ANNUAL REVIEWS

Annual Review of Ecology, Evolution, and Systematics

Evaluating Model Performance in Evolutionary Biology

Jeremy M. Brown[1] and Robert C. Thomson[2]

[1]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, Louisiana 70803, USA; email: jembrown@lsu.edu

[2]Department of Biology, University of Hawai'i, Honolulu, Hawai'i 96822, USA; email: thomsonr@hawaii.edu

Annu. Rev. Ecol. Evol. Syst. 2018. 49:95–114

First published as a Review in Advance on July 11, 2018

The *Annual Review of Ecology, Evolution, and Systematics* is online at ecolsys.annualreviews.org

https://doi.org/10.1146/annurev-ecolsys-110617-062249

Keywords

model fit, phylogenetics, comparative methods, posterior prediction, parametric bootstrap, cross-validation

Abstract

Many fields of evolutionary biology now depend on stochastic mathematical models. These models are valuable for their ability to formalize predictions in the face of uncertainty and provide a quantitative framework for testing hypotheses. However, no mathematical model will fully capture biological complexity. Instead, these models attempt to capture the important features of biological systems using relatively simple mathematical principles. These simplifications can allow us to focus on differences that are meaningful, while ignoring those that are not. However, simplification also requires assumptions, and to the extent that these are wrong, so is our ability to predict or compare. Here, we discuss approaches for evaluating the performance of evolutionary models in light of their assumptions by comparing them against reality. We highlight general approaches, how they are applied, and remaining opportunities. Absolute tests of fit, even when not explicitly framed as such, are fundamental to progress in understanding evolution.

1. INTRODUCTION

1.1. What Are Models and How Do We Use Them?

Random variable: a variable that takes on different values based on the outcome of a random process

Modeling is an exercise in explanation. A good model provides an accessible simplification of reality that both offers insight into a complex world and, at the same time, makes predictions that allow different explanations to be weighed against one another. The best models walk an idealized, if abstract, line. They capture enough reality to offer new insights, while discarding extraneous detail.

All sciences rely on models to generate hypotheses and testable predictions. Evolutionary biology is no exception. Our history began with some of the most insightful and poetic verbal models ever put forward (Darwin 1859). But Darwin (1887, p. 46) recognized the limitations of strictly verbal models: "I have deeply regretted that I did not proceed far enough at least to understand something of the great leading principles of mathematics, for men thus endowed seem to have an extra sense." Beginning with the Modern Synthesis (Mayr 1982), mathematics has been applied with great effect to formalize and extend Darwin's ideas. There is hardly an area of evolutionary biology that does not now rely on mathematical models.

Given the complexity of the processes involved, evolutionary models are nearly always stochastic. Some portion of these models relies on random variables to define predictions. Depending on a researcher's goals, the parameters that define these random variables can be considered focal or nuisance. Parameters with direct relevance to the hypotheses under consideration are focal, and those necessary for distinguishing among, but not directly related to, the hypotheses are nuisance. The same analysis could assign these labels to precisely opposite parts of a model, depending on the circumstance. For instance, phylogenetic inference often considers the details of sequence evolution to be nuisance and the tree topology to be focal [e.g., Ren et al. (2005), who inferred trees using codon models], whereas studies of molecular evolution may do the reverse [e.g., Yang & Nielsen (2002), who inferred sites under positive selection along particular lineages of a tree also using codon models].

As a simple example, imagine that researchers are interested in understanding the average size of members of a species, perhaps for comparison to the average of another species. In this instance, they may not have much interest in the variability of individual size within the species, but they may still employ a model that accommodates such variation (**Figure 1**). Other researchers, however, may not care about the average size of individuals, but they may be interested in within-species variation, perhaps to better understand the strength of stabilizing selection. If we assume a normal distribution of unknown mean and standard deviation as a model, the first researchers would consider the mean to be focal and the standard deviation nuisance, whereas the second would do the reverse. Both may also have implicitly considered the number of normal distributions to be a nuisance parameter, but fixed this value at one (**Figure 1*a***). For the example data in **Figure 1**, the use of a one-normal model is clearly problematic on the basis of a simple visual inspection, but for the sake of explication, we assume that these researchers have not performed such an inspection.

1.2. Fitting, Comparing, and Evaluating Models

Models help us learn in a variety of ways. For instance, models with unknown parameter values actually represent an entire family of models, with each specific set of parameter values corresponding to a different model. By inferring the values of these parameters, we are choosing one specific model out of this family. Or, if we infer the uncertainty in parameter values, we reduce the set of acceptable models within the family. For simplicity, we continue to refer to such model families simply as models with unknown (or free) parameters.

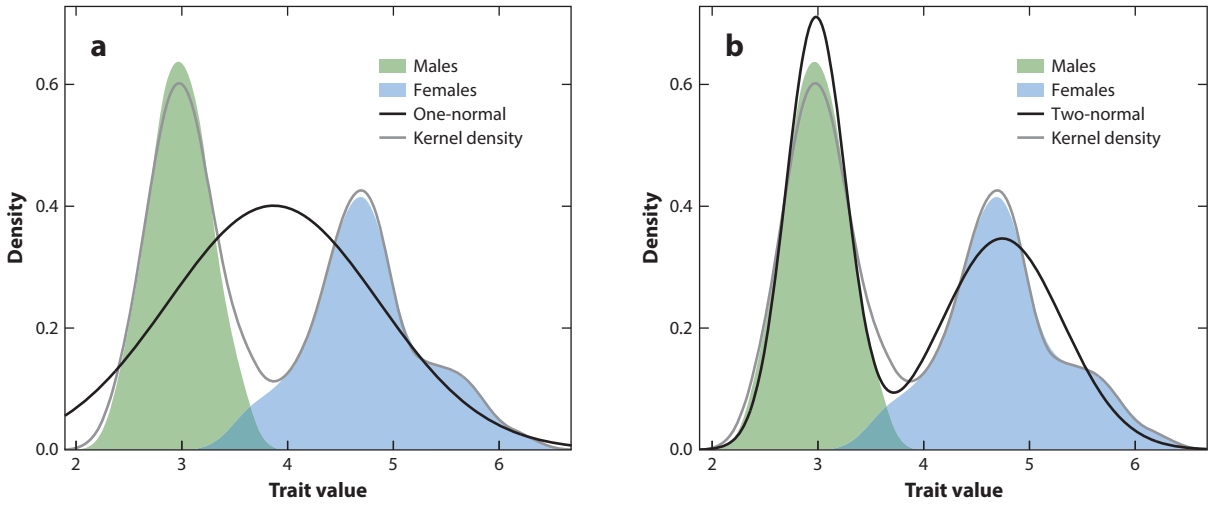

Figure 1

Hypothetical example of sexual dimorphism. Male (*green*) and female (*blue*) trait values are drawn from two normal distributions with different means and standard deviations. The light gray line is a kernel density estimate of the overall trait variation across the entire population. The black line in panel *a* shows a model assuming one-normal distribution with mean and standard deviation fit by maximum likelihood. The black line in panel *b* shows a mixture model of two underlying normal distributions with six parameters fit by maximum likelihood—two means, two standard deviations, and two mixture weights.

Models may also be compared with one another in terms of their relative fit to the data. In this context, fit often corresponds to the probability (or probability density) that a model assigns to observations (i.e., the model likelihood). If we infer the single set of model parameter values that provide the best fit, and then compare them against one another, we are using the method of maximum likelihood. However, we may also compare models by calculating the average fit across many reasonable parameter values. This approach is used in Bayesian inference, where the average fit (i.e., marginal likelihood) is weighted according to the prior probability assigned to different parameter values. In either case, if the fit of one model is substantially better than another, we have evidence that the preferred model is better able to predict some features of the empirical observations. If our models are thoughtfully constructed, their differences can help us understand the importance of some biological pattern or process. In this way, we learn through the comparison of relative model fit (Sullivan & Joyce 2005). Note, however, that biological hypotheses do not always (or even usually) have a one-to-one relationship with statistical models (discussed in McElreath 2016, pp. 5–7). Hypotheses and models must be carefully aligned for statistical results to inform biological understanding. For our example, we might wish to compare our original one-normal model (**Figure 1*a***) to a mixture model of two underlying normal distributions (**Figure 1*b***), perhaps because we suspect the possibility of sexual dimorphism. In this case, the observed trait values are 366,439,767,643 times more probable under a model with two normal distributions instead of one, when using maximum-likelihood parameter estimates. Although some correction is necessary because the two-normal model has more free parameters, and therefore more flexibility, it remains an overwhelmingly better explanation of these data. By selecting the two-normal model on the basis of its relatively better likelihood compared with that of the one-normal model, we have used a comparison of relative model fit to support hypotheses that can explain this distribution. Methods of relative model fit have received extensive attention in evolutionary biology (e.g., Posada & Crandall 2001, Minin et al. 2003, Posada & Buckley 2004, Sullivan & Joyce 2005).

Relative model fit: the ability of a model to predict or replicate important features of the data, in comparison to other available models

Although both parameter inference and evaluation of relative model fit are valuable tools for learning, they require the models defined by the researcher (or, perhaps more often, the software developer) to be able to account for all salient features in the data. If important aspects of the data-generating process are poorly understood, the available model set may not contain them. This situation is problematic for two reasons. First, if these shortcomings pertain to focal parts of the models, we have not given the data the ability to show us support for the true hypothesis. We have constrained our ability to learn by examining only models that inadequately describe the process that generated the data. Second, even if these shortcomings are not directly focal, nuisance portions can still be important in how we use our models to determine support for our hypotheses. Returning to our example, we may not be directly interested in the fact that our species is sexually dimorphic (**Figure 1**), but failing to consider a model that includes dimorphism can mislead us about the nature of trait variation in our population and, perhaps in turn, the strength of stabilizing selection. Therefore, a key feature in any framework for model-based learning is the ability to assess the fit of our models in an absolute sense and thus reject even the best model under consideration.

Absolute model fit: the ability of a model to replicate important features of the data as created by the true data-generating process

Parametric bootstrap: a frequentist approach to assessing absolute model fit by comparing an observed data set to data sets simulated using maximum-likelihood parameter estimates

Posterior prediction: Bayesian approach to assessing model fit by comparing observed data to data simulated using parameter values from the posterior distribution

1.3. How Do We Critically Evaluate Absolute Fit?

The importance of model realism has been widely appreciated for as long as stochastic models have been applied, and many approaches for assessing absolute model fit exist. Some rely on researcher intuition, for instance, regarding inferred parameter values (Gelman et al. 2013). Intuition may also be compared against data sets simulated from a model (i.e., through parametric bootstrapping or posterior prediction; described below). Do they have the features one would expect for a new empirical data set? Although these approaches provide useful sanity checks, by themselves they are subjective and prone to reinforcing researcher biases. Applied even to our simple example (**Figure 1**), these two approaches may not be very effective. If a simple visual summary was unavailable (often true for more complicated data sets and models) and we examined only the inferred parameter values or data sets simulated from them, all may appear well unless we have prior information about the expected distribution of trait values in this, or closely related, species.

The most straightforward formal statistical approaches to assessing absolute fit rely on analytical expectations. When available, one can apply standard tests (e.g., normally distributed residuals for linear regression) to compare features of the empirical data set to these expectations. For our example (**Figure 1**), a variety of tests of normality may be employed. The Shapiro-Wilk test (Shapiro & Wilk 1965) rejects the one-normal model ($P < 0.001$) (**Figure 1*a***). This approach is both useful and powerful, but it has some limitations. Analytical expectations may be unavailable for complex models or applicable to only a small set of assumptions. By limiting tests to only certain aspects of models, there is a danger of overconfidence if these tests accept the fit as reasonable.

More general approaches for assessing absolute model fit employ Monte Carlo simulation to define expectations (**Figure 2**). They use the model, along with fitted parameter values, to generate replicate data sets that are as similar as possible to the empirical data. The empirical data are then compared with the simulations to see if they represent a reasonable draw from the distribution defined by the model. When the single set of maximum-likelihood parameter values is used, this approach is commonly known as parametric bootstrapping (Efron & Tibshirani 1993). When parameter values are drawn from Bayesian posterior distributions, the approach is known as posterior predictive simulation (Rubin 1984; Meng 1994; Gelman et al. 1996, 2013; Gelman 2003). Despite deep philosophical divides between the statistical frameworks that underpin these two methods (a distinction that is beyond the scope of this article, but see Kass 2011 and Bayarri & Berger 2004 for modern perspectives on their interplay), they are procedurally similar and often

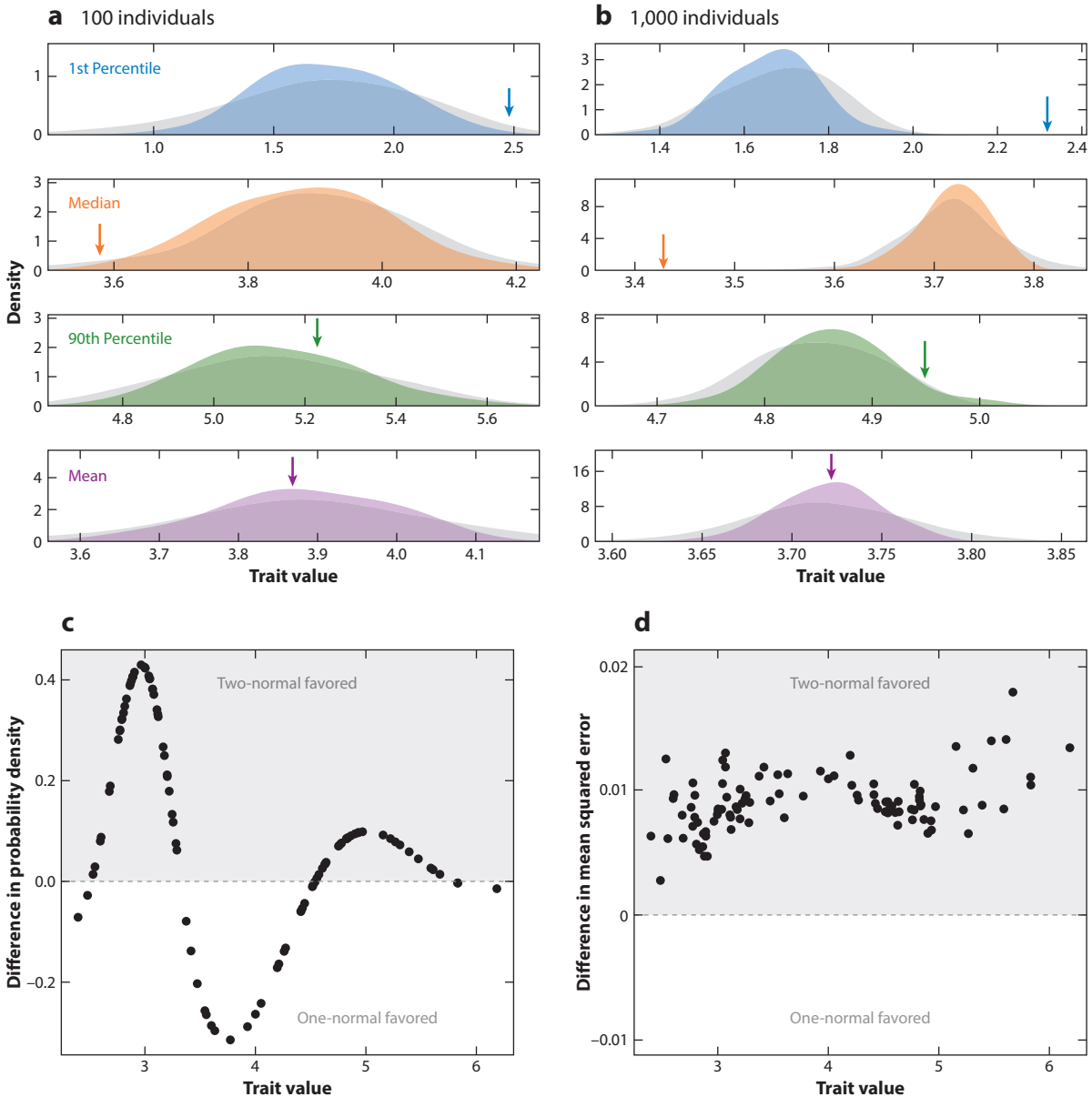

Figure 2

Example assessments of fit for the models shown in **Figure 1**. (*a*,*b*) Distributions of test statistic values generated by parametric bootstrapping (*color*) or posterior prediction (*gray*) for the one-normal model with a population size of either (*a*) 100 or (*b*) 1,000 individuals. Vertical arrows show empirical values. Rows contain different test statistics, each of which summarizes some aspect of the overall distribution of trait values in a population: (*top to bottom*) 1st percentile, median, 90th percentile, and mean. (*c*,*d*) Examples of leave-one-out cross-validation, in which one individual measurement is removed from the data set. Both the one- and two-normal models are then fit to the remaining trait values using maximum likelihood, and each fitted model is used to predict the single removed value. This process is repeated for each individual measurement in the population. (*c*) Difference in probability density between the two models for each measurement. (*d*) Difference in mean squared error between the models for each measurement. Each estimate of the mean squared error is based on 10,000,000 predicted values.

Test statistic: a numerical value used to characterize and compare data sets

Cross-validation: a model validation approach that involves fitting a model to data and evaluating the model's ability to predict independent data

produce similar distributions of simulated data (**Figure 2**). In both cases, however, the best method of comparing simulated and empirical data is not always obvious. Test statistics must be defined to quantify meaningful features of each data set and allow numeric comparison. The definition of these statistics, much like model design, relies on user intuition, experience, and goals. Once test statistics are defined, we can summarize results by calculating *P*-values as the percentage of simulations that resulted in test statistic values more extreme than the empirical (Meng 1994). When empirical test statistic values fall in the tails of the simulated distributions, *P*-values will be small, indicating poor absolute fit of the model to the data.

For illustrative purposes, we performed both parametric bootstrapping and posterior prediction for our example using four test statistics that summarize different parts of the trait distribution in each data set: the first percentile, median, 90th percentile, and mean (**Figure 2*a*,*b***). These results illustrate several general features of these tests. First, and most importantly, we are able to reject the fit of the one-normal model even without having defined a better alternative. This feature becomes increasingly important as the complexity of models and data increase, and good alternative models may not be obvious. Second, posterior predictive distributions tend to be broader than corresponding parametric bootstrapping distributions. By considering uncertainty in model parameter values, the variance of posterior predictive distributions will almost always be greater, rendering these tests more conservative. Third, the difference between statistical frameworks tends to be much less influential than the choice of test statistic. Whereas differences between posterior predictive and parametric bootstrapping *P*-values for each combination of data set and test statistic are minor, the differences between test statistics are large. For instance, in the 100-individual data set (**Figure 2*a***), the first percentile and median statistics decisively reject the fit of the one-normal model, whereas the 90th percentile and mean test statistics do not. Clearly, test statistic definition requires thought and a clear definition of analytical goals. For instance, the researcher focused on learning only about the trait mean might be comfortable with these results, while the other interested in within-species variation might be deeply concerned. Lastly, and not surprisingly, the power of these tests increases with data set size. Deviations between observed and expected test statistic values are larger when examining 1,000 individuals (**Figure 2*b***) than when looking at only 100 (**Figure 2*a***). Note, however, that the mean test statistic still suggests that the empirical data are entirely reasonable irrespective of sample size. This statistic would clearly be a poor choice if we want to ensure that our model does a good job explaining the entire distribution of trait values. In our experience, these four lessons generalize to many other applications of parametric bootstrapping and posterior prediction.

Cross-validation tests are another class of approaches for assessing model fit, but they focus on predictive ability (James et al. 2013). Rather than using the entire data set to fit the model, cross-validation divides the data into training and validation sets. The model is fit to the training set and then asked to predict some feature of the validation set. Many variants of this general procedure exist, defined by how much of the original data set is assigned to each set and how predictive ability is measured. In our example, we applied leave-one-out cross-validation (a data set of size n is divided into a training set of size $n - 1$ and a validation set of size 1) to each observation separately, using two measures of prediction: probability density and mean squared error (**Figure 2*c*,*d***). The two-normal model does a better job of assigning high probability near the two modes in the empirical distribution, whereas the one-normal model does better at predicting those values near its single mode (**Figure 2*c***). Overall, the two-normal model does better because there are more observations in the data set near its modes. When using mean squared error to measure predictive ability (which we estimated by predicting 10,000,000 trait values for each observation and averaging the squared error), the pattern of predictive ability is more equivocal. The two-normal model is always better than the one-normal model, but only slightly so. This result

highlights an important lesson about how choices in the metric of model assessment can affect conclusions. Even though the one-normal model is clearly not the model that generated the data, it tends not to predict values that are especially far away from the observed, so its mean squared error remains approximately the same as that of the two-normal model. If prediction of individual future observations is the goal, both models should perform about the same. A pragmatic challenge of cross-validation tests is the requirement to fit a separate model for each training set, although clever analytical solutions can avoid this burden in some circumstances (e.g., Lewis et al. 2014).

Below we review the application of these techniques in some areas of evolutionary biology. We primarily focus at or above the species level: molecular evolution, phylogenetics, species delimitation, divergence-time estimation, and comparative methods for traits and diversification. We do so largely because the models typically used in these fields are unified and there is an identifiable history of development for the approaches we discuss. However, we note that these same approaches are broadly applicable and have been (or could be) used successfully in many other fields, including population genetics, quantitative genetics, hybrid zone analyses, evolutionary ecology, and epidemiology.

2. SOME AREAS OF APPLICATION IN EVOLUTIONARY BIOLOGY

2.1. Gene Trees and Molecular Evolution

Models of sequence evolution have arguably received the most focused attention in terms of model evaluation of any application area in evolutionary biology (Felsenstein 2004, Yang 2014). Studies have repeatedly demonstrated that accurate phylogenetic inference requires realistic models of sequence evolution (e.g., Huelsenbeck 1995, Huelsenbeck & Rannala 2004, Lemmon & Moriarty 2004, Yang & Rannala 2005, Brown & Lemmon 2007). Developed nearly 50 years ago (Jukes & Cantor 1969), oversimplifications of the earliest sequence evolution models had some obvious disconnects with biological reality (e.g., assumptions of equal nucleotide frequencies), leading to a period of rapid model development as new extensions better accommodated the complexity observed in empirical data sets (reviewed in Felsenstein 2004). The general model structure for sequence evolution that is still employed in most modern phylogenetic analyses (the general time-reversible family of models) was largely developed prior to the mid-1990s (e.g., Tavaré 1986, Yang 1994). Since that time, various methods for accommodating mixtures of these models have received considerable attention (e.g., Lartillot & Philippe 2004, Nylander et al. 2004, Pagel & Meade 2004, Brown & Lemmon 2007, Zhou et al. 2010, Lanfear et al. 2012).

Following several early efforts (e.g., Ritland & Clegg 1987, Kishino & Hasegawa 1990, Navidi et al. 1991), Goldman (1993a) developed the first general and firmly grounded statistical approach to assess the absolute fit of sequence evolution models (see also Reeves 1992). Goldman's (1993a) approach had several advantages over previous efforts: (*a*) It tested the overall adequacy of the model to produce the data, rather than specific model assumptions; (*b*) it did not require the tree or branch lengths to be known ahead of time; (*c*) it did not rely on an analytical distribution against which to compare; and (*d*) by employing a likelihood framework, it was able to use the full data set, rather than a summary, which gave greater power.

Goldman's (1993a) approach, which is a form of parametric bootstrapping and based on a more general procedure given by Cox (1961), involves generating replicate data sets from the composite null hypothesis formed by a phylogeny and specified model of sequence evolution. Neither the phylogeny nor the model parameters are assumed known. Instead, maximum-likelihood estimates are used, and the test statistic is the difference in log-likelihood between the phylogenetic and an unconstrained, multinomial model. Essentially, this statistic gives the loss in likelihood required

to explain the data as having evolved along a bifurcating tree under the assumed model of sequence evolution.

Site pattern: states of a character for all taxa in a phylogeny, usually one column in an aligned set of molecular sequences

Stochastic map: a history of character change along a tree sampled from the distribution of such histories defined by a statistical model

Despite its advantages and statistical grounding, Goldman's (1993a) approach did not become standard practice in empirical phylogenetic studies. Several reasons probably contributed to this lack of adoption, especially given the time of its proposal. First, generation of the null distribution requires both Monte Carlo simulation of replicate data sets and independent maximum-likelihood inference of parameter values for each replicate. These requirements impose a substantial computational burden. Second, data were expensive and time-consuming to gather. The logical next step after rejecting the adequacy of a model was not clear. The data were too valuable to discard, and no adequate model may have been available. Developing a new model was infeasible for most practitioners, and the test did not naturally suggest how existing models should be extended or replaced.

Shortly after proposing his general test of model fit, Goldman (1993b) proposed the use of three more specific test statistics that aimed to provide guidance about how models fit poorly—the number of invariable sites, the number of unique site patterns, and the number of site patterns that may have arisen owing to parallel evolution. These statistics were chosen because they could reveal specific ways in which model assumptions are violated. For instance, the number of invariable sites and the number of unique site patterns should be closely connected to the way that substitution rates vary across sites. This rate variation is presumably controlled by variation in selective constraint across sites, with highly constrained sites experiencing low rates of substitution and vice versa. We now recognize the importance of accommodating among-site rate variation, and standard practice in phylogenetics nearly always employs models that account for it (e.g., gamma-distributed rate models) (Yang 1994). Goldman's (1993b) parallel evolution test statistic was intended to highlight how positive selection could create misleading signal.

Soon after Bayesian methods of inference began to be adopted in phylogenetics (Huelsenbeck et al. 2001, 2002), Bollback (2002) proposed a posterior predictive test of overall model adequacy. As with the example in **Figure 2**, this posterior predictive test tends to be more conservative than its parametric bootstrapping counterpart (Ripplinger & Sullivan 2010). Bollback's (2002) proposal took inspiration for its choice of test statistic from Goldman (1993a) and employed the multinomial likelihood owing to its generality. However, the multinomial likelihood has since been criticized for being highly sensitive in some circumstances (e.g., to incorrect branch-length priors) (Brown 2014) and having low power in other circumstances (e.g., when the number of taxa is large) (Waddell et al. 2009, Duchêne et al. 2015). Waddell et al. (2009) proposed increasing the power of the test by binning site patterns, and Koch & Holder (2012) developed a novel algorithm to speed up these calculations.

Roughly coincident with Bollback's (2002) posterior predictive approach, Nielsen (2002) developed a method for sampling stochastic mappings of nucleotide changes along a phylogeny from the corresponding posterior distribution. He also simulated new mappings from the posterior predictive distribution and compared them with the posterior distributions conditioned on the observed data. To illustrate the potential of this approach, he compared both the variance in number of substitutions across sites and the ratio of nonsynonymous to synonymous substitutions. Framed primarily in terms of molecular evolutionary questions, this approach is closely related to the sequence-based test statistics proposed by Goldman (1993b) for phylogenetics. In addition, Nielsen (2002) outlined how inference of mappings for simulated data sets could be used to generate posterior predictive *P*-values, which were used by Nielsen & Huelsenbeck (2002) to detect positively selected sites.

Approaches for assessing fit have been applied periodically, although not frequently, since Bollback's (2002) proposal and mostly in a Bayesian framework. One of the earliest applications

was to test the assumption of constant base composition across a tree. Violations of this assumption have been implicated in spurious phylogenetic results (e.g., Foster & Hickey 1999). Huelsenbeck et al. (2001) described a posterior predictive approach using a χ^2 statistic to test models of homogeneous base composition, which was later extended by Foster (2004) to models with varying base composition. More recently, Duchêne et al. (2017) explored the relationship between the results of this test and misleading phylogenetic inferences.

Over several years, Lartillot, Philippe, and colleagues developed multiple new tests. Lartillot & Philippe (2004) used an inference-based approach in testing their CAT mixture model for accommodating among-site heterogeneity in amino acid frequencies. Comparing the number of mixture categories inferred from empirical data with those inferred from posterior predictive data, the authors found that the model stably recovers a large number of heterogeneous categories in both cases. Lartillot et al. (2007) used both cross-validation and posterior predictive approaches to assess the relative and absolute performance for their CAT model compared with a more standard WAG model (Whelan & Goldman 2001) when inferring metazoan phylogeny. For cross-validation, they divided their phylogenomic data set into halves and assessed the predictive ability of the two models to assign high likelihoods to each half as a validation set when trained on the other half. They also used posterior prediction to assess the ability of each model to replicate realistic numbers of substitutions and degrees of homoplasy [termed measures of saturation and calculated using Nielsen's (2002) method], as well as the sets of different amino acids found in each alignment column. In their tests, CAT consistently outperformed WAG. Rodrigue et al. (2007) focused on site-interdependent models of protein evolution and also used posterior prediction with test statistics that capture among-site rate variation as well as residue-specific exchangeabilities. They found that, despite some improvements in absolute fit when modeling site interdependence, elements of existing independent models (gamma-distributed rate variation and realistic exchangeabilities) could not be replaced by considering interactions alone.

More recently, Zhou et al. (2010) implemented a mixture model for covarion processes (a form of heterotachy where sites switch between on and off states) and developed three discrepancy variables to assess the fit of their model with respect to rate variation across sites and branches in a posterior predictive framework. These variables also rely on substitution mapping. Compared with standard test statistics, discrepancy variables are functions of both the data and the parameter(s) (Gelman et al. 1996). Zhou et al. (2010) applied these tests to five empirical data sets and demonstrated both support for and the utility of covarion processes. This study is one of the few in phylogenetics to employ discrepancy variables for posterior prediction. Future work could investigate how discrepancy variables in general perform relative to test statistics based only on data.

As an approach to assessing specifically when inadequate models of sequence evolution might compromise inference of reliable gene trees, Brown (2014) proposed a set of posterior predictive test statistics that were inference based. Rather than using site mappings, Brown's (2014) approach used statistics based on inferred posterior distributions of trees (e.g., the dispersion of trees in tree space). One advantage of this approach is its direct dependence on focal inferences, meaning that a model will be assessed as inadequate only if it has a noticeable influence on the resulting inference. Estimation of posterior distributions for all the simulated data sets can be time-consuming, but faster approximations (e.g., likelihood ratios based on a few targeted topologies) may lower the computational burden. One application of these tests compared model fit for different genes in each of two phylogenomic data sets and concluded that focusing on genes with better absolute fit resulted in trees that are more consistent and more accurate (Doyle et al. 2015). This study also suggested the use of effect sizes, rather than posterior predictive *P*-values, as a way to compare absolute model fit in a more meaningful way. Posterior predictive effect sizes are calculated by

Covarion: a model of molecular sequence evolution where codons can switch between on (changing) and off (unchanging) states across a tree

Discrepancy variable: a numerical value used to characterize data sets in a Bayesian analysis, whose value depends on data and corresponding parameter values

Heterotachy: character-specific changes in rate of evolution across branches in a phylogenetic tree

Gene tree: a tree that represents the history of allelic descent within a population (or populations)

examining how far away an empirical test statistic is from the median of the posterior predictive distribution, normalized by the standard deviation of the distribution. In this way, tests that all give *P*-values near zero can be distinguished by the degree of mismatch between the empirical and expected test quantities.

Species tree: a phylogenetic tree representing relationships among species, which may differ from gene trees for genes sampled from those species' genomes

Lewis et al. (2014) proposed a novel cross-validation approach for Bayesian phylogenetics, by generalizing a technique known as conditional predictive ordinates (CPOs) (e.g., Chen et al. 2000). CPO values represent the probability of site i conditioned on the rest of the data, $p(\mathbf{y}_i|\mathbf{y}_{(-i)})$, where $\mathbf{y}$ is the alignment. Conveniently, CPO values can be estimated as the harmonic mean of the site likelihoods resulting from Markov chain Monte Carlo (MCMC) analysis of the entire data set. Lewis et al. (2014) advocated the use of these values as an exploratory technique to suggest modeling strategies, although CPO values tend to be strongly influenced by a site's rate of evolution (i.e., sites with higher rates are harder to predict). CPO scores can also be combined into a pseudomarginal likelihood that is easily calculated from a single MCMC run and can be compared across models.

2.2. Species Trees and Species Delimitation

All the studies mentioned above focus on the fit of models of sequence evolution either to infer accurate gene trees or because they are testing molecular evolutionary hypotheses. However, methods for inferring species-level phylogenies now recognize that gene trees are expected to vary, at a minimum due to variation in patterns of coalescence (Maddison 1997). Several software packages now explicitly model coalescent variation when inferring species histories (*BEAST, MrBayes, RevBayes) (Edwards et al. 2007, Heled & Drummond 2010, Ronquist et al. 2012, Höhna et al. 2016). These models are hierarchical in nature. Probability distributions on species trees influence the distribution of gene trees, which in turn inform the interpretation of sequence data. Assumed distributions at each level of this hierarchy can influence the overall species tree inference.

Joly et al. (2009) proposed a test of absolute model fit for multispecies coalescent models, with the goal of identifying introgression. This test is an example of assessing a specific violation of model fit to provide information about focal hypotheses. Their chosen test statistic is the minimum sequence divergence between two species, and they used posterior prediction to generate the expected distribution of this statistic under incomplete lineage sorting alone. This approach seems to have good statistical properties. However, the ability to reliably infer hybridization when the coalescent model is rejected assumes that gene trees are well estimated and other coalescent assumptions are met.

Reid et al. (2014) also used a posterior predictive framework to investigate the multispecies coalescent model, but with the aim of testing the reliability of species tree inference, rather than detecting introgression. They designed two sets of test statistics to examine fit for two levels of the hierarchical model. One set tests the fit of gene trees to the distribution imposed by the species tree, and the other set examines fit between sequence alignments and gene trees. At this lower level, the statistics draw heavily from Goldman (1993a,b). Applying these tests to a series of empirical data sets, Reid et al. (2014) found some evidence of poor fit at the level of coalescent genealogies, but much more evidence of poor fit using the sequence-based tests. Interestingly, they showed that in some cases removing loci with evidence of poor fit changed the inferred species tree. Additionally, some evidence of poor fit at the sequence level seemed to be driven by coalescent assumptions. When gene trees were inferred independently, sequence-based tests no longer showed evidence of poor fit. On the basis of these results, Reid et al. (2014) recommended removing loci that are inferred to violate assumptions of the coalescent process and rerunning species-tree analyses.

They also called for increased development of approaches that incorporate multiple sources of gene tree variation. These tests are now available in an R package (Gruenstaeudl et al. 2015).

The multispecies coalescent model has also been extended to delimit species (Yang & Rannala 2010, Yang & Rannala 2014, Rannala & Yang 2017). Barley et al. (2017) proposed posterior predictive tests specific to species delimitation. Their test statistics are inference based and closely related to those proposed by Brown (2014). Application of these new tests to a series of empirical data sets showed a range of model fit, with poorest fit for those data sets that had previously been suggested to violate coalescent assumptions. Barley & Thomson (2016) also applied both data- and inference-based test statistics to species delimitation in the context of DNA barcoding, where decisions about species boundaries are primarily based on levels of sequence divergence. They showed that the commonly assumed K2P (Kimura 1980) model often shows poor fit to barcoding sequence data, whereas other models from the general time-reversible family provide much better fit. Importantly, different models also change the inferred number of species.

Data-based test statistic: a numerical value used to characterize data sets, whose value depends on the data alone

Inference-based test statistic: a numerical value used to characterize data sets, whose value depends on both the data set and the model assumed for analysis

2.3. Inference of Divergence Times

Molecular clock models provide a description of how substitution rates vary (or not) across the branches of a phylogenetic tree. These methods began with the molecular clock hypothesis of Zuckerkandl & Pauling (1962, 1965), which posits that all branches in a tree share a single evolutionary rate. This so-called strict clock allows for the estimation of branch lengths in terms of relative time, or if independent information about absolute time is incorporated (e.g., by calibrating node ages using fossils), it is possible to convert branch lengths from relative to absolute time. Rates of molecular substitution are now known to vary across the tree of life, which has motivated the development of a wide range of relaxed clock models that allow rates of substitution to vary in diverse ways (reviewed in Heath & Moore 2014, Ho & Duchêne 2014). This development has been driven by the desire to develop models that provide a closer description of biological reality. However, the diversity of available models, coupled with the observation that substitution rates may vary both across an alignment and across branches, makes selection and assessment of clock models a challenge.

Duchêne et al. (2015) introduced approaches to assess the adequacy of clock models. They used posterior predictive simulation and inference-based statistics to identify branch lengths that may be implausible under a particular clock model. The approach begins by sampling from the joint posterior distribution for a substitution and clock model. This provides samples from the marginal posterior distributions for branch-specific rates and time. For each branch, samples are uniformly chosen from the marginal distribution of these two parameters and multiplied, yielding (for each branch) a product in units of substitutions per site. These products are taken as the branch lengths of a phylogram and used, along with samples from the remaining parameters of the substitution model and the tree, to simulate new data sets. Finally, a phylogram is inferred for each of these simulated data sets as well as for the initial empirical data set.

This procedure allows adequacy of the clock model to be assessed in multiple ways. For example, the length of each branch in the empirical phylogram can be compared with the distribution of lengths for the corresponding branch in the posterior predictive phylograms, allowing calculation of a *P*-value for each branch. Using simulations, Duchêne et al. (2015) identified bias that arises from several forms of model violation including overly simplistic substitution models, clock models, and other aspects of the analysis such as misleading node age calibrations. In empirical analyses, these authors detected substantial variation in model fit across four different data sets. The method was able to highlight particular branches in the empirical trees that may be reliable even when there was overall evidence of bias. Several extensions to these methods should be

possible, including integration across tree topologies, application to the large and growing set of available clock models, and further exploration of possible test statistics.

The fit of clock models has also been explored using cross-validation methods analogous to those developed for substitution models (Duchêne et al. 2016). This approach divides an empirical alignment into training and validation sets, uses MCMC to collect samples from the joint posterior distribution for the model conditioned on the training data set, and then uses these samples to compute the phylogenetic likelihood of the validation set. This approach can be used for model selection, choosing the model that yields the best average likelihood for the validation sets, or to assess the absolute fit of the model, by asking how well the trained model can predict the features of the validation data. Both the posterior predictive and cross-validation methods highlight that model underparameterization, in particular, is likely to result in biased inferences (Duchêne et al. 2015, 2016), similar to what has been observed for models of substitution (Huelsenbeck & Rannala 2004).

2.4. Model Adequacy for Comparative Methods

Explicit assessments of model adequacy are increasingly being applied to phylogenetic comparative methods. Inferences about how diversification rates vary, how traits evolve, or how the two may be linked depend critically on stochastic models, and all may be biased when the corresponding models are inadequate. For example, stochastic models of the branching process are well known to fit empirical trees poorly. Phylogenies estimated from empirical data tend to be more imbalanced and contain deeper branching times than expected under commonly used homogenous branching process models (Heard 1992, Mooers 1995, Blum & François 2006, Etienne & Rosindell 2012, Stadler et al. 2016). This mismatch has prompted several studies into potential causes investigating possible deficiencies in diversification models as a way to guide model elaboration (e.g., Heard & Mooers 2002) or potential biases that arise in how data sets are assembled and analyzed (e.g., Heath et al. 2008).

Recently, researchers have developed explicit assessments of model adequacy that allow for measurement of the fit of diversification models to individual empirical data sets. For example, Höhna et al. (2015) implemented a set of flexible models that allow diversification rates to vary continuously or episodically through time. In addition, they implemented posterior prediction using test statistics based on tree shape or other aspects of the inference. Although these methods have yet to be widely applied to analyses of diversification rates, they hold promise for enhancing the accuracy and reliability of studies of diversification rate variation.

For models of trait evolution, recognition of the importance of model adequacy dates to some of the earliest papers in comparative methods. Mere paragraphs after describing the justification and procedure for calculating phylogenetically independent contrasts, Felsenstein (1985) raised a series of reasonable concerns about the adequacy of Brownian motion as a model for the evolution of continuous traits. A large set of later studies explored the absolute fit of models applied to traits (see Pennell et al. 2015 for a summary) either by using analytical tests based on specific assumptions (e.g., Garland et al. 1992) or through more general Monte Carlo–based tests (e.g., Slater & Pennell 2014). Pennell et al. (2015) explored posterior prediction to assess the adequacy of trait models, both in simulation and with 337 empirical data sets for angiosperm functional traits. They failed to detect some form of bias in only 133 of the 337 sets and found model violations far more frequently than in simulation (where violations are relatively simple and known as part of the study design). Intuition suggests that larger phylogenies contain a more complicated and heterogeneous mixture of evolutionary processes, making the analysis of traits on such phylogenies more difficult. Accordingly, Pennell et al. (2015) found that model violations were more frequently

detectable for larger clades. This finding could be driven by increased power in larger data sets, but further development and application of these tests is clearly warranted, particularly as the field moves to analyses of larger segments of the tree of life.

One of the most conspicuous areas of recent progress in comparative methods includes models that recognize that traits may influence the probabilities of speciation and extinction (Maddison 2006). In these cases, attempts to model the evolution of traits or diversification dynamics without accounting for their interdependence are prone to being seriously misled. A growing family of state-dependent speciation and extinction (SSE) models that allow diversification rates to vary as a function of traits or geography has been proposed (Maddison et al. 2007; Fitzjohn et al. 2009; Fitzjohn 2010, 2012; Goldberg et al. 2011). Though specifically proposed to address a major potential source of bias in empirical analyses, SSE models also highlight the importance of ongoing evaluations of model performance. Rabosky & Goldberg (2015) pointed out that SSE models assume all diversification rate heterogeneity in a tree arises in association with the trait being modeled. If diversification rates vary at all (regardless of whether linked to a trait), an SSE model that allows for diversification rate variation will often be selected over simpler homogenous rate models. This may lead researchers to the erroneous conclusion that diversification is mediated by a trait when it is not.

Recognition of these issues has led to SSE model extensions that seek to (at least partially) alleviate the problem (Beaulieu et al. 2013, Beaulieu & O'Meara 2016, Rabosky & Goldberg 2017). Beaulieu & O'Meara (2016) extended the simplest binary SSE model to allow diversification rates to vary in ways not related to the trait under study. This hidden SSE model helps alleviate the inability of binary SSE models to capture complicated changes in diversification rates and provides a useful set of related null models that allow true character-dependent diversification to be distinguished from false positives (Beaulieu & O'Meara 2016, Caetano et al. 2018).

The growing availability of complex and realistic models of character-dependent diversification provides an example of the role that evaluating model fit plays in evolutionary biology. These studies have not always been carried out as explicit assessments of model adequacy (using, e.g., cross-validation or posterior prediction). However, each major step of progress arose from the recognition that available models did not capture biological reality in some important way, demonstrated the bias that resulted, and then led to model elaboration that sought to alleviate the bias.

3. PROSPECTS AND OPPORTUNITIES

3.1. How Can We Learn About Evolution by Assessing Absolute Model Fit?

Evaluation of model performance has a long history in many areas of applied statistics and deserves a place in the standard tool kit of evolutionary biologists. Assessments of absolute model fit can lead to greater biological insight or improve modeling efforts in a number of ways. Precisely how to best use these results depends on the nature of the tests and the goals of the study. Generally, we can categorize the tests as assessing overall or specific fit, as applying to the entire data set (e.g., parametric bootstrapping and posterior prediction) or a specific subset of the data (e.g., cross-validation), and we can categorize their application as pertaining to either the focal or nuisance portions of a model.

Ideally, we would have available for our model of interest an entire suite of tests that have been thoughtfully chosen in light of our research goals. As one example, imagine that we are conducting a multispecies, coalescent-based phylogenetic analysis to infer a species tree. The alignments are informative, and tests of fit indicate that our models of sequence evolution seem

to be sound, meaning that our gene tree estimates are informative and reliable. However, tests of fit for the coalescent process indicate that alleles at some loci consistently coalesce earlier than expected given inferred species boundaries. Cross-validation tests could also identify the loci that are the strongest outliers relative to coalescent expectations. Even without an explicit model of gene flow, we might conclude that introgression has occurred and highlight the loci that probably introgressed. Alternatively, we can extend our modeling framework to accommodate our new insight. In this case, we could build models of gene flow that would accommodate the type of early coalescences seen when rejecting the coalescence-only model.

Such applications are now feasible for many types of evolutionary analyses. An increasingly rich and flexible set of tools is being developed (e.g., Brown & ElDabaje 2009; Zhou et al. 2010; Höhna et al. 2015, 2017; Duchêne et al. 2018) both to conduct tests of model fit and to extend the available set of evolutionary models (Höhna et al. 2016). Coupled with the increasing size and diversity of available data, these tools provide an opportunity to increase both the depth and rigor of the conclusions we draw. Some considerations to bear in mind when applying such approaches include the matching between the test statistic and the question of interest (i.e., will the outcome of this test narrow the range of plausible biological hypotheses or suggest unexplored directions), the power of the test, whether the test is of overall fit or a specific assumption, and whether the primary use of the model is inference based on current observations or prediction of future observations. If the goal is inference from current observations, one may wish to use analytical tests of assumptions (when available), parametric bootstrapping, or posterior prediction to evaluate model fit. If prediction is more important, cross-validation may be more useful.

3.2. Remaining Opportunities and Research Priorities

Assessment of absolute model fit has the potential to be much more broadly and creatively applied than is currently the norm. Here, we outline some ideas that strike us as particularly timely.

Some analyses benefit from including different types of data (e.g., sequences and morphology for phylogenetic inference). Different models will often be necessary for different data types, and model fit may vary substantially. By assessing fit for all data types independently, we can identify those that are most likely to produce misleading inferences. This insight into varying model fit may lead us to downweight certain conclusions or to prioritize improvements for those combinations of models and data types that match most poorly.

Some evolutionary processes may be difficult to model explicitly, but they may exhibit consistent and recognizable departures from a simpler model. By employing tests of model fit that look specifically for these recognizable signatures, we can begin to understand how common and strong these processes might be. For instance, convergent molecular evolution can be difficult to model, since we often do not understand when taxa will experience similar selective pressures at the molecular level or, when they do, how similar the evolutionary response will be. However, we can anticipate that convergence will produce more similarity in distantly related taxa than expected by chance (e.g., Goldman 1993b), and these patterns are clearly identifiable in some circumstances (e.g., Castoe et al. 2009). By applying genome-wide tests of model fit that look specifically for an excess of these types of sites, we can gain a better understanding of how often convergence occurs without explicit models.

A third application of these tests is to attempt to resolve long-standing conflicts where different methods or data sets consistently and strongly support different answers (e.g., the debate over whether sponges or ctenophores are sister to the remaining animals) (Lartillot et al. 2007, Lartillot & Philippe 2008; reviewed in Dunn et al. 2014). Continuing attempts to address many

of these questions rely primarily on generating new or larger data sets, but with little change in the underlying models. Especially for challenging questions, overly simplistic models seem likely to drive much of the conflict. Critical and creative evaluation of model fit with existing data may offer new insights and more rapidly bring about consensus.

Several aspects of model performance tests are also understudied, and additional, focused investigation offers the potential to expand the scope and application of these tests. When model extension is difficult, what is the best way to use these tests? One possibility is to focus attention on subsets of the data that already match model assumptions well (e.g., Doyle et al. 2015). This strategy should work best when we expect our observations to result from the same process. Examples include genes that we expect to have evolved along the same tree topology (e.g., Richards et al. 2017) or fossils from the same fossilized birth-death process. Outlying genes or fossils may have been subject to unique evolutionary processes and could bias inferences if used to try to understand a process from which they were not drawn (e.g., Brown & Thomson 2017). Identification of these outliers also presents the opportunity to understand the underlying processes that led to their unique pattern.

Another area of opportunity relates to the evaluation of hierarchical models (e.g., Reid et al. 2014, Duchêne et al. 2015), most common in Bayesian inference. The levels of these models (e.g., see **Figure 1** of Reid et al. 2014) are coupled, such that changes at one level (e.g., different prior distributions) can have cascading effects at others. Reid et al. (2014) investigated species-tree inference under a multispecies coalescent model using tests that focused on different levels of the model hierarchy. In their tests, poor fit between coalescent expectations and the actual distribution of gene trees could appear at the level of sequence evolution, if the coalescent prior had a strong influence on the inferred gene tree. Conversely, a poorly specified model of sequence evolution could produce an erroneous gene tree that then affects the inferred species tree. At present, it is unclear when we expect such cascading effects to occur and how best to incorporate them. Further investigation into hierarchical model evaluation offers the potential to simultaneously explore different scales of biological processes.

Lastly, how do we know if poor fit really matters? Simplifying assumptions require that our models will never precisely replicate reality, so testing the "truth" of our model is not productive. We should focus on assessing fit with respect to model utility. Brown (2014) employed posterior prediction with test statistics designed around model-based inferences (i.e., inference-based statistics) in an attempt to directly identify cases where model reliability is compromised. Duchêne et al. (2015) also used such an approach to identify problems with molecular clock models and the inference of divergence times. In both of these studies, the inference-based test statistics outperformed a general test based on the multinomial likelihood (Bollback 2002) in many circumstances. However, the trade-offs between data- and inference-based tests in terms of computation time, power, and relevance need further exploration.

3.3. The Future of Evaluating Model Fit in Evolutionary Biology

Despite advocacy for their adoption as an integral component of applied statistical modeling (Gelman et al. 2013), tests of absolute model fit are not currently standard practice in our field. We recommend that assessment of absolute model fit become standard practice in evolutionary studies, since most of our models do not lend themselves to intuitive visual inspection and have potentially restrictive assumptions. Establishment of these approaches as standard practice will depend on researcher motivation and reviewer diligence. Because of the flexibility and diversity of model evaluation tests in general, precise prescription is difficult. Their use will always require judgement, perhaps even more than most statistical procedures.

Many historical constraints on the adoption of these approaches (e.g., limited computing power, lack of implementation, small data sets) are no longer major barriers. For instance, many researchers are now able to design their own models in flexible statistical languages (e.g., R, RevBayes). In addition, the computational demands of these approaches are naturally accommodated by highly parallel computing systems. Modern data sets (particularly genetic data sets) are often enormous and offer unprecedented opportunities to find subsets that are not well suited to model assumptions, highlighting interesting and potentially novel biological processes.

Both the opportunity and need for tests of absolute model fit have never been greater. Because of the size of many modern data sets, resulting statistical power is enormous. In addition, the heterogeneity inherent to such large data sets makes them challenging to understand. Paradoxically, also because of their size, many modern analyses have reverted to the use of simpler models. This situation creates the potential for bias when precision is greatest and may explain ongoing debates that are not easily resolved by adding more data. Conversely, there is also now great opportunity for using such tests to identify particularly interesting exceptions to model assumptions, whether they are nucleotides, fossils, traits, or individuals. By identifying these exceptions, careful evaluation of model fit offers opportunities for new insights. Ultimately, by explaining these residual effects (Stigler 2016), evolutionary science will advance.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

We thank the National Science Foundation for grants DEB-1355071 to J.M.B. and DEB-1354506 to R.C.T., which funded original research related to the topic of this review. We also thank members of the Brown and Thomson laboratories as well as the Phyleaux Discussion Group at Louisiana State University and the Ecology and Evolution Supergroup at the University of Hawai'i, for many thoughtful discussions related to this topic. G. Mount, L. Coghill, J. Esselstyn, and B. Shaffer provided comments that greatly improved this manuscript.

LITERATURE CITED

Barley AJ, Brown JM, Thomson RC. 2017. Impact of model violations on the inference of species boundaries under the multispecies coalescent. *Syst. Biol.* 67:269–84

Barley AJ, Thomson RC. 2016. Assessing the performance of DNA barcoding using posterior predictive simulations. *Mol. Ecol.* 25:1944–57

Bayarri MJ, Berger JO. 2004. The interplay of Bayesian and frequentist analysis. *Stat. Sci.* 19:58–80

Beaulieu JM, O'Meara BC. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. *Syst. Biol.* 65:583–601

Beaulieu JM, O'Meara BC, Donoghue MJ. 2013. Identifying hidden rate changes in the evolution of a binary morphological character: the evolution of plant habit in campanulid angiosperms. *Syst. Biol.* 62:725–37

Blum MGB, François O. 2006. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Syst. Biol.* 55:685–91

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.* 19:1171–80

Brown JM. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Syst. Biol.* 63:334–48

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics* 25:537–38

Brown JM, Lemmon AR. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Syst. Biol.* 56:643–55
Brown JM, Thomson RC. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* 66:517–30
Caetano DS, O'Meara BC, Beaulieu JM. 2018. Hidden state models improve the adequacy of state-dependent diversification approaches using empirical trees, including biogeographical models. bioRxiv 302729. **https://doi.org/10.1101/302729**
Castoe TA, de Koning APJ, Kim H-M, Gu W, Noonan BP, et al. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. *PNAS* 106:8986–91
Chen M-H, Shao Q-M, Ibrahim JG, editors. 2000. *Monte Carlo Methods in Bayesian Computation*. New York: Springer
Cox DR. 1961. Tests of separate families of hypotheses. In *Proc. Fourth Berkeley Symp. Math. Stat. Probab.*, pp. 105–23. Berkeley: Univ. Calif. Press
Darwin C. 1859. *On the Origin of Species by Means of Natural Selection, Or, the Preservation of Favoured Races in the Struggle for Life*. London: John Murray
Darwin F, ed. 1887. *The Life and Letters of Charles Darwin, Including an Autobiographical Chapter*, Vol. 1. London: John Murray
Doyle VP, Young RE, Naylor GJP, Brown JM. 2015. Can we identify genes with increased phylogenetic reliability? *Syst. Biol.* 64:824–37
Duchêne DA, Duchêne S, Ho SYW. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Mol. Biol. Evol.* 34:1529–34
Duchêne DA, Duchêne S, Ho SYW. 2018. PhyloMAd: efficient assessment of phylogenomic model adequacy. *Bioinformatics* 34:2300–1
Duchêne DA, Duchêne S, Holmes EC, Ho SYW. 2015. Evaluating the adequacy of molecular clock models using posterior predictive simulations. *Mol. Biol. Evol.* 32:2986–95
Duchêne S, Duchêne DA, Di Giallonardo F, Eden J-S, Geoghegan JL, et al. 2016. Cross-validation to select Bayesian hierarchical models in phylogenetics. *BMC Evol. Biol.* 16:115
Dunn CW, Giribet G, Edgecombe GD, Hejnol A. 2014. Animal phylogeny and its evolutionary implications. *Annu. Rev. Ecol. Evol. Syst.* 45:371–95
Edwards SV, Liu L, Pearl DK. 2007. High-resolution species trees without concatenation. *PNAS* 104:5936–41
Efron B, Tibshirani RJ. 1993. *An Introduction to the Bootstrap*. Boca Raton, FL: Chapman & Hall
Etienne RS, Rosindell J. 2012. Prolonging the past counteracts the pull of the present: Protracted speciation can explain observed slowdowns in diversification. *Syst. Biol.* 61:204–13
Felsenstein J. 1985. Phylogenies and the comparative method. *Am. Nat.* 125:1–15
Felsenstein J. 2004. *Inferring Phylogenies*. Sunderland, MA: Sinauer
FitzJohn RG. 2010. Quantitative traits and diversification. *Syst. Biol.* 59:619–33
FitzJohn RG. 2012. Diversitree: comparative phylogenetic analyses of diversification in R. *Methods Ecol. Evol.* 3:1084–92
FitzJohn RG, Maddison WP, Otto SP. 2009. Estimating trait-dependent speciation and extinction rates from incompletely resolved phylogenies. *Syst. Biol.* 58:595–611
Foster PG. 2004. Modeling compositional heterogeneity. *Syst. Biol.* 53:485–95
Foster PG, Hickey DA. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* 48:284–90
Garland T, Harvey PH, Ives AR. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst. Biol.* 41:18–32
Gelman A. 2003. A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *Int. Stat. Rev.* 71:369–82
Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2013. *Bayesian Data Analysis*. Boca Raton, FL: CRC Press. 3rd ed.
Gelman A, Meng X-L, Stern H. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Stat. Sin.* 6:733–807
Goldberg EE, Lancaster LT, Ree RH. 2011. Phylogenetic inference of reciprocal effects between geographic range evolution and diversification. *Syst. Biol.* 60:451–65

Goldman N. 1993a. Statistical tests of models of DNA substitution. *J. Mol. Evol.* 36:182–98
Goldman N. 1993b. Simple diagnostic statistical tests of models for DNA substitution. *J. Mol. Evol.* 37:650–61
Gruenstaeudl M, Reid NM, Wheeler GL, Carstens BC. 2015. Posterior predictive checks of coalescent models: P2C2M, an R package. *Mol. Ecol. Res.* 16:193–205
Heard SB. 1992. Patterns in tree balance among cladistic, phenetic, and randomly generated phylogenetic trees. *Evolution* 46:1818–26
Heard SB, Mooers AØ. 2002. Signatures of random and selective mass extinctions in phylogenetic tree balance. *Syst. Biol.* 51:889–97
Heath TA, Moore BR. 2014. Bayesian inference of species divergence times. In *Bayesian Phylogenetics: Methods, Algorithms, and Applications*, ed. M-H Chen, L Kuo, PO Lewis, pp. 487–533. Sunderland, MA: Sinauer
Heath TA, Zwickl DJ, Kim J, Hillis DM. 2008. Taxon sampling affects inferences of macroevolutionary processes from phylogenetic trees. *Syst. Biol.* 57:160–66
Heled J, Drummond AJ. 2010. Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.* 27:570–80
Ho SYW, Duchêne S. 2014. Molecular-clock methods for estimating evolutionary rates and timescales. *Mol. Ecol.* 23:5947–65
Höhna S, Coghill LM, Mount GG, Thomson RC, Brown JM. 2017. P^3: phylogenetic posterior prediction in RevBayes. *Mol. Biol. Evol.* 35:1028–34
Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, et al. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* 65:726–36
Höhna S, May MR, Moore BR. 2015. TESS: an R package for efficiently simulating phylogenetic trees and performing Bayesian inference of lineage diversification rates. *Bioinformatics* 32:789–91
Huelsenbeck J. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44:17–48
Huelsenbeck JP, Larget B, Miller RE, Ronquist F. 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51:673–88
Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53:904–13
Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294:2310–14
James G, Witten D, Hastie T, Tibshirani R. 2013. *An Introduction to Statistical Learning*. New York: Springer
Joly S, McLenachan PA, Lockhart PJ. 2009. A statistical approach for distinguishing hybridization and incomplete lineage sorting. *Am. Nat.* 174:E54–70
Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, ed. HN Munro, pp. 21–132. New York: Academic
Kass RE. 2011. Statistical inference: the big picture. *Stat. Sci.* 26:1–9
Kimura M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–20
Kishino H, Hasegawa M. 1990. Converting distance to time: application to human evolution. *Methods Enzymol.* 183:550–70
Koch JM, Holder MT. 2012. An algorithm for calculating the probability of classes of data patterns on a genealogy. *PLOS Curr.* 4:e4fd1286980c08
Lanfear R, Calcott B, Ho SYW, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* 29:1695–701
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.* 21:1095–109
Lartillot N, Philippe H. 2008. Improvement of molecular phylogenetic inference and the phylogeny of Bilateria. *Philos. Trans. R. Soc. B* 363:1463–72
Lemmon AR, Moriarty EC. 2004. Importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53:265–77
Lewis PO, Xie W, Chen M-H, Fan Y, Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. *Syst. Biol.* 63:309–21

Maddison WP. 1997. Gene trees in species trees. *Syst. Biol.* 46:523–36
Maddison WP. 2006. Confounding asymmetries in evolutionary diversification and character change. *Evolution* 60:1743–46
Maddison WP, Midford PE, Otto SP. 2007. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* 56:701–10
Mayr E. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge, MA: Belknap
McElreath R. 2016. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Boca Raton. FL: CRC Press
Meng X-L. 1994. Posterior predictive *p*-values. *Ann. Stat.* 22:1142–60
Minin V, Abdo Z, Joyce P, Sullivan J. 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* 52:674–83
Mooers A. 1995. Tree balance and tree completeness. *Evolution* 49:379–84
Navidi WC, Churchill GA, von Haeseler A. 1991. Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.* 8:128–43
Nielsen R. 2002. Mapping mutations on phylogenies. *Syst. Biol.* 51:729–39
Nielsen R, Huelsenbeck JP. 2002. Detecting positively selected amino acid sites using posterior predictive p-values. *Pac. Symp. Biocomput.* 7:576–88
Nylander JAA, Ronquist F, Huelsenbeck JP, Nieves-Aldrey JL. 2004. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* 53:47–67
Pagel M, Meade A. 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.* 53:571–81
Pennell MW, Fitzjohn RG, Cornwell WK, Harmon LJ. 2015. Model adequacy and the macroevolution of angiosperm functional traits. *Am. Nat.* 186: E33–50
Posada D, Buckley T. 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
Posada D, Crandall K. 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50:580–601
Rabosky DL, Goldberg EE. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.* 64:340–55
Rabosky DL, Goldberg EE. 2017. FiSSE: a simple nonparametric test for the effects of a binary character on lineage diversification rates. *Evolution* 71:1432–42
Rannala B, Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. *Syst. Biol.* 66:823–42
Reeves JH. 1992. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.* 35:17–31
Reid NM, Hird SM, Brown JM, Pelletier TA, McVay JD, et al. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. *Syst. Biol.* 63:322–33
Ren F, Tanaka H, Yang Z. 2005. Empirical examination of the utility of codon-substitution models in phylogeny reconstruction. *Syst. Biol.* 54:808–18
Richards EJ, Brown JM, Barley AJ, Chong RA, Thomson RC. 2018. Variation across mitochondrial gene trees provides evidence for systematic error: How much gene tree variation is biological? *Syst. Biol.* 67:847–60
Ripplinger J, Sullivan J. 2010. Assessment of substitution model adequacy using frequentist and Bayesian methods. *Mol. Biol. Evol.* 27:2790–803
Ritland K, Clegg MT. 1987. Evolutionary analysis of plant DNA sequences. *Am. Nat.* 130:S74–100
Rodrigue N, Philippe H, Lartillot N. 2007. Assessing site-interdependent phylogenetic models of sequence evolution. *Mol. Biol. Evol.* 23:1762–75
Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61:539–542
Rubin DB. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Stat.* 12:1151–72
Shapiro SS, Wilk MB. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52:591–611
Slater GJ, Pennell MW. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. *Syst. Biol.* 63:293–308

Stadler T, Degnan JH, Rosenberg NA. 2016. Does gene tree discordance explain the mismatch between macroevolutionary models and empirical patterns of tree shape and branching times? *Syst. Biol.* 65:628–39

Stigler SM. 2016. *The Seven Pillars of Statistical Wisdom*. Cambridge, MA: Harvard Univ. Press

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36:445–66

Tavaré S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. In *Some Mathematical Questions in Biology: DNA Sequence Analysis*, ed. Miura RM, pp 57–86. Providence, RI: Am. Math. Soc.

Waddell PJ, Ota R, Penny D. 2009. Measuring fit of sequence data to phylogenetic model: gain of power using marginal tests. *J. Mol. Evol.* 69:289–99

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol.* 18:691–99

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39:306–14

Yang Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford: Oxford Univ. Press

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19:908–17

Yang Z, Rannala B. 2005. Branch-length prior influences Bayesian posterior probability of phylogeny. *Syst. Biol.* 54:455–70

Yang Z, Rannala B. 2010. Unguided species delimitation using DNA sequence data from multiple loci. *Mol. Biol. Evol.* 31:3125–35

Yang Z, Rannala B. 2014. Bayesian species delimitation using multilocus sequence data. *PNAS* 107:9264–69

Zhou Y, Brinkmann H, Rodrigue N, Lartillot N, Philippe H. 2010. A Dirichlet process covarion mixture model and its assessments using posterior predictive discrepancy tests. *Mol. Biol. Evol.* 27:371–84

Zuckerkandl E, Pauling L. 1962. Molecular disease, evolution and genetic heterogeneity. In *Horizons in Biochemistry*, ed. M Kasha, B Pullman, pp. 189–225. New York: Academic

Zuckerkandl E, Pauling L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, ed. V Bryson, H Vogel, pp. 97–166. New York: Academic