Appendix 1

Choosing the correct Akaike penalty term

The AICc is a corrected AIC for small sample size; it is supposed to better estimate Kullback-Leibler distance. However, there remains an unaddressed question in phylogenetics regarding what exactly sample size is. For comparative methods, it is taken as number of taxa, as when comparing Ornstein-Uhlenbeck models (Butler and King, 2004; O'Meara et al. 2006; Beaulieu et al. 2012) or discrete character models (Beaulieu et al. 2013). For evaluating molecular models, it is taken as number of sites (Posada and Buckley, 2004). These cannot both be true; especially for discrete models, where the underlying model is ultimately the same between DNA and other discrete traits.

Jhwueng et al (2014) investigated AIC and related measures of KL distance and found no consistent best option. They used a two stage simulation approach to estimate KL distance. This has since been implemented in the KLchecker package. We used this approach to examine a simple example where data were generated under an HKY model fit to the Laurasiatherian dataset of phangorn (Schliep, 2017) and then fit using HKY and JC models. These are fast enough to analyze to confirm our suspicion about both number of characters and number of taxa mattering for dataset size.

The RunKLchecker.R and ParseKL2.R scripts contain the code used for these analyses. In brief, parameters estimated from the Laurasiatherian dataset were used to simulate datasets on random trees made from TreeSim (Stadler, 2017). Number of sites used were 1, 2, 3, 4, 5, 10, 20, 50, 100, 200, 500, and 1000 and number of taxa were 5, 10, 25, 50, 100, and 250. For every simulation, the KL distance between the truth and the HKY and JC models was calculated, as well as the likelihood for each of the models. A global model was constructed:

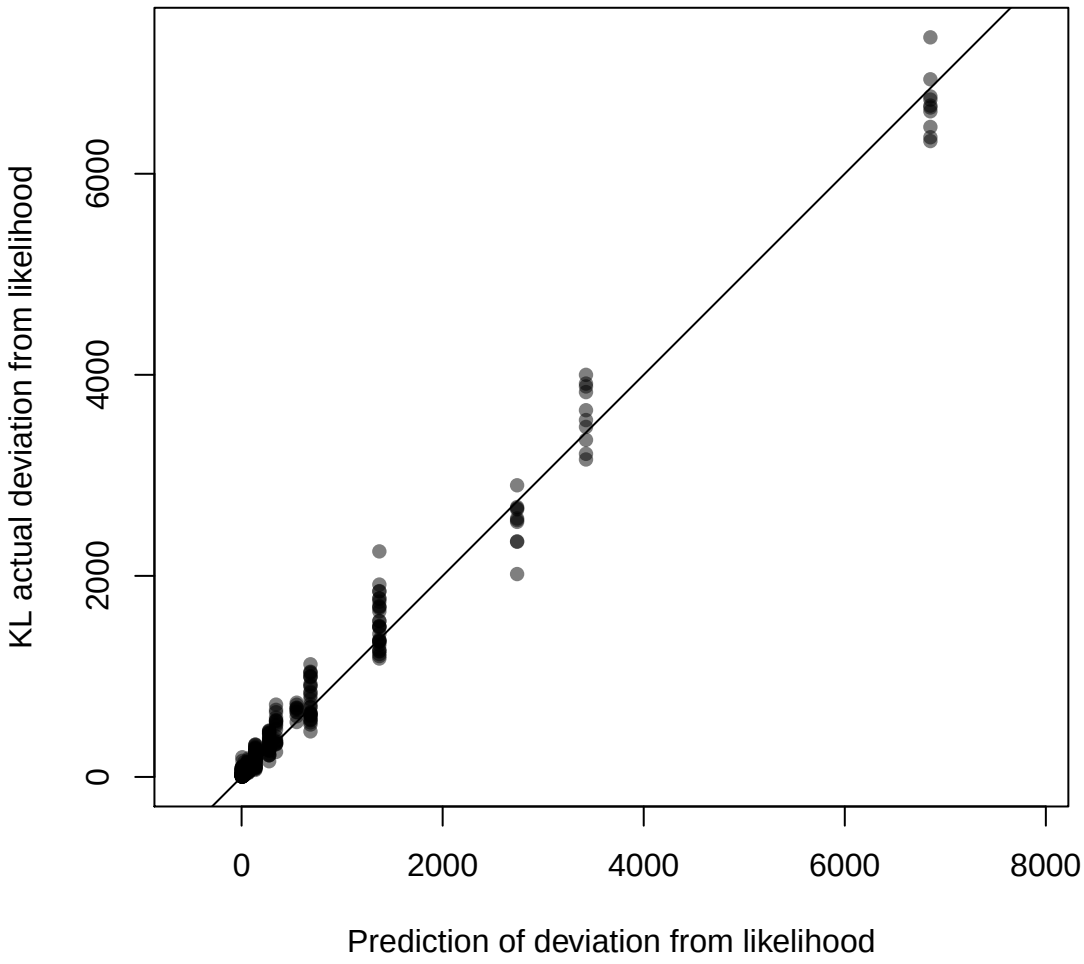
```
global.model <- lm(KLminusNeg2lnL ~ AICc_ntax_penalty + AICc_nchar_penalty
  + AICc_ncharTimesntax_penalty + nchar + ntax + logNtax + logNchar
  + LogNcharTimesNtax + NcharTimesNtax + NcharTimesLogNtax
  + LogNcharTimesLogNtax - 1, data=results.for.model)
```

AIC and relatives are typically of the form `-2 lnL + X`, where `X` can be various penalty terms. The above sums a wide range of penalty terms. The MuMIn package (Bartoń, 2016) is then used to dredge a series of simpler models from this, limiting each to using only one penalty term.

The best model (by far; next one had a delta AICc of 1722) was

```
##
## Call:
## lm(formula = KLminusNeg2lnL ~ 0 + NcharTimesNtax, data = results.for.model)
##
## Coefficients:
## NcharTimesNtax
##        0.02741
```

Plotting the predictions versus the estimated KL, with a 1:1 line.



References

Bartoń, Kamil. 2016. MuMIn: Multi-Model Inference. https://CRAN.R-project.org/package=MuMIn. Beaulieu, Jeremy M, Dwueng-Chwuan Jhwueng, Carl Boettiger, and Brian C O'Meara. 2012. "Modeling stabilizing selection: expanding the Ornstein-Uhlenbeck model of adaptive evolution." Evolution 66 (8) 2369–83. Beaulieu, Jeremy M., Brian C O'Meara, and Michael J Donoghue. 2013. "Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms." Systematic Biology 62(5): 725-737. Butler, Marguerite, and Aaron A. King. 2004. "Phylogenetic comparative analysis: A modeling approach for adaptive evolution." American Naturalist 164(6): 683-695. Jhwueng, Dwueng-Chwuan, Snehalata Huzurbazar, Brian C O'Meara, Liang Liu. 2014. "Investigating the performance of AIC in selecting phylogenetic models." Statistical Applications in Genetics and Molecular Biology 13(4): 459-475. O'Meara, Brian C, Cécile Ané, Michael J Sanderson, and Peter C. Wainwright. 2006. "Testing for different rates of continuous trait evolution using likelihood." Evolution 60(5): 922-933. Posada, David, and T.R. Buckley. 2004. "Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and Bayesian approaches over likelihood ratio tests." Systematic Biology 53(5): 793-808. Schliep, Klaus, Emmanuel Paradis, Leonardo de Oliveira Martins, Alastair Potts, and Tim W. White. 2017. Phangorn: Phylogenetic Analysis in R. https://CRAN.R-project.org/package=phangorn. Stadler, Tanja. 2017. TreeSim: Simulating Phylogenetic Trees. https://CRAN.R-project.org/package= TreeSim.