P^3: Phylogenetic Posterior Prediction in RevBayes

Sebastian Höhna,*[,1,2] Lyndon M. Coghill,[3] Genevieve G. Mount,[3] Robert C. Thomson,[4] and Jeremy M. Brown[3]

[1]Division of Evolutionary Biology, Ludwig-Maximilians-Universität, München, Germany

[2]Department of Integrative Biology, University of California, Berkeley, CA

[3]Department of Biological Sciences and Museum of Natural Science, Louisiana State University, Baton Rouge, LA

[4]Department of Biology, University of Hawai'i, Honolulu, HI

***Corresponding author:** E-mail: sebastian.hoehna@gmail.com.

Associate editor: Keith Crandall

Abstract

Tests of absolute model fit are crucial in model-based inference because poorly structured models can lead to biased parameter estimates. In Bayesian inference, posterior predictive simulations can be used to test absolute model fit. However, such tests have not been commonly practiced in phylogenetic inference due to a lack of convenient and flexible software. Here, we describe our newly implemented tests of model fit using posterior predictive testing, based on both data- and inference-based test statistics, in the phylogenetics software RevBayes. This new implementation makes a large spectrum of models available for use through a user-friendly and flexible interface.

Key words: **model testing, Bayesian inference, phylogenetics.**

Brief Communication

Introduction

Statistical models are central to nearly all modern phylogenetic analyses (Huelsenbeck et al. 2001) and the accuracy of inferred phylogenies depends on the use of models that fit the data well (Yang 1994; Huelsenbeck and Rannala 2004; Sullivan and Joyce 2005; Brown 2014a). To fit well, a phylogenetic model must capture the salient features of evolution while avoiding unnecessary parameters. To balance these goals, typical model selection procedures choose the best fitting member of a set of candidate models (Posada and Crandall 2001; Darriba et al. 2012). However, model selection alone cannot guarantee that the chosen model captures the dynamics of evolution sufficiently to provide an unbiased phylogenetic estimate (Doyle et al. 2015).

In contrast to model selection, testing absolute model fit requires asking whether a given model is plausible in light of the data. One way to answer that question is to compare data that could be generated by the model to empirical observations (Bollback 2002). In a Bayesian framework, this comparison is conducted using posterior prediction, which simulates new data sets by drawing parameter values from the posterior distribution conditioned on the model (Gelman 1996; Brown 2014b). However, evaluation of absolute model fit is still relatively rare in phylogenetic studies and this disconnect is due in part to a lack of broadly available and easily usable software. Here, we describe a new implementation of phylogenetic posterior prediction (**P^3**) in the software package RevBayes (Höhna et al. 2016). This implementation makes posterior predictive tests of model fit broadly available in the same robust and flexible software used for inference. In addition, it lays the groundwork for the method to be applied to many different types of models (e.g., nucleotide evolution, continuous trait evolution, and lineage diversification).

RevBayes Implementation of P^3 Workflow

RevBayes provides an ideal framework to implement posterior predictive tests of absolute model fit. RevBayes is designed around probabilistic graphical models that divide a statistical model into a set of conditional probability distributions, thereby making simulation under any model straightforward (Höhna et al. 2014). After a simulation function is implemented once for each conditional probability distribution, such as a continuous-time Markov process to simulate character evolution, it can be (re-)used in any model. In principle, any model implemented in RevBayes that is used for parameter inference can also be used for tests of absolute model fit.

The first step in posterior prediction is to generate draws from the joint posterior distribution of parameters. In RevBayes, samples from the posterior distribution are drawn using a standard implementation of Markov chain Monte Carlo (MCMC) sampling (fig. 1, `mcmc.run()`; Rannala and Yang 1996) or Metropolis-Coupled Markov chain Monte Carlo (MCMCMC) sampling (Geyer 1991; Altekar et al. 2004). For each sample from the posterior distribution, values of all stochastic variables are stored in a single tab-delimited file using a new Stochastic-Variable-Monitor. In the second step, RevBayes reads the trace of values written by the Stochastic-Variable-Monitor and simulates new data sets using the original model, by setting the values of the stochastic variables to the values drawn from the posterior

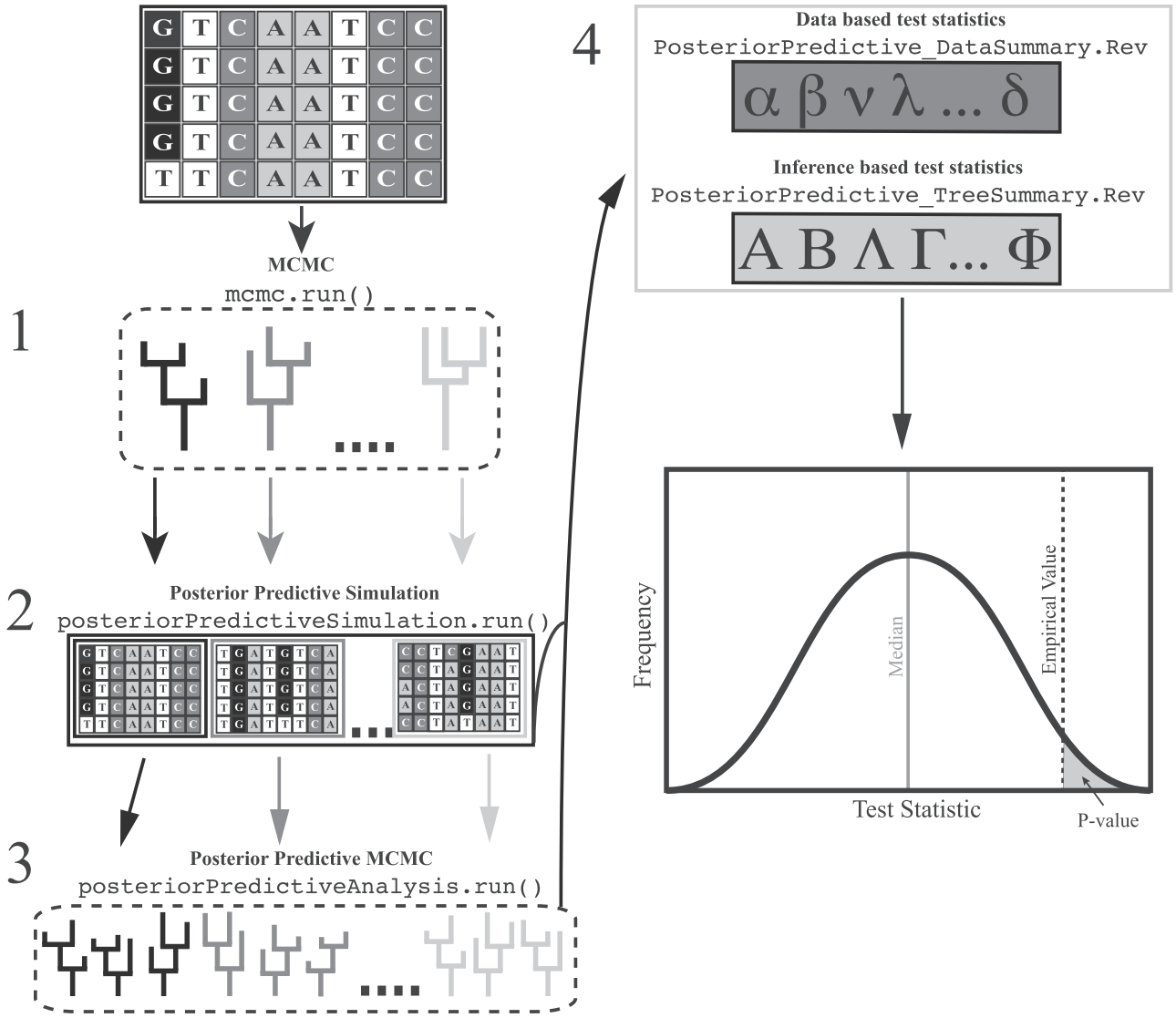

FIG. 1. Overview of the P[3] workflow as implemented in RevBayes together with the specific commands necessary in each step. Step 1 involves sampling parameters, for example, tree topology and branch lengths, from the posterior distribution using MCMC simulation. Step 2 simulates new data sets given the parameter samples from step 1. Step 3 estimates posterior distributions of the same parameters but from the simulated data sets. This third step is optional and is only needed for inference-based test statistics. Next, step 4 involves computing data- and/or inference-based test statistics and comparing the distribution of the test statistic from the simulated data with the test statistic value from the observed data. Finally, data sets and models can be rejected or ranked based on the posterior predictive *p*-values or the posterior predictive effect size (PPES), which is the difference between the median of the posterior predictive distribution and the empirical value, normalized by the distribution's SD.

(fig. 1, `posteriorPredictiveSimulation.run()`). Simulated data sets are stored in standard file formats, for example, in Nexus file format for sequence alignments and in Newick file format for trees. We provide the additional option to thin the samples by using only every *j*-th sample from the trace. In the third step, RevBayes performs an MCMC simulation for each simulated data set after reading it in from file (fig. 1, `posteriorPredictiveAnalysis.run()`). All three steps use the same model script, which simplifies usage. Furthermore, we parallelized our implementation using the Message-Passing-Interface (MPI) so that MCMC analysis of both the empirical and simulated data sets can be conducted efficiently on large, parallel computer architectures. Specifically, step 1 is parallelized by computing the likelihood, heated chains and/or independent replicates distributed over many CPUs (see Höhna et al. 2017) and step 3 is parallelized by distributing MCMC runs of simulated data sets over all available CPUs. The two final steps compute data- and/or inference-based test statistics across all data sets or sets of posterior samples, respectively (fig. 1, `PosteriorPredictive_DataSummary.Rev` and `PosteriorPredictive_TreeSummary.Rev`). Finally, the posterior predictive distribution of the test statistic and the observed test statistic value provide the *p*-values and the Posterior Predictive Effect Size (PPES) to assess absolute model fit (fig. 1, `PosteriorPredictive_DataSummary.Rev` and `PosteriorPredictive_TreeSummary.Rev`).

Test Statistics and Interpretation

The ability to assess model fit depends strongly on the test statistics used. Here, we provide an overview with a description of some test statistics implemented in RevBayes. Note that this list is only a subset of test statistics available in RevBayes and users can easily design their own. Moreover, the number of preimplemented test statistics is continuously expanding. We divide these test statistics into two main categories: 1) data-based and 2) inference-based. We conclude with a brief presentation and discussion on computing *p*-values.

Data-Based Test Statistics

Number of Invariant Sites

The number of invariant sites (partially) captures the characteristic of a sequence alignment where some sites are variable and other sites are not, that is, evolve under different rates of evolution. Let us denote the sequence length by L and let inv(j) be true if the jth site (column) of an alignment is invariant. We compute the number of invariant sites by $\sum_{j=1}^{L} 1_{\text{inv}(j)}$. Note that the number of invariant sites as a test statistic is directly related to Wattersons θ which computes the number of segregating sites (i.e., L minus the number of invariant sites) divided by the harmonic mean of $N-1$.

Maximum GC Content

The maximum GC content test statistic aims to detect outlier sequences with high GC content. We simply find the sequence i that has the highest fraction of GC characters by computing:

$$\max_i \left(\frac{1}{L_i} \sum_{j=1}^{L_i} 1_{c_{ij}=C||c_{ij}=G} \right),$$

where c_{ij} refers to the jth character of the ith sequence in the alignment.

Minimum GC Content

Similarly, the minimum GC content test statistic detects outlier sequences that have a very low GC content. The computation of the minimum GC content test statistic is

$$\min_i \left(\frac{1}{L_i} \sum_{j=1}^{L_i} 1_{c_{ij}=C||c_{ij}=G} \right).$$

Mean GC Content

The mean GC content test statistic computes the average GC content over all sequences. Therefore, the mean GC content test statistic mainly captures the nucleotide composition of an alignment and can be used to detect if a model would produce similar frequencies of GC content, that is, if the stationary frequencies are modeled adequately. We compute the mean GC content test statistic by

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{1}{L_i} \sum_{j=1}^{L_i} 1_{c_{ij}=C||c_{ij}=G} \right).$$

Variance of GC Content

Lastly, several studies have reported that high variation in GC content biases phylogeny inference (Romiguier et al. 2013). Here, we introduce the variance in GC content across sequences as a test statistic that can directly test if the variation in GC content between sequences is modeled plausibly. If not, we can detect possibly problematic alignments for phylogeny inference due to high GC variation. We compute the variation in GC content by

$$\frac{1}{N-1} \sum_{i=1}^{N} \left[\frac{1}{L_i} \sum_{j=1}^{L_i} 1_{c_{ij}=C||c_{ij}=G} - \frac{1}{N} \sum_{k=1}^{N} \left(\frac{1}{L_k} \sum_{j=1}^{L_k} 1_{c_{kj}=C||c_{kj}=G} \right) \right]^2.$$

Maximum Pairwise Distance

The maximum pairwise distance test statistic is intended to be sensitive to the model of rate-variation among site and/or among branches. We find the pair of sequences i and j for which the number of mismatched characters is largest and report the scaled number of mismatches (pairwise-distance),

$$\max_{i,j} \left(\frac{1}{L_i} \sum_{k=1}^{L} 1_{c_{ik} \neq c_{jk}} \right).$$

Minimum Pairwise Distance

Similarly, the minimum pairwise distance test statistic finds the pair of sequences i and j that has the smallest pairwise distance,

$$\min_{i,j} \left(\frac{1}{L_i} \sum_{k=1}^{L} 1_{c_{ik} \neq c_{jk}} \right).$$

Multinomial Likelihood

The multinomial likelihood of a sequence alignment was originally introduced to test for model fit in posterior predictive simulations (Goldman 1993; Bollback 2002; Brown and ElDabaje 2009). Here, we include it as a benchmark and for completeness. Let $\mathcal{P}_i$ denote the frequency of the ith site-pattern and $|\mathcal{P}|$ the number of site-patterns. The log-likelihood of the multinomial model is given by

$$\sum_{i=1}^{|\mathcal{P}|} (\mathcal{P}_i \ln(\mathcal{P}_i)).$$

Inference-Based Test Statistics

The inference-based test statistics are motivated by the work of Brown (2014a). These test statistics are applied on the posterior distribution estimated from the original data set (fig. 1, step 1) and additionally on the posterior distributions estimated for each simulated data set (fig. 1, step 3).

Table 1. Runtimes of the Different Steps in the Full **P[3]** Pipeline for Our Test Example Given Seconds.

Statistic\Model	#CPUs	JC	GTR	GTR + Inv	GTR + Gamma	GTR + Gamma + Inv
MCMC	1	328.6	399.8	465.1	945.0	1,057.6
	20	113.6	171.0	189.5	252.8	278.0
PP-Simulation	1	58.12	61.2	49.4	58.2	50.6
	20	4.0	4.0	3.3	4.5	3.3
PP-Analysis	1	48,740.1	73,177.1	65,383.7	143,306.6	149,284.8
	20	2,677.5	3,955.1	3,506.6	13,854.4	15,353.5
Tree-Summary	1	1,260.3	1,402.1	2,105.2	1,590.4	1,875.2
	20	1,825.3	2,079.3	6,286.5	2,419.4	4,691.4
p-values (PP-Analysis)	1	0.2	0.3	0.3	0.3	0.3
	20	0.3	0.3	0.3	0.3	0.4
Data-Summary	1	830.2	840.7	840.6	857.6	846.1
	20	902.9	892.9	892.0	888.6	897.4
p-values (PP-Simulation	1	0.6	0.5	0.5	0.6	0.6
	20	0.6	0.5	0.6	0.6	0.7

NOTE.—PP-Simulation, posterior predictive simulation; PP-Analysis, MCMC analysis of posterior predictive data sets.

Mean Robinson–Foulds Distance

The mean Robinson–Foulds-Distance test statistic captures the spread of the posterior distribution on trees. Specifically, we compute the Robinson–Foulds distance (RF, Robinson and Foulds 1981), or symmetric-distance, for any pair of trees Ψ_i and Ψ_j from the sample of trees in the posterior distribution. Let K be the number of samples from the posterior distribution. The mean RF-Distance is compute by

$$\frac{2}{K(K-1)}\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} RF(\Psi_i,\Psi_j).$$

Quantiles of Robinson–Foulds Distance

Using the kth quantile of the ordered vector of RF distances as a test statistic provides a similar measure as the mean RF-Distance test statistic. Previously, the 1st quartile, median, 3rd quartile, and 99th percentile have been used as choices for k (Brown 2014a), although the performance of test statistics based on these quantiles has not yet been thoroughly explored across data sets. In principle, any quantile of any set of values computed from posterior distributions can be used. We chose the RF distance as an example. Assume that RF^S represents the $\frac{K(K-1)}{2}$ sorted, pairwise RF distances. The k-th quantile (e.g., where $k = 0.99$ for the 99th percentile and $q = \frac{K(K-1)}{2}k$) is given by

$$\begin{cases} \frac{1}{2}(RF^S_{\lfloor q \rfloor} + RF^S_{\lceil q \rceil}) & \text{if} q \notin \mathbb{N} \\ RF^S_q & \text{otherwise} \end{cases}.$$

Mean Tree-Length

The mean tree-length test statistic focuses on inferred branch lengths and thus the expected number of substitutions used to explain the observed data. Let us define the tree-length, TL, as the sum of branch lengths bl, $\text{TL} = \sum_{i=1}^{2N-3} bl_i$. We compute the mean of the tree-lengths of all sampled trees from the posterior distribution by

$$\frac{1}{K}\sum_{i=1}^{K}\text{TL}_i.$$

Variance in Tree-Length

The variance in tree-length test statistic captures the uncertainty in the posterior distribution of branch lengths. As defined before, TL_i is the sum of branch lengths for the ith sampled tree from the posterior distribution. Thus, the variance in tree-length is defined as

$$\frac{1}{K-1}\sum_{i=1}^{K}\left[\text{TL}_i - \left(\frac{1}{K}\sum_{j=1}^{K}\text{TL}_j\right)\right]^2.$$

Entropy

Finally, the entropy test statistic captures the information gain when comparing the prior to the posterior distribution of phylogenetic tree topologies. Again, the entropy could be computed for any variable of interest in addition to tree topologies, although continuous variables must be discretized to compute the entropy. Let $B(N)$ be the total number of possible tree topologies for N taxa. The entropy is then given by (see Brown 2014a, eq. 4)

$$\ln(B(N)) + \sum_{i=1}^{B(N)}[p(\Psi_i|\mathbf{X}) \times \ln(p(\Psi_i|\mathbf{X}))].$$

Computing *P*-Values

Several different methods exist to compute p-values based on comparing empirical and posterior predictive test statistic values. Let us denote the empirical data as e, the simulated data as d, the test statistic function as s, and M as the number simulated data sets. First, we can compute the one-tailed lower p-value, p_L, as

$$p_L = P(s(d) \le s(e)) = \frac{1}{M}\sum_{i=1}^{M} 1_{s(d_i)\le s(e)},$$

which simply counts the number of times the test statistics for the simulated data was smaller than or equal to the test statistic of the observed data. Similarly, we can compute the one-tailed upper p-value, p_U, as

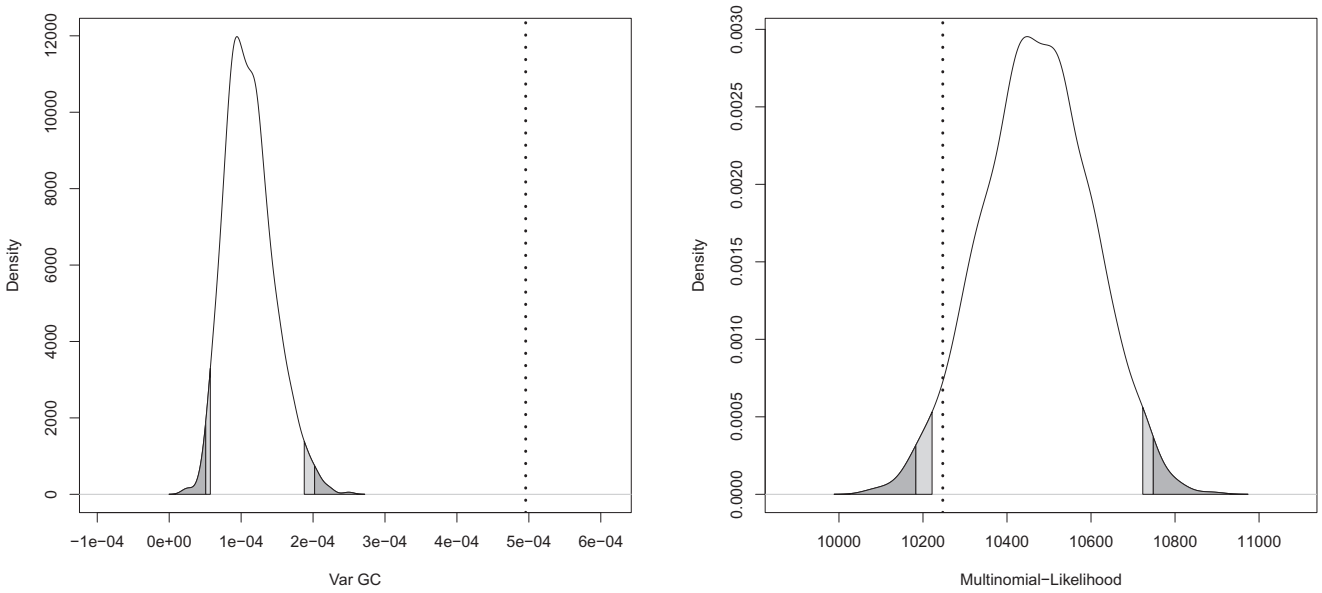

FIG. 2. Estimated posterior predictive distribution for the variance of GC content (left) and multinomial likelihood test statistics. These posterior predictive distributions were estimated under the GTR substitution model with gamma distributed rate-variation among sites and invariant sites (GTR + Gamma + Inv). The dark gray areas show the 1st and 99th percentiles and the light gray areas show the 2.5 and 97.5 percentiles. The dotted blue lines show the value of the test statistic for the empirical data.

$$p_U = P(s(d) \geq s(e)) = \frac{1}{M}\sum_{i=1}^{M} 1_{s(d_i)\geq s(e)}.$$

Note that the upper one-tailed *p*-value, p_U, does not necessarily equal $1-p_L$, because the test statistic value for some simulated data sets might be exactly equal to the test statistic of the empirical data. Both *p*-values are conservative because they include equal values in the calculation of tail-area probabilities (for a discussion and further details see Chapter 6, Gelman et al. 2004). Given the two one-tailed *p*-values, we can compute the two-tailed *p*-value as $p_T = 2\times\min(p_L, p_U)$. Again, this *p*-value is conservative because both one-tailed *p*-values are conservative. Another alternative is to use the mid-point one-tailed lower *p*-value, p_M, which is computed by

$$p_M = P(s(d) \lesssim s(e)) = \frac{1}{M}\sum_{i=1}^{M} 1_{s(d_i)<s(e)} + \frac{1}{2M}\sum_{i=1}^{M} 1_{s(d_i)=s(e)}.$$

This mid-point one-tailed lower *p*-value, p_M, is equivalent to one minus the mid-point one-tailed upper *p*-value (i.e., its upper counterpart) and therefore we can omit separate computation of the upper *p*-value. As above, we can compute the corresponding two-tailed *p*-value as $p_{MT} = 2\times\min(p_{ML}, p_{MU})$.

Example Analysis of Model Fit Using P^3

To demonstrate the use of **P^3** we selected a small example data set consisting of 23 cytochrome subunit B (cyt-b) sequences from the order Primates. We performed five different analyses on the same data set using: 1) a Jukes–Cantor (JC) substitution model (Jukes and Cantor 1969), 2) a general-time-reversible (GTR) substitution model (Tavaré 1986), 3) a GTR substitution model with invariant sites, 4) a GTR substitution model with gamma distributed rate-variation among sites (Yang 1994), and 5) a GTR substitution model with gamma distributed rate-variation among sites and invariant sites. For each model we performed the full **P^3** pipeline (see fig. 1). Example scripts are provided in the Supplementary Material online. We provide the runtimes for each step of the analyses (table 1). The runtimes show the performance when we ran **P^3** on a cluster using a single core or using all 20 cores available, thus, demonstrating the improvement due to our MPI implementation. The slowest part of the full **P^3** pipeline is the posterior predictive MCMC analysis which can be sped-up up to 18.1-fold when using 20 CPUs (see table 1). The summary of the trees and data are not parallelized and thus does not gain from additional CPUs. Future work will attempt to improve computational speed when summarizing data and trees as this is currently a weakness of our implementation.

The resulting *p*-values for the test statistics described in the previous section are given in tables 2 and 3. As expected, we observed that the fit of simpler models was worse. The traditionally used test statistic, the multinomial likelihood, does not detect model inadequacy for the GTR + Gamma + Inv model, whereas both the variance in GC content, Var(GC), and the variance in tree-length, Var(TL), test statistics do detect poor model fit and thus a potential bias in the inferred phylogeny even for our most complex substitution model (fig. 2). This small example highlights the need for new test statistics, improved phylogenetic models, and shows that our new **P^3** pipeline provides a framework for future development.

These results also highlight the need to more fully explore the relationship between poor model fit, as detected by particular test statistics, and the potential for biased inferences. Although we should always have some level of concern about

Table 2. Lower One-Tailed Midpoint *p*-values for Inference-Based Test Statistics.

Statistic\Model	JC	GTR	GTR + Inv	GTR + Gamma	GTR + Gamma + Inv
Mean(RF)	*1*	0.974	*0.981*	0.931	0.772
q(RF, 0.25)	0.5	0.499	0.9335	0.755	0.4905
q(RF, 0.5)	*0.999*	0.4795	0.9625	0.829	0.5795
q(RF, 0.75)	*1*	0.868	*0.9835*	0.8885	0.6685
q(RF, 0.99)	*1*	*0.9995*	*0.995*	*0.9925*	0.755
q(RF.0.999)	*1*	1	*0.9985*	0.969	0.823
mean(TL)	0.426	0.3855	0.51	0.945	0.887
var(TL)	0.391	*0.998*	0.911	*1*	*1*
Entropy	*0.999*	*1*	*0.99*	*0.992*	0.972

P-values are significant (shown in italics) if they are smaller than 0.025 or greater than 0.975.

Table 3. Lower 1-Tailed Midpoint *p*-values for Data-Based Test Statistics.

Statistic\Model	JC	GTR	GTR + Inv	GTR + Gamma	GTR + Gamma + Inv
#Invariant Sites	*1*	*1*	0.903	0.8535	0.224
Max(GC)	*0*	*1*	*0.9815*	0.683	0.9355
Max(PD)	*0*	*0*	*0*	0.0885	0.3075
Min(GC)	*0*	*0.984*	0.5455	0.088	0.341
Min(PD)	0.3685	0.8665	0.554	0.7205	0.544
Mean(GC)	*0*	*0.999*	0.7485	0.158	0.5385
Var(GC)	*1*	*1*	*1*	*1*	*1*
Multinomial-Likelihood	*1*	*1*	0.879	0.535	0.043

P-values are significant (shown in italics) if they are smaller than 0.025 or greater than 0.975.

our results when any test statistic detects poor fit, we should be more concerned when posterior predictive effect sizes are large and many test statistics show this pattern. Inference-based statistics may offer insights about which aspects of our inferences are most likely to be affected by unrealistic assumptions, whereas data-based statistics may help in identifying ways in which models can be improved. In our example analysis we should be concerned about biased inference because variance in GC content has be shown to bias phylogenetic inference (Romiguier et al. 2013; Duchêne et al. 2017) and has most likely affected our estimation of the tree length.

Comparison to PuMA

PuMA (Brown and ElDabaje 2009) is an application that, similar to **P^3**, implements a posterior predictive simulation approach to assess model fit. Although there are similarities in the conceptual basis of these two packages, there are substantial differences in their scope and implementation. First, PuMA employs only a single data-based test statistic, the multinomial likelihood, which serves as a very general assessment of model adequacy. **P^3** currently has nine inference-based and eight data-based test statistics (tables 1 and 2). These additional test statistics offer a much more robust and nuanced assessment of model fit. Second, PuMA relies on other software for inference (e.g., MrBayes, Ronquist et al. 2012) and simulation (e.g., seg-gen, Rambaut and Grass 1997), limiting the types of models and analyses to which posterior prediction can be applied. For example, PuMA can only assess fit for the inference of unrooted, nonclock phylogenies. **P^3** is fully contained within RevBayes, and therefore requires no additional software to perform inference or simulation. Furthermore, **P^3** can harness the full flexibility of RevBayes to provide many additional models for testing. Third, PuMA is implemented in Java and available as a native OS X application, whereas **P^3** is implemented in C++ and natively available for OS X, Windows, and Linux. Additionally, our implementation of **P^3** uses the MPI technology to run easily on parallel computer architectures, such as large computer clusters.

Conclusion

The combined implementation of posterior prediction and inference in RevBayes provides a general and flexible framework that allows the use of more models and test statistics than all previous methods and software (Bollback 2002; Brown and ElDabaje 2009; Brown 2014a). Furthermore, our implementation is easily extensible and will readily accommodate any new model specified in the Rev language, as well as any new test statistics. Future work will focus on improving computational efficiency and assessing test statistic performance.

Availability

P^3 is implemented in RevBayes v1.0.3. RevBayes is freely available from http://revbayes.com, last accessed November 8, 2017 and https://github.com/revbayes/revbayes, last accessed November 8, 2017. Examples and tutorials about P^3 are available from http://revbayes.github.io/tutorials.html, last accessed November 8, 2017.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank participants in a workshop on P^3 in RevBayes at the 2017 Society of Systematic Biologists meeting in Baton Rouge for helpful feedback. This work was supported by the Miller Institute for Basic Research in Science (to S.H.) and the US National Science Foundation under DEB-1355071 (to J.M.B.) and DEB-1354506 (to R.C.T.).

References

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics* 20(3):407–415.

Bollback JP. 2002. Bayesian model adequacy and choice in phylogenetics. *Mol Biol Evol*. 19(7):1171–1180.

Brown JM, ElDabaje R. 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. *Bioinformatics* 25(4):537–538.

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

Brown JM. 2014b. Predictive approaches to assessing the fit of evolutionary models. *Syst Biol*. 63:289–292.

Darriba D, Taboada GL, Doallo R, Posada D. 2012. jModelTest 2: more models, new heuristics and parallel computing. *Nat Methods* 9(8):771–772.

Doyle VP, Young RE, Naylor GJP, Brown JM. 2015. Can We Identify Genes with Increased Phylogenetic Reliability? *Syst Biol*. 64(5):824–837.

Duchêne DA, Duchêne S, Ho SY. 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Mol Biol Evol*. 34(6):1529–1534.

Gelman A. 1996. Posterior predictive assessment of model fitness via realized discrepancies. *Stat Sin*. 6:733–807.

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

Geyer CJ. 1991. Markov chain Monte Carlo maximum likelihood. In Keramidas EM, editor. Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface. Fairfax Station: Interface Foundation. p. 156–163.

Goldman N. 1993. Statistical tests of models of DNA substitution. *J Mol Evol*. 36(2):182–198.

Höhna S, Heath TA, Boussau B, Landis MJ, Ronquist F, Huelsenbeck JP. 2014. Probabilistic graphical model representation in phylogenetics. *Syst Biol*. 63(5):753–771.

Höhna S, Landis MJ, Heath TA, Boussau B, Lartillot N, Moore BR, Huelsenbeck JP, Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst Biol*. 65(4):726–736.

Höhna S, Landis ML, Huelsenbeck JP. 2017. Parallel power posterior analyses for fast computation of marginal likelihoods in phylogenetics. bioRxiv. doi:10.1101/104422.

Huelsenbeck J, Rannala B. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol*. 53(6):904–913.

Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* 294(5550):2310–2314.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro MN, editor. Mammalian protein metabolism. NewYork: Academic Press. p. 21–132.

Posada D, Crandall KA. 2001. Selecting the best-fit model of nucleotide substitution. *Syst Biol*. 50:580–560.

Rambaut A, Grass NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci*. 13(3):235–238.

Rannala B, Yang Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J Mol Evol*. 43(3):304–311.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Math Biosci*. 53(1–2):131–147.

Romiguier J, Ranwez V, Delsuc F, Galtier N, Douzery EJ. 2013. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol Biol Evol*. 30(9):2134–2144.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61(3):539–542.

Sullivan J, Joyce P. 2005. Model selection in phylogenetics. *Annu Rev Ecol Evol Syst*. 36(1):445–466.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. *Lect Math Life Sci*. 17:57–86.

Yang Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol*. 39(3):306–314.