Syst. Biol. 63(3):289–292, 2014

DOI:10.1093/sysbio/syu009
Advance Access publication February 21, 2014

Predictive Approaches to Assessing the Fit of Evolutionary Models

JEREMY M. BROWN*

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, USA
**Correspondence to be sent to: Department of Biological Sciences, Louisiana State University, 202 Life Science Building, Baton Rouge, LA 70803, USA; E-mail: jembrown@lsu.edu*

Received 13 January 2014; reviews returned 30 January 2014; accepted 10 February 2014
Associate Editor: Frank (Andy) Anderson

Evolutionary inference is now an overwhelmingly model-based endeavor, allowing biological hypotheses to be tested in formal statistical frameworks. Stochastic models are employed to describe a broad range of evolutionary processes (e.g., coalescence within populations, long-term trends in diversification rates, and the evolution of both genomes and phenotypic traits). The shift to statistical methods for inference in evolution has expanded the power and consistency of inferences (e.g., Nielsen 2005; Yang 2006), provided a hierarchical framework to combine processes occurring within and between species (e.g., Liu and Pearl 2007; Heled and Drummond 2010; Knowles and Kubatko 2010), and allowed biologically interpretable insights into evolutionary processes (e.g., Nielsen and Yang 1998; Hey and Nielsen 2004).

Traditional approaches to evaluating fit between model and data rely primarily on likelihoods, either maximized or marginalized. Equivalently read as the probability of the data given the model, a model's likelihood tells us how well that model is able to predict our exact observations. Models with the highest likelihoods, after any necessary adjustment for additional parameters in more general models, are preferred for use in inference because they balance errors caused by failing to account for important evolutionary processes (bias) with errors caused by asking too much of the information in a finite data set (variance; Sullivan and Joyce 2005). However, such approaches only allow assessment of relative model fit with no ability to reject all models as adequate descriptions of relevant evolutionary processes, focus on only one aspect of model performance, and usually require the ability to analytically calculate likelihoods. These limitations restrict the complexity of models that can be explored and the ways in which their performance is evaluated, while not providing a mechanism to promote skepticism when all available models are unable to explain important patterns in the observed data.

Complementary and alternative approaches to model fitting and assessment can be achieved by expanding the judgment of a model's predictive ability so that it no longer relies solely on the likelihood. This more general predictive framework provides additional avenues for comparing what one might expect to observe under a given model to the data that have actually been collected. In so doing, researchers can assess the absolute fit of a model, evaluate relative fit using a wider variety of criteria, and explore models for which the formulation of a likelihood function is intractable. These advantages can translate to the use of more biologically realistic models and increased confidence in inferences of evolutionary history and process.

One increasingly popular class of predictive techniques in evolutionary biology is known as approximate Bayesian computation (ABC; Beaumont 2010). Developed largely in response to the challenges posed by complex population genetic models (e.g., Beaumont et al. 2002), ABC obviates the need to formulate a likelihood function in order to perform inference under complicated scenarios. Instead, replicate data sets are simulated from the prior predictive distribution (the distribution that specifies the probability of any data set based on what is believed *a priori* about appropriate values for model parameters) and compared to observations. Simulations producing data that are sufficiently similar to the observed data are "accepted" while all others are "rejected". The collection of parameter values associated with the accepted simulations provides a simulated draw from the posterior distribution. When the simulate-accept/reject procedure is performed sufficiently many times, this collection of parameter values mimics the shape of the posterior distribution with increasing accuracy.

Another predictive Bayesian approach that has seen some use in evolutionary inference is known as posterior prediction (PP). This general approach can assess model fit in an absolute sense (Bollback 2002; Brown 2014; Reid et al. 2014; Slater and Pennell 2014) as well as provide alternative perspectives on relative model fit (Lewis et al. 2014). Assessing the fit of different models to empirical data in such a way that the data can reject the fit of all models should be a fundamental step in the inference process (Penny et al. 1992; Gelman et al. 2004), but it is often neglected in studies of evolutionary history. Models that fit the observed data poorly can lead to unsound and erroneous biological conclusions. By simulating replicate data sets using fitted model parameter values (i.e., drawn from the posterior distribution), PP approaches can reject all available models if the observed data do not seem to be a reasonable draw from the simulated replicates. The predictions of different models can also be quantified

and compared on a relative scale, providing a route to model choice not based on likelihoods. PP-based model choice judges a model's performance not only on how well it specifically predicts the observed data, but also on the consistency and bias inherent to all of its predictions.

The typical mechanics of PP resemble ABC, in the sense that both involve drawing parameter values from a probability distribution and using them to simulate replicate data sets (but see the discussion of conditional predictive ordinates by Lewis et al. (2014) for a PP approach that does not require the simulation of replicate data sets). The goals of PP and ABC are distinct, however. PP draws parameter values from the posterior distribution (after taking into account the observed data), while ABC draws values from priors. Both typically summarize the "appearance" of a data set with a (set of) summary statistic(s), but ABC uses similarity in summary statistic values between simulated and observed data to find the parameter values that provide the best fit between model and data. PP uses differences in summary statistic values between simulated and observed data to quantify the ability of a model to reproduce chosen features of observed data after model fitting.

The additional tools provided by an expanded predictive framework are not without their own drawbacks. Both ABC and PP often require the simulation and summary of hundreds (or hundreds of millions) of replicate data sets. This process can be tedious, time-consuming, and the appropriate number of replications is not always clear. Additionally, tests of absolute model fit can never guarantee freedom from bias. Even if the method chosen to compare simulated and observed data sets fails to detect any differences, relevant differences may still exist but be detectable only under alternate criteria. Predictive approaches to assessing relative model fit suffer from some of the same drawbacks as likelihood-based approaches (e.g., an inability to reject all models). Also, they may prefer different models to likelihood-based approaches by putting increased weight on a model's ability to extend predictions to other data sets. In such cases, researchers will need to think carefully about their goals. Nonetheless, recent work suggests that predictive approaches offer substantial benefits in broadening the scope and rigor of evolutionary inference, by facilitating the use of new models and expanding our view of model performance.

At the 2012 annual meeting of the Society of Systematic Biologists (part of the First Joint Congress on Evolutionary Biology) in Ottawa, Ontario, Canada, I organized and participated in a symposium titled "Predictive Approaches to Assessing the Fit of Evolutionary Models." This symposium brought together researchers who either develop or apply predictive approaches in evolutionary biology, with ABC and PP as common themes. Papers based on several of these talks follow in this special issue of *Systematic Biology*. In addition, the following presentations (with presenting authors listed first) were also included in the symposium: Mark Beaumont, Heather Battey, and Dan Lawson, "Kernel-based Approximate Bayesian Computation for inferring the history of recently diverged taxa"; Jeffrey Wall, Laurie Stevison, August Woerner, and Michael Hammer, "Ancient admixture in human evolution". Further information on the four symposium papers included in this special issue is provided below.

Slater and Pennell (2014) present new approaches for assessing the absolute fit of continuous trait models. They illustrate these approaches by applying them to the search for early bursts of trait evolution. Such bursts are expected to be associated with adaptive radiations under a Simpsonian view (Simpson 1944, 1953). While expected, the statistical evidence favoring early bursts from comparative analyses of extant taxa has often been weak (Harmon et al. 2010). Slater and Pennell argue convincingly that the previous lack of evidence for early bursts may be the result of low statistical power and outline two novel approaches to testing for this pattern. The first of their novel proposals, closely tied to the theme of this special issue, employs PP simulations to characterize a model's ability to describe observed patterns of trait disparity across clades at different times. Their PP approaches show greater power than existing approaches to model comparison based on likelihood ratio tests and information theory. Their second proposal is designed to reduce the masking influence of "outlier" taxa on the ability to detect underlying, clade-wide early burst patterns. Such outliers may result from some subset of taxa jumping to a new adaptive zone and may profoundly influence the power of a method to detect an underlying early burst pattern. By employing a robust regression technique, Slater and Pennell are able to simultaneously identify those parts of the tree with outlying evolutionary rates and downweight the influence of these morphological comparisons when trying to recover broader, clade-wide trends. The authors apply these approaches to a data set of cetacean body lengths, which support a clade-wide early burst pattern. This pattern was partially obscured by outlying evolutionary rates (e.g., convergence) in some parts of the cetacean tree and its existence has been the subject of previous debate (Slater et al. 2010, Venditti et al. 2011). In aggregate, Slater and Pennell's methods improve both the power and robustness of tests for particular patterns of continuous trait evolution. In their study, the pattern of interest was an early burst, but the methods are more generally applicable.

Lewis et al. (2014) describe two distinct PP approaches to model selection in phylogenetics. Their first approach extends a method proposed by Gelfand and Ghosh (1998), which Lewis et al. term GG. The GG approach, as with many other PP approaches, involves the simulation of replicate data sets using trees and parameter values drawn from the posterior distribution. The GG criterion compares the predictive ability of different models based on two components: one measuring the variance in the PP distribution and one measuring the goodness-of-fit of the model to the empirical data. Models that

both fit empirical data well and have low variance in their predictions will be favored. The GG approach is flexible in both the relative weight it gives to each of these components and the loss function it uses to quantify goodness-of-fit and variance. Loss functions provide a real number, ≥ 0, to associate with the degree of difference between data sets. Lewis et al. explore the use of a deviance loss function, appropriate for molecular data sets with discrete categories of site patterns, and equal weights between the goodness-of-fit and variance components. The second approach that they explore is termed the conditional predictive ordinate (CPO) method. In contrast to GG, CPO is a site-specific measure of model fit and does not require the simulation of replicate data sets. CPO also differs from GG, and the other PP methods outlined in the papers of this special issue, in that it is a cross validation method. The fit of a model to a particular site in an alignment is calculated conditional on the information in all other sites. Conveniently, this value can be calculated simply as the posterior harmonic mean of the likelihood for the site of interest. Site-specific CPO values can also be combined into a pseudomarginal likelihood, providing a measure of model fit across an entire data set. CPO values tend to be sensitive to among site variation in rates of evolution, since site patterns at slowly evolving sites are inherently easier to predict than those at rapidly evolving sites, so Lewis et al. suggest that this approach might be particularly useful for exploratory analyses using models that do not include site-specific estimates of rate. The authors then demonstrate the utility of both approaches with four examples based on two empirical data sets (four protein-coding genes from 28 algal plastids and the *rps*11 protein-coding gene from 5 angiosperms). These examples demonstrate the diversity of questions that can be addressed using GG and CPO, including topics of much recent interest: the sensitivity of analyses to chosen branch-length priors, the identification of variation in rates of evolution across genes, the identification of chimaeric sequences generated by horizontal gene transfer, and the choice of partitioned models.

Reid et al. (2014) apply a PP test of absolute model fit that, like the proposals of Slater and Pennell (2014; above) and Brown (2014; below), is based on the use of posterior predictive *P*-values first introduced to systematics by Bollback (2002). Reid et al. are specifically interested in assessing whether observed variation across gene trees can be explained solely by coalescent stochasticity, as assumed in implementations of the multispecies coalescent model such as *BEAST (Heled and Drummond 2010). As a hierarchical model, the multispecies coalescent offers different levels on which to compare simulated and empirical data. Reid et al. focus on two components of the model: the distribution of inferred coalescent genealogies and the distribution of DNA sequence alignments. Multiple test statistics are employed at each level in an attempt to increase the sensitivity of the tests to poor fit between model and data. While the authors are primarily interested in whether coalescent stochasticity can explain variation across gene trees, inappropriate assumptions (or poorly specified priors) at one level of a hierarchical model can manifest themselves in poor fit at other levels—hence the use of tests based on DNA sequence alignments as well as inferred coalescent genealogies. After developing these tests, Reid et al. broadly applied them to 25 empirical data sets drawn from recently published studies. Interestingly, model fit at the level of coalescent genealogies was rejected for only 4/25, while model fit at the level of DNA sequence alignments was rejected for 20/25. Despite a much higher rejection rate for model fit at the level of alignments, Reid et al. provide evidence that misfit detected at both levels may be driven by coalescent assumptions. For two data sets where a single locus was identified to fit the model particularly poorly at the genealogical level, removal of this locus led to acceptance of overall model fit for the remaining loci. In the case of two other data sets, where detection of model misfit was confined to the alignment level but more widespread across loci, phylogenetic inference allowing independent gene trees showed no evidence of model misfit. While intriguing, these results also highlight a challenge of any test assessing the fit of hierarchical models. Since different levels of the model are interdependent, accurately identifying the level that fits poorly can be quite challenging. Nonetheless, the tools developed by Reid et al. should prove useful in identifying outlier loci, suggesting when caution is warranted in interpreting the results of coalescent analyses, and providing insight into the prevalence and strength of other factors driving gene tree discordance.

Brown (2014) develops a suite of novel test statistics for use in assessing the plausibility of phylogenetic inferences. Also employing a PP *P*-value framework, the newly proposed tests compare the phylogenetic information content of empirical data to PP replicates. These tests are agnostic about the nature of poor fit between model and data and aim to reject the plausibility of inferences specifically when fit is so poor that it influences inferences. To accomplish this goal, the phylogenetic information in different data sets must be quantified. Brown's proposal begins with an estimate of the joint posterior distribution of tree topologies and branch lengths for each data set and then summarizes the phylogenetic information that it contains in various ways. One topological summary statistic employs statistical entropy (Shannon and Weaver 1949) to summarize the marginal distribution of posterior probabilities across trees. Another set of topological statistics uses the positions of quantiles in the ordered vector of all pairwise tree-to-tree distances from the posterior distribution. The continuous nature of branch lengths allows the use of simpler summaries, namely the posterior mean and variance. In a proof-of-principle simulation study where data were simulated on trees of varying length and analysed with oversimplified models or poorly specified priors, the proposed test statistics rejected inferential plausibility with increasing frequency as errors became worse. Importantly, the

plausibility of specific inferences (e.g., branch lengths) could be rejected when problematic while still accepting the plausibility of other inferences (e.g., tree topology). Brown also applied these tests to three empirical data sets, demonstrating that poor fit between model (or priors) and data does affect a variety of empirical inferences. However, such effects can be identified on a data set- and inference-specific basis. While focusing on the fit of single-locus models of molecular evolution and resulting impacts on inferred tree topologies and branch lengths, the same general framework outlined by Brown could easily be extended to other types of evolutionary models and inferences.

In aggregate, the work discussed during this symposium and further elaborated in the papers below demonstrates the wide applicability of predictive approaches to assessing the fit of evolutionary models. Many of these advances are natural outgrowths of maturing statistical disciplines within evolutionary biology, some are driven by the new challenges posed by massive data sets, and nearly all are facilitated by the increasing availability of large computing resources. In the future, I expect that we will see increasingly diverse and creative applications of such approaches to new studies of evolution.

Acknowledgments

I would like to thank the Society of Systematic Biologists for supporting this symposium. I would also like to thank Mark Beaumont, Paul Lewis, Noah Reid, Graham Slater, and Jeff Wall for presenting excellent talks during the symposium. F. Anderson, B. Carstens, and V. Doyle provided valuable comments on this manuscript.

References

Beaumont M.A. 2010. Approximate Bayesian computation in evolution and ecology. Annu. Rev. Ecol. Evol. Syst. 41:379–406.

Beaumont M.A., Zhang W., Balding D.J. 2002. Approximate Bayesian computation in population genetics. Genetics 162: 2025–2035.

Bollback J.P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Brown J.M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Syst. Biol. 63:334–348.

Gelfand A.E., Ghosh S. 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85:1–11.

Gelman A., Carlin J.B., Stern H.S., Rubin D.B. 2004. Bayesian data analysis. 2nd ed. New York: Chapman & Hall/CRC.

Harmon L.G., Losos J.B., Davies T.J., Gillespie R.G., Gittleman J.L., Jennings W.B., Kozak K.H., McPeek M.A., Moreno-Roark F., Near T.J., Purvis A., Ricklefs R.E., Schluter D., Schulte II J.A., Seehausen O., Sidlauskas B.L., Torres-Carvajal O., Weir J.T., Mooers A.Ø. 2010. Early bursts of body size and shape evolution are rare in comparative data. Evolution 64:2385–2396.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hey J., Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167:747–760.

Knowles L.L., Kubatko L.S., editors. 2010. Estimating species trees: practical and theoretical aspects. Hoboken: Wiley-Blackwell.

Lewis P.O., Xie W., Chen M.-H., Fan Y., Kuo L. 2014. Posterior predictive Bayesian phylogenetic model selection. Syst. Biol. 63:309–321.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Nielsen R., editor. 2005. Statistical methods in molecular evolution. New York: Springer.

Nielsen R., Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.

Penny D, Hendy M.D., Steel M.A. 1992. Progress with methods for constructing evolutionary trees. Trends Ecol. Evol. 7:73–79.

Reid N.M., Hird S.M., Brown J.M., Pelletier T.A., McVay J.D., Satler J.D., and Carstens B.C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Syst. Biol. 63:322–333.

Shannon C.E., Weaver W. 1949. The mathematical theory of communication. Urbana: Univ. of Illinois Press.

Simpson G.G. 1944. Tempo and mode of evolution. New York: Columbia University Press.

Simpson G.G. 1953. Major features of evolution. New York: Columbia University Press.

Slater G.J., Pennell M.W. 2014. Robust regression and posterior predictive simulation increase power to detect early bursts of trait evolution. Syst. Biol. 63:293–308.

Slater G.J., Price S.A., Santini F., Alfaro M.E. 2010. Diversity versus disparity and the radiation of modern cetaceans. Proc. Roy. Soc. B 277:3097–3104.

Sullivan J., Joyce P. 2005. Model selection in phylogenetics. Annu. Rev. Ecol. Evol. Syst. 36:445–466.

Venditti C., Meade A., Pagel M. 2011. Multiple routes to mammalian diversity. Nature 479:393–396.

Yang Z. 2006. Computational molecular evolution. New York: Oxford University Press.