**Lab assignments #3 and #4 – BIOINF702, Comparative Bioinformatics**

**Due:** the last day of class, Friday, June 1 – email to Nick Matzke, [n.matzke@auckland.ac.nz](mailto:n.matzke@auckland.ac.nz)

In our paper readings, we have been focusing on model-based inference in phylogenetic comparative methods (PCMs), and statistical model comparison. We have seen that the likelihood function is key to these methods, as well as much of the rest of statistics (both frequentist and Bayesian approaches).   
  
In lab, we have been learning R. In these assignments, we will develop our skills in R and in likelihood functions, and think further about the readings in some short-answer questions.  
  
I will be available to help with these assignments in the computer lab, and in email/by appointment (Commerce A, room 108).

**Lab assignment #3 – Likelihood functions**

**Marks: 35% for each of the two models, 30% for the questions in #5)**

**1.** Pick two of these models, and write R functions that calculate the likelihood of the data, given inputs (parameters, and tree/data). We will work on the Birth-Death model in computer lab, so you might want to pick that as one of your models

* Birth-death model of a phylogeny
* 2-rate Markov model for a discrete binary character
* Brownian motion model for a continuous character
* Binary state-dependent speciation/extinction (BiSSE) model
* Some other model of interest to you (as long as it involves genetic or phylogenetic data, and produces a likelihood)

You are welcome and encouraged to help each other, and to reverse-engineer R functions that already exist in various R packages (including looking at their source code).

**Data:** The easiest data to access are datasets available inside the various R packages, but you may choose any you like.

For both models, make sure that:

* Your function is doing the actual calculation on the phylogeny, not merely calling a function from an R package
* If you copy lines of a previously-published function, that is fine, but re-name the variables so they make sense to you, and comment each line as below.
* Write the function as if you were teaching it to a beginner – so, comment each major part of the calculation, and say what it is doing
* Write documentation (e.g., as comments before the function).
  + Explain what the function does,
  + explain each input (e.g. what format),
  + cite any packages/functions that inspired your function, and
  + make a citation to an article/book/page/equation number(s) that gives the equations used in your function (these may well be mentioned in the published R packages you are looking at)

**2.** Save the functions as .R files, and post them to GitHub or Gist (you will have to create an account). Most work in phylogenetic comparative methods is open-source and in R, so I encourage you to post it publicly. If you post it publicly, paste the links here for me to look at. If you post it privately, screenshot the posted code and paste the screenshot here. Also copy/paste the function text here, or email the code files to me.

**3.** Demonstrate that your function works by writing a script that

* Reads in data/tree, and prints out the log-likelihood
* Does the same for an already-available version of the function
* Show that both your function and the published one return the same log-likelihood
* I will run this script to double-check that you’ve got it working!
* Again, post the scripts to GitHub or Gist and give me either links or screenshots/pasted code.

**4.** For each of the two models, calculate the data likelihood for a large number of parameter values (change just one parameter value, if there are several possible inputs), and plot both the likelihood curve, and the log-likelihood curve.

Print out the curves (make sure they are properly labeled, have good axis labels, etc.). You will draw on them in question 5.

**5.** Answer these questions (short answers, e.g. a few sentences at most)

Briefly describe the datasets you chose to use. Find the original source (usually a publication), and read it to find out how the data was originally gathered by the authors. Summarize that here.

(E.g., if they are example datasets from R, usually, the R package will have a citation).

Looking at the likelihood curves, what, approximately, are the best estimates of parameters that you varied? Draw and label this point on the curves.

Approximately what confidence would you have in that parameter estimate? Again, draw your intuition onto the likelihood curves.

(the point of the previous two questions is not to get exact answers, but to think about likelihood curves, and reading such curves)

How would you expect the likelihood curve to change if you had a bigger dataset? If you had a smaller dataset? Draw your expectation by hand (assume that the bigger and smaller datasets support the same best parameter estimate).

What did you learn from writing your own R functions and drawing likelihood curves?

What surprised you the most?

What was the most challenging?

How did you overcome this challenge?