Today's agenda:

- Absolute goodness of fit/Model adequacy
- Hypothesis testing
- Tying up loose ends
- Exam review

Syst. Biol. 63(3):334–348, 2014

DOI:10.1093/sysbio/syu002
Advance Access publication January 11, 2014

Detection of Implausible Phylogenetic Inferences Using Posterior Predictive Assessment of Model Fit

JEREMY M. BROWN*

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
**Correspondence to be sent to: E-mail: jembrown@lsu.edu.*

Received 4 April 2013; reviews returned 10 June 2013; accepted 30 December 2013
Associate Editor: Mark Holder

Abstract.—Systematic phylogenetic error caused by the simplifying assumptions made in models of molecular evolution may be impossible to avoid entirely when attempting to model evolution across massive, diverse data sets. However, not all deficiencies of inference models result in unreliable phylogenetic estimates. The field of phylogenetics lacks a direct method to identify cases where model specification adversely affects inferences. Posterior predictive simulation is a flexible and intuitive approach for assessing goodness-of-fit of the assumed model and priors in a Bayesian phylogenetic analysis. Here, I propose new test statistics for use in posterior predictive assessment of model fit. These test statistics compare phylogenetic inferences from posterior predictive data sets to inferences from the original data. A simulation study demonstrates the utility of these new statistics. The new tests reject the plausibility of inferred tree lengths or topologies more often when data/model combinations produce biased inferences. I also apply this approach to exemplar empirical data sets, highlighting the value of the novel assessments. [Bayesian; Markov chain Monte Carlo; model fit; phylogenetic; posterior predictive distribution; sequence evolution; simulation.]

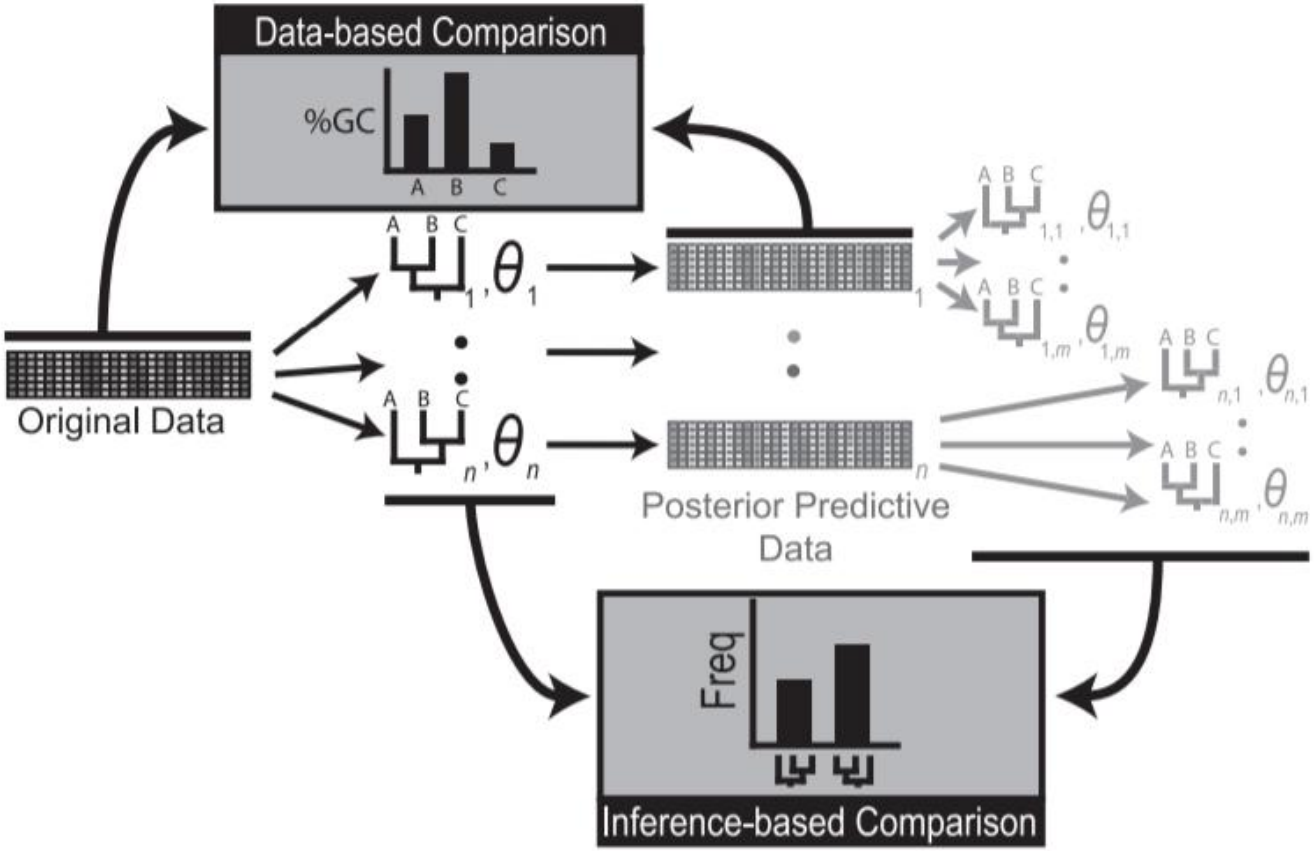
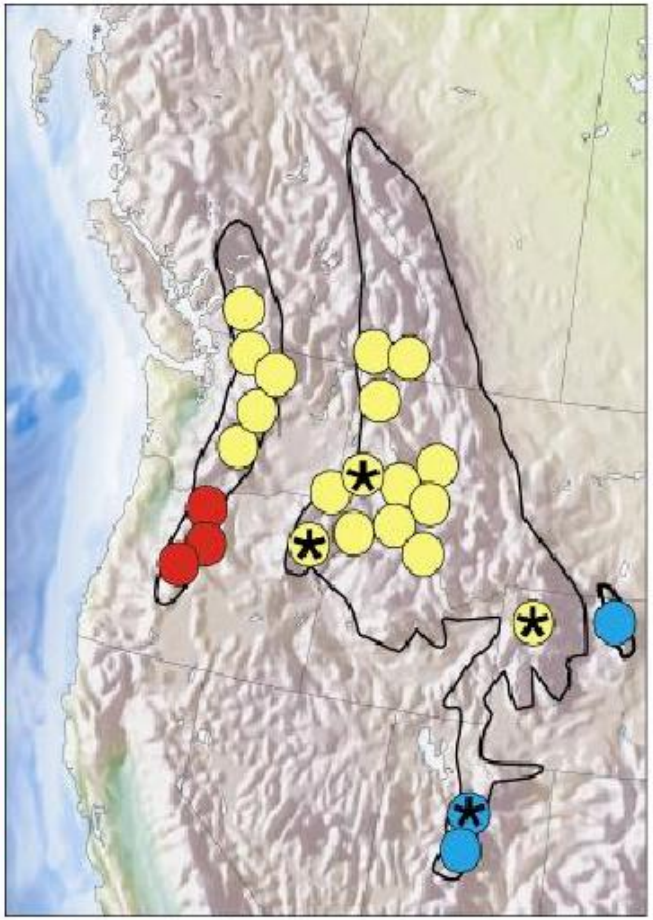


FIGURE 1. A schematic representation of data- versus inference-based approaches to assessing model plausibility with posterior predictive simulation. Most statistics proposed for testing model plausibility compare data-based characteristics of the original data set to the posterior predictive data sets (e.g., variation in GC-content across species). This study proposes and implements test statistics that compare the inferences resulting from different data sets (e.g., the distribution of posterior probability across topologies). Multiple sequence alignments (MSAs) are represented as shaded matrices and arrows originating from MSAs point to the MCMC samples of tree topologies and scalar model parameters (θ) resulting from Bayesian analysis of that MSA. Subscripts of MCMC samples taken during analysis of the original data index the samples (1, ..., n). Subscripts for each posterior predictive data set indicate which MCMC sample was used in its simulation. Subscripts for MCMC samples resulting from analysis of a posterior predictive data set first indicate the posterior predictive data set that was analyzed and next index the MCMC samples from analysis of that particular data set (1, …, m). Two other approaches to assessing model fit that are not explicitly outlined in this schematic involve comparing (i) the posterior distribution derived from the empirical data to prior expectations about the model or (ii) the posterior predictive data sets to prior expectations about the data (see the text for more details).

Example: Constraining trees

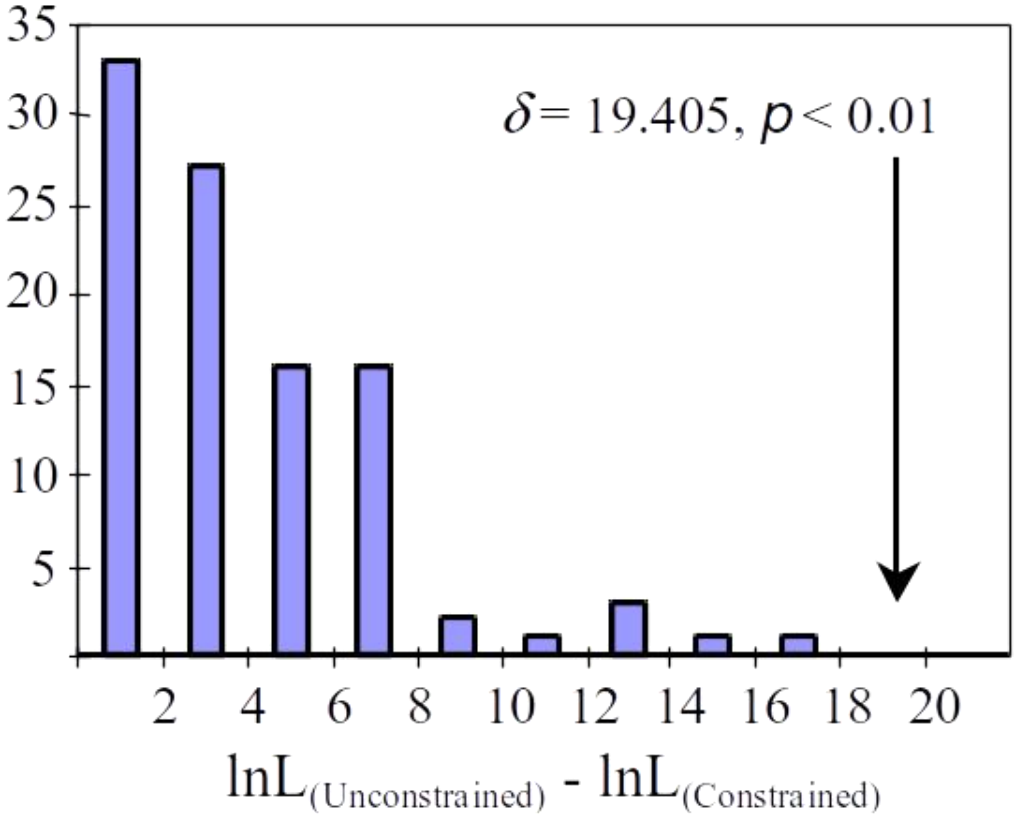
2.8% divergence

ML Tree:
GTR+I
n = 85



These data support a northern dispersal hypothesis, although from the Rockies to the Cascades.

Run parametric bootstrap



Bayesian version...

$P(\theta \mid \text{Data})$ = Posterior distribution

Hypothesis testing often just a simple matter of checking the posterior distribution!

Can also do Bayes Factors

Estimating Bayes Factors

Denominator of Bayes formula (marginal likelihoods of the model) difficult to estimate, but possible with new methods

RJMCMC

Path-sampling methods (e.g. Stepping-stone or thermodynamic integration)

- Computationally intensive (like running an MCMC on your data, but at least 50 times)

Summary

Model selection using relative goodness of fit (hLRT, AIC, BIC, DT) common, useful

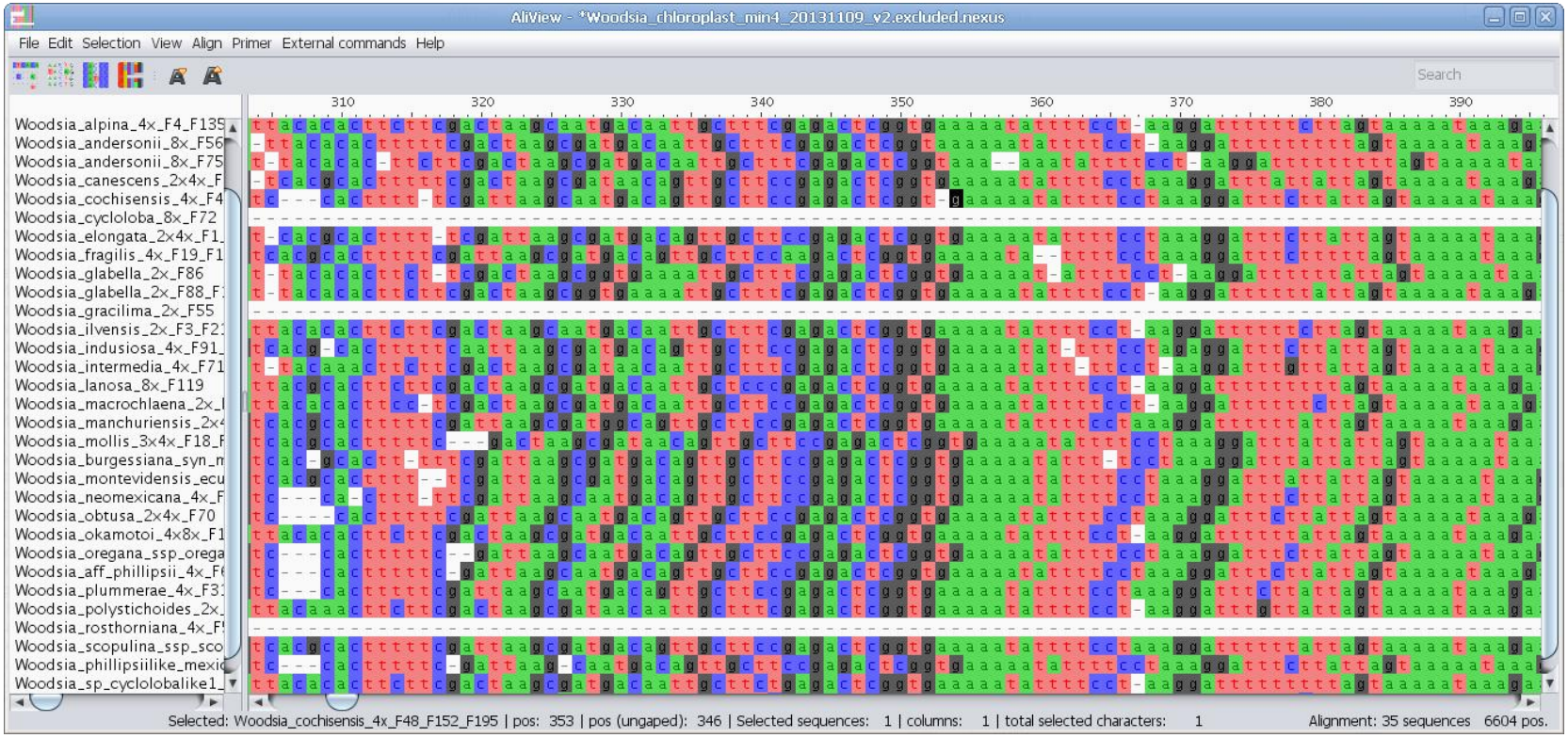
Often, we want absolute goodness of fit, which can be evaluated with simulations (e.g. parametric bootstrap, posterior predictive simulation), cross-validation etc.

Bayesian vs. Likelihood approaches can be very similar, but ask different questions, leading to differences in response to model complexity.

All models are wrong, targeting our inferences with hypothesis testing is useful.

Loose ends

Rewind: Aligning DNA/Protein sequences



Rewind: Aligning DNA/Protein sequences

Aligning can be challenging

all-individual dataset. The 16S fragments (563 bp) were aligned using Clustal W algorithm (THOMPSON et al. 1994) as is implemented in BioEdit (HALL 1999). Alignments were checked by eye and low quality ends trimmed. Ambiguously aligned region/gaps were ignored for the subsequent analysis. We used a network approach (POSADA & CRAN-

for all protein-coding genes. Alignments were checked by eye using MEGA 6.06 (Tamura *et al.*, 2013) and SEAVIEW 4 (Gouy *et al.*, 2010). Obvious errors or frame shifts were corrected manually. We subsequently added sequences from ingroup species

These regions generally represent areas of lower sequencing coverage, and thus more sequencing error. These regions are also more difficult to align due to variation in sequence lengths across taxa. In the past, alignments were checked by eye to remove any potentially mis-aligned regions that could include non-homologous characters. However, this is not possible with a genome-scale dataset, and thus we rely on computational methods to scan alignments for us. Potential

Progressive alignment

First align all pairwise sequences according to some cost function

Rank them by their similarity to generate a guide tree

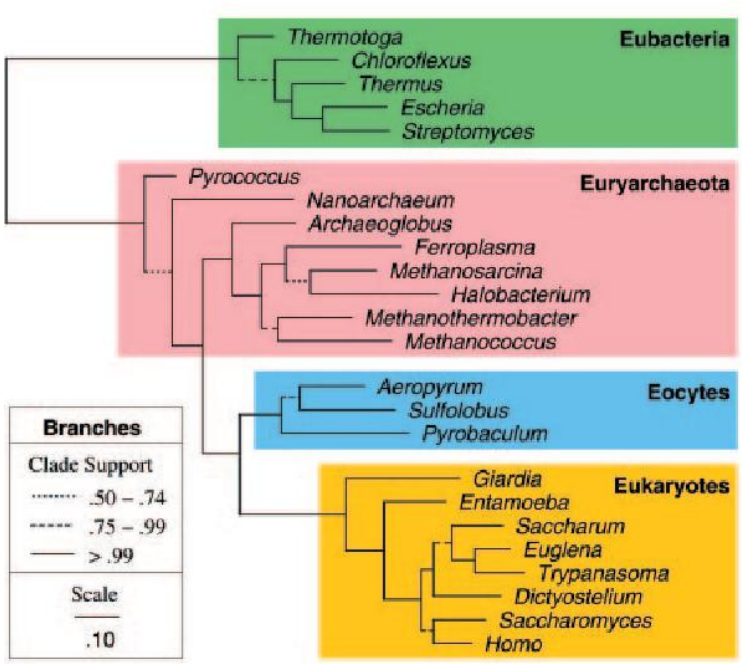
Start with the most similar pair of sequences and begin adding sequences to the alignment (“multiple alignment”) until all sequences are aligned

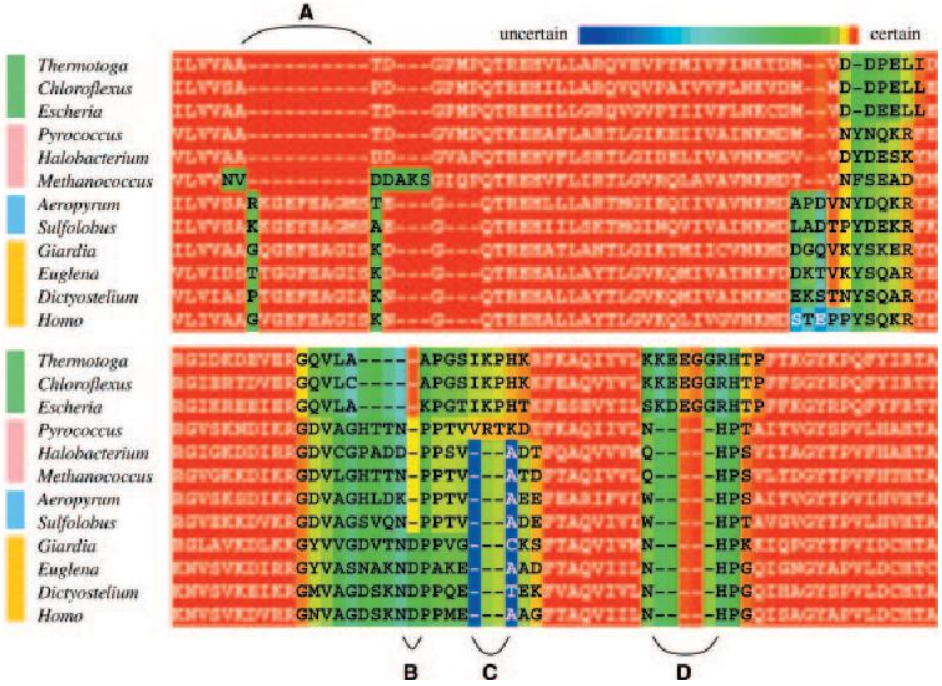
Repeat with distance matrix with evolutionary model

We typically align *before* phylogenetic analysis

It would be be better to infer the tree and alignment simultaneously

BALi-Phy (Redelings & Suchard 2005) - joint estimation of tree and alignment





POY (Wheeler) does the same for Maximum Parsimony

Other loose ends

Data partitioning

- partition finder (Rob Lanfear)
- autoparts (Brian Moore)
- “autopartition” command in PAUP*

Consensus trees

- Majority-rule, strict consensus etc.
- Suitability for analyses?