**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 #4 – Statistical model comparison**

For assignment 4, we will modify your likelihood functions/parameters to specify multiple models.

Then, we will statistically comparing how the models fit a dataset.

Start with the same two models that you chose for assignment #3.

* For a birth-death model, we will compare Yule versus Birth-Death (total: two models)
* For a discrete character model, convert the binary character model to a 3-or 4-state model, and compare several models (all rates equal, all rates different, and several in between – at least 4 models total)
* For a Brownian motion model, we will compare BM to delta, lambda, and kappa (see <https://lukejharmon.github.io/pcm/chapter6_beyondbm/>).
* For BiSSE, choose several submodels of BiSSE (e.g. parameters set to be equal, or to be 0) -- At least 4 models total.
* For other models, chose several submodels -- at least 4 models total.

**Tasks (35% for each of the two model groups):**

1. For each submodel, maximize the likelihood, using an ML optimizer available in R
   1. Post your modified functions and your optimization scripts to GitHub or Gist, and give me either (a) a public link, or (b) a screenshot of the private repository, and email me the code.
2. Calculate the Likelihood Ratio Test statistic and p-value for each pair of *nested* models.   
     
   See this page for general advice on model comparison, and example calculations in Excel (see “Files” at the bottom of the page: <http://phylo.wikidot.com/advice-on-statistical-model-comparison-in-biogeobears> )
3. For all models, calculate AIC, AICc, and BIC, and calculate the corresponding model weights.
4. Assemble a nicely-formatted table for each of your two model classes, like you might do for a publication. (These tables will also include log likelihoods, number of data, number of free parameters, and the parameter estimates, in addition to AIC etc.) Print these out and turn them in.

**Discussion questions (30% total):**

5. Interpret your results tables. For the datasets you used, did your models show any big differences in model fit? Can you make any interpretation of this result? Both scientific and statistical interpretations are valid – e.g., it might be that your dataset is too small to show big differences between models, it might be that some model shows a big improvement in model fit, suggesting that the data was produced by a particular process, etc.

6. Discuss the advantages and disadvantages of LRT, AIC, AICc, and BIC. Often, researchers will just choose one for use in a scientific study; which one would you choose here and why? (Feel free to look up advice in other publications, just cite the advice.)

7. What did you find challenging about this assignment, and how did you overcome it?

8. If you were to do a fully Bayesian version of comparison of these models, what additional decisions, algorithms, and statistics would be needed?