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.]

> We do not like to ask, 'Is our model true or false?', since probability models in most data analyses will not be perfectly true … The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'
>
> — A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin

Model-based approaches to phylogenetic inference have become an integral part of the phylogenetic toolkit for a variety of reasons, including the fact that they allow statements of statistical support with explicit definition of the underlying assumptions. In particular, the use of Bayesian inference has grown rapidly in recent years, in large part because it provides a natural framework for accommodating uncertainty as well as an intuitive measure thereof: the posterior probability. Statements of statistical support, such as posterior probabilities, remain conditional on the assumptions of chosen models, even when those assumptions are explicitly defined. Such statements can be inaccurate when model assumptions poorly reflect reality (Huelsenbeck and Hillis 1993; Yang et al. 1994; Swofford et al. 2001; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004; Brown and Lemmon 2007).

Phylogeneticists have been interested in the degree to which assumed models of character evolution adequately describe evolutionary processes (Kelchner and Thomas 2007). Many studies have developed new models that relax previously held assumptions (e.g., Pagel and Meade 2004; Lartillot and Philippe 2004; Whelan 2008). Others have developed a general understanding about which assumptions may be problematic when violated (e.g., Huelsenbeck and Hillis 1993; Yang et al. 1994; Swofford et al. 2001; Huelsenbeck and Rannala 2004; Lemmon and Moriarty 2004; Brown and Lemmon 2007; Holder et al. 2008). Still others have devised methods for choosing the best model from an available pool (e.g., Minin et al. 2003; Posada and Buckley 2004; Sullivan and Joyce 2005; Fan et al. 2011; Xie et al. 2011). However, these approaches do not assess whether a model is adequate for inference with a particular data set.

Assessing the plausibility of an assumed model should be a fundamental step in any Bayesian analysis (Gelman et al. 1996), yet is rarely practiced in phylogenetics (but see Huelsenbeck et al. 2001; Bollback 2002; Foster 2004; Lartillot et al. 2007; Rabeling et al. 2008; and Rodrigue et al. 2009 for exceptions). To address this shortcoming, we need to assess more than the performance of our chosen model relative to other available models. We must ask about model performance relative to the actual processes generating phylogenetic data, by comparing our data to expectations under our assumed model. Ideally, this comparison should be performed with an eye toward the reliability of the inferences we are drawing.

Approaches to Bayesian Assessment of Model Plausibility

Assessment of model plausibility in a Bayesian framework generally relies on the use of two closely related distributions: the posterior and posterior predictive distributions. The posterior distribution is the usual target of a Bayesian analysis and represents uncertainty in models and parameter values taking into account both prior beliefs and the data. The posterior predictive distribution represents the range of plausible data sets that could have been observed (or would be predicted to be observed if the process that generated the data was repeated). In practice, a sample from the

posterior predictive distribution can be obtained by sampling parameter values from the posterior and using them to simulate replicate data sets (Fig. 1).

Gelman et al. (2004) outline three approaches to model assessment: (1) assessment of the posterior distribution's plausibility through comparison to prior expectations about the model; (2) assessment of the posterior predictive distribution's plausibility through comparison to prior expectations about the data; and (3) assessment of the posterior predictive distribution's plausibility through comparison to the data that have been analyzed. Approach (1) is, presumably, already practiced in phylogenetics. If a posterior is strongly at odds with biological expectations (e.g., it suggests that first and second codon positions of conserved genes evolve more quickly than third codon positions, or transversions occur much more frequently than transitions), we should be suspicious. Approach (1) is also practiced when comparing posteriors from different data sets.

Approach (2) is not, to my knowledge, regularly applied in phylogenetics. Such a check would involve simulating data sets using samples from the posterior and comparing them to biological expectations. This technique may have received little attention because biological expectations regarding a "typical" data set are poorly defined. Both approaches (1) and (2) rely on prior knowledge regarding the biology of the characters used in the analysis.

Approach (3), comparing the posterior predictive distribution to the data at hand, has been proposed for use in phylogenetics (Bollback 2002; Nielsen 2002; Bollback 2005), although it is rarely applied (but see Huelsenbeck et al. 2001; Foster 2004; Rabeling et al. 2008; and Rodrigue et al. 2009 for examples of notable exceptions). This approach is the most "statistical" of the three (Gelman et al. 2004). Application of this posterior predictive check in a phylogenetic context involves (i) simulating replicate data sets using draws from the posterior, (ii) choosing a test statistic to quantify some relevant aspect of each data set, (iii) calculating the value of this test statistic for all simulated and empirical data, and (iv) comparing the empirical test statistic value to the posterior predictive distribution of such values (Fig. 1). This comparison is often summarized by calculating the tail-area probability of the empirical test statistic value relative to the simulated posterior predictive distribution, a quantity known as the posterior predictive *P*-value. Freedom of test statistic choice is essentially limited only by a researcher's creativity, although not all choices perform equally well. Phylogenetically relevant test statistics have been proposed for assessing the overall plausibility of an alignment (e.g., the multinomial likelihood, Bollback 2002), as well as specific violations of model assumptions (e.g., nonstationarity of base composition: Huelsenbeck et al. 2001; Foster 2004; unequal synonymous vs. nonsynonymous substitution rates: Nielsen 2002; Rodrigue et al. 2009). Whereas each of the proposed statistics is useful for identifying deficiencies that make the model unable to replicate particular aspects of the observed data, the connection between these deficiencies and the robustness of inferences is unclear. For instance, if we can reject the homogeneity of base composition across taxa, is the observed heterogeneity sufficient to make inferences implausible? Certain differences between the original and posterior predictive data sets may indicate little about the phylogenetic performance of the model.

Here, I propose the use of test statistics that directly assess the plausibility of phylogenetic inferences. This approach can alternately be thought of as assessing the plausibility of the phylogenetic information contained in a data set, where information is conditional on the assumed model. Such an assessment is accomplished by comparing inferences from empirical data to the same inferences from posterior predictive data (Fig. 1). The important advantage of this approach is its focus on the plausibility of specific inferences drawn from individual data sets. Note that I have chosen to use the term "plausibility," rather than "adequacy," because the tests indicate whether empirical inferences can be considered surprising under the assumed model. The tests cannot guarantee that an unsurprising result is correct. Similarly, not every surprising result is incorrect. Even when our model and priors are properly specified, relatively low probability events are expected to occur occasionally. However, frequent surprise should serve as a strong indication that our expectations are not accurate. Below, I outline a sample of inference-based statistics, test their performance on analyses of simulated data, highlight their utility through analyses of empirical data, and suggest a range of other inference-based statistics.

Methods

Posterior Predictive Assessment of Inferential Plausibility

In the Bayesian framework, a model is composed of both the parameters it uses to describe the process of data generation as well as the prior distributions placed on values of those parameters. For the purposes of this article, we will assume that priors are chosen to reflect a researcher's expectations about the probability that any particular parameter value gave rise to an observed data set before having actually observed said data. When the distribution from which parameter values were drawn for data generation is known (as in a simulation study), this distribution best reflects this expectation and will be referred to as the "true" prior. Any other prior will be "incorrect" to the extent that it does not reflect "true" probabilities. For empirical analyses, the existence of a true prior will be assumed, although any prior chosen by a researcher will surely differ from this truth to some extent. In this sense, the use of an incorrect prior does not necessarily indicate that a researcher has done a poor job of choosing a prior, but rather that they have done so with limited or imperfect information about the true distribution. Each of the test statistics that I

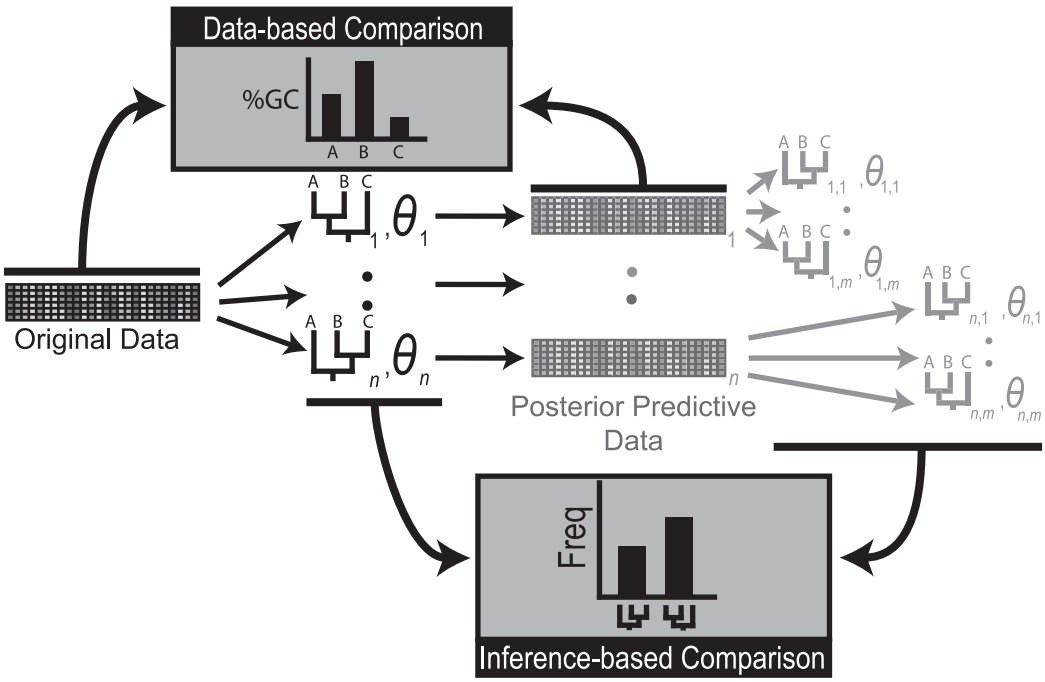

FIGURE 1. A schematic representation of data- versus inference-based approaches to assessing model plausibility with posterior predictive simulation. Most statistics proposed for testing model plausibility compare data-based characteristics of the original data set to the posterior predictive data sets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

propose here relies on comparing marginal distributions from analyses of posterior predictive data to the same distribution from analysis of the original data, but note that this approach could easily be extended to other summaries of a data set's information content (e.g., marginal- or maximum-likelihood ratios).

Two aspects of the marginal tree-length distribution were used as test statistics: the mean and variance. If M_c is the chosen model and $\mathbf{X}$ is a multiple sequence alignment, these test statistics will be denoted $T_l(\mathbf{X},M_c)$ and $T_v(\mathbf{X},M_c)$ for the mean and variance, respectively. M_c is included in the test statistic designation to emphasize the model-dependent nature of the statistics. A variety of possible reasons exist for discord in the mean and variance of posterior tree lengths between empirical and posterior predictive data. For example, differences in tree lengths preferred by the prior and the empirical data (as expressed in the likelihood) will result in posterior estimates that are intermediate. Posterior predictive data sets will then be simulated with such intermediate values. When those posterior predictive data sets are analyzed, estimated tree lengths will be closer to the prior than the empirical estimates. Such a pattern would result in a low posterior predictive *P*-value since the empirical estimate would not be a plausible draw from the posterior predictive distribution. Tension between the prior and likelihood could similarly affect posterior predictive assessment of the variance. Differences in variance may also reflect differences in the amount of information about tree length that each data set contains.

One approach to assessing the plausibility of overall topological inference employs test statistics based on the distribution of symmetric differences (i.e., unweighted Robinson-Foulds distances; Robinson and Foulds 1981) between all trees in a posterior sample. The use of these statistics is based on the supposition that support across trees can differ when a model is used to analyze data generated by a process that differs from the process it assumes, as compared to data generated by the model itself. The distribution of support across trees can be characterized by the position of particular quantiles in the ordered vector of all symmetric differences, x (Fig. 2). If l is the length of the ordered symmetric-distance vector x, then for the k-th q-quantile with $k<q$, let $j=\lfloor lk/q \rfloor$ and $g=(lk/q)-j$. The k-th q-quantile test statistic, $T_{k.q}(\mathbf{X},M_c)$, is then defined as

$$T_{k.q}(\mathbf{X},M_c)=\begin{cases}\frac{1}{2}(x_j+x_{j+1}), & g=0\\ x_{j+1}, & g>0\end{cases}. \quad (1)$$

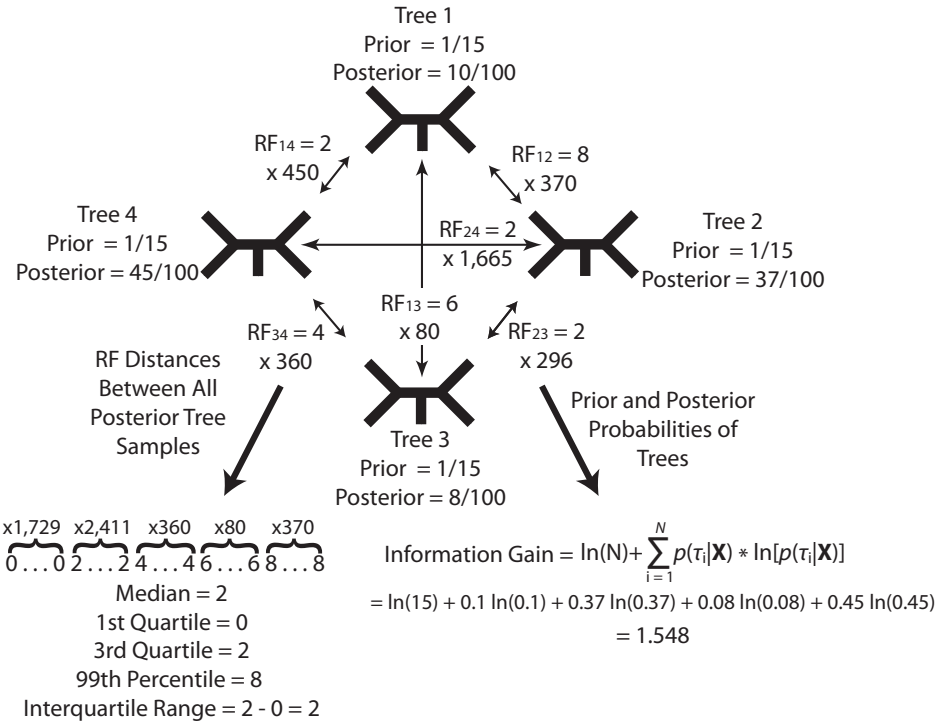

FIGURE 2. Example topological test statistic calculations from a posterior distribution. In this hypothetical scenario, four unique topologies were found in 100 MCMC samples from a Bayesian analysis. The prior (uniform) and estimated posterior probabilities are given next to each unique topology. Bidirectional arrows are labeled with the symmetric (Robinson–Foulds) distance between topologies, along with the number of times this distance will be included in the vector of all pairwise distances between posterior samples. Note that the comparison of samples with the same topology (RF = 0) is not explicitly shown, but the number of such pairwise comparisons is easily calculated as $\sum_{i=1}^{B(N)} \binom{p(\tau_i|\mathbf{X})\times 100}{2}$ where all notation is as in the text. The bottom left depicts the ordered vector of symmetric distances between all posterior samples and example summary statistics resulting from this vector. The bottom right shows an example calculation of the gain in topological information that results from analyzing the data.

A series of different quantile positions may be used to probe various parts of this distribution. Several quantiles were tested in this study, including the 1st quartile (1st 4-quantile), median (1st 2-quantile), 3rd quartile (3rd 4-quantile), 99th percentile (99th 100-quantile), 999th permillage (999th 1000-quantile), and the 9999th 10,000-quantile.

Another test statistic that summarizes support across topologies, but without reference to topological similarity, employs the difference in statistical entropy (i.e., the information gain) between the marginal prior and posterior distributions of tree topologies. The statistical entropy of the prior is defined as

$$H[p(\tau)] = -\sum_{i=1}^{B(N)} p(\tau_i)\ln[p(\tau_i)], \quad (2)$$

where H is entropy, $p(\tau)$ is the marginal prior distribution of tree topologies, $B(N)$ is the total number of bifurcating tree topologies for N taxa, and $p(\tau_i)$ is the probability of drawing the i-th topology from the prior (Shannon and Weaver 1949; Reza 1961). The entropy of the posterior can be calculated in an exactly analogous manner, by replacing $p(\tau)$ with $p(\tau|\mathbf{X})$. Statistical entropy represents the amount of uncertainty associated with a draw from either the posterior or the prior. A uniform distribution across topologies has maximal entropy. As the data provide more information, and the posterior probabilities of different topologies become more uneven, the entropy of the posterior will decrease and the difference in entropy between the posterior and a uniform prior will increase. The test statistic, $T_e(\mathbf{X},M_c)$, representing this difference is then

$$T_e(\mathbf{X},M_c) = H[p(\tau)] - H[p(\tau|\mathbf{X})] = \sum_{i=1}^{B(N)} p(\tau_i|\mathbf{X})\ln[p(\tau_i|\mathbf{X})] - \sum_{i=1}^{B(N)} p(\tau_i)\ln[p(\tau_i)], \quad (3)$$

where $p(\tau_i|\mathbf{X})$ is the posterior probability of the i-th topology. If the prior on topologies is uniform, this test statistic simplifies to

$$T_e(\mathbf{X},M_c) = \ln[B(N)] + \sum_{i=1}^{B(N)} p(\tau_i|\mathbf{X})\ln[p(\tau_i|\mathbf{X})]. \quad (4)$$

This statistic provides information about the overall amount of topological information in a data set, conditional on the assumed model (Fig. 2).

For comparison, a general data-based assessment of plausibility was performed in PuMAv0.905 (Brown and ElDabaje 2009) using the multinomial likelihood test statistic proposed by Bollback (2002)

$$T_m(\mathbf{X}) = \ln\left(\prod_{i=1}^{s}\left(\frac{l_{\xi(i)}}{L}\right)^{l_{\xi(i)}}\right), \quad (5)$$

where s is the number of observed, unique site patterns, $\xi(i)$ is the i-th unique site pattern, $l_{\xi(i)}$ is the number of instances of $\xi(i)$ in the multiple sequence alignment, and L is the total number of sites. The multinomial likelihood statistic assesses the ability of a model to explain the frequencies of different site patterns. No explicit connection exists between this quantity and any particular inference under the assumed model. The power of tests based on site-pattern frequencies may be increased through various binning strategies (e.g., Waddell et al. 2009), although I did not explore such alternatives here.

AMP

All of the inference-based test statistics outlined above can be calculated from the corresponding posterior distributions using the software AMP: Assessing Phylogenetic Model Fit with Posterior Prediction. AMP is written in Python and relies on the DendroPy Phylogenetic Computing Library (Sukumaran and Holder 2010). It is freely available from http://code.google.com/p/phylo-amp and is distributed under the GNU General Public License version 3.

Data Simulation

Data sets were simulated using a 29-taxon tree topology and model parameters derived from empirical data (Brandley et al. 2005), which have been used to parameterize simulations in previous studies (Brown and Lemmon 2007; Brown et al. 2010). The tree topology used in the simulations was identical to figure 1 in Brown and Lemmon (2007), but with branch lengths drawn from exponential distributions to match the branch-length priors assumed in data analysis. Fifty data sets were simulated under each of three branch-length distributions. Branch lengths used to simulate the first group of 50 were drawn from a distribution adjusted to give, on average, the same tree length as that of Tree B in Brown and Lemmon's (2007) figure 1 (λ = 442.44). The exact branch lengths were drawn independently for each simulated data set. This group is referred to as 1x. Two additional groups of 50 data sets were simulated in exactly the same manner, but with branch lengths drawn from exponential distributions with larger means (10x or 50x larger). The increased number of substitutions on these trees creates data sets that are more likely to mislead insufficient models. Parameter values for the general time reversible (GTR; Tavaré 1986) model of sequence evolution were drawn from empirical estimates (Brandley et al. 2005; table 1 in Brown and Lemmon 2007). For each set of parameters (e.g., equilibrium base frequencies, exchangeabilities, and those describing rate variation across sites or RAS), the estimated values that most strongly violated the assumptions of a Jukes–Cantor (JC; Jukes and Cantor 1969) model were chosen. For instance, of the nine available sets of equilibrium base frequencies, the set with the highest variance across the four character states was chosen. Similarly, the set of exchangeabilities with the highest variance and the set of parameters describing RAS (α—the shape parameter of the Γ distribution with mean 1, Yang (1994); I—the proportion of invariable sites) that gave the highest variance in site rates were chosen independently. The composite model forms a strong, but empirically grounded, violation of the assumptions of the JC model.

Empirical Data

Three empirical data sets were analyzed with the newly proposed inference-based test statistics to illustrate their utility. The first was taken from the study of Regier et al. (2008) on arthropod phylogeny. I selected all genes (27) with complete taxon sampling (13 taxa) and analyzed each separately using an unpartitioned GTR+I+Γ model of sequence evolution and an exponential branch-length prior (λ = 10; the default value in MrBayes). Details of the Bayesian analysis are given below. Inferential plausibility was assessed using the topological and tree-length test statistics introduced in this study, as well as the multinomial likelihood (Bollback 2002; Brown and ElDabaje 2009) for comparison. Since the multinomial likelihood is calculated using the frequency of different site patterns, the presence of missing or ambiguous data is problematic. Therefore, I removed all sites with such character states before the analysis. Given the depth of divergences between taxa in this study and the relatively sparse taxon sampling, some degree of topological error due to model inadequacy may be expected.

The second data set was taken from a phylogeographic study of two frog species in the genus *Acris* (Gamble et al. 2008). These sequences come from four genes (one mitochondrial and three nuclear). Taxa and sites were deleted from the original data matrix to eliminate missing and ambiguous character states while retaining as much data as possible. The final matrix contained 53 sequences and 909 characters. As with the arthropod data, I assumed an unpartitioned GTR+I+Γ model of sequence evolution with an exponential branch-length prior (λ = 10) and assessed plausibility using the inference-based test statistics, as well as the

multinomial likelihood. This data set has relatively shallow divergences and good taxon sampling, but is known to give biased branch-length estimates with the branch-length prior assumed here (Brown et al. 2010). Therefore, implausibility of branch-length estimates is expected. Each gene in this data set was also analyzed separately using the same test statistics for comparison with the concatenated results.

The third data set consists of complete mitochondrial genomes from 16 mammals and was originally collected for a study focusing on the phylogenetic position of guinea pigs relative to other rodents (D'Erchia et al. 1996). Based on their results, D'Erchia et al. (1996) suggested that these data strongly refuted the hypothesized position of guinea pigs within rodents. Following up on this study, Sullivan and Swofford (1997) showed that the strong and surprising conclusions of D'Erchia et al. (1996) likely stemmed from their use of an insufficient model that assumed equal rates of evolution across sites. This implausible assumption led to erroneously strong support for rodent nonmonophyly. Following Sullivan and Swofford (1997), I only analyzed first and second codon positions of mitochondrial protein-coding genes due to known nonstationarity of base frequencies at third positions and to maintain comparability with previous authors' results. I performed three sets of analyses with these data, assuming (i) a GTR model with equal rates across sites, (ii) a GTR+I+Γ model, and (iii) independent GTR+I+Γ models for the two codon positions with additional position-specific rate multipliers. Sullivan and Swofford (1997) considered the first and second of these three models in a maximum-likelihood framework, but not the third. All analyses assumed an exponential branch-length prior (λ =10).

Bayesian Phylogenetic Analyses

All Bayesian phylogenetic analyses were performed using MrBayes v3.1.2 (Ronquist and Huelsenbeck 2003) with default priors, except for specified alterations of the branch-length prior. For each analysis, four independent runs were used (each with four Metropolis-coupled chains) and convergence was assessed according to the criteria outlined by Brown and Lemmon (2007) as implemented in MrConverge v1b2 (written by A.R. Lemmon). Runs were considered to have converged once the widest 95% confidence interval for the posterior probability of any bipartition fell below 0.1. Markov chain Monte Carlo (MCMC) samples were saved every 1000 generations.

For all simulated data sets in each group of 50 (1x, 10x, and 50x), the posterior distribution was estimated twice: once assuming a JC model and once assuming a GTR+I+Γ model. Comparison of these analyses is used to investigate the relationship between assessment of model plausibility and bias in topological inference for an oversimplified model. In both cases, assumed branch-length priors were identical to the exponential distributions from which branch lengths were drawn for simulation.

For the 1x data sets, two additional analyses were performed in which a GTR+I+Γ model was assumed, but the mean of the branch-length prior was adjusted up ($\lambda = 50$) or down ($\lambda = 1200$) until inferred 95% credible intervals on tree length no longer included the true tree length. These additional analyses were performed to assess the ability of tree-length test statistics to detect biased tree-length inference in the absence of topological bias, as branch-length estimates are known to sometimes be sensitive to assumed branch-length priors (Brown et al. 2010; Marshall 2010).

For each analysis that assumed an incorrect model or prior, the extent of topological error was assessed by comparing support for true bipartitions when assuming the generating model to the same support when assuming an incorrect model, normalized by the maximum possible support for the true tree. Thus, error was calculated as:

$$\varepsilon_{\mathrm{BPP}} = \frac{1}{N-3}\sum_{i=1}^{N-3} p(B_i|\mathbf{X}, M_T) - p(B_i|\mathbf{X}, M_I), \qquad (6)$$

where $N-3$ is the number of internal bipartitions in a bifurcating tree with N taxa, B_i is the i-th true bipartition, $\mathbf{X}$ is the observed data, M_T is the true, generating model, and M_I is an incorrect model. Error was calculated this way in order to differentiate between too much versus too little support for the true tree. However, inflated support for some branches and deflated support for others may cause the underparameterized model to appear to have reduced error.

For each of the test statistics (T), the posterior predictive P-value for a lower one-tailed test is defined as the proportion of samples in the posterior predictive distribution with a test statistic value less than or equal to the observed value,

$$\begin{aligned} p_l &= p\left(T\left(\mathbf{X}_{\mathrm{rep}}\right) \leqslant T(\mathbf{X})|\mathbf{X}\right) \\ &= \sum_{\tau}\int\sum_{\mathbf{X}_{\mathrm{rep}}} I_{T(\mathbf{X}_{\mathrm{rep}})\leqslant T(\mathbf{X})} p(\mathbf{X}_{\mathrm{rep}}|\theta_\tau, \tau) p(\theta_\tau, \tau|\mathbf{X}) d\theta_\tau, \end{aligned} \qquad (7)$$

where p_l is the lower one-tailed posterior predictive P-value, $\mathbf{X}_{rep}$ is a replicate multiple sequence alignment from the posterior predictive distribution, $\mathbf{X}$ is the original alignment, θ_τ is a vector of model parameter values corresponding to topology τ, I is the indicator function, and the sums are taken across all possible topologies or replicate data sets (Gelman et al. 2004). In practice, this posterior predictive P-value is approximated by simulating posterior predictive data sets ($\mathbf{X}_{\mathrm{rep}}$) using trees and model parameters drawn from the posterior distribution conditioned on $\mathbf{X}$, evaluating the test statistic for each replicate, $T(\mathbf{X}_{\mathrm{rep}})$, and comparing each posterior predictive test statistic value to the empirical value, $T(\mathbf{X})$. The P-value for an

upper one-tailed test, p_u, is simply the converse of the lower one-tailed test

$$p_u = p\left(T\left(\mathbf{X}_{\text{rep}}\right) \geqslant T(\mathbf{X}) \mid \mathbf{X}\right). \quad (8)$$

The two-tailed posterior predictive *P*-value, p_2, is twice the minimum of the corresponding one-tailed tests

$$p_2 = 2\min(p_l, p_u). \quad (9)$$

Use of two-tailed tests is recommended because discrepancies from model expectations in either direction should be cause for concern. Posterior predictive simulation of data sets was performed using PuMAv0.905 (Brown and ElDabaje 2009) and Seq-Gen v1.3.2 (Rambaut and Grassly 1997) with 100–200 parameter values and trees drawn uniformly from postburn-in MCMC samples.

Results

Simulated

When inference was conducted using the generating model, support for the true topology varied among data sets. The greatest variation was found among replicates simulated along the shortest tree (Fig. 3). Topological error (i.e., the difference between JC and GTR+I+Γ in support for true bipartitions) for analyses assuming JC increased with the length of the tree used for simulation (Fig. 4), consistent with expectations. Error in estimated bipartition support due to underparameterization was generally small at shorter tree lengths (Fig. 5). As tree length increased, several branches strongly supported by the generating model were estimated to have no support by the incorrect model (Fig. 5). Assuming the generating model with a marginally biased branch-length prior induced very little topological error (Supplementary Fig. S1; http://dx.doi.org/10.5061/dryad.nc866), as suggested previously (Brown et al. 2010).

Posterior predictive tests using the multinomial likelihood test statistic, $T_m(\mathbf{X})$, strongly rejected the plausibility of all analyses assuming an underparameterized model (JC) or an incorrect branch-length prior, regardless of the induced topological error (Supplementary Fig. S2). This behavior, although perhaps desirable in situations where the ability of a model to accurately reproduce specifics of the underlying data structure (e.g., site pattern frequency) is important, provides motivation for pursuing test statistics that directly assess the plausibility of inferences.

When analyses of simulated data were conducted under the generating model and true branch-length prior, newly proposed tests of topological plausibility assessed almost all inferences to be reasonable, whether using directional or nondirectional alternative hypotheses (Fig. 6 and Supplementary Fig. S3). Only a single extreme *P*-value (<0.05) was detected, despite testing 150 data sets with 10 test statistics and

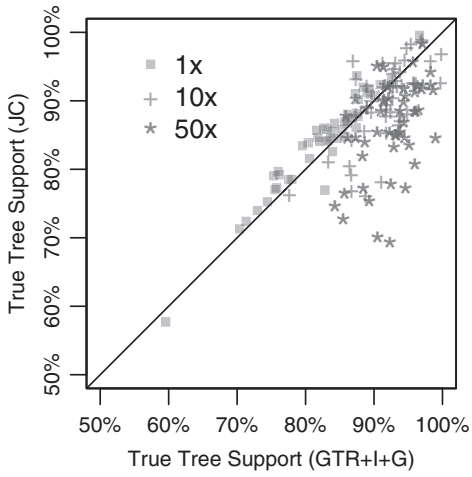

FIGURE 3. Support for the true tree when simulated data are analyzed with the generating model (GTR+I+Γ) or an incorrect, underparameterized model (JC). Topological support is measured as the sum of the posterior probabilities for all true internal bipartitions divided by their count (i.e., the average support for an internal bipartition). Points with different shapes represent data sets simulated with different expected branch lengths (1x expected branch length (EBL) = 0.002, 10x EBL = 0.023, 50x EBL = 0.113). The solid line has a slope of 1 and represents equal support for the true tree under the generating and underparameterized models. Points above the line indicate more support for the true tree when assuming the underparameterized model, whereas points below the line indicate more support for the true tree when assuming the generating model. Note that as the simulated tree length increases, the difference in support between the correct and incorrect models (i.e., deviation from the 1:1 line) increases.

employing both two- and one-tailed (both directions) tests. However, simulations were conducted with fixed parameters for the model of sequence evolution, in order to consistently generate data sets that violated the underparameterized model's assumptions. The frequency of extreme *P*-values when simulation parameter values are drawn from assumed priors remains to be determined, but is expected to remain low. The problem of multiple testing may not be as substantial as it first seems, because different test statistics probe the same set of posterior predictive distributions in related ways and tests are not independent. Additionally, even if the occasional small posterior predictive *P*-value is expected, we might still be interested in knowing which inferences are least plausible under the assumed model (Gelman et al. 2004).

The mean tree-length test statistic, $T_l(\mathbf{X}, M_c)$, performed well in detecting use of slightly biased branch-length priors (Fig. 7 and Supplementary Fig. S4). Both the proper directional test and the two-tailed test rejected analyses in which the mean of the assumed prior was either too small or too large. Analyses assuming the true model and branch-length prior have mean posterior predictive *P*-values close to 0.5 (directional tests; Supplementary Fig. S4) or higher (two-tailed tests; Fig. 7). The plausibility of mean tree lengths from analyses assuming an underparameterized model (JC) with the true branch-length prior was never

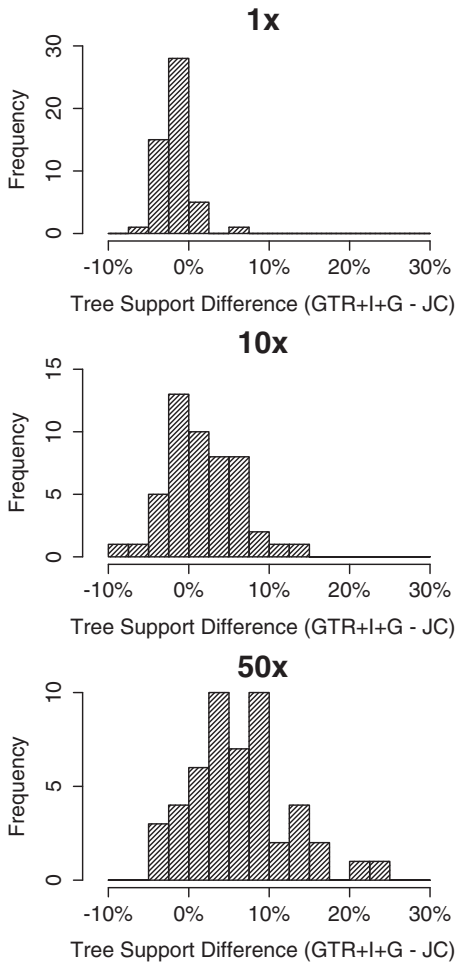

FIGURE 4. Differences in support for the true tree between the generating and underparameterized models across simulations with different expected tree lengths. Note that the generating model increasingly outperforms the underparameterized model as the expected tree length increases. These results are also presented in Figure 3, but this plot more clearly shows the frequencies of deviations in support across different simulation conditions.

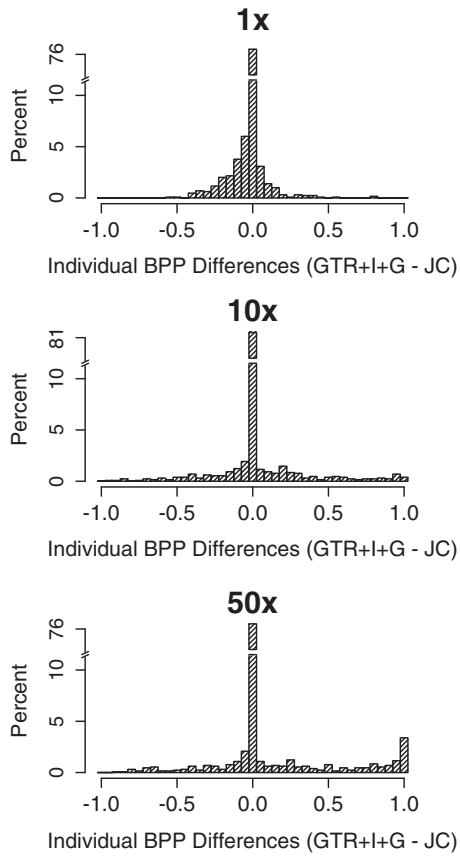

FIGURE 5. Differences in bipartition posterior probabilities (BPPs) when assuming the generating model or an underparameterized model. Examining support on a bipartition-specific basis more clearly indicates how the reduction in support for the true tree (Figs. 3 and 4) arises when assuming an underparameterized model. Note that the frequency of large errors for individual bipartitions increases as tree length is increased. Also note the discontinuous *y*-axis.

rejected, although mean posterior predictive *P*-values for the proper directional test do deviate more from 0.5 as the simulated tree length increases. This behavior is expected because overly simplistic models will tend to underestimate the amount of sequence evolution as multiple substitutions take place at the same sites (Fitch and Beintema 1990; Sanderson 1990).

Topological test statistics varied widely in the relationship between their posterior predictive *P*-values and the degree of topological error (Supplementary Figs. S5–S11). Quantile-based test statistics had the strongest relationship between posterior predictive *P*-values and topological error when positioned in the far right tail of the distribution (Fig. 8, Supplementary Figs. S5–S10). Visual inspection of symmetric differences showed that when the underparameterized model (JC) most often induced topological error (50x simulations), symmetric differences from analyses of original data sets usually went as high as 2 or 4. However, analyses of posterior predictive data sets nearly always sampled only a single topology. Test statistics that probe the upper extremes of the distribution of symmetric differences are most likely to detect such contrasts. As quantile-based test statistics, $T_{k.q}(\mathbf{X},M_c)$, were positioned more toward the center or lower extreme of the distribution, their power tended to decrease dramatically (Supplementary Figs. S5–S10). The relative performance of different statistics may depend on the original data and the consequent manner in which an incorrect model misinterprets its phylogenetic information.

The statistical entropy test statistic, $T_e(\mathbf{X},M_c)$, performed quite well (Supplementary Fig. S11), although with slightly less power than quantile-based test statistics positioned in the upper extreme (Fig. 8, Supplementary Fig. S10). This result might be explained by the fact that the statistic is measuring topological information gain, but only with regard to the relative probabilities of different topologies and not their distribution in tree space. The statistical entropy test statistic does have the desirable feature, unlike the

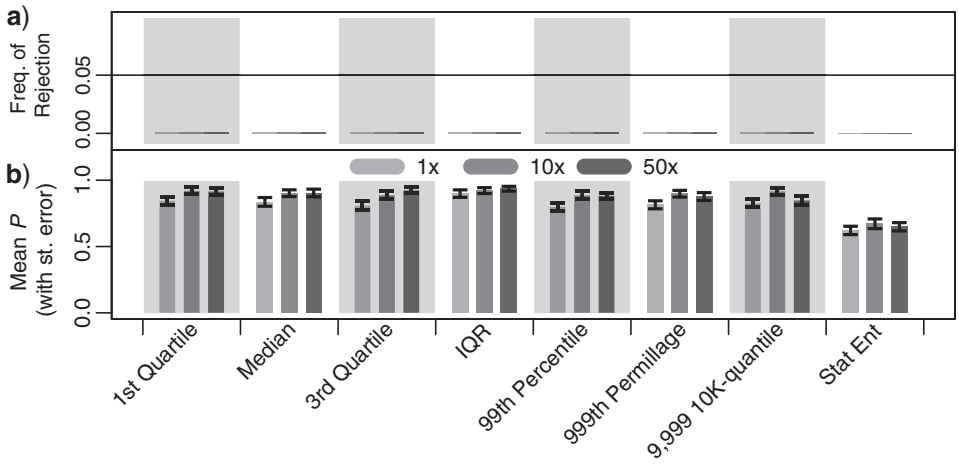

FIGURE 6. Performance of topological test statistics when the generating model is assumed during data analysis. a) Frequency with which inferences are assessed to be implausible when using a range of topological test statistics. Note that topological plausibility is never rejected with the two-tailed tests shown here. b) Mean posterior predictive *P*-values (± standard error) for analyses assuming the generating model and branch-length prior (GTR+I+Γ with the correct mean for the exponential branch-length prior). Each set of three bars corresponds to a different test statistic. All test statistics other than statistical entropy ("Stat Ent") are based on the position (or relative positions) of quantiles in the ordered vector of symmetric tree differences drawn from the posterior distribution. "IQR" is the inter-quartile range. "Stat Ent" is the topological information gained when moving from the prior to the posterior (see text for details). Only two-tailed test results are shown here. Results from upper and lower one-tailed tests are given in Supplementary Figure S3. Bar shading denotes the expected length of the tree along which data were simulated.

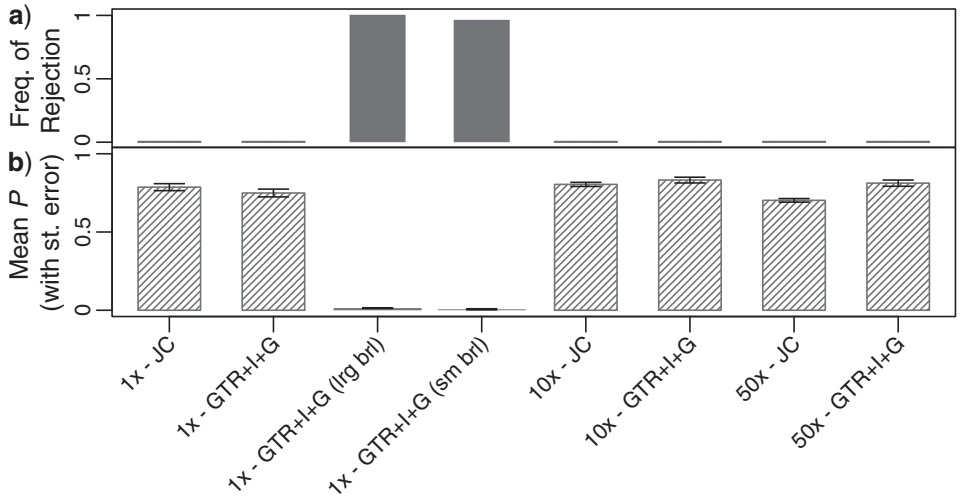

FIGURE 7. Performance of the mean tree-length test statistic. a) Frequency with which tree-length plausibility is rejected when assuming the generating model and true branch-length prior (GTR+I+Γ), the generating model and incorrect branch-length prior [GTR+I+Γ (lrg or sm brl)], or an oversimplified model (JC). b) Mean posterior predictive *P*-values (± standard error) for all analyses. Only two-tailed test results are given here. Results from upper and lower one-tailed tests are given in Supplementary Figure S4. Labels of 1x, 10x, or 50x denote the relative expected length of the tree on which data were simulated. GTR+I+Γ or JC denote the model assumed in the analyses. All analyses assumed the true branch-length prior unless denoted by lrg brl (mean of assumed branch-length prior is larger than the truth) or sm brl (mean of assumed branch-length prior is smaller than the truth). Results from analyses assuming the true branch-length prior are displayed as hatched bars, whereas results from analyses assuming an incorrect, informative branch-length prior are displayed as solid bars. Note that only the tree-length inferences from analyses with incorrect, informative branch-length priors are rejected as plausible by this test statistic.

quantile-based test statistics, of summarizing the entire distribution in a single value.

Empirical

Arthropods.—Using the multinomial likelihood test statistic, model plausibility was not rejected for any of the 27 arthropod genes (Supplementary Table S1). Posterior predictive *P*-values across genes ranged from 0.23 to 0.73. Topological test statistics that performed well in simulations (e.g., statistical entropy, $T_e(\mathbf{X},M_c)$, or upper extreme quantiles, $T_{k.q}(\mathbf{X},M_c)$) rejected inferential plausibility for many genes ($T_e(\mathbf{X},M_c)$: 6 genes; $T_{9.10}(\mathbf{X},M_c)$: 16 genes; $T_{99.100}(\mathbf{X},M_c)$: 13 genes;

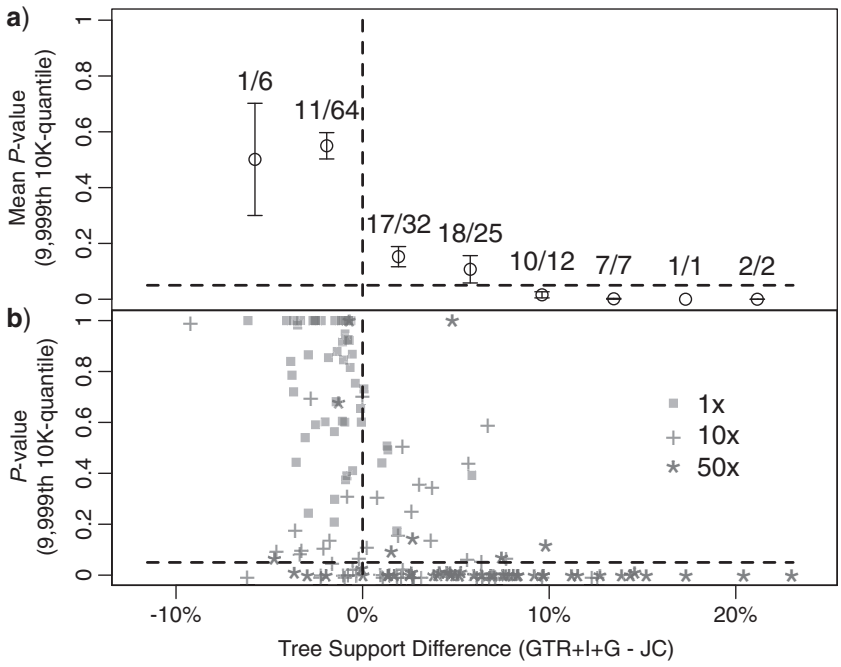

FIGURE 8. Relationship between posterior predictive *P*-values for topological plausibility [using $T_{9999.10,000}(\mathbf{X},M_c)$] from an oversimplified model (JC) and topological error induced by the oversimplified model. a) Analyses are binned by the difference in support between the generating and oversimplified model. The mean (± standard error) *P*-value is calculated for each bin. The frequency of rejection (with an arbitrarily chosen cutoff of 0.05) for analyses in each bin is given above the corresponding mean. b) The posterior predictive *P*-value and difference in support between models is plotted for each analysis individually. Points with different shapes and shading represent analyses of data sets simulated with different expected branch lengths. The horizontal dashed line indicates the conventional, frequentist *P*-value cutoff of 0.05. This cutoff is plotted merely for comparison and not due to an expectation that posterior predictive *P*-values should follow frequentist expectations. The vertical dashed line represents equal support for the true tree when assuming either the generating or oversimplified model. Topological error (tree support difference) is calculated as the difference in posterior probability between analyses assuming the generating and oversimplified models, summed across all bipartitions, and normalized by the maximum possible support for the true tree (i.e., the number of internal bipartitions). Note that the frequency with which inferential plausibility from the oversimplified model is rejected increases as the topological error induced by the oversimplified model increases.

$T_{999.1000}(\mathbf{X},M_c)$: 14 genes; $T_{9999.10,000}(\mathbf{X},M_c)$: 14 genes). Topological inference was assessed to be plausible across all test statistics for only 6 genes. Tree-length plausibility was never rejected using the mean tree-length test statistic.

The true arthropod phylogeny is not unambiguously known, so a gene's topological plausibility could not be compared with its ability to infer the true phylogeny. However, 100 trees were sampled from the posterior distribution of each gene and multidimensional scaling was used to plot their relative positions in a two-dimensional projection of tree space (Supplementary Fig. S12; Hillis et al. 2005). Although there was substantial overlap, genes producing plausible topological inferences (using $T_{9999.10,000}(\mathbf{X},M_c)$) did seem to sample different parts of tree space than those producing implausible inferences. Topologically plausible genes and topologically implausible genes were then concatenated separately and each data set was analyzed assuming a model partitioned by gene. The consensus topologies and bipartition posteriors from these two sets were relatively similar, although they differed strongly in the placement of two taxa (*Thulina stephaniae* and *Speleonectes tulumensis*).

Cricket frogs.—For *Acris*, plausibility of the concatenated data set was strongly rejected by the multinomial likelihood test statistic and plausibility of inferences was strongly rejected by the mean tree-length test statistic, as well as most of the topological test statistics (Supplementary Table S2). Tree-length implausibility was expected based on previous work (Brown et al. 2010). Interestingly, the plausibility of topological inferences was also rejected. This result is surprising given the shallow divergences among sampled sequences. Either unconsidered heterogeneity in the evolutionary process (i.e., differing models across subsets of sites in the concatenated alignment) or unreasonable branch-length priors may have significantly affected topological inference, although these possibilities were not specifically explored further.

Assumed constancy of the tree topology across genes may also have led to the rejection of plausible topological inference for the concatenated data set. If individual genes vary in topology, no single tree can adequately describe all of their evolutionary histories. To investigate the effect of concatenated analysis on the outcome of these tests, the plausibility of inferences was assessed for each gene individually. Although all inferences from

TABLE 1. Results of posterior predictive tests of plausibility for the mammalian mitochondrial data set of D'Erchia et al. (1996), reanalyzed by Sullivan and Swofford (1997)

Models	9.10	99.100	999.1000	9999. 10,000	IQR	Stat.Ent	TL.mean	MN
	Posterior predictive *P*-values							
GTR	1.00[a]	0.00[a]	0.00[a]	0.00[a]	1.00[a]	0.00[a]	0.80	0.00
GTR+I+Γ	0.00	0.00	0.00	0.00	0.00	0.02	0.86	0.26
GTR+I+Γ (cod.parts)	0.64	0.00	0.00	0.00	1.00	0.47	0.99	0.27
	Posterior predictive effect sizes							
GTR	IND	IND	IND	IND	IND	IND	0.22	24.33
GTR+I+Γ	6.11	5.05	4.71	5.87	4.57	3.06	0.22	0.65
GTR+I+Γ (cod.parts)	2.14	5.18	4.64	3.93	0.00	1.06	0.09	0.61

Note: This data set consists of the first and second codon positions from all protein-coding genes in the mitochondrion, with sites containing missing or ambiguous data removed. In the maximum-likelihood framework, D'Erchia et al. (1996) only analyzed these data using a GTR model. Sullivan and Swofford (1997) extended these analyses to consider a GTR+I+Γ model. Neither previous study partitioned models by codon position (cod.parts). The upper half of the table gives relevant posterior predictive *P*-values for each test statistic based on analyses assuming the models given in the rows. The lower half gives the effect sizes, or the distance between the empirical test statistic value and the median of the posterior predictive distribution standardized by its standard deviation. The first 4 columns show results from test statistics based on topological distance quantile positions. For the *k*-th *q*-quantile, these are written "*k.q*." "IQR" stands for inter-quartile range. "Stat.Ent" is the statistical entropy test statistic. "TL.mean" corresponds to the mean tree-length test statistic. "MN" corresponds to the multinomial likelihood. "IND" stands for invariant null distribution.

[a] *P*-value calculated from an invariant null distribution.

beta-crystallin were plausible, topological plausibility was rejected for the other three genes individually under at least two test statistics (Supplementary Table S2). In particular, topological plausibility was rejected under all quantile-based test statistics other than the interquartile range for the only mitochondrial gene in this data set, cytochrome *b*. The plausibility of inferred tree lengths was not rejected for any gene using the mean tree-length test statistic, despite the fact that this same statistic rejected plausibility for the concatenated data set. Therefore, to further test plausibility of the marginal tree-length distribution, its variance was used. Plausibility was rejected for cytochrome *b* when using the marginal tree-length variance test statistic, despite no rejection of plausibility for the concatenated data set when using this test statistic. The disconnect in test results pertaining to tree length between concatenated and individual gene analyses may be due simultaneously to an increase in the power of these tests when concatenating, as well as accommodation of heterogeneity in signals from the underlying genes when analyzing them separately.

Mammals.—For the mammalian mitochondrial data, the multinomial likelihood only rejected data set plausibility for analyses assuming GTR, not those assuming a homogeneous or partitioned GTR+I+Γ model (Table 1). In contrast, at least half of the topological test statistics rejected the plausibility of inferences under all three models. Although two of the topological test statistics (ninth decile and interquartile range) had two-tailed *P*-values of 1 for analyses that assumed GTR, the null distribution for these tests was invariant, suggesting that the lack of rejection was due to a lack of sensitivity. Plausibility of the mean tree length was not rejected for any analyses.

Although topological plausibility was strongly rejected for at least some test statistics under each model, the sizes of the effects might vary substantially. To explore the possibility that increasingly realistic models move inferences closer to plausibility, effect sizes were calculated as the difference between the empirical test statistic value and the median of the posterior predictive distribution, normalized by its standard deviation (Table 1). Because null distributions were frequently invariant for GTR, effect sizes could not always be calculated. Effect sizes for the partitioned GTR+I+Γ model were generally the same or smaller than corresponding effect sizes for the unpartitioned GTR+I+Γ model. Nonetheless, some empirical test statistic values (e.g., 99th percentile) were still up to 5 standard deviations away from the median, suggesting large differences in the spread of topological distributions between empirical and posterior predictive data sets.

DISCUSSION

Test statistics based on the marginal distributions of topology and branch lengths have the desirable property, in these simulations, of rejecting model plausibility more frequently in analyses with biased inferences. The multinomial likelihood does not share this property. Rather than relying on the appearance of data sets to test model plausibility, inferential test statistics directly examine marginal posterior distributions. These statistics are currently the only avenue available to systematists for directly testing if inferences are biased on a data set-specific basis. Instead of indirect arguments about whether a particular data set is affected by factors believed to bias inference based on properties of the data,

individual data–model combinations can be tested to see if the resulting inferences are plausible. However, these simulations are far from exhaustive. Much about the performance of these statistics across a wider range of parameter space remains to be understood. Nonetheless, the results presented here are promising and suggest that further work may be fruitful.

The performance of the newly proposed test statistics is likely an underestimate of their power to detect biased inference. Although similarity between models used to generate and analyze data is often seen as a weakness in simulation studies, this similarity actually makes detecting model implausibility more difficult. Realistic complexities of the evolutionary process should make biased inference, and rejection of a model's inferential plausibility, more likely. Tests of this intuition with more complicated simulations are desirable, particularly with regard to the strength of the relationship between model plausibility *P*-values and the degree of bias caused by different model violations. Unfortunately, because the goal of these statistics is to detect inferences biased relative to the generating model, rigorous benchmarking requires that the generating model can be used to analyze the data. This constraint may exclude many interesting models (e.g., Holder et al. 2008), which are likely to be more representative of empirical data.

Below I address some of the merits and drawbacks of these statistics relative to those in current, albeit rare, use. In particular, I highlight the new insights that these statistics provide, discuss the computational effort required to implement such tests, and suggest other statistics that are worthy of investigation.

Advantages of Inferential Test Statistics

Perhaps the biggest advantage of using marginal distributions as a basis for defining test statistics is that the test is based directly on the inference. The burden need not be on the researcher to decide if the implausibility of a data set's appearance (e.g., its multinomial likelihood) under the assumed model is sufficient reason to be suspicious of the resulting inferences. As is clear from the simulations performed in this study, certain types of model violations will result in data sets with very different distributions of site patterns, yet have little effect on estimation of the quantities of interest to most systematists: topology and branch lengths (Supplementary Figs. S1 and S2). In particular, biased branch-length priors can have a strong effect on posterior predictive *P*-values based on site pattern frequencies when the error in the resulting branch-length estimates is only a few percent, since branch lengths define the overall probability of change on the tree. However, even analyses with strongly biased branch-length estimates can produce accurate topological estimates (Brown et al. 2010; Marshall 2010).

By assessing plausibility using a range of inferential test statistics, researchers may gain greater insight into those specific inferences that may be compromised by poor fit between model and data. The empirical examples highlight this advantage. For the arthropod data, tests employing the multinomial likelihood never reject model plausibility, perhaps because its power depends strongly on the number of taxa (Bollback 2002). By employing test statistics based on the marginal distributions of topologies and tree lengths, we can see that tree lengths are plausible but topological inferences are implausible for many of the genes. For the cricket frog (*Acris*) data, the multinomial likelihood strongly rejects the plausibility of the data set. However, inferential tests show us that both the distributions of branch lengths and topologies are unexpected. Gene-specific analyses implicate cytochrome *b* as the primary source of branch-length misfit, whereas topological implausibility is more widespread. These analyses also indicate that topological problems are not simply caused by the concatenation of data with incongruent underlying topologies. The mammalian mitochondrial data (D'Erchia et al. 1996) are a classic example of poor fit between model and data leading to erroneous phylogenetic conclusions. Sullivan and Swofford (1997) demonstrated that modeling rate variation across sites (RAS) greatly reduced support for the placement of guinea pigs outside rodents. Results of model plausibility tests using the multinomial likelihood seem to corroborate Sullivan and Swofford's (1997) results, strongly rejecting plausibility of the homogeneous equal-rates model (GTR) but not a model that includes rate variation (GTR+I+Γ; Table 1, upper panel). Application of inferential test statistics suggests that mean tree lengths are plausible across all models, whereas topological distributions are not plausible for any of the assumed models (Table 1, upper panel). Comparing all three analyses, effect sizes for posterior predictive tests generally get smaller as model complexity increases, suggesting improved performance (Table 1, lower panel). Effect sizes for the multinomial likelihood highlight its strong sensitivity to RAS. Implausibility of topological distributions under all models indicates that some assumption(s) in common to these models is violated. Further, this violation is strong enough to bias the resulting inferences of topology but not branch lengths. None of the above statements regarding the reliability of particular conclusions would be possible using only relative model fit or posterior prediction with data-based comparisons.

Drawbacks of Inferential Test Statistics

Topological test statistics will be most useful when there is a reasonable amount of support for multiple topologies. These test statistics may lose a substantial amount of power when support for one topology becomes very strong. For instance, note the high variance in topological *P*-values resulting from analysis of the mammalian mitochondrial data using GTR (Table 1, upper panel). All of the corresponding null distributions were invariant and the two *P*-values of 1 were likely due

to a loss of power rather than genuine disagreement among test statistics. At least two possible remedies to this problem exist. The first is to test the fit of subsets of sites drawn from the original data. If the model is sufficient for inference across all sites, it should also perform properly when applied to a subset. However, this approach will not capture poor model fit caused by data concatenation. The second approach is to use a measure of topological support other than the posterior probability, since MCMC is not effective at estimating very small posteriors (Larget 2013). Marginal likelihood ratios (e.g., Bayes factors) comparing well-chosen topological hypotheses could provide such a measure. This proposal and other possibilities are discussed below.

Posterior predictive tests may also be conservative in detecting poor model fit (Bollback 2005; Ripplinger and Sullivan 2010; but see Waddell et al. (2009) for approaches to increasing the power of multinomial tests through binning). Because the tests I outline seek to avoid rejecting models when they do not result in biased inference, they may suffer from this same problem. For instance, when the generating model was used for data analysis, mean posterior predictive *P*-values were often greater than the value of 0.5 expected from an unbiased frequentist test (Supplementary Fig. S3). Additionally, for the best performing topological test statistics used in this study, mean posterior predictive *P*-values did not consistently fall below 0.05 until the error induced by the incorrect model was over 8–10% of the possible support for the true tree. This performance may not be powerful enough to satisfy some users. However, the simulations used in this study may give an overly conservative view of the power of these statistics, as discussed above. Additionally, the number of approaches currently used to assess model fit in an absolute sense is quite limited. Prudence would suggest application of a test with lower power than application of no test at all. More work needs to be done to quantify performance over a wider range of parameter space.

Computational Effort

For most practitioners, the greatest drawback of using statistics based on inferences will be the required computation time. Often the original analysis is a nontrivial undertaking and the prospect of repeating it at least 100 more times (per gene) is daunting. However, there are several reasons to believe that the overall computational burden is not so high as it first appears.

First, MCMC analyses of simulated data sets often converge much faster than analyses of corresponding empirical data sets. In fact, the initial motivation for using marginal distributions as test statistics arose from this observed difference in convergence behavior. When models of simulation and analysis are closely matched, posterior distributions tend to be conveniently unimodal. Empirical data sets exhibiting convergence difficulties will likely show the greatest speedup in this regard (sometimes an order of magnitude improvement, in my experience) and are also the most likely to poorly fit the assumed model.

Second, identification of data–model combinations with very poor fit should not require precise posterior estimation. Highly implausible empirical distributions will differ fundamentally from those estimated from simulated data. Such strong mismatches should be detectable with lower precision posterior estimates than would be standard for an analysis from which biological conclusions would be drawn.

Finally, posterior predictive analysis is highly parallelizable. Posterior estimation for each simulated data set can simply be allocated to an independent processor. Therefore, this approach is readily amenable to the use of high-performance computing clusters.

Other Inferential Test Statistics

The approach outlined here allows a great deal of flexibility in tailoring tests of fit to model components of most interest. Many other possible statistics exist for assessing the plausibility of phylogenetic inferences. For instance, the mean value for some metric of tree shape could be used to further query topological distributions. The position of quantiles in an ordered vector of tree-to-tree differences could also be used in conjunction with other tree difference metrics. Since accurate estimation of posterior distributions may be computationally expensive, other test statistics may provide useful approximations with a lower computational cost (e.g., the relative maximum or marginal likelihoods of a few, well-chosen topologies). Comparisons based on likelihood ratios also have the desirable property of being unbounded, which may increase their power.

Test statistics based on overall tree length assess the plausibility of inferred divergence across the entire tree, not the relative amounts of divergence expected across different branches. However, many downstream analyses (e.g., divergence time estimation) rely primarily on relative branch lengths. Future test statistics could be designed around relative measures of divergence or, better yet, could be based directly on inferences from the downstream analyses themselves.

Researchers are often primarily interested in one, or a small number of, specific relationships within a larger phylogenetic tree. A bipartition-specific analog of the statistical entropy statistic could be used to quantify the distribution of support across those topologies that contain a specific bipartition and those that do not. An even more sensitive approach might be to use the Bayes factor in favor of a bipartition.

Test statistics could also be designed around marginal distributions of parameters. For instance, statistics based on parameters describing RAS may prove useful in comparing models applied to the mammalian mitochondrial data set. Such an approach might also be useful when inferring nonsynonymous to synonymous rate ratios.

Site-specific inferences may also be of utility when testing for variation in fit across sites. Possible site-specific statistics might utilize likelihood scores in some fashion, either by drawing them from the posterior distribution or finding maximum site likelihoods for topological hypotheses chosen *a priori*. Other site-specific statistics might be based on sampled character histories (e.g., Nielsen 2002).

Conclusions

Model checking through posterior predictive simulation has the potential to meet an important need in phylogenetics. By employing the phylogenetic information contained in a data set as the basis for defining test statistics, researchers can gain a sense for the plausibility of various phylogenetic inferences. Absolute tests of model fit should not replace model choice tests, which may be more sensitive to violations of model assumptions. Rather, such tests should allow biologists to decide if the best-fit model is sufficient to provide plausible inferences. If fit between the chosen model and the data at hand is found to be poor, phylogenetic results need to be interpreted with caution or analyses with better fit between model and data should be preferred.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.nc866.

Funding

Financial support for portions of this work was provided by National Science Foundation graduate research and postdoctoral fellowships (DBI-0905867) to J.M.B., as well as start-up funds from the College of Science and Department of Biological Sciences at Louisiana State University.

Acknowledgments

I am grateful to B. Boussau, T. Heath, S. Höhna, J. Huelsenbeck, T. Keller, B. Moore, C. Nasrallah, D. Rabosky, F. Ronquist, S. Scarpino, M. Smith, and C. Wilke for discussions about appropriate test statistics. I thank the Center for Computational Biology and Bioinformatics at the University of Texas, in particular C. Wilke, and High-Performance Computing at Louisiana State University (HPC@LSU) for the use of computational resources. J. Andersen, F. Anderson, B. Boussau, V. Doyle, D. Hillis, M. Holder, A. Lemmon, E.J.B. McTavish, B. Nelson, B. Moore, J. Thorne, and one anonymous reviewer provided comments that significantly improved this article. S. Hird provided valuable guidance in the creation of Figure 1. J. Sullivan kindly provided mammalian mitochondrial data.

References

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.
Bollback J.P. 2005. Posterior mapping and posterior predictive distributions. In: R. Nielsen, editor. Statistical methods in molecular evolution. New York: Springer. p. 439–462.
Brandley M.C., Schmitz A., Reeder T.W. 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. Syst. Biol. 54:373–390.
Brown J.M., ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25:537–538.
Brown J.M., Lemmon A.R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. Syst. Biol. 56:643–655.
Brown J.M., Hedtke S.M., Lemmon A.R., Lemmon E.M. 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. Syst. Biol. 59:145–161.
D'Erchia A.M., Gissi C., Pesole G., Saccone C., Arnason U. 1996. The guinea-pig is not a rodent. Nature 381:597–600.
Fan Y., Wu R., Chen M.-H., Kuo L., Lewis P.O. 2011. Choosing among partition models in Bayesian phylogenetics. Mol. Biol. Evol. 28: 523–532.
Fitch W.M., Beintema J.J. 1990. Correcting parsimonious trees for unseen nucleotide substitutions: the effect of dense branching as exemplified by ribonuclease. Mol. Biol. Evol. 7:438–443.
Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.
Gamble T., Berendzen P.B., Shaffer H.B., Starkey D.E., Simons A.M. 2008. Species limits and phylogeography of North American cricket frogs (*Acris*: Hylidae). Mol. Phylogenet. Evol. 48:112–125.
Gelman A., Meng X.-L., Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. Statist. Sinica 6:733–807.
Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2004. Bayesian data analysis. 2nd ed. New York: Chapman & Hall/CRC.
Hillis D.M., Heath T.A., St. John K. 2005. Analysis and visualization of tree space. Syst. Biol. 54:471–482.
Holder M.T., Zwickl D.J., Dessimoz C. 2008. Evaluating the robustness of phylogenetic methods to among-site variability in substitution processes. Phil. Trans. R. Soc. B. 363:4013–4021.
Huelsenbeck J.P., Hillis D.M. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.
Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.
Huelsenbeck J.P., Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. Syst. Biol. 53: 904–913.
Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro M.N., editor. Mammalian protein metabolism. New York: Academic Press. p. 21–132.
Kelchner S.A., Thomas M.A. 2007. Model use in phylogenetics: nine key questions. Trends Ecol. Evol. 22:87–94.
Larget, B. 2013. The estimation of tree posterior probabilities using conditional clade probability distributions. Syst. Biol. 62:501–511.
Lartillot N., Brinkmann H., Philippe H. 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. BMC Evol. Biol. 7(Suppl. 1):S4.
Lartillot N., Philippe H. 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol. Biol. Evol. 21:1095–1109.
Lemmon A.R., Moriarty E.C. 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53:265–277.
Marshall D.C. 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: lost in the land of long trees. Syst. Biol. 59:108–117.

Minin V., Abdo Z., Joyce P., Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. Syst. Biol. 52:674–683.
Nielsen R. 2002. Mapping mutations on phylogenies. Syst. Biol. 51: 729–739.
Pagel M., Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst. Biol. 53:571–581.
Posada D., Buckley T.R. 2004. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53:793–808.
Rabeling C., Brown J.M., Verhaagh M. 2008. Newly discovered sister lineage sheds light on early ant evolution. Proc. Natl Acad. Sci. USA 105:14913–14917.
Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Regier J.C., Shultz J.W., Ganley A.R.D., Hussey A., Shi D., Ball B., Zwick A., Stajich J.E., Cummings M.P., Martin J.W., Cunningham C.W. 2008. Resolving arthropod phylogeny: exploring phylogenetic signal within 41 kb of protein-coding nuclear gene sequence. Syst. Biol. 57:920–938.
Reza F. 1961. An introduction to information theory. New York: McGraw-Hill.
Ripplinger J., Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. Mol. Biol. Evol. 27:2790–2803.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Rodrigue N., Kleinman C.L., Philippe H., Lartillot N. 2009. Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. Mol. Biol. Evol. 26:1663–1676.
Ronquist F. and J.P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.
Sanderson M.J. 1990. Estimating rates of speciation and evolution: a bias due to homoplasy. Cladistics 6:387–391.
Shannon C.E., Weaver W. 1949. The mathematical theory of communication. Urbana, IL: University of Illinois Press.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.
Sullivan J., Swofford D.L. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mol. Evol. 4:77–86.
Swofford D.L., Waddell P.J., Huelsenbeck J.P., Foster P.G., Lewis P.O., Rogers J.S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst. Biol. 50:525–539.
Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life Sci. 17:57–86.
Waddell P.J., Ota R., Penny D. 2009. Measuring fit of sequence data to phylogenetic model: gain of power using marginal tests. J. Mol. Evol. 69:289–299.
Whelan S. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. Mol. Biol. Evol. 25:1683–1694.
Xie W., Lewis P.O., Fan Y., Kuo L., Chen M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. Syst. Biol. 60:150–160.
Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol.39:306–314.
Yang Z., Goldman N., Friday A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.