all your estimates ok?

Step 4. Figure out how much **burn-in** :

Because of the way MCMC works, we generally start out with terrible parameter estimates based on

random guessing and initial settings, and slowly converge towards good estimates (jargon: “reach stationarity”).

The samples we take from the period before we reach stationarity is called `burn-in`. We want to throw these away as they’re not good estimates of the real parameter values.

To identify how much burn-in we need, we look at where the traces all stabilise. A common choice is 10% (it should be the default in Tracer). Is this enough for our data? You can change it in the top left of the Tracer window to try out different amounts.

[] What burn-in do you think is appropriate here?

Step 5. Check out interesting parameters:

Let’s see how old the age of Eastern Polynesian is. Remember that there’s a debate between the ‘short’ and ‘long’ chronologies (~800-1000 years vs ~1000-1200 years ago).

Select the `mrcatime(EastPolynesian)` Parameter and choose the `Estimates` tab. This is the time to the `Most Recent Common Ancestor` (MRCA) of the East Polynesian group.

This will show you the posterior probability distribution of the age of East Polynesian according to these data and model. The key bits of information are the mean and 95% Highest Posterior Density Interval (HPD) – the Bayesian equivalent of a Confidence Interval.

The mean age of Eastern Polynesian is – in my analysis – about 1283 years, with a 95% H.P.D. of 1153-1414 years.

Your turn. Let’s see how old we are estimating Central Pacific to be. The archaeological record suggests that the settlement of Fiji happened around 3200 B.P. (Nunn et al ‘04).

Select the `TreeHeight` Parameter and choose the `Estimates` tab.

[] What's the mean age and 95% HPD of Central Pacific?

[] Is this close to the age of around ~3200 B.P.

View the Trees:

Remember that Bayesian phylogenetics produces a posterior tree distribution which contains lots of trees. We can do a couple of things.

Visualise the entire posterior using Densitree:

A `densitree` (Bouckaert '10) draws each tree in the posterior on top of each other. It's a great way to visualise the whole posterior at once.

Open the tree in `densitree` (in the `BEAST` folder) and then load the tree file that you generated (e.g. `cpacific-ctmc-strict.trees`).

Central Pacific is quite a reticulate language zone with lots of dialect chains and borrowing between languages (e.g. Walworth '12, Wilson '12, Salisbury '10, Gray et al '10). We should expect to see some conflicting signal where different trees show support for different groupings.

[] What regions of high conflicting signal are there?

Construct a summary “maximum clade credibility tree”:

The M.C.C. Tree is a good *single* tree representation of the posterior. Here we calculate the product of all the clade probabilities on each tree in the posterior to find the 'best' one. See

<http://beast.bio.ed.ac.uk/summarizing-posterior-trees>.

1. Open `TreeAnnotator`
2. Set `Burn-in Percentage` to 10 (percent) or whatever you thought appropriate from the `Tracer` results.
3. Select `Input Tree File` by clicking 'Choose File' and select your tree file (`cpacific-ctmc-strict.trees`)
4. Select the `Output File` by clicking 'Choose File' and typing a good file name for the resulting trees in. Personally I'd choose something like “`cpacific-ctmc-strict.mcct.trees`”.
5. Make the MCC tree.

[] Successfully make an MCC Tree

Open this file in `FigTree`. Because this is a Bayesian analysis with time-dated trees, let's turn on a few options.

1. Click the checkbox by `scale axis` to turn on a time-scale.
2. Turn on Node Labels. Open the `Node Labels` section on the left
3. Click the check box
4. Change the `Display` value to `posterior`.
5. Change `Sig. Digits` to 2 to make it easier to read.

These values give us the number of times each of the subgroups was found in all trees in the posterior. It is the strength of support, in these data, given this model, for that subgrouping. We want high numbers – A value of 1.00 means that this subgrouping was always found. Anything below about 0.7 is weak.

[] What subgroups have very high support (=1.00)

☐ Are there subgroups with low support (<0.70)

Remember that in our data exploration we saw that Mangareva and Nukuoro in particular had lots of missing data.

☐ Do Mangareva and Nukuoro behave here, or should we remove them?

Which is better?

What visualisation should you prefer? It all depends on how well attested your trees are and how much conflicting signal there is. Densitree shows you everything, but by showing you everything it can be hard to see the forest for the trees. On the other hand, the MCC tree shows the majority signal but loses all the more complex history.

As always, don't rely on one tree. If you go on to use comparative methods or do other subsequent things with your trees, make sure you do it across **all** the trees in the posterior and not just one.

Your Turn (if we have time):

Go back to BEAUTi (remember that you left it running, right?) and change the clock model to a relaxed clock (log normal).

Make sure you change the `tracelog` and `treelog` filenames on the `MCMC` tab to something like `cpacific-ctmc-relaxed.log` and `cpacific-ctmc-relaxed.trees`.

☐ Generate an XML for the relaxed clock analysis

☐ Run the relaxed clock analysis

If you have time, you could try setting up a *covarion* analysis. The Covarion is a nice model that allows each cognate in our data to switch 'on' and 'off' at different points on the tree (Penny et al. '01). It's often found to be the best fitting model for language data in my experience, however, it can be a bit trickier to stabilise (especially the hidden frequencies which control the on/off switching).

To do this, use the `BinaryCovarion` template in `BEAUTi` rather than `BinaryCTMC`, and ask me for tips.