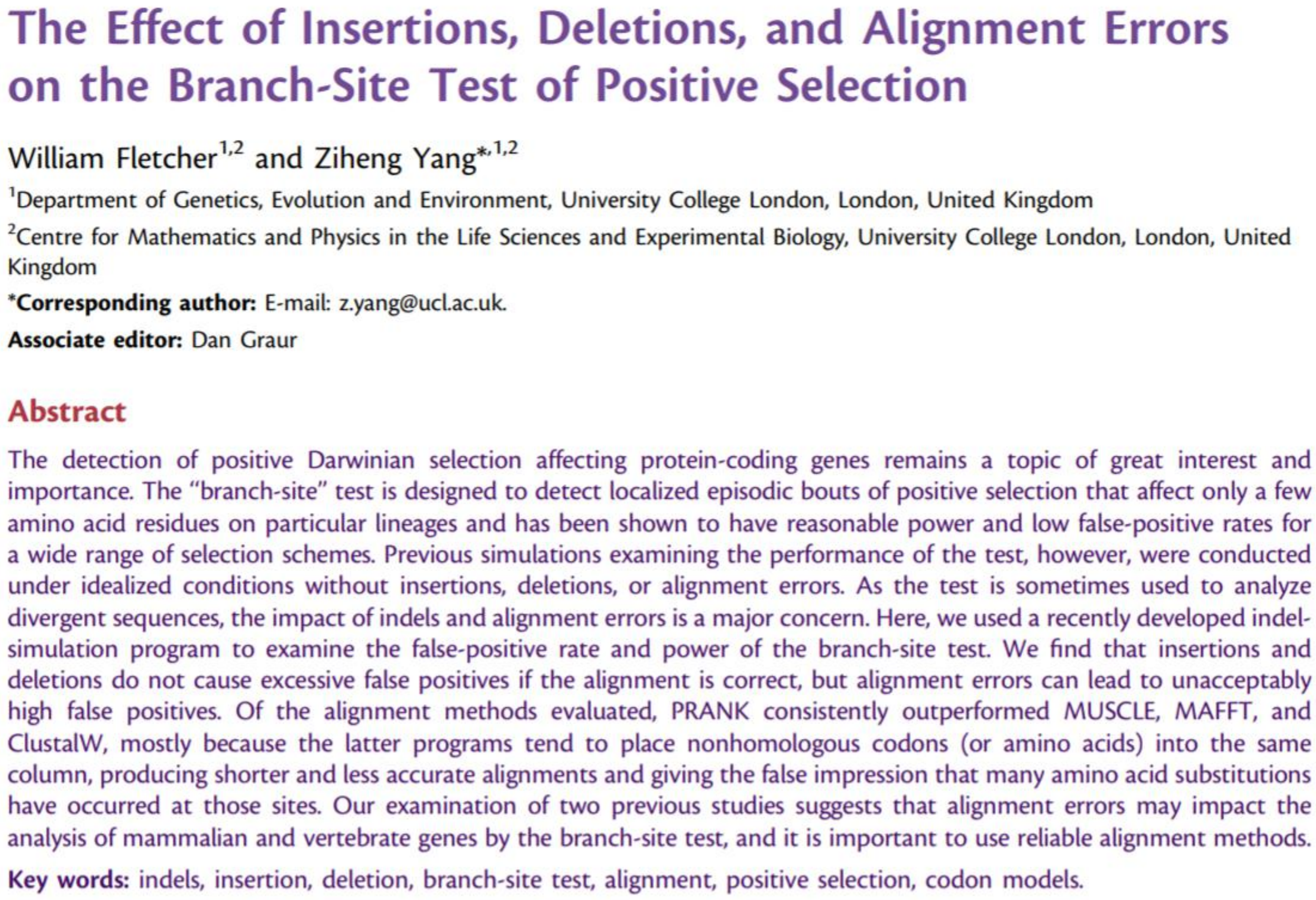
Alignments

Another (often) hard problem!

```
1    ACCGAATATTAGGCTC
2    ACATAGTGGGATC
3    AAAGGCATTAGGATC
4    GAAGGCATTAGCATC
5    CACGAAGGCATTGGGCTC
```

```
1    ACCGAAT--ATTAGGCTC
2    AC---AT--AGTGGGATC
3    AA---AGGCATTAGGATC
4    GA---AGGCATTAGCATC
5    CACGAAGGCATTGGGCTC
```

What if you get it wrong?



The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection

William Fletcher[1,2] and Ziheng Yang*,[1,2]

[1]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

[2]Centre for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London, United Kingdom

***Corresponding author:** E-mail: z.yang@ucl.ac.uk.

Associate editor: Dan Graur

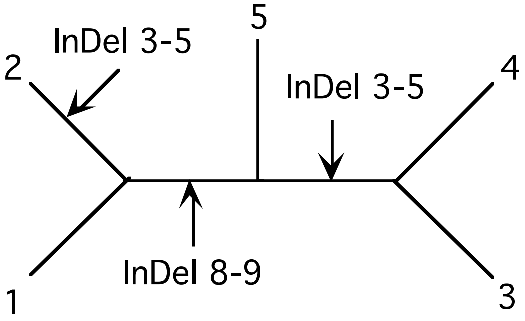
Abstract

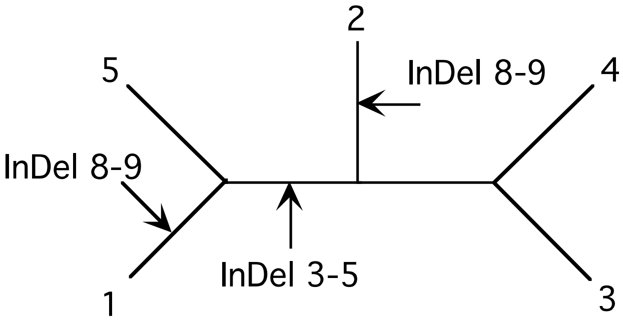
The detection of positive Darwinian selection affecting protein-coding genes remains a topic of great interest and importance. The "branch-site" test is designed to detect localized episodic bouts of positive selection that affect only a few amino acid residues on particular lineages and has been shown to have reasonable power and low false-positive rates for a wide range of selection schemes. Previous simulations examining the performance of the test, however, were conducted under idealized conditions without insertions, deletions, or alignment errors. As the test is sometimes used to analyze divergent sequences, the impact of indels and alignment errors is a major concern. Here, we used a recently developed indel-simulation program to examine the false-positive rate and power of the branch-site test. We find that insertions and deletions do not cause excessive false positives if the alignment is correct, but alignment errors can lead to unacceptably high false positives. Of the alignment methods evaluated, PRANK consistently outperformed MUSCLE, MAFFT, and ClustalW, mostly because the latter programs tend to place nonhomologous codons (or amino acids) into the same column, producing shorter and less accurate alignments and giving the false impression that many amino acid substitutions have occurred at those sites. Our examination of two previous studies suggests that alignment errors may impact the analysis of mammalian and vertebrate genes by the branch-site test, and it is important to use reliable alignment methods.

Key words: indels, insertion, deletion, branch-site test, alignment, positive selection, codon models.

```
1      ACCGAAT--ATTAGGCTC
2      AC---AT--AGTGGGATC
3      AA---AGGCATTAGGATC
4      GA---AGGCATTAGCATC
5      CACGAAGGCATTGGGCTC
```

CIRCULAR LOGIC WORKS BECAUSE



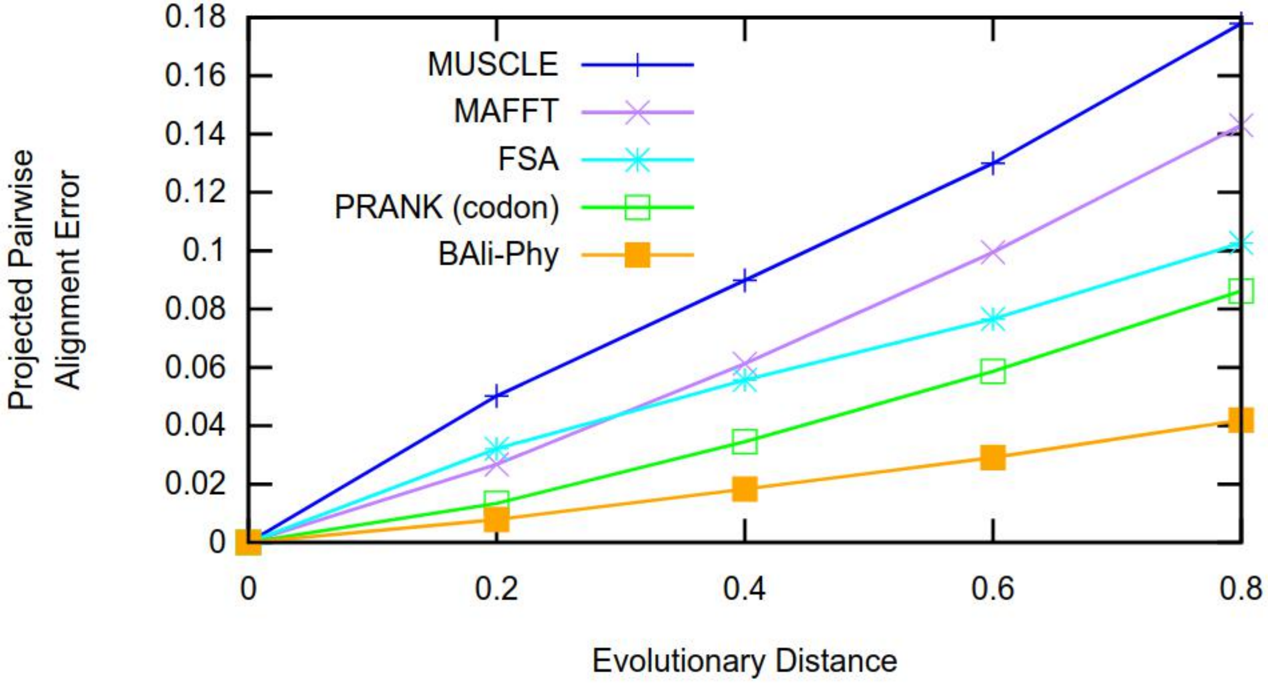


Knowing the tree will help you figure out the most likely alignment

Best approach: Estimate alignment & phylogeny simultaneously

Hard problem x Hard problem = :'-(

BAli---Phy



Measuring support: Bootstrap, jacknife, posterior probability

Bootstrap: Sample sites in your alignment with replacement, reanalyze each time

Jacknife: Randomly delete sites in your alignment, reanalyze

Posterior probability: Falls out naturally from a Bayesian analysis, probability of a split occuring given the data

Model selection,
Model adequacy &
Hypothesis testing

Model selection & hypothesis testing:

- Of my set of models, which one is best?

- Can I reject one hypothesis in favor of another?

Model adequacy:

- Is this model a good description of the data

Model selection

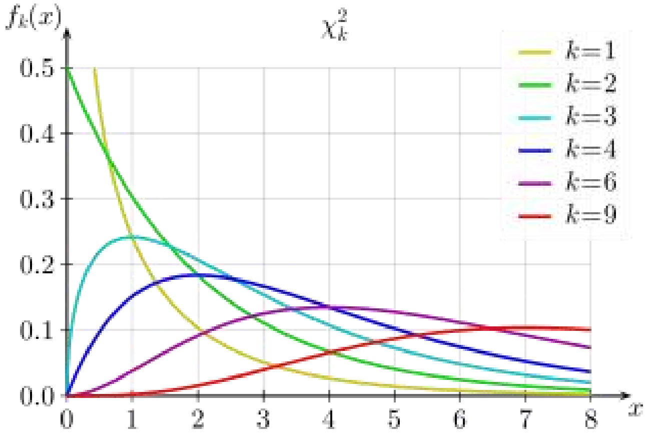
Likelihood ratio tests

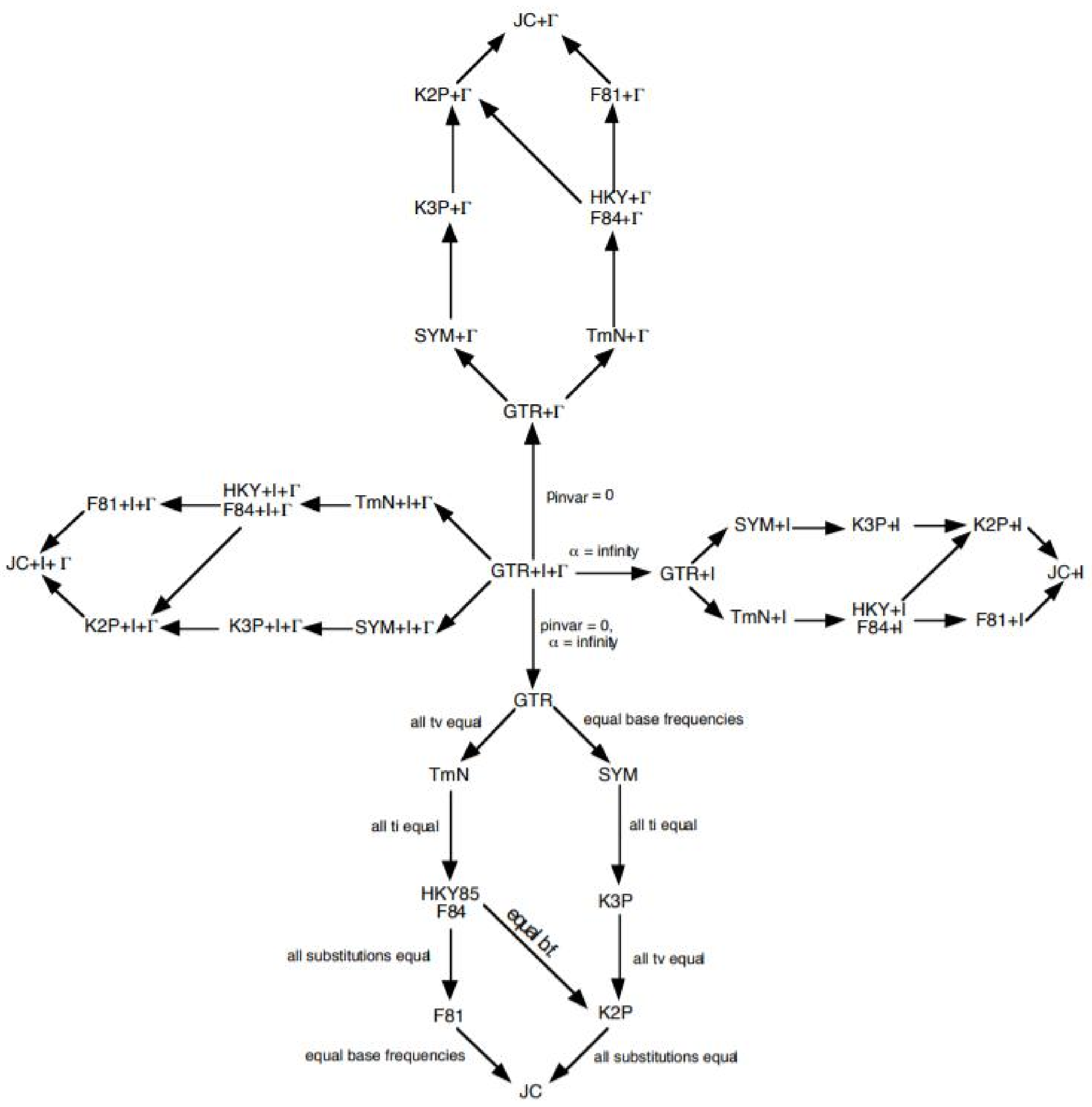
Models must be nested (hierarchical)

As sample size approaches infinity:

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

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




JC+Γ
K2P+Γ
F81+Γ
K3P+Γ
HKY+Γ
F84+Γ
SYM+Γ
TmN+Γ
GTR+Γ
pinvar = 0
F81+I+Γ
HKY+I+Γ
F84+I+Γ
TmN+I+Γ
JC+I+ Γ
GTR+I+Γ
α = infinity
GTR+I
SYM+I
K3P+I
K2P+I
JC+I
TmN+I
HKY+I
F84+I
F81+I
K2P+I+Γ
K3P+I+Γ
SYM+I+Γ
pinvar = 0,
α = infinity
GTR
all tv equal
equal base frequencies
TmN
SYM
all ti equal
all ti equal
HKY85
F84
equal bf
K3P
all substitutions equal
all tv equal
F81
K2P
equal base frequencies
all substitutions equal
JC

Information theory

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

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

Bayesian Information Criterion

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

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 ± 2.2
AIC	8.4 ± 1.8
BIC	6.7 ± 1.7
DT	6.7 ± 1.4

What is sample size in phylogenetic datasets?

ML hypothesis testing

Hypothesis: Are archaea monophyletic?

H1: No constraints on topology
H2: Constrain archaea to be monophyletic

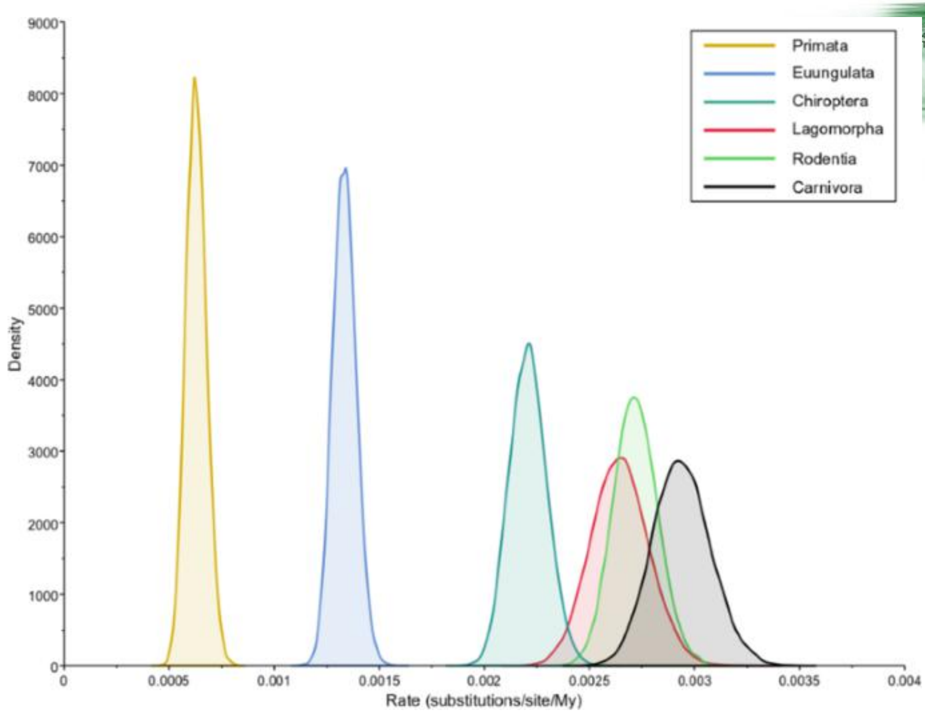
Compare with a model selection (AIC, BIC, hLRT)

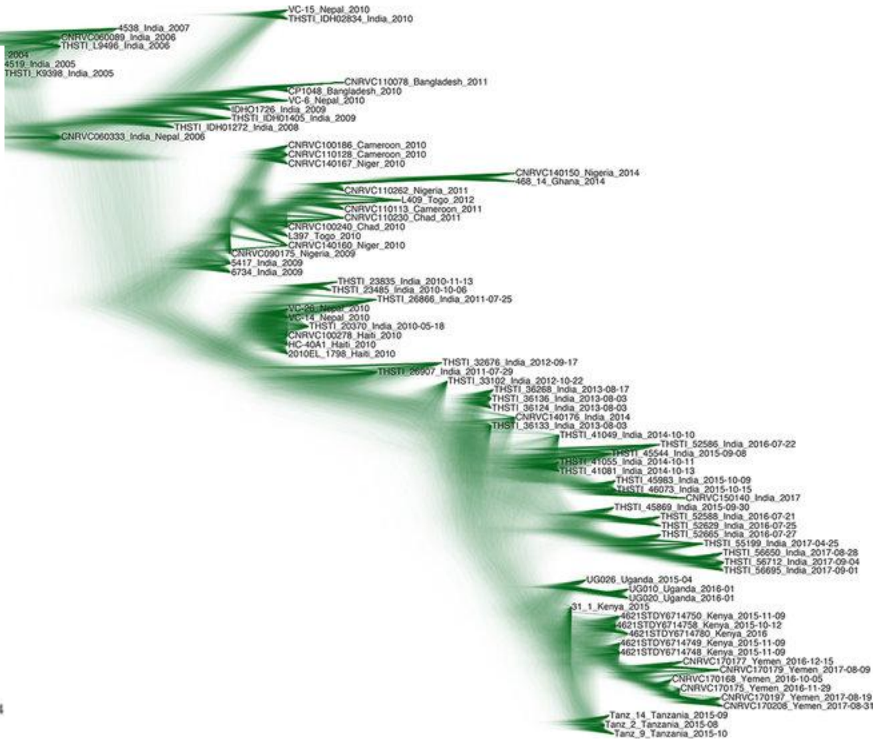
Reward ↑ likelihood, penalize ↑ # of parameters

Bayesian hypothesis testing

Often stupidly easy! Just check your posterior distribution.

Posterior probability is the hypothesis test.




VC-15_Nepal_2010
THSTI_IDH02834_India_2010
4538_India_2007
CNRVC060089_India_2006
THSTI_L9496_India_2006
2004
4519_India_2005
THSTI_K9398_India_2005
CNRVC110078_Bangladesh_2011
CP1048_Bangladesh_2010
VC-6_Nepal_2010
IDHO1726_India_2009
THSTI_IDH01405_India_2009
THSTI_IDH01272_India_2008
CNRVC060333_India_Nepal_2006
CNRVC100186_Cameroon_2010
CNRVC110128_Cameroon_2010
CNRVC140167_Niger_2010
CNRVC140150_Nigeria_2014
468_14_Ghana_2014
CNRVC110262_Nigeria_2011
L409_Togo_2012
CNRVC110113_Cameroon_2011
CNRVC110230_Chad_2011
CNRVC100240_Chad_2010
L397_Togo_2010
CNRVC140160_Niger_2010
CNRVC090175_Nigeria_2009
5417_India_2009
6734_India_2009
THSTI_23835_India_2010-11-13
THSTI_23485_India_2010-10-06
THSTI_26866_India_2011-07-25
VC-26_Nepal_2010
VC-14_Nepal_2010
THSTI_20370_India_2010-05-18
CNRVC100278_Haiti_2010
HC-40A1_Haiti_2010
2010EL_1798_Haiti_2010
THSTI_32676_India_2012-09-17
THSTI_26907_India_2011-07-29
THSTI_33102_India_2012-10-22
THSTI_36268_India_2013-08-17
THSTI_36136_India_2013-08-03
THSTI_36124_India_2013-08-03
CNRVC140176_India_2014
THSTI_36133_India_2013-08-03
THSTI_41049_India_2014-10-10
THSTI_52586_India_2016-07-22
THSTI_45544_India_2015-09-08
THSTI_41055_India_2014-10-11
THSTI_41081_India_2014-10-13
THSTI_45983_India_2015-10-09
THSTI_46073_India_2015-10-15
CNRVC150140_India_2017
THSTI_45869_India_2015-09-30
THSTI_52588_India_2016-07-21
THSTI_52629_India_2016-07-25
THSTI_52665_India_2016-07-27
THSTI_55199_India_2017-04-25
THSTI_56650_India_2017-08-28
THSTI_56712_India_2017-09-04
THSTI_56695_India_2017-09-01
UG026_Uganda_2015-04
UG010_Uganda_2016-01
UG020_Uganda_2016-01
31_1_Kenya_2015
4621STDY6714750_Kenya_2015-11-09
4621STDY6714758_Kenya_2015-10-12
4621STDY6714780_Kenya_2016
4621STDY6714749_Kenya_2015-11-09
4621STDY6714748_Kenya_2015-11-09
CNRVC170177_Yemen_2016-12-15
CNRVC170179_Yemen_2017-08-09
CNRVC170168_Yemen_2016-10-05
CNRVC170175_Yemen_2016-11-29
CNRVC170197_Yemen_2017-08-19
CNRVC170208_Yemen_2017-08-31
Tanz_14_Tanzania_2015-09
Tanz_2_Tanzania_2015-08
Tanz_9_Tanzania_2015-10

Marginal vs. joint estimation

Example: Ancestral state reconstruction

Joint reconstruction: Set of internal nodes that together have the single highest likelihood.

Marginal reconstruction: For a focal node, *marginalize* over all possible reconstructions and find the mostly likely state of each node individually.

-> *Nuisance parameters*

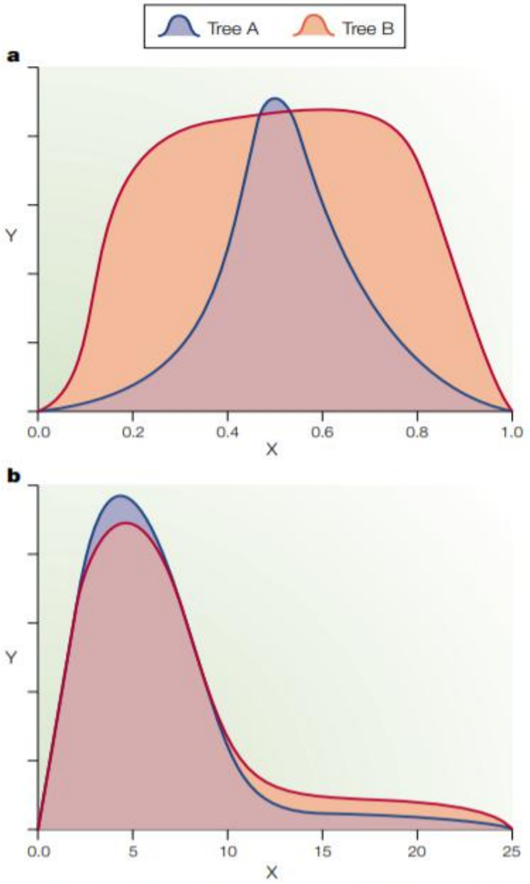


Figure 1 | **Contrast between marginal and joint estimation.** Panels **a** and **b** depict the likelihood profile for two trees versus a hypothetical parameter *x*. The *x* axis represents some nuisance parameter (for example, the ratio of the rate of transitions to the rate of transversions). The *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).

Holder & Lewis 2003

Frequentist model selection: Which model has the *highest likelihood* for one possible scenario that the model allows.

Bayesian model selection: Which model has the highest *marginal likelihood* over all possible worlds the model allows -> prior sensitive because prior determines what worlds are allowed!

Bayesian Model Selection

A job needs done, and I want to know what tools are needed to do it as efficiently as possible

Model 1: I give Joel a hammer
Model 2: I give Frank a screwdriver
Model 3: I give Ignacio an entire toolshed

Who will find a better tool for the job?

Bayesian Model Selection

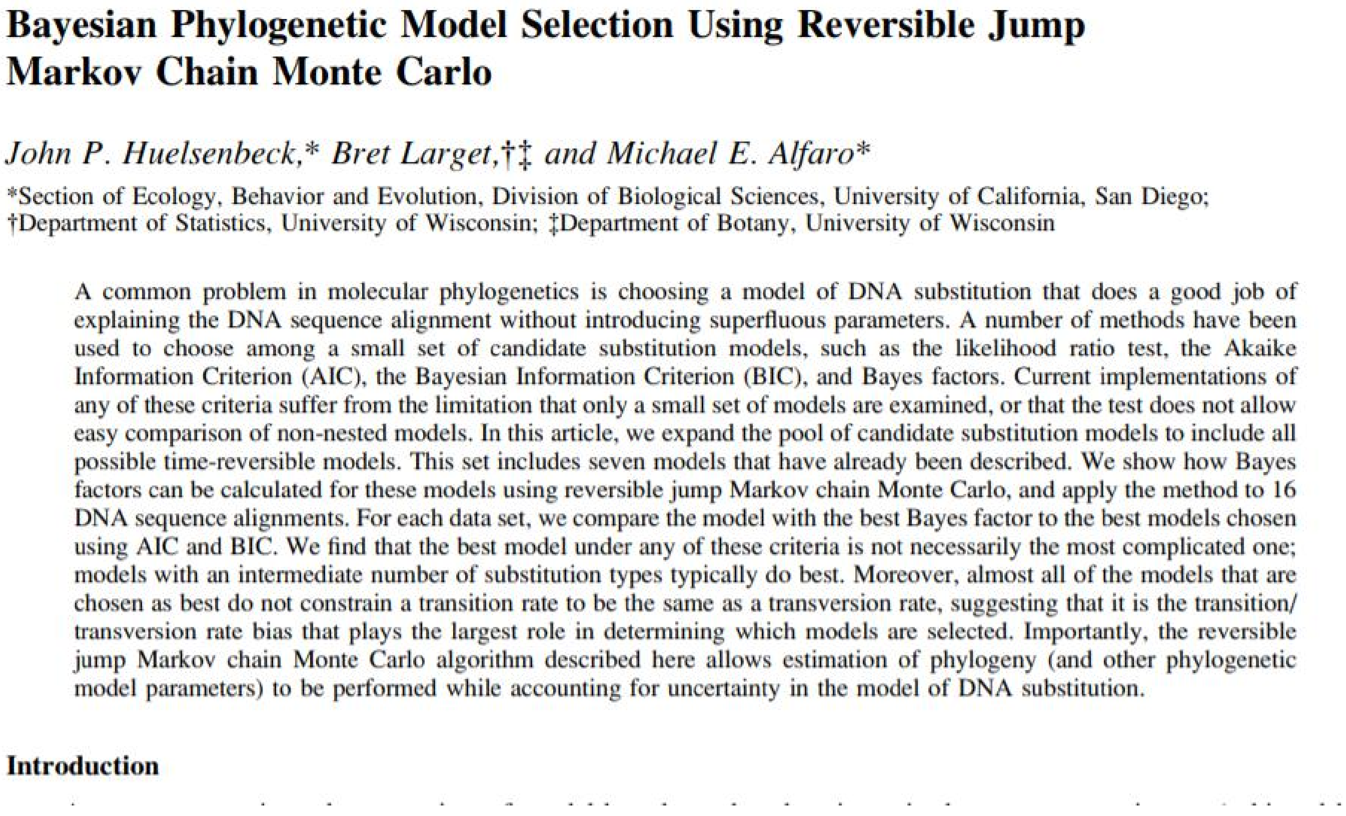
A job needs done, and I want to know what tools are needed to do it as efficiently as possible

Model 1: I give Joel a hammer
Model 2: I give Frank a screwdriver
Model 3: I give Ignacio an entire toolshed

Who will finish the job first?

Because the questions are different, we aren't as concerned about overfitting in Bayesian models, because we are marginalizing over parameters, rather than searching them for the best possible combination

Reversible-Jump MCMC



Bayesian Phylogenetic Model Selection Using Reversible Jump Markov Chain Monte Carlo

John P. Huelsenbeck, Bret Larget,†‡ and Michael E. Alfaro**

*Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California, San Diego;
†Department of Statistics, University of Wisconsin; ‡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

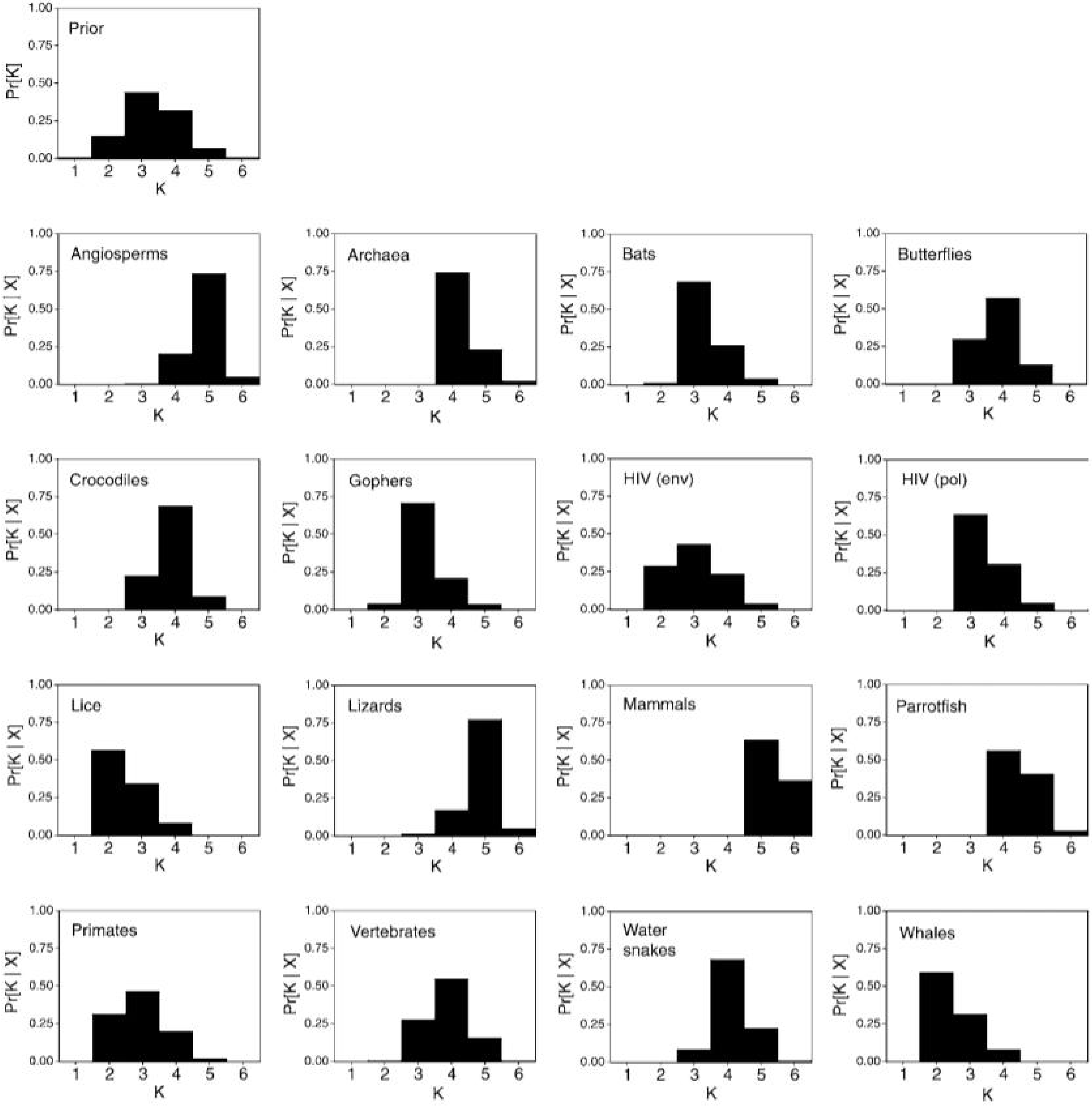


FIG. 1.—The posterior probability of *K*, the number of substitution types, for each DNA sequence alignment. The first figure shows the prior probability distribution of *K*, calculated as the number of models with *K* substitution parameters divided by the total number of substitution models (203).

Model Adequacy

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

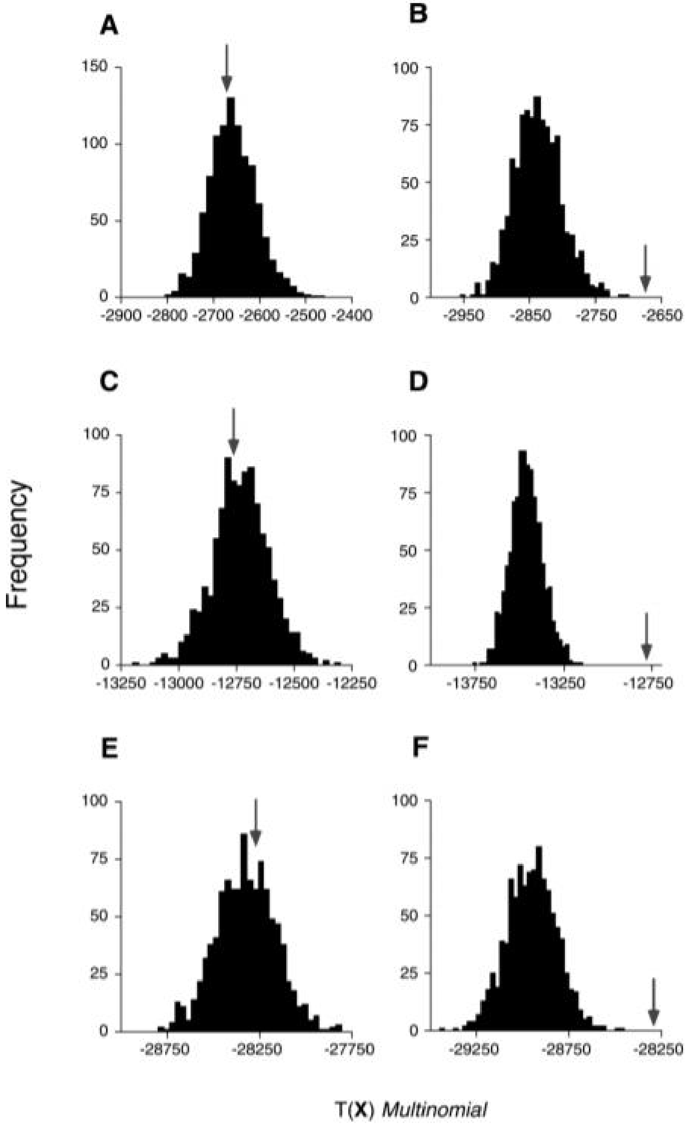
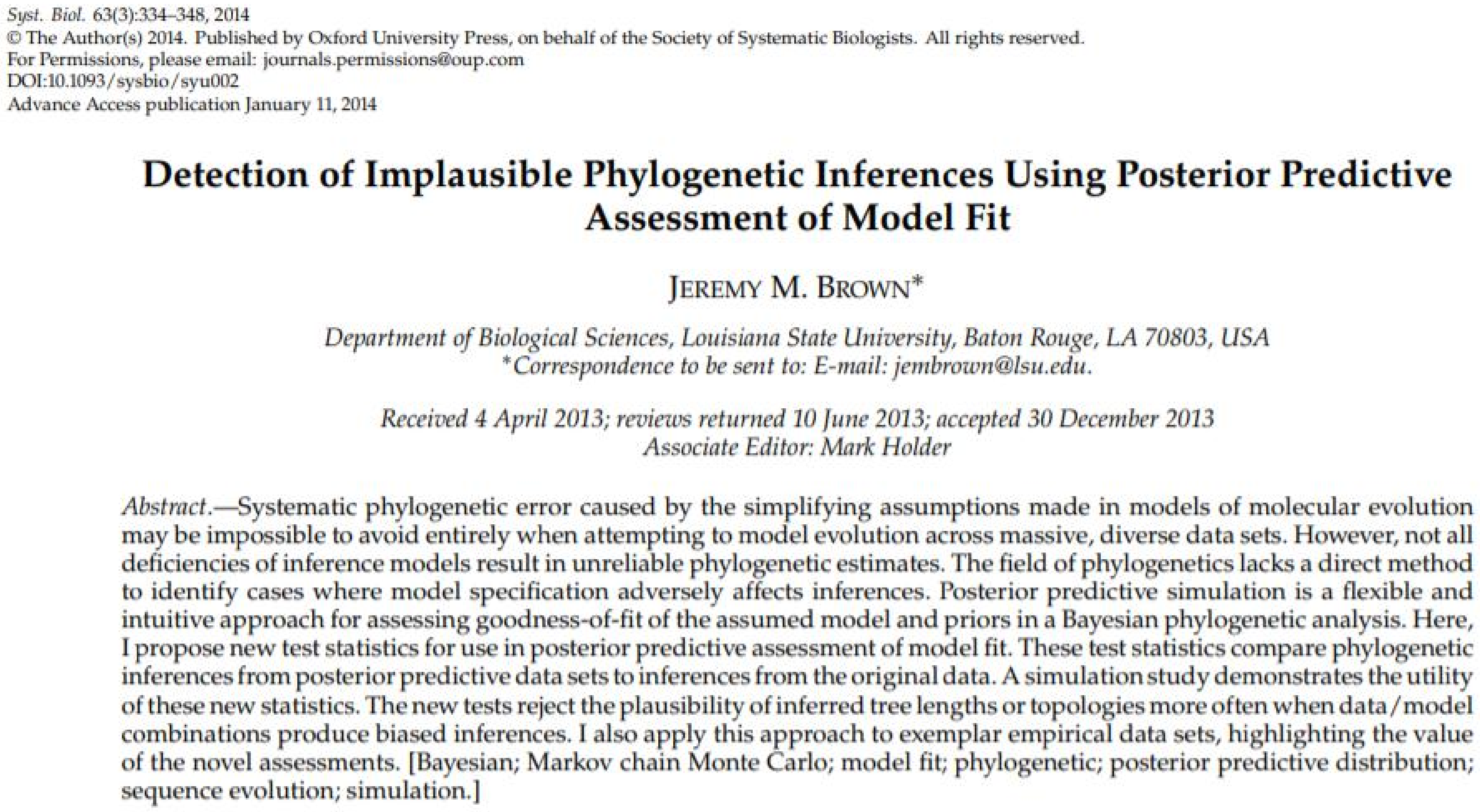


FIG. 2.—Illustration of the method comparing GTR versus JC69. Data sets of $c = 500$ (A, B), $c = 2{,}000$ (C, D), and $c = 4{,}000$ (E, F) sites were simulated under the GTR model. Predictive distributions were simulated under the GTR (A, C, E) and JC69 (B, D, F) models. Arrows indicate the values for 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?

good question...



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.]

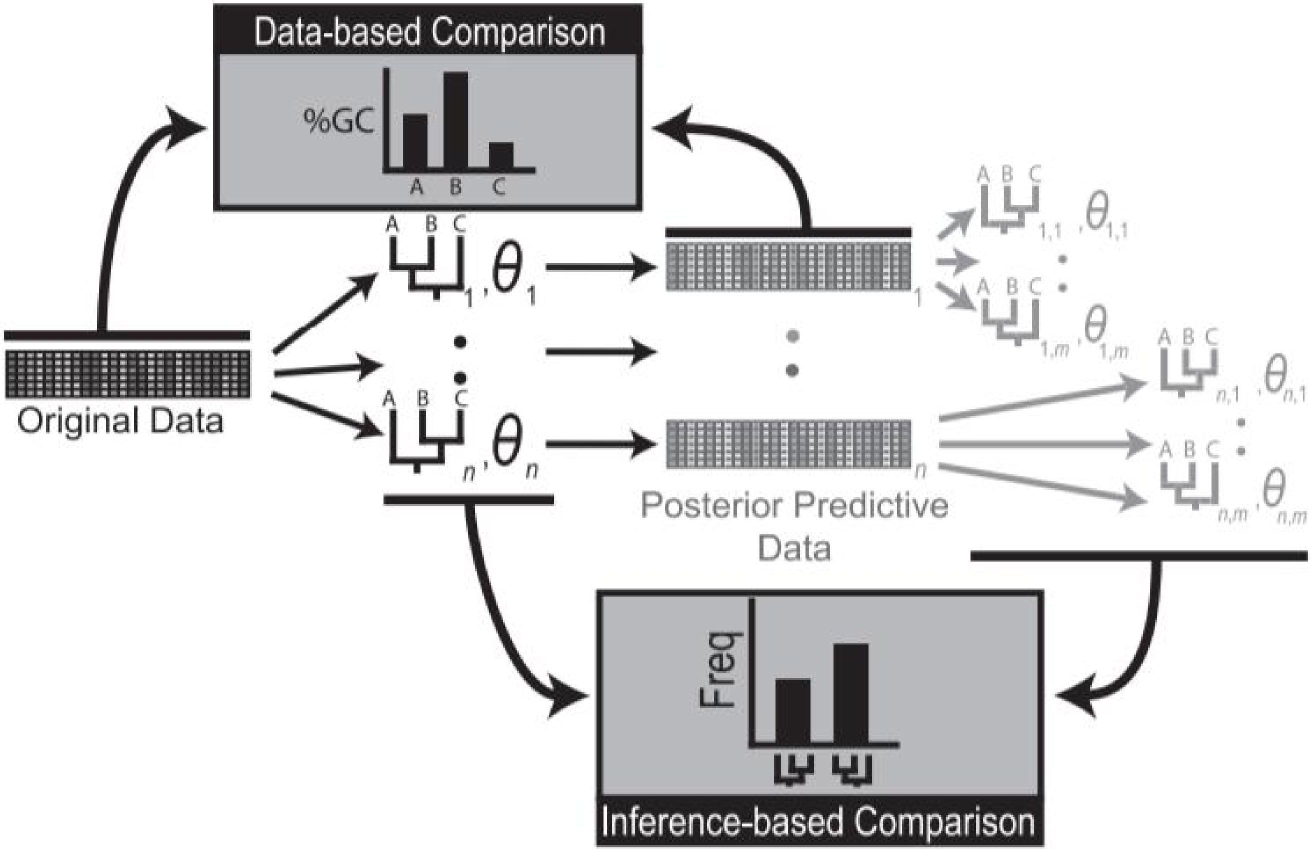
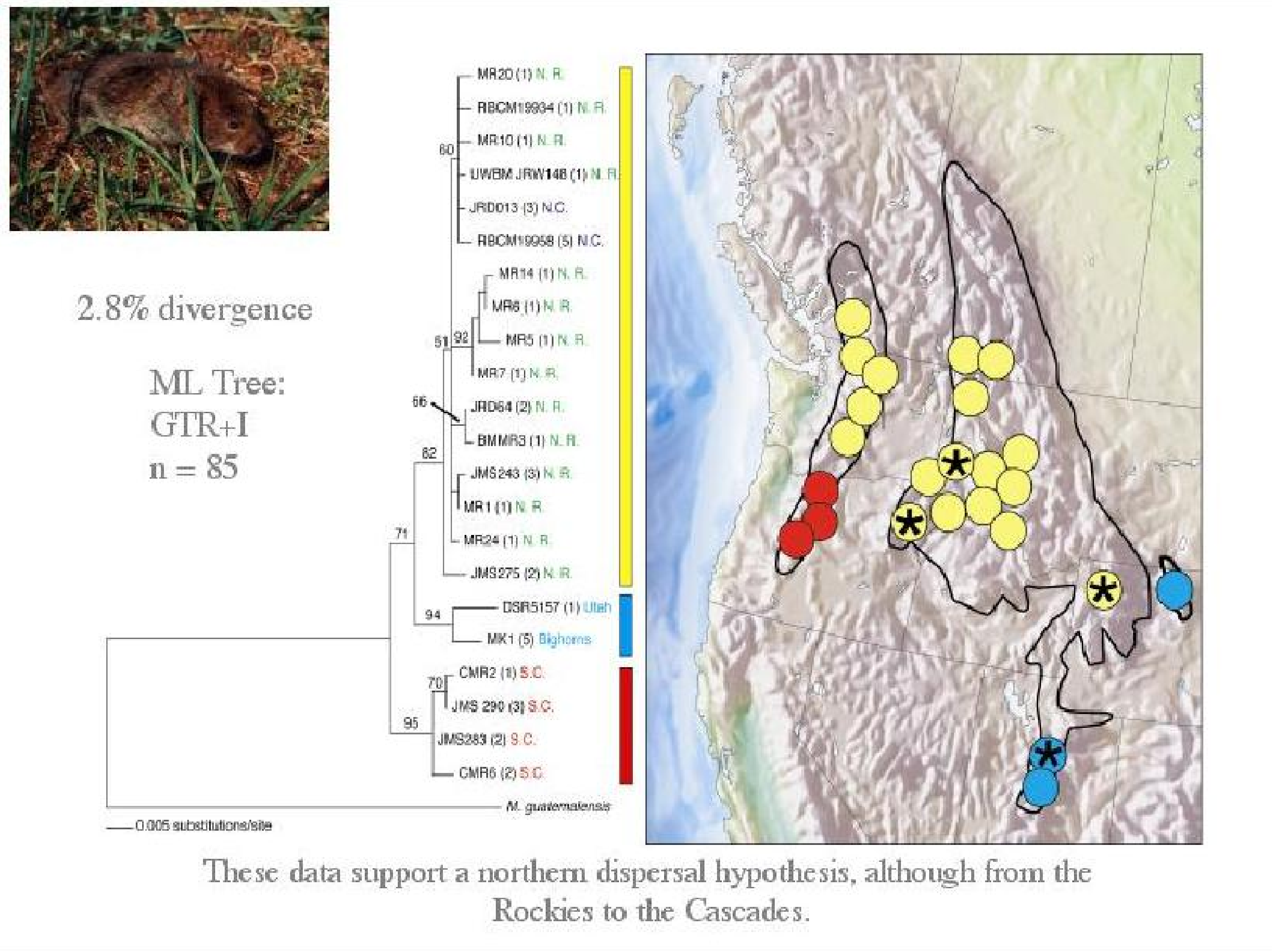


FIGURE 1. A schematic representation of data- versus inference-based approaches to assessing model plausibility with posterior predictive simulation. Most statistics proposed for testing model plausibility compare data-based characteristics of the original data set to the posterior predictive data sets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

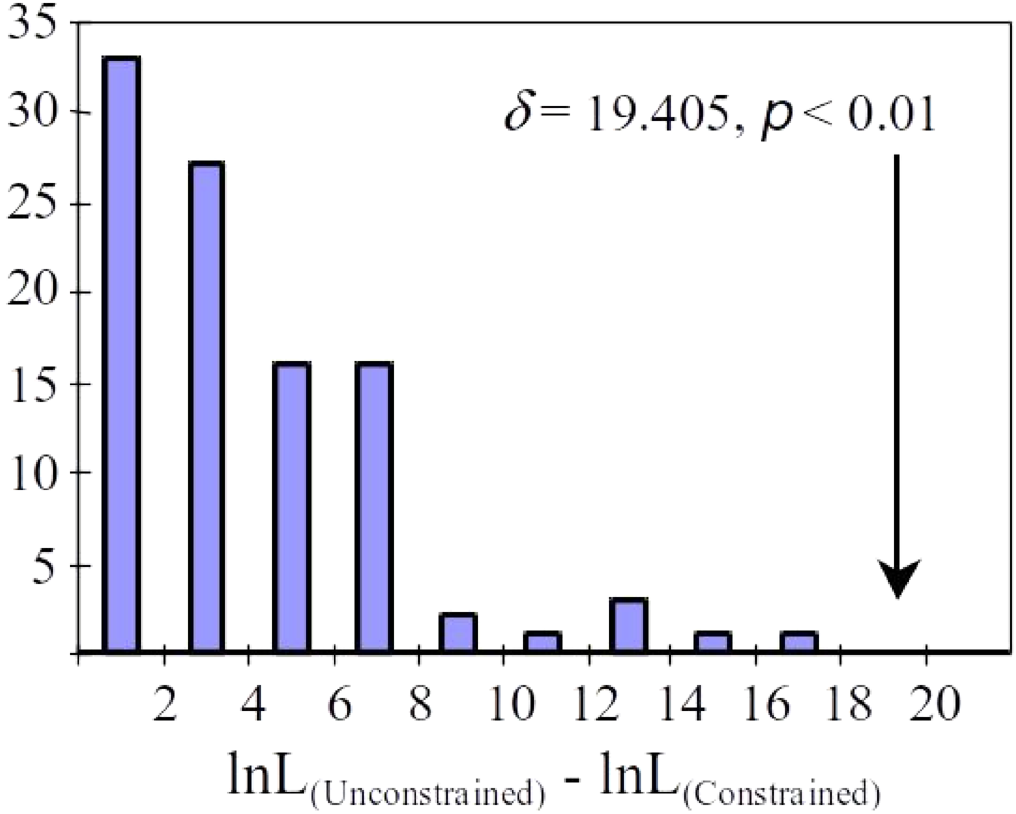


2.8% divergence

ML Tree:
GTR+I
n = 85

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

Run parametric bootstrap



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.