Lab 13:
Estimating Diversification Rates: Heterogenous and Character Dependent Processes

By Will Freyman

1 Before you begin

Today's lab will use R and the software BAMM. Please install R:
`https://www.r-project.org/`
Please install BAMM:
`http://bamm-project.org/`

2 Introduction to heterogenous birth-death processes

In labs 6 and 11 we used birth-death processes as tree priors in Bayesian analyses. These were birth-death processes with constant speciation and extinction rates, and so were homogenous diversification processes. In this lab we'll look at heterogenous processes in which the rates of diversification change over the tree. The changes in diversification rates could be caused by a change in a character state (e.g. a key innovation) or they could be due to some other dynamic such as climate change or the availability of new niches.

3 Character dependent birth-death processes

In the first exercise we'll use the **binary-state speciation and extinction** [BiSSE; Maddison et al., 2007] model using the R package **diversitree** [FitzJohn, 2012]. BiSSE models the evolution of a binary character and a birth-death process jointly, and uses a separate set of speciation and extinction rates for each state of the character (see Figure 1). Under BiSSE, the diversification process is dependent on the character's state.

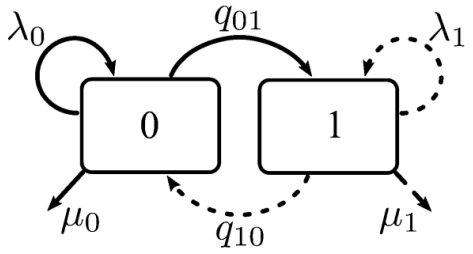

Figure 1: States and transitions in the BiSSE model. BiSSE has 6 rate parameters: speciation (λ) and extinction (μ) for character states 0 and 1 and transition rates (q) for changing state. Image from Goldberg et al. [2011].

3.1 Install and setup diversitree

Start up R and install the diversitree package:

```
install.packages("diversitree")
```

Now lets load the package:

```
library(diversitree)
```

3.2 Simulate constant rate birth-death process

Let's first simulate a constant rate birth-death process. We'll simulate a tree where speciation λ is 0.1 and extinction μ is 0.03. We'll save our diversification rates in a vector, and simulate the tree for 30 time units:

```
div_rates = c(0.2, 0.03)
tree = tree.bd( div_rates, max.t=30.0 )
```

Let's plot the tree:

```
plot(tree)
axisPhylo()
```

Try simulating and plotting a few trees with different values of λ and μ.

3.3 Simulate an independent binary character over the tree

Now let's simulate a character evolving neutrally over the tree. During the evolution of this character the tree is fixed – the state of the character is independent of the diversification process. We'll use the **Mk2** model which simply allows for two separate transition rates between character states, q_{01} and q_{10}.

```
char_rates = c(.1, .2)
character = sim.character(tree, char_rates, x0=0, model="mk2")
```

Let's plot our character so we can see the way it evolved over the tree:

```
plot(tree, show.tip.label=FALSE)
axisPhylo()
colors=c("lightblue", "blue")
tiplabels(col=colors[character+1], pch=19, adj=1)
nodelabels(col=colors[attr(character, "node.state")+1], pch=19)
```

Try simulating and plotting with different values of q_{01} and q_{10}.

3.4 Simulate under BiSSE

Now we will simulate a tree and a binary character jointly using BiSSE. Whenever the character is in state 1 the speciation rate will be triple that of state 0. First set up our rates vector in the order λ_0, λ_1, μ_0, μ_1, q_{01}, q_{10}. We'll also set the random number generator seed to 13 so that we all get the same results.

```
set.seed(13)
rates = c(0.1, 0.3, 0.03, 0.03, 0.01, 0.01)
```

Now let's simulate our tree and character and then plot it:

```
tree = tree.bisse(rates, max.t=30, x0=0)
plot(history.from.sim.discrete(tree, 0:1), tree, col=colors, show.tip.label=FALSE)
```

You should see that at the base of the tree the root state is 0 (light blue) and it diversifies at a medium rate. At about 17 time units ago one lineage transitions to state 1 (dark blue) and undergoes a rapid radiation (Figure 2).

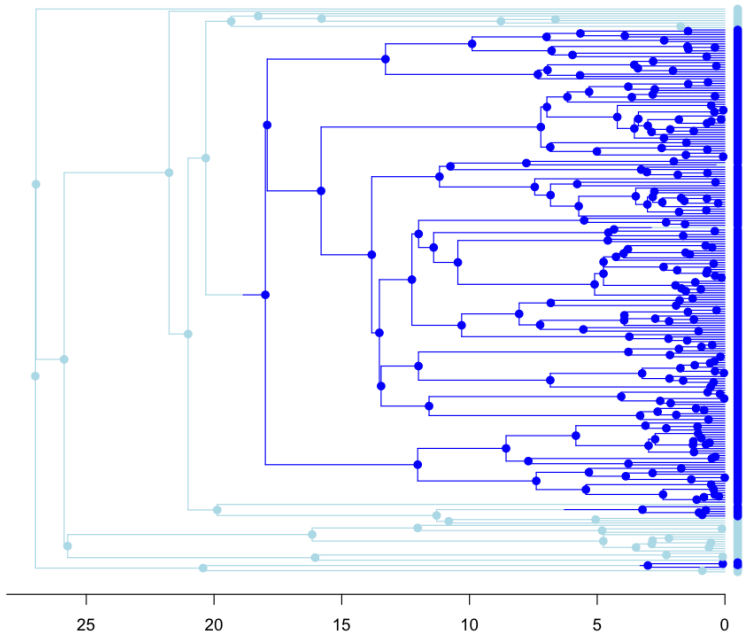

Figure 2: Tree and binary character simulated under BiSSE. State 1 (dark blue) has a speciatian rate three times that of state 0 (light blue).

3.5 Testing for character dependent diversification

We know the tree and character shown in Figure 2 evolved jointly under a BiSSE process; the diversification process was dependent on the character. However, in an actual empirical study usually all we have is a tree and the observed tip states. How do we test for character dependent diversification? Let's treat our simulated tree and tip data as if this was an empirical study and see if we can successfully detect character dependent diversification.

First, let's fit a BiSSE model to the simulated data. Since the diversification rate is dependent on the character state, the BiSSE likelihood function takes both the tree and the tip data:

```
bisse_lik = make.bisse(tree, tree$tip.state)
```

We'll find the maximum likelihood estimate for all parameters. We use a helper function to find starting values of the parameters for the search heuristic.

```
p = starting.point.bisse(tree)
fit_bisse = find.mle(bisse_lik, p)
```

What were the parameter estimates and log-likelihood?

```
coef(fit_bisse)
fit_bisse$lnLik
```

Ok, now let's estimate diversification rates under a constant rate model, where the diversification process is independent of the character states. We'll use the same BiSSE likelihood function as above, but constrain the diversification rates to be equal in both character states:

```
constant_lik = constrain(bisse_lik, lambda0 ~ lambda1, mu0 ~ mu1)
```

Again, find the maximum likelihood estimates of our parameters:

```
fit_constant = find.mle(constant_lik, p)
coef(fit_constant)
fit_constant$lnLik
```

Now let's use AIC to select the best model:

```
AIC(fit_constant, fit_bisse)
```

Question 1:

1. Which model did the AIC select? Can we successfully test for character dependent diversification?
2. Compare the estimated parameter values to the true values you used to simulate the data. How close were the speciation rates? What about the extinction rates?

3.6 Extensions and caveats

After the groundbreaking Maddison et al. [2007] introduced BiSSE, an entire class of models were built upon it. These models have provided an exciting new quantitative framework to test macroevolutionary patterns/processes such as adaptive radiations and other ecological and geographic factors that affect the shape of the tree of life. They include MuSSE (Multiple State Speciation and Extinction), QuaSSE (Quantitative State Speciation and Extinction), GeoSSE (Geographic State Speciation and Extinction), and BiSSE-ness (BiSSE-Node Enhanced State Shift). All these models are implemented in the excellent **diversitree** R package.

However, these models have been found to be prone to high Type 1 error rates, where characters that evolved independent of the diversification process are inferred to have statistically significant associations with diversification rates; the diversification process is erroneously inferred to be character dependent [Rabosky and Goldberg, 2015]. Novel extensions of these models that deal with these problems include the HiSSE (Hidden State Speciation and Extinction) model, in which an unobserved "hidden" character is allowed to affects the diversification rates instead of the observed character [Beaulieu and O'Meara, 2015]. If you plan to use BiSSE or other similar models, be sure to test for false positives!

Run this code and then answer Question 2. It might take 5 to 10 minutes, so you may want to go on to the next section while you wait.

```
fp = 0
n = 20
char_rates = c(.1, .2)
for (i in 1:n) {
    print(paste("Beginning replicate", i, "out of", n))
    ind_character = sim.character(tree, char_rates, x0=0, model="mk2")
    dep_bisse_lik = make.bisse(tree, ind_character)
    fit_dep_bisse = find.mle(dep_bisse_lik, p)
    ind_constant_lik = constrain(dep_bisse_lik, lambda0 ~ lambda1, mu0 ~ mu1)
    fit_ind_constant = find.mle(ind_constant_lik, p)
    aic_results = AIC(fit_ind_constant, fit_dep_bisse)
    if (aic_results$AIC[[2]] < aic_results$AIC[[1]])
        fp = fp + 1
}
fp/n
```

Question 2:

1. What did this code do? Describe what was being tested and how. All the major functions were described above.
2. What were the results? Put the results into context; what do they mean for analyses using BiSSE type models?

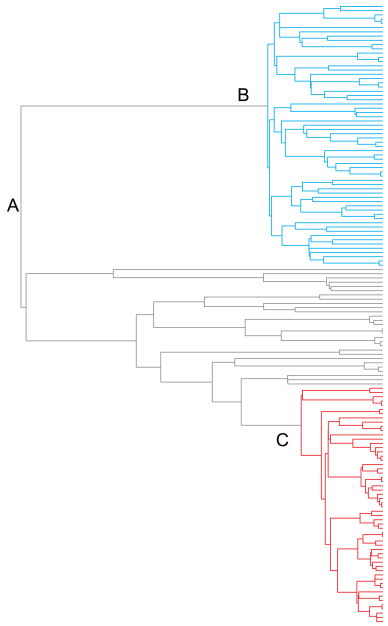

Figure 3: Example of tree simulated under mixture of three distinct evolutionary processes. (A) Clade diversification under constant-rate "background" diversification process with $\lambda = 0.032$ and $\mu = 0$. (B) Shift to new adaptive zone with subsequent diversity-dependent regulation of speciation and diversity-independent extinction (blue branches; $\lambda_0 = 0.395; K = 66; \mu = 0.041$). (C) Another lineage shifts to diversity-dependent speciation regime (red branches; $\lambda_0 = 0.21; K = 97; \mu = 0.012$). Text and image from Rabosky [2014]

4 Inferring diversification rate shifts

The BiSSE class of models test whether a character's state is associated with different diversification rates. But how do we detect shifts in diversification rates without character information or without specifying the locations of rate shifts ahead of time? The program BAMM [Bayesian Analysis of Macroevolutionary Mixtures Rabosky, 2014] models time-varying and heterogenous diversification processes on phylogenies (see Figure 3). The BAMM model assumes that distinct diversification regimes occur across the branches of phylogenetic trees under a compound Poisson process. This allows for complex mixtures of time-dependent, diversity-dependent, and constant-rate diversification processes through time and among lineages.

Here we will run a BAMM analysis over an example whale phylogeny. Our goal is to identify the set of diversification rate shifts that best fits the data (see Figure 4). The BAMM

model can approximate continuous variation in evolutionary rates by inferring multiple discrete rate shifts.

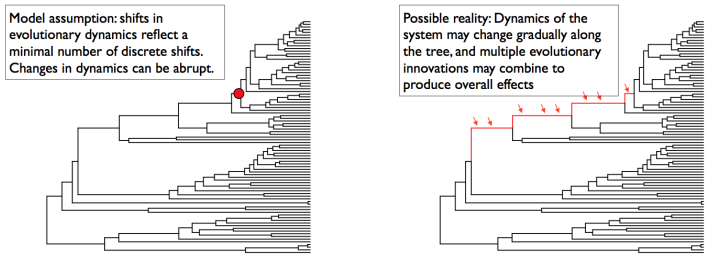

Figure 4: The tree on the left depicts the maximum shift credibility configuration identified by BAMM. This is simply the shift configuration sampled during simulation of the posterior with the highest marginal probability, analogous to the "maximum clade credibility tree" sampled during a Bayesian phylogenetic analysis (see below). The tree on the right gives an alternative interpretation of the results that would be invisible in the BAMM framework. Here, a relatively minor set of evolutionary shifts combine to produce - over several branches (red) - a major shift in evolutionary dynamics. Text and image from `http://bamm-project.org/rateshifts.html`

4.1 Setting up a BAMM analysis

Download the `whales.tre` and `template_diversification.txt` files from here:

- `http://ib.berkeley.edu/courses/ib200/labs/13/whales.tre`
- `http://ib.berkeley.edu/courses/ib200/labs/13/template_diversification.txt`

Modify the following lines of `template_diversification.txt`:

```
treefile = whales.tre
simulatePriorShifts = 0
numberOfGenerations = 100000
mcmcWriteFreq = 1000
eventDataWriteFreq = 1000
printFreq = 100
acceptanceResetFreq = 1000
```

Run the analysis:

```
bamm -c template_diversification.txt
```

4.2 Check for MCMC convergence

We'll use R to check for MCMC convergence and summarize the output from our BAMM analysis. Be sure to set your R working directory to the same directory you ran BAMM in.

First, let's check convergence with a quick and dirty plot of the MCMC log-likelihood trace:

```
mcmcout <- read.csv("mcmc_out.txt", header=T)
plot(mcmcout$logLik ~ mcmcout$generation)
```

Discard the first %50 of samples as burnin:

```
burnstart <- floor(0.5 * nrow(mcmcout))
postburn <- mcmcout[burnstart:nrow(mcmcout), ]
```

Now calculate the effective sample size (ESS) for the log-likelihood values and for the parameter we are most interested in – the number of diversification rate shifts. To calculate

ESS we'll use the R package **coda**, which was mentioned in earlier labs though not required. Install it if you need to.

```
library(coda)
effectiveSize(postburn$N_shifts)
effectiveSize(postburn$logLik)
```

Question 3:
Using both the log-likelihood trace plot and the ESS values, did the MCMC converge?

4.3 Summarize the results

For the sake of time, we are going to go ahead and summarize our results. Obviously, if the MCMC has not converged none of these results should be taken seriously!

We'll use the R package **BAMMtools** [Rabosky et al., 2014] to summarize the results. First read in data:

```
install.packages("BAMMtools")
tree = read.tree("whales.tre")
edata = getEventData(tree, eventdata = "event_data.txt", burnin=0.5)
summary(edata)
```

Question 4:
What is maximum a posteriori (MAP) number of diversification rate shifts?

Now let's visualize the mean, model-averaged speciation rates along the entire phylogeny:

```
plot.bammdata(edata, lwd=2, legend=T)
```

Looks cool. This averaged over all the models (different shift configurations) proportionately to their posterior distribution. One clade – the dolphins – is inferred to be undergoing much high speciation rates compared to the rest of the whales.

What if we want to visualize the individual rate shift configurations? Using this we can visualize the credible set of shift configurations over the tree:

```
css <- credibleShiftSet(edata, expectedNumberOfShifts=1, threshold=5, set.limit = 0.95)
css$number.distinct
summary(css)
plot.credibleshiftset(css)
```

4.4 Caveats

The BAMM model assumes that diversification rate shifts do not occur on the unobserved branches that went extinct. This means the model assumes some lineages can effectively predict the future – if the lineage will eventually go extinct it cannot undergo a rate shift. This is highly unrealistic and will bias the calculation of extinction probabilities. Simulations have shown that rate shifts on lineages going extinct do not strongly influence estimates, however the consequences for empirical data are unknown. Another approach that solves this problem is implemented in RevBayes. The approach used in RevBayes is to make the diversification rates drawn from a discrete (as opposed to continuous) distribution. This allows

RevBayes to numerically integrate over all possible rate categories and correctly calculate probabilities.

A forthcoming paper in PNAS has shown that BAMM analyses can be highly sensitive to the prior on the expected number of rate shifts. If you use BAMM or other similar models you should consider experimenting with different priors and testing whether your results are robust to different choices on the prior.

Question 5:
How could incorporating fossils into a BAMM analysis affect the outcome?

Please email me the following:

1. Your answers to questions 1-5.

References

Jeremy M Beaulieu and Brian C O'Meara. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *bioRxiv*, page 016386, 2015.

Richard G FitzJohn. Diversitree: comparative phylogenetic analyses of diversification in r. *Methods in Ecology and Evolution*, 3(6):1084–1092, 2012.

Emma E Goldberg, Lesley T Lancaster, and Richard H Ree. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Systematic Biology*, 60(4):451–465, 2011.

Wayne P Maddison, Peter E Midford, and Sarah P Otto. Estimating a binary character's effect on speciation and extinction. *Systematic biology*, 56(5):701–710, 2007.

Daniel L Rabosky. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PloS one*, 9(2):e89543, 2014.

Daniel L Rabosky and Emma E Goldberg. Model inadequacy and mistaken inferences of trait-dependent speciation. *Systematic Biology*, 64(2):340–355, 2015.

Daniel L Rabosky, Michael Grundler, Carlos Anderson, Jeff J Shi, Joseph W Brown, Huateng Huang, Joanna G Larson, et al. Bammtools: an r package for the analysis of evolutionary dynamics on phylogenetic trees. *Methods in Ecology and Evolution*, 5(7):701–707, 2014.