Title: One size to fit them all: On calculating dataset size for AICc

Abstract: For those using model selection approaches in phylogenetics, the small-size corrected Akaike Information Criterion (AICc) has become a standard approach. However, the field is unconsciously but deeply split over how to compute the size of a dataset: for tree inference the size is the number of sites, while for trait evolution the size is the number of taxa.

Body:

Choosing the least bad model is key for getting good parameter estimates (\_\_\_\_), whether we want to estimate a tree, rates of evolution, or ancestral states. While likelihood ratio tests were at one point commonly used, criteria based on information theory have largely replaced them due to their ability to compare non-nested models and, to a lesser degree, ease of using multiple models rather than fixing on a single one. Possibly the most commonly used criterion is the Akaike Information Criterion (AIC), which is based on just the number of free parameters and the likelihood. However, for small sample sizes, the Bayesian Information Criterion (also known as the Schwartz Information Criterion) as well as the sample size corrected Akaike Information Criterion (AICc) use the number of free parameters, likelihood and sample size.

However, what *is* the sample size? One of the most frequently used approaches in model selection in phylogenetics is ModelTest and its descendants (\_\_\_cite\_\_\_). (Posada and Buckley, 2004), in describing sample size for both AICc and BIC, mention that while they use sample size of total number of characters, effective sample size remains poorly understood. They explicitly mention that sample size depends on the number of taxa, though this remains unused in their software (which leads to smaller sample sizes being used, which favors simpler models). In contrast, in (Butler and King, 2004)’s software OUCH, for AICc and SIC sample size is the number of characters times the number of taxa. In Geiger \_\_\_\_, for univariate characters the AICc is based on number of taxa. Papers will often use number of characters with AICc to pick a nucleotide model, and then use number of taxa with AICc to pick a model for discrete traits of biological interest, even though these models are ultimately the same (O'Meara, 2012).

Intuitively, it seems that both number of taxa and number of sites should be relevant for sample size. When trying to estimate parameters of molecular evolution for the same model across all sites and taxa, for example, a dataset of 10,000 sites probably has more information than one of 100 sites when estimating the transition/transversion rate of an HKY (\_\_\_) model; a dataset of 400 taxa probably has more information than one of four taxa. However, it may not be a simple extension: imagine if three taxa are on long branches and the remaining 397 are a polytomy with zero terminal length: in that case, only 1 of the 397 taxa actually adds any data.

Butler MA, King AA, 2004. Phylogenetic comparative analysis: A modeling approach for adaptive evolution. American Naturalist 164:683-695.

O'Meara BC, 2012. Evolutionary Inferences from Phylogenies: A Review of Methods. In: Annual Review of Ecology, Evolution, and Systematics, Vol 43; 267-285.

Posada D, Buckley TR, 2004. Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. Systematic Biology 53:793-808.