

# Model selection, Model adequacy & Hypothesis testing

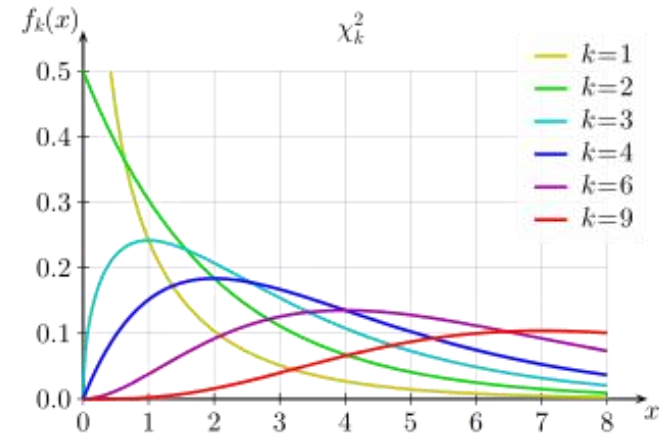
# Likelihood ratio tests

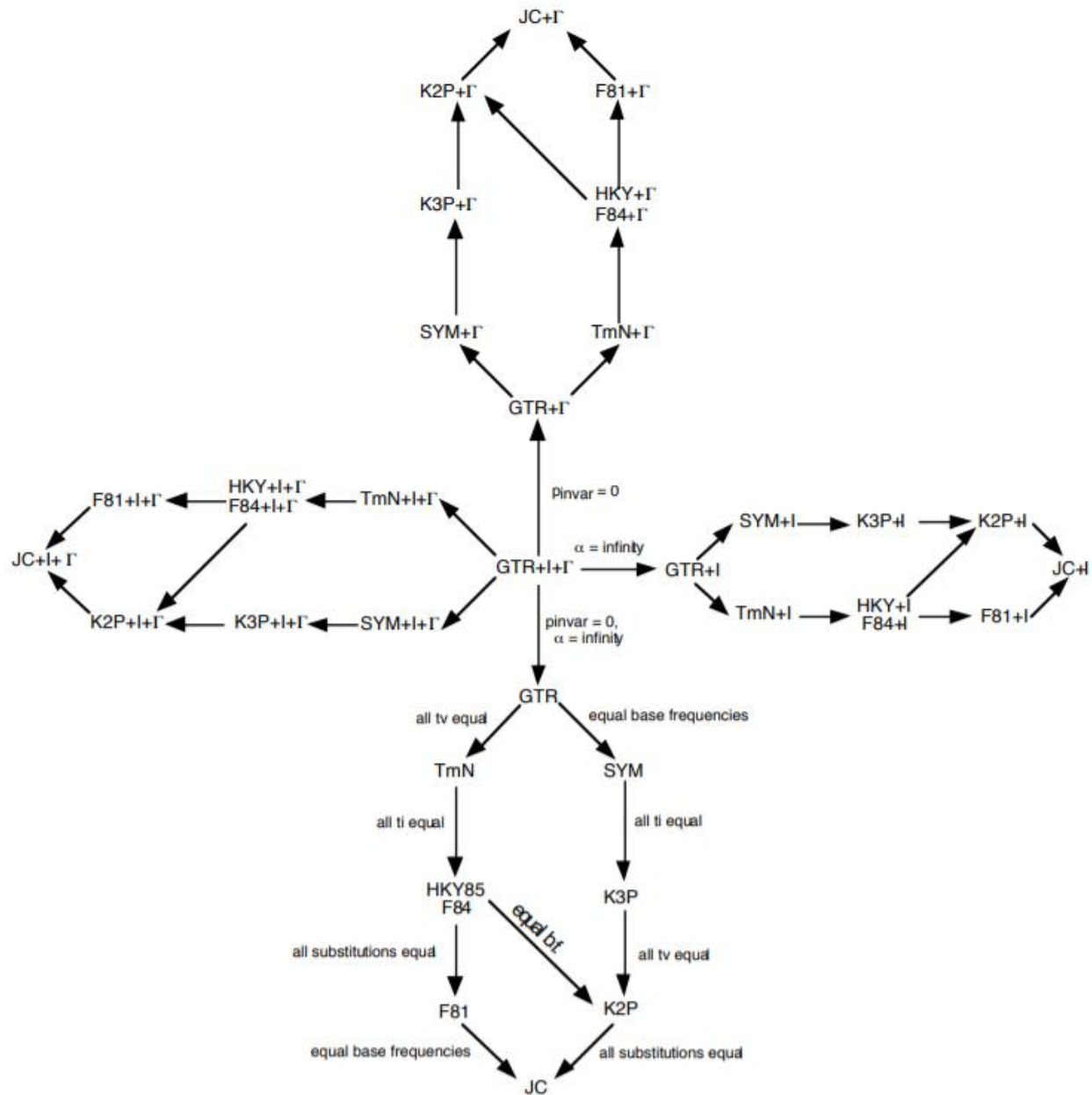
Models must be nested (hierarchical)

As sample size approaches infinity:

$-2 \ln \lambda \sim \text{Chi-Square}(p_2 - p_1)$

$$\log \lambda = \log \frac{L(H_1)}{L(H_2)}$$





# Information theory

Akaike Information Criterion (Approximates Kullback-Leibler Divergence asymptotic result, small sample size correction available.)

$$\text{AIC} = 2k - 2 \ln(\hat{L})$$

Bayesian Information Criterion

$$\text{BIC} = \ln(n)k - 2 \ln(\hat{L}).$$

Relative fit for a given dataset

Others: Deviance Information Criterion, Focused Information Criterion etc.

# Information theory

Minimize loss of information by maximizing relative goodness of fit while penalizing model complexity

BIC penalizes more heavily, AIC will usually find more heavily parameterized models

Jennifer Ripplinger addressed this in the first chapter of her dissertation by downloading 250 phylogenetic data sets from TreeBASE and selecting model using hLRT, AIC, BIC and DT.

All four picked same model in 51 data sets.

Two models were selected in 123 data sets.

Three were selected in 70 data sets.

All picked different models in 6 data sets.

Approach	avg # param
----------	-------------

hLTR*	$6.9 \pm 2.2$
AIC	$8.4 \pm 1.8$
BIC	$6.7 \pm 1.7$
DT	$6.7 \pm 1.4$

What is sample size in  
phylogenetic datasets?

# A word about Bayesian analyses and Bayes Factors...

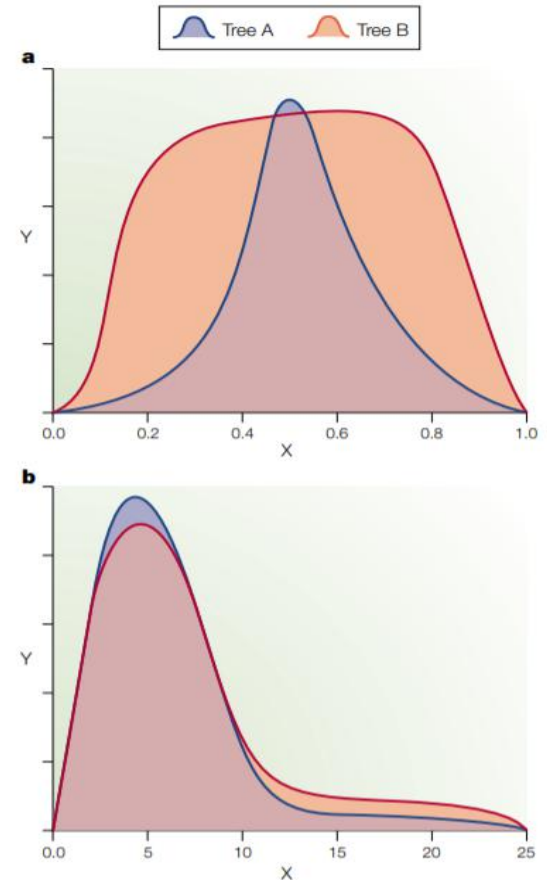


Figure 1 | **Contrast between marginal and joint estimation.** Panels **a** and **b** depict the likelihood profile for two trees versus a hypothetical parameter  $x$ . The  $x$  axis represents some nuisance parameter (for example, the ratio of the rate of transitions to the rate of transversions). The  $y$  axis represents the likelihood in the case of ML, or the posterior-probability density in a Bayesian approach. The area under the likelihood curve for tree A is shown in light blue, the area for tree B is shown in orange. Mauve regions are under the curve for both trees. In both cases, jointly estimating  $x$  and the tree favours tree A (that is, the highest peak is blue in both cases), but marginalizing over  $x$  favours tree B (that is, the orange area is greater than the blue area).



# Reversible-Jump MCMC

## Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo

*John P. Huelsenbeck,\* Bret Larget,<sup>†‡</sup> and Michael E. Alfaro\**

\*Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego;

<sup>†</sup>Department of Statistics, University of Wisconsin; <sup>‡</sup>Department of Botany, University of Wisconsin

A common problem in molecular phylogenetics is choosing a model of DNA substitution that does a good job of explaining the DNA sequence alignment without introducing superfluous parameters. A number of methods have been used to choose among a small set of candidate substitution models, such as the likelihood ratio test, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), and Bayes factors. Current implementations of any of these criteria suffer from the limitation that only a small set of models are examined, or that the test does not allow easy comparison of non-nested models. In this article, we expand the pool of candidate substitution models to include all possible time-reversible models. This set includes seven models that have already been described. We show how Bayes factors can be calculated for these models using reversible jump Markov chain Monte Carlo, and apply the method to 16 DNA sequence alignments. For each data set, we compare the model with the best Bayes factor to the best models chosen using AIC and BIC. We find that the best model under any of these criteria is not necessarily the most complicated one; models with an intermediate number of substitution types typically do best. Moreover, almost all of the models that are chosen as best do not constrain a transition rate to be the same as a transversion rate, suggesting that it is the transition/transversion rate bias that plays the largest role in determining which models are selected. Importantly, the reversible jump Markov chain Monte Carlo algorithm described here allows estimation of phylogeny (and other phylogenetic model parameters) to be performed while accounting for uncertainty in the model of DNA substitution.

### Introduction

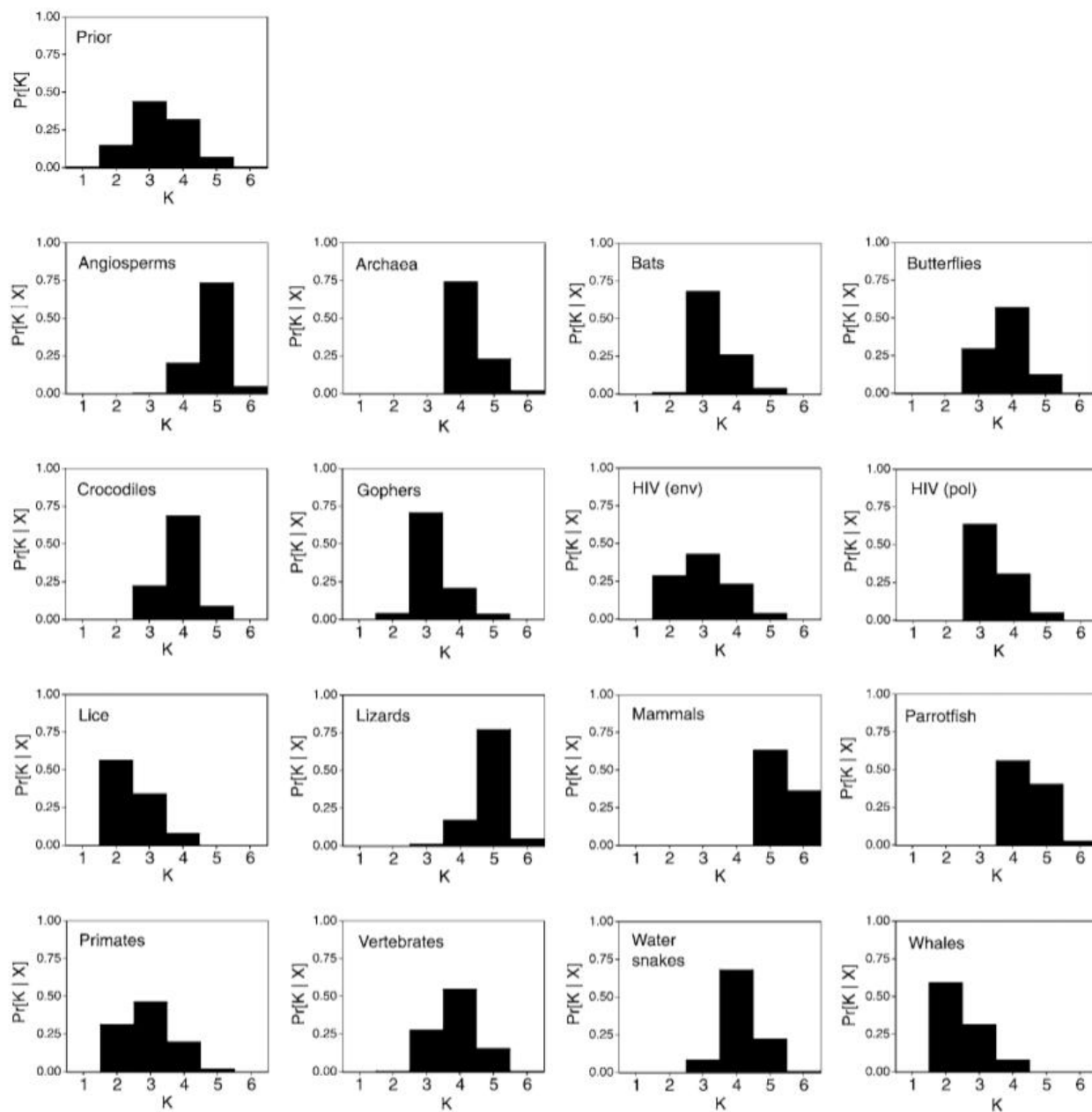


FIG. 1.—The posterior probability of  $K$ , the number of substitution types, for each DNA sequence alignment. The first figure shows the prior probability distribution of  $K$ , calculated as the number of models with  $K$  substitution parameters divided by the total number of substitution models (203).

# Relative vs. absolute goodness of fit

Best model may still be bad!

We want a model that *adequately describes the data generating process*.

Frequentist solution: Parametric bootstrap

Bayesian solution: Posterior predictive simulation

# Basic idea:

Estimate parameters (either ML estimate or posterior distribution)

Simulate data using parameters

Calculate a test statistic seeing if the model predicts “similar” datasets to your original

Compare to summary statistic measured on empirical data

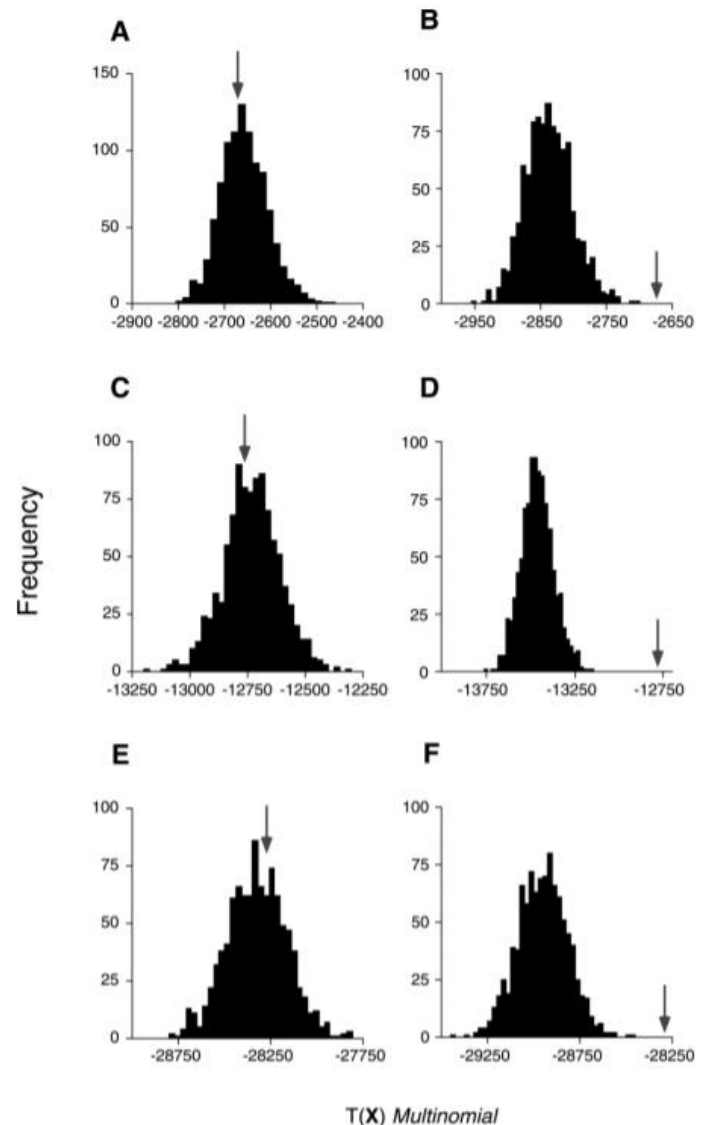


FIG. 2.—Illustration of the method comparing GTR versus JC69. Data sets of  $c = 500$  (A, B),  $c = 2,000$  (C, D), and  $c = 4,000$  (E, F) sites were simulated under the GTR model. Predictive distributions were simulated under the GTR (A, C, E) and JC69 (B, D, F) models. Arrows indicate the values for the realized statistic from the original data. In all cases the GTR model, as expected, produced an adequate fit to the data, whereas the JC69 did not ( $P_T = 0.000$ ).

**What test statistic should be used?**

## 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.]



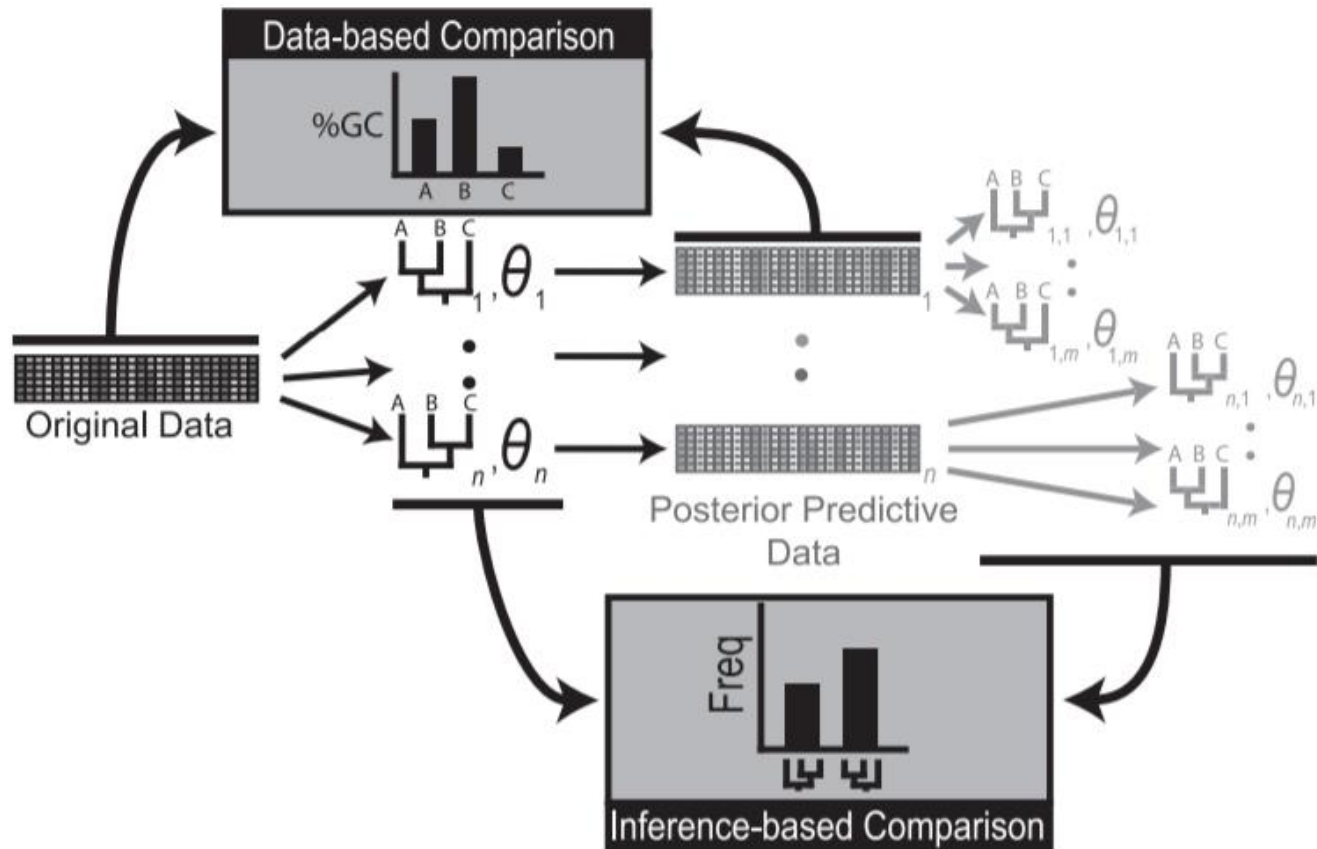


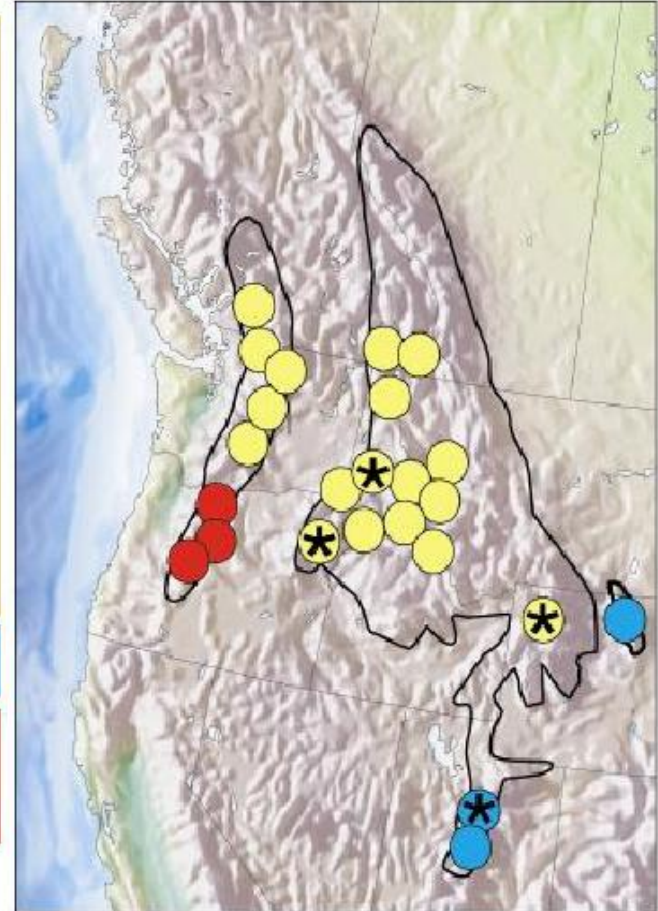
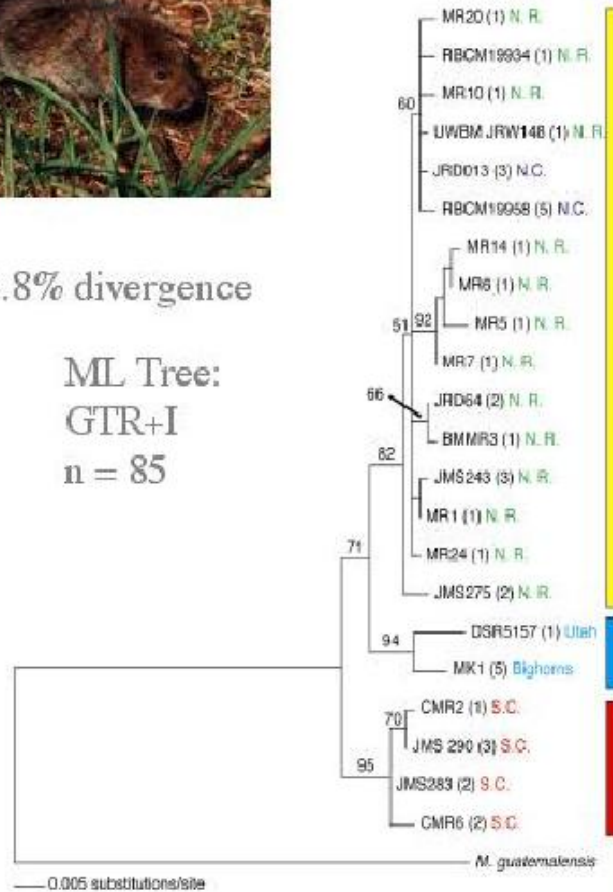
FIGURE 1. A schematic representation of data- versus inference-based approaches to assessing model plausibility with posterior predictive simulation. Most statistics proposed for testing model plausibility compare data-based characteristics of the original data set to the posterior predictive data sets (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 ( $\theta$ ) 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:



2.8% divergence

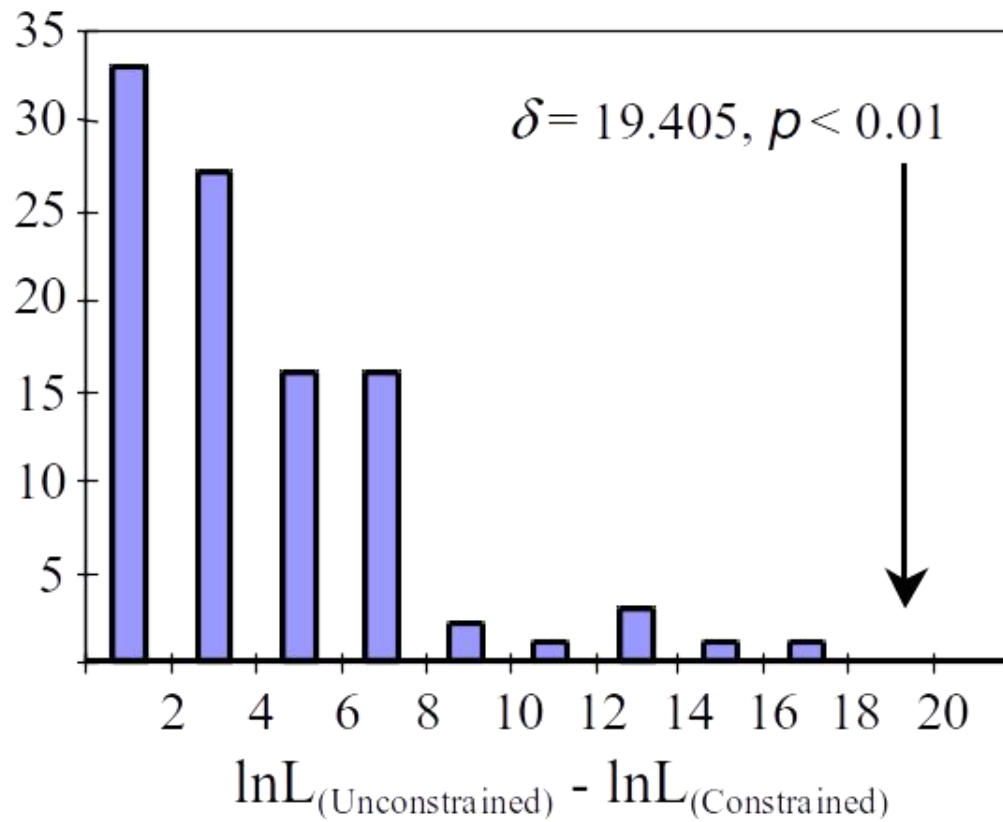
ML Tree:  
GTR+I  
n = 85



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

# 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.