Computational macroevolution with BAMM and BAMMtools

Dan Rabosky

June 20, 2017

1 BAMM exercises

Work in pairs for these sets of exercises. Try to ensure that at least one member of your group has experience with R and is comfortable with loading and plotting data, performing basic statistical analyses, etc. Choose an example dataset for which your group wishes to analyze rates of species diversification. You may also choose one of your own datasets if you have one. Choose a dataset that is relatively small (say, < 300 species) if possible, such that you will be able to obtain convergence in the brief time we have to work on this as a group. The idea is to explore a number of basic BAMM and BAMMtools analyses.

All groups should do a BAMM analysis on some example dataset. You should feel free to perform this analysis on multiple computers and to compare results as a partial check for convergence. Once you've obtained a decent sample, choose several of the *post-BAMM* analysis and visualization exercises listed below to work on. For help, please look at the relevant the BAMMtools documentation files. To obtain help on a particular BAMMtools function, you can simply type the name of the function, preceded by a question mark. For example, for help on the function `plot.bammdata`, you would run the following command in your R console:

```
> ?plot.bammdata
```

Be sure to check the `Examples` section at the end of the help file, which illustrate many of the types of analyses that a typical user may wish to do. Extensive documentation is also available on the project website (`www.bamm-project.org`).

1.1 Convergence-checking notes

If you have an MCMC output file `mcmc_out.txt`, the following commands illustrate how you can load the file and discard a particular fraction as burnin:

```
> xx <- read.csv("mcmc_out.txt")
> burnfrac <- 0.10;
```

```
> start <- floor(burnfrac * nrow(xx))
> xxpostburn <- xx[start:nrow(xx), ]
```

xxpostburn is now a set of samples after discarding the first 10% as burnin. Some specific things you may wish to do:

- Plot the trace of your sampled log-likelihoods as a function of the number of generations to determine an appropriate burnin threshold.
- If you have performed multiple runs, compare the log-likelihoods and numbers of rate shifts across runs. For example, in the example above, `median(xxpostburn$logLik)` and `median(xxpostburn$N_shifts)` will give you the median log-likelihood and shift counts across your simulated posterior distribution.
- load the `coda` package and check the effective sample sizes of the number of rate shifts and the log-likelihoods
- If you are not achieving convergence, consider what changes you would make to your MCMC settings to obtain a better sample. It is quite possible that you will not be able to obtain convergence in the amount of time available for the workshop, so at some point you should just move forward to the `BAMMtools` exercises below, even if your convergence statistics could be better...

2 Postprocessing with `BAMMtools`

2.1 Core activities

- Test sensitivity to the prior using the `plotPrior` function
- Create a `bammdata` object using the function `getEventData`
- Advanced convergence checks with `BAMMtools`
- Visualize overall rates using `plot.bammdata`
- Bayesian model selection with `computeBayesFactors`
- Construct the 95% credible set of macroevolutionary rate shift configurations, using the function `credibleShiftSet`; plot with `plot.credibleshiftset`
- Extract the best overall rate shift configuration with `getBestShiftConfiguration`
- Rate through time analyses with `getRateThroughTimeMatrix` and `plotRateThroughTime`
- Clade and tip-specific macroevolutionary rates with `getCladeRates` and `getTipRates`
- Using `subtreeBAMM` to prune a `BAMM` analysis to a smaller subset of taxa

2.2 Model selection

Identify the best diversification model by summarizing model posterior probabilities. Conduct Bayesian model selection using Bayes factors if possible. Consider using the function `computeBayesFactors` for this. See also the function I distributed (`plotBFcolorMatrix.R`), which allows you to plot a nice pairwise color matrix of Bayes factors. Consider in particular the support for models with k rate shifts relative to a null model with 0 rate shifts. At a minimum, evaluate the relationship between your prior and posterior densities on the number of rate shifts. To extract the posterior probabilities of a given number of rate shifts from your post-burnin MCMC data (see above), you can do something like this:

```
> postprobs <- table(xxpostburn$N_shifts) / nrow(xxpostburn)
```

If your prior or posterior probabilities are not estimated well for some particular model (such as the null model with zero shifts), you may not be able to compute Bayes factors. This follows immediately from the definition of a Bayes factor. If $Prob(M_i)$ and $\pi(M_i)$ are the posterior and prior probabilities of model M_i, the Bayes factor evidence for some model M_j relative to model M_k can be computed as:

$$BF_{j,k} = \frac{\frac{Prob(M_j)}{\pi(M_j)}}{\frac{Prob(M_k)}{\pi(M_k)}} = \frac{Prob(M_j)}{Prob(M_k)} \frac{\pi(M_k)}{\pi(M_j)} \quad (1)$$

2.3 Advanced convergence tests with **BAMMtools**

The convergence checks you performed above were pretty simplistic: you tested effective sample sizes in log likelihoods and the number of shifts. With `BAMMtools`, there are many advanced convergence checks that will give you better insight into whether independent `BAMM` runs are converging to the same target distribution.

One good option is to compare the marginal shift probabilities for two BAMM runs. If independent runs have converged on the same distribution of macroevolutionary rate shift configurations, they should have similar posterior probabilities of rate shifts on individual branches. You can compute branch-specific marginal shift probabilities with the function `marginalShiftProbsTree`.

As another option, you can compare tip-specific speciation rates for two individual BAMM runs, using the `BAMMtools` function `getTipRates` . If runs have converged, they should yield similar rate estimates for individual taxa. You can plot these rates in pairwise fashion to see if you have a 1:1 correspondence.

2.4 Phylorate plots

Create a `bammdata` object using the function `getEventData`. Summarize patterns of rate variation using a phylorate plot (`plot.bammdata`). When you read in your event data file, set the `nsamples` argument to be no greater than 1000; you will possibly run into

memory issues if you do not do this. Make plots with several different color schemes (see the detailed help on this function). Plot a few random shift configurations from your posterior by first extracting a sample from the posterior using the function `subsetEventData`. You can then plot this with plot.bammdata, and you can add the precise location of the rate shifts with `addBAMMshifts`. For example, to extract and plot the 10^{th} sample from your posterior, you might do something like this:

```
> edata <- getEventData(mytree, "event_data.txt",
+          nsamples=500, burnin=0.1)
> zz <- 10
> edatasub <- subsetEventData(edata, index = zz)
> plot.bammdata(edatasub)
> addBAMMshifts(edatasub, cex=1.5)
```

2.5 Clade-specific rates

Find a clade of interest within your dataset for which you would like to examine evolutionary rates. Using the `getCladeRates` function, extract the posterior distribution of rates for the clade. You will need to identify the `ape` format node number for your focal clade. You can find this node number in one of several ways. One, you can plot the tree and then plot the node numbers:

```
> plot.phylo(mytree)
> nodelabels(cex=0.5)
```

I sometimes find it helpful to output a large pdf with my tree with the node numbers written small to reduce overprinting (then open the file in Preview or another application to find specific node numbers):

```
> pdf(file = "mytree.pdf", height=15, width=10)
> plot.phylo(mytree, cex=0.3)
> nodelabels(cex=0.2)
> dev.off()
```

Finally, you can do this interactively using the `identify.phylo` function:

```
> data(whales)
> plot.phylo(whales)
> identify(whales)
```

The `cex` option controls the relative scaling of the node and tip label font size, so you may have to adjust this to achieve readability. A better option is to use the function `getMRCA` to access the specific node number that is the common ancestor of two taxa. For example, if I had a complete phylogenetic tree of mammals, I could find the node number of the common ancestor by specifying the species *Homo_sapiens* and *Tachyglossus_aculeatus* (short-beaked echidna).

```
> sp1 <- Tachyglossus_aculeatus
> sp2 <- Homo_sapiens
> getMRCA(mytree, tip = c(sp1, sp2))
```

This should give you the relevant node number. Look at the rate distribution for this clade (hint: see help on `getCladeRates`). The base R function quantile can be very useful: if you want the 0.05 and 0.95 quantiles on some distribution, you could do:

```
> myrates <- getCladeRates(myEventData)
> lambda <- myrates$lambda
> median(lambda)
> quantile(lambda, c(0.05, 0.95))
```

Now, compare this to the distribution of rates across the remainder of the tree, after excluding the clade above (see the `nodetype` argument to `getCladeRates`).

2.6 Rates through time

Generate a rate-through-time plot for speciation, extinction, and/or trait evolution. Make a rate-through-time plot for just a single subtree from your dataset (e.g., by including or excluding specific nodes). See the previous section for more information on identifying particular nodes.

2.7 Node-specific shift evidence

Analyze the evidence for rate shifts at specific nodes in your phylogeny. First, use the function `marginalShiftProbsTree` to compute the marginal shift probabilities for each node in your phylogeny. See help on this function for details. Plot a copy of your phylogeny, but where the branch lengths are equal to the (marginal) posterior probability that a rate shift occurred on this branch. Which nodes have the highest probability of a rate shift? Is your phylogeny dominated by one or several rate shifts?

2.8 Distinct shift configurations

Identify the distinct shift configurations in your BAMM data object. How many distinct configurations are in the 95% credible set? The object returned by credible shift set contains many useful things. You can use the function `summary` on your credible shift set object to obtain some information. However, the best thing to do is to access attributes of the credible set object directly. See the help on `credibleShiftSet`, and look in the section `Value`. These are all attributes that can be accessed with the dollar-sign operator, e.g.,

```
> css <- credibleShiftSet(myBAMMdata, myPrior)
> css$frequency
```

Find the MAP probability shift configuration using `getBestShiftConfiguration`.

2.9 Macroevolutionary cohorts

Generate and interpret a macroevolutionary cohort matrix for your data. You first need the function `getCohortMatrix` to compute pairwise probabilities that any two taxa share a common macroevolutionary rate regime. You can then use the function `cohorts` to generate the actual cohort matrix plot. The function will plot your reference phylogeny in the left-hand and upper margins.

2.10 "Pruning" a BAMM analysis to a subset of taxa

Imagine that you are interested in speciation rates for just a handful of taxa - perhaps the set of all birds that regularly occur in your geographic region, for example. Because incomplete taxon sampling can bias diversification analyses, you can't just analyze this set of taxa in BAMM - it is (almost always) best to analyze the most complete phylogenetic tree possible. In this case, you might analyze a phylogeny for all birds, which should minimize the bias attributable to incomplete taxon sampling. However, once you've finished, you might want to just pull out the BAMM results for the set of taxa that you care about. You can do this with the function `subtreeBAMM` . This function prunes out all the taxa you don't care about, while retaining all the diversification information from the full BAMM analysis that applies to the branches you do care about.

```
> special_tips <- c("Pied_butcherbird", "Grey_currawong", "Black_kite")
> subt <- subtreeBAMM(myBAMMdata, special_tips)
> plot.bammdata(subt)
```

3 And most importantly....

Provide an interpretation of the macroevolutionary dynamics you have observed in the dataset you analyzed. You may need to use multiple lines of evidence to fully understand things. How many rate shifts overall do you observe, and how strong is the evidence? How many distinct shift configurations, and are there one or several rate shifts (or, perhaps, regions of the tree) that are consistently and strongly associated with rate shifts? Do those clades tend to undergo accelerations or decelerations in diversification? Which lineages tend to have correlated macroevolutionary dynamics?