Is the best model good enough? Assessing the absolute fit of phylogenetic models via posterior predictive sampling

Gerhard Jäger ***Tübingen University***

Bayesian phylogenetic inference has revolutionized historical linguistics over the past twenty years, adding a statistically sound, data-driven approach to the field's toolbox. It is still occasionally met with some skepticism though. One of the reasons for this is arguably the mind-boggling multitude of choices a researcher has to make when setting up an analysis. This involves the choice of likelihood models for (a) character evolution, (b) rate variation across sites, (c) rate variation across branches (aka "molecular clock"), (d) the mechanism generating trees, to mention just a few. Each of these choice comes with a collection of parameters for which prior distributions have to be chosen. This immense number of degrees of freedom conveys the impression that the choice of models and priors is highly subjective, making Bayesian inference less rigorous than one might wish in a data science context.

A reliable way to counter this line of criticism is to conduct a sensitivity analysis. If inference results are largely invariant across modeling choice, they can be considered reliable. However, given the huge computational effort required by a single Bayesian analysis of realistic data sets in phylogenetic linguistics, this approach is impractical.

Extant work mostly applies *model selection*, e.g., via Bayes Factor comparison, to justify modeling choices. This only tells us, however, which model *out of a predefined collection of models* explains the data best. If all models considered are off the mark, we might still end up with a poor model and thus with unreliable inference.

The technique of *posterior predictive sampling* (cf. Gelman et al. 2014, subsection 6.3) is a method to test whether a model is a plausible explanation for the data in absolute terms. In such an analysis, samples are simulated from the distribution $P(\tilde{\theta}, \tilde{y}|y_{\text{obs}}, M)$, where M is the model, y_{obs} the observed data, $\tilde{\theta}$ the posterior distribution of model parameters, and $\tilde{y}$ the posterior distribution of the quantities sampled in the observed data. If M provides a reasonable explanation for y_{obs}, we expect a sample of the distribution $(\tilde{y}, \tilde{\theta})$ to be similar to $(y_{\text{obs}}, \tilde{\theta})$—which contains observed data. A comparison of these distributions can tell us which aspects of M are correct and which are wrong. This can inform our decision how much to trust inferences based on M.

As a proof of concept, I conducted such a study using cognate class data from `ielex.mpi.nl`. A sample of 30 living languages were sampled at random and Bayesian phylogenetic inference was performed using a GTR model, Γ-distributed rates, a lognormal uncorrelated relaxed clock and a uniform distribution over ultrametric trees. The analysis was carried out using the software *RevBayes* (Höhna et al., 2016), which also implements posterior predictive sampling. For 800 samples from the posterior distribution, a character matrix was sampled, and the *Homoplasy Index* (HI; see `en.wikipedia.org/wiki/Cladogram#Retention_index`) both for the observed and the simulated character matrix with respect to the sampled phylogeny was computed. The HI for the observed data was between 0.606 and 0.610 (95% HPD), while it was between 0.681 and 0.725 for the simulated data. From this it can be concluded that the observed cognate class data show a much lower degree of homoplasy than predicted by the continuous-time Markov chain (CTMC) model used to model character evolution. This does not necessarily invalidate all inferences based on cognate-class data and CTMC, but it points to fruitful directions of possible model improvement.

References

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. CRC Press, Boca Raton.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, 65(4):726–736.