Validating Bayesian inference tools for phylogenetics

Fábio K. Mendes[1†*], Remco Bouckaert[2†*],
Luiz M. Carvalho[3†], Alexei J. Drummond[4]

[1]Department of Biology, Washington University in St. Louis

[2]School of Computer Science, The University of Auckland

[3]Escola de Matemática Aplicada, Fundação Getulio Vargas

[4]School of Biological Sciences, The University of Auckland

[*]Corresponding authors: f.mendes@auckland.ac.nz; remco@cs.auckland.ac.nz

[†]Authors contributed equally to this work

March 21, 2022

Abstract

Biology has become a highly mathematical discipline in which probabilistic models play a central role, and as a result research in the biological sciences is now dependent on computational tools capable of carrying out complex analyses. These tools must not only be efficient, but also correctly implemented. Both goals are difficult to achieve for several reasons, such as the multidisciplinary nature of method development, and a still embrionic literature on good software development and statistical practices aimed at professionals from disparate fields. Here we provide guidelines for the validation of probabilistic model implementations, focusing on Bayesian approaches. This manuscript summarizes good practices for assessing the correctness of simulation and inference procedures under a model, and is available in the traditionally static print version as well as in a reproducible and executable form.

[Probabilistic models, Bayesian models, model validation, coverage]

Introduction

The last two decades have seen the biological sciences undergo a major revolution. Critical technological innovations such as the advent of massive parallel sequencing and the accompanying improvements in computational power and storage have flooded biology with unprecedented amounts of data ripe for analysis. Not only has intraspecific data from multiple individuals allowed progress in fields like medicine and epidemiology (e.g., The 1000 Genomes Project Consortium, 2015; Human Microbiome Project Consortium, 2012; Neafsey et al., 2015), population genetics (e.g., Lynch, 2007; Lack et al., 2016; de Manuel et al., 2016) and disease ecology (e.g., Rosenblum et al., 2013; Bates et al., 2018), but now a large number of species across the tree of life have had their genomes sequenced, furthering our understanding of species relationships and diversification (e.g., Martin et al., 2013; Brawand et al., 2014; Jarvis et al., 2014; Novikova et al., 2016; Pease et al., 2016; Kawahara et al., 2019; Upham et al., 2019). However, as the old adage goes, with great power comes great responsibility: never has the data available to the average biologist been so abundant, but also never has one been so aware of both its complexity and the necessary care needed to analyze it. Almost on par with with data accumulation is the rate at which new computational tools are being proposed, as evidenced by journals entirely dedicated to method advances, methodological sections in biological journals, and computational biology degrees being offered by institutions around the world.

One extreme case is the discipline of evolutionary biology (on which we focus our attention). While it could be said that many decade-old questions and hypotheses in evolutionary biology have aged well and stood up the test of time (e.g., the Red Queen hypothesis, Van Valen 1973; Lively 1987; Morran et al. 2011; Gibson and Fuentes-González 2015; the Bateson-Dobzhansky-Muller model, Dobzhansky 1936; Muller 1940; Hopkins and Rausher 2012; Roda et al. 2017), data analysis practices have changed drastically in recent years, to the point they would likely seem exotic and obscure to an evolutionary biologist active forty years ago. In particular, evolutionary biology has become highly statistical, with the development and utilization of models now being commonplace.

Models are employed in the sciences for many reasons, and fall within a biological abstraction continuum (Servedio et al., 2014), going from fully verbal, highly abstract models (e.g., Van Valen 1973), through proof-of-concept models that formalize verbal models (e.g., Maynard Smith 1978; Reinhold et al. 1999; Mendes et al. 2018), to models that interact directly with data through explicit mathematical functions (Yule, 1924; Felsenstein, 1973; Hasegawa et al., 1985; Hudson, 1990). Within the latter category, probabilistic models have seen a sharp surge in popularity within evolutionary biology, in conjunction with computational tools implementing them.

Despite the increasing pervasiveness of probabilistic models in the biological sciences, tools implementing such models show large variation not only with respect to code quality (from a software engineering perspective), but also correctness (Darriba et al., 2018). This is unsurprising given the multidisciplinary nature of model and method development, and the challenges inherent to software research funding (Siepel, 2019). The bioinformatics community is thus in dire need of resources that provide guidance for code improvement and validation.

Here, we summarize best practices in probabilistic model validation for method developers, with an emphasis on Bayesian methods. Scripts for reproducing our validation protocols and figures are available on [Link to DeveloperManual]. This repository also hosts tools for model validation within the BEAST 2 platform (Bouckaert et al., 2019).

Probabilistic models

Probabilistic models mathematically formalize natural phenomena having an element of randomness or uncertainty. This is done through probability distributions describing both the observed empirical data – seen as the result of one or more random instantiations of the modeled process – as well the model parameters, which abstract relevant, but usually unknown aspects of the phenomenon at hand. The historical, stochastic, and highly dimensional nature of evolutionary processes makes the utility of probabilistic models in evolutionary biology self-evident.

The central component of a probabilistic model, $\text{Pr}(D = d|\Theta = \theta)$, allows us to describe the probability distribution over the data (d) given the model parameters (θ). This probability mass function – pmf; or its continuous countepart, the probability density function, pdf, $f_D(d|\Theta = \theta)$ – is sometimes referred to as the likelihood function. Whenever there is no risk of confusion, we will simplify notation and drop variable subscripts, e.g., $f_D(d|\Theta = \theta) \rightarrow f(d|\theta)$. We will also assume variables are always continuous. As illustrated in the next sections, probabilistic models can be hierarchical, in which case there may be several likelihood functions. In a frequentist statistical framework, $f(d|\theta)$ is the sole component of a probabilistic model, and is maximized across parameter space during parameter estimation and model comparison.

In the present study we focus on Bayesian inference, however, where a probabilistic model $\mathcal{M}$ defines a posterior probability distribution for its parameters, $f(\theta|d) = \frac{f(d|\theta)f(\theta)}{f(d)}$. Here, our prior inferences or beliefs about the natural world – represented by the prior distribution $f(\theta)$ – are confronted with and updated by the data through the likelihood function – or multiple likelihood functions. The probability of the data, $f(d) = \int_\Theta f(d|\theta)f(\theta)d\theta$, is also known as the marginal likelihood or the model evidence and is of central importance in hypothesis testing and model comparison, but its computation can usually be circumvented if the main goal is parameter estimation. Crucially, a Bayesian model includes a prior, $f(\theta)$: when models are compared, for example, $f(\theta)$ needs to be taken into account when computing the model evidence $f(d)$.

Models routinely used in evolutionary biology are often characterized by continuous parameters, and are normally complex enough to preclude analytical solutions for the posterior density $f(\theta|d)$, mainly due to the intractability of the integral appearing in the marginal likelihood. In those cases, one can make use of the fact that $f(d)$ is a constant that can be ignored (i.e., $f(\theta|d) \propto f(\theta|d)f(\theta)$), and use techniques like Markov chain Monte Carlo (MCMC), to obtain approximate samples from the posterior distribution. This is because many MCMC algorithms, such as the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), only require the posterior to be evaluated up to a constant. In practice, the Metropolis-Hastings algorithm samples the posterior distribution (also referred to as the "target" distribution) by means of a transition mechanism. If the proposal distribution generated by this mechanism is irreducible, positive recurrent, and aperiodic, and the resulting chain is long enough, then the sampled posterior distribution will approximate the target distribution $f(\theta|d)$ (Smith and Roberts, 1993; Tierney, 1994; Gelman et al., 2013).

We note that the transition mechanism is not part of the model, however. It is possible in fact to sample $f(\theta|d)$ with other techniques such as importance sampling, Hamiltonian Monte Carlo (Duane et al., 1987), or even by converting the sampling problem into an optimization one (e.g., Zhang and Matsen, 2019). Below, we focus on practices for verifying model *correctness* – while MCMC can be a critical component in model evaluation, and is a technique we employ here, we do not pay attention to this sampling machinery. Tests to ensure an MCMC analysis is converging to its target as desired exist, but are outside the scope of this work. @Fabio: maybe some citations here? I wonder if more phylogenetics-oriented ones would go better...

Validating the machinery

Validating the simulator, S$[\mathcal{M}]$

When a probabilistic model $\mathcal{M}$ is implemented for the first time, a simulator S$[\mathcal{M}]$ must be devised and itself validated before we can validate an inferential engine I$[\mathcal{M}]$, which will be ultimately employed by users in empirical analyses. A simulator conventionally requires a parameter value as input (i.e., a θ value, where $\boldsymbol{\theta}$ might represent more than one parameter), or a prior distribution on those values, $f(\boldsymbol{\theta})$. The simulator then outputs a sample of random variable(s), which for hierarchical models will include not only an instantiation of data D, but also of a subset of the parameters in $\boldsymbol{\theta}$.

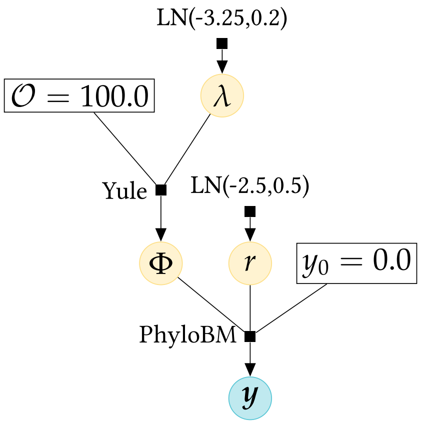

Figure 1: The graphical representation of a simple Bayesian phylogenetic model. Parameters are represented by yellow circles, observed data in blue circles, and fixed values are shown within boxes. Sampling distributions are represented by filled squares. LN(μ, σ^2) denote log-normal sampling distributions with mean μ and variance σ^2 in log-space. Parameter descriptions can be found in the main text.

In the case of hierarchical models, it is sometimes useful to consider S$[\mathcal{M}]$ as a collection of component simulators, each characterized by a different sampling distribution. In figure X, for example, S$[\mathcal{M}]$ can be seen as an ensemble comprised by (i) S$[f(\boldsymbol{\theta})]$ (where $\boldsymbol{\theta} = \{\lambda, r, y_0\}$), which jointly simulates all parameters in $\boldsymbol{\theta}$, (ii) S$[f(\Phi|\lambda)]$, which simulates a Yule tree Φ given a value of λ simulated in (i), and (iii) S$[f(\boldsymbol{y}|\Phi, r)]$, which simulates an array of n continuous-trait values, $\boldsymbol{y}$, given a phylogeny Φ with n species and an evolutionary rate r (simulated in [i] and [ii], respectively). Being able to isolate the building blocks of a hierarchical model simulator helps divide and conquer the validation task, especially when some, but not all of the sampling distributions are well-known parametric distributions, or when they result from well characterized stochastic processes (see below).

One way to validate a probabilistic model simulator is by using it to sample a large number of data sets given a set of parameters. These parameters can be seen as characterizing a “population” of the entities being modeled. For each data set, one can then construct $100 \times \alpha\%$-confidence intervals (where $\alpha \in (0, 1)$ denotes the credibility level) for certain summary statistics (e.g., mean, variance, covariance). If the simulator is behaving as expected, one should be able to verify that the (population) summary statistic is contained approximately $\alpha\%$ of the time within their α-confidence intervals. An example is the Yule model (also known as the pure-birth model; Yule 1924), a continuous-time Markov process that has been classically employed in phylogenetics to model the number of species in a clade (Yule, 1924; Aldous, 2001). Under a Yule process with a species birth rate of λ, the expected tree height, $\text{E}[t_{\text{root}}]$, for a tree with n tips is:

Box 1: Models characterized by well-known parametric distributions

One commonly used model in macroevolution for the study of continuous traits is the phylogenetic Brownian motion model (“PhyloBM” in Fig. X; Felsenstein 1973). The pdf characterizing this model’s sampling distribution is in fact the pdf of the multivariate normal (MVN) probability distribution:

$$\begin{aligned}\log f(\boldsymbol{y} \mid \boldsymbol{y_0}, r, \boldsymbol{T}) = &-\frac{1}{2}\Big[n\log(2\pi) + \log|r\boldsymbol{T}|\Big] \\ &-\frac{1}{2}\Big[(\mathbf{y} - \boldsymbol{y_0})^T(r\boldsymbol{T})^{-1}(\mathbf{y} - \boldsymbol{y_0})\Big],\end{aligned} \tag{1}$$

where $\boldsymbol{y}$ corresponds to the observed values of a trait scored for n species, $\boldsymbol{y_0}$ is the trait value at the root of the tree, r is the instantaneous rate of change (i.e., the evolutionary rate, and sometimes represented by σ^2), and $r\boldsymbol{T}$ is the variance-covariance matrix. $\boldsymbol{T}$ is a matrix whose elements are deterministically defined by tree Φ’s topology and branch lengths; see Fig. 1 below).

The probability density function in equation (1) describes the distribution that would result from an infinite number of BM “experiments” (each experiment being non-mean-reverting, and representing an independent evolutionary trajectory). Under this model, $\boldsymbol{\theta} = \{\boldsymbol{y_0}, r, \boldsymbol{T}\}$ and $d = \{\boldsymbol{y}\}$ (but note that sometimes researchers treat Φ and consequently $\boldsymbol{T}$ as data).

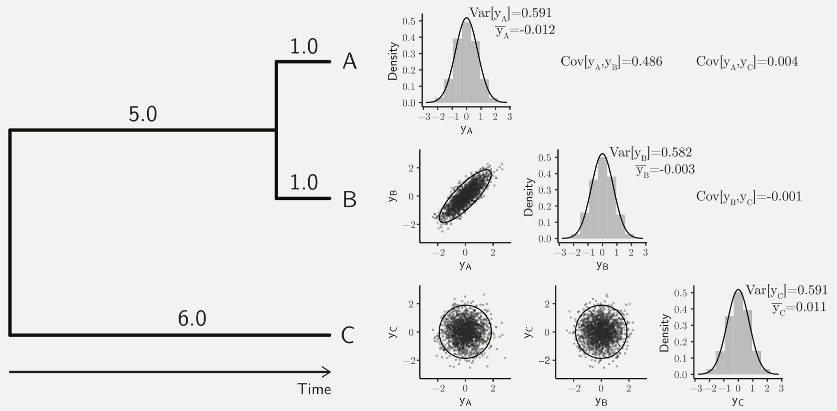

Figure 2: A sample of 1000 draws from a MVN distribution, each representing the evolutionary trajectory of one continuous trait along the species tree on the left. The root trait value, $\boldsymbol{y_0}$, and the evolutionary rate of the process, r, were set to 0.0 and 0.1, respectively. The panel on the right shows histograms of 1000 trait values sampled from the MVN for each species, as well as their covariation.

Validating a phylogenetic BM simulator

The MVN is a well-characterized parametric distribution. When used as the sampling distribution of the phylogenetic BM process, it explicitly defines the expected trait value for each species ($\boldsymbol{y_0}$), as well as their trait value variances and covariances. The latter comes from the variance-covariance matrix; for the tree shown in figure 1 and with $r = 0.1$, this matrix is:

$$r\boldsymbol{T} = 0.1 \begin{bmatrix} 6 & 5 & 0 \\ 5 & 6 & 0 \\ 0 & 0 & 6 \end{bmatrix} \tag{2}$$

Together, variance-covariance matrix $r\boldsymbol{T}$ and $\boldsymbol{y_0} = [0.0, 0.0, 0.0]$ characterize a population of phylogenetically related species trait values whose means are 0.0, variances are 6.0, and co-variances are 5.0 (between species "A" and "B") and 0.0 (between species "C"and either "A" or "B").

Figure 1 shows the distributions of trait values and their variances and covariances for one sample of X independent realizations of phylogenetic BM processes. One can see that the sample's average trait value and the variances and covariances approach their expectations. In order to be igorous, one can follow the method described in the main text and verify that those expectations fall within their 95% confidence intervals 95% of the time, as calculated from a large number of samples (Supplementary Fig. 1 and Supplementary Table 1).

$$\mathrm{E}[t_{\mathrm{root}}] = \sum_{i=2}^{n} \frac{1}{1\lambda}. \tag{3}$$

One can then verify if $\mathrm{E}[t_{\mathrm{root}}]$ is 95% of the time within ±1.96 standard errors of the average Yule-tree height (from each sampled data set). Confirming that this is the case indicates $\mathrm{S}[f(\Phi|\lambda)]$ is correctly implemented (Fig. 3). Other, similar checks are discussed in the supplementary material. In Box 1, we illustrate this procedure for the (parametric) sampling distribution underlying the phylogenetic Brownian motion model ("PhyloBM"; Felsenstein, 1973). Protocols for validating $\mathrm{I}[\mathcal{M}]$ (see below) will also automatically validate $\mathrm{S}[\mathcal{M}]$.

We note that we have so far used $\mathrm{S}[\mathcal{M}]$ to represent a *direct* simulator under model $\mathcal{M}$ (Table 1), meaning each and every sample generated by $\mathrm{S}[\mathcal{M}]$ is independent. This is contrast with other simulation strategies, such as conducting MCMC under model $\mathcal{M}$ with no data, given specific parameter ($\boldsymbol{\theta}$) values. This latter approach may be the only option if $\mathrm{S}[\mathcal{M}]$ has not been yet implemented, and it is predicated upon the existence of correct implementations of both an inferential engine $\mathrm{I}[\mathcal{M}]'$ and of proposal functions. We distinguish $\mathrm{I}[\mathcal{M}]'$ from $\mathrm{I}[\mathcal{M}]$ because simulations are being carried out precisely to validate $\mathrm{I}[\mathcal{M}]$. Unless MCMC simulations are done with $\mathrm{I}[\mathcal{M}]'$ – an independent

(a)

Birth rate (λ)	E[t_{root}] $\in$ 95% CI (%)
0.5	94
0.6	91
0.7	98
0.8	95
0.9	96
1.0	92

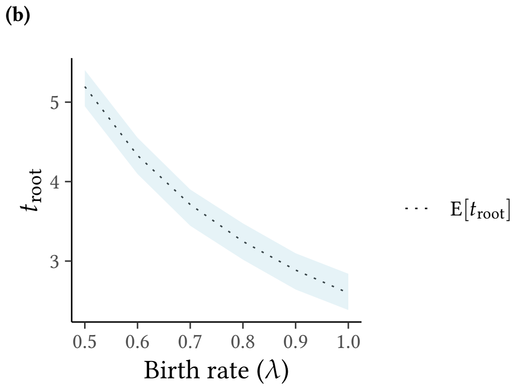

Figure 3: Validation of Yule tree simulator. (a) Number of simulated data sets (out of 100) for which the expected tree height (t_{root}) was inside the 95% CI about its sample average. Each data set consisted of 50 twenty-taxon simulated Yule trees. (b) The area shaded in light blue represents the 95% confidence interval about the average tree height, obtained from the 5,000 Yule trees simulated in (a). Simulations were carried out with the `TreeSim` R package (Stadler, 2011).

implementation of I[$\mathcal{M}$] – they can introduce circularity to the validation task.

Table 1: A non-exhaustive list of direct simulation software commonly used in evolutionary biology analyses.

Software package	Model type	Platform	Reference
Seq-Gen	Molecular sequence evolution models	Standalone	Rambaut and Grass, 1997
ms	Coalescent model	Standalone	Hudson, 2002
SLiM	Population genetic models	Standalone	Haller and Messer, 2019
TreeSim	Birth-death models	R	Stadler, 2011
mvMORPH	Continuous trait evolution models	R	Clavel et al., 2015
phytools	Several phylogenetic models	R	Revell, 2012
MASTER	Continuous-time Markov (tree) models	BEAST 2	Vaughan and Drummond, 2013

Validating the inferential engine, I[$\mathcal{M}$]

The more complex the natural phenomenon under study, the more difficult it will be to strike a good balance between model practicality and realism. The popular aphorism rings true: “all models are wrong but some are useful” (Box, 1979). Very simple models are easier to implement in efficient inference tools, but will commonly make assumptions likely to be broken by the data (e.g., Sullivan and Swofford, 1997; Mendes and Hahn, 2018; Mendes et al., 2019). Conversely, complex models will fit the data better (e.g., Ogilvie et al., 22), but may become unwieldly with increasing levels of realism. A large number of parameters can cause overfitting and unidentifiability, and highly complex models might lead to prohibitively slow inferential analyses (e.g., Lartillot and Poujol, 2011).

Deciding on the utility of a model for real-world problems is a daunting task (Brown and Thomson, 2018; Shepherd and Klaere, 2018), and is a challenge we do not address in the present contribution. Such model appraisals are normally carried out after a model is published, often in multiple contribution bouts, and are critical for a model's longevity. Analyses of model fit against data are normally accompanied by discussions on assumption validity, and more rarely by benchmarking and scrutinization of model behavior and implementation (e.g., Maddison et al., 2007; Rabosky and Goldberg, 2015; Rabosky et al., 2013; Moore et al., 2016; Stadler, 2010; Luo et al., 2020).

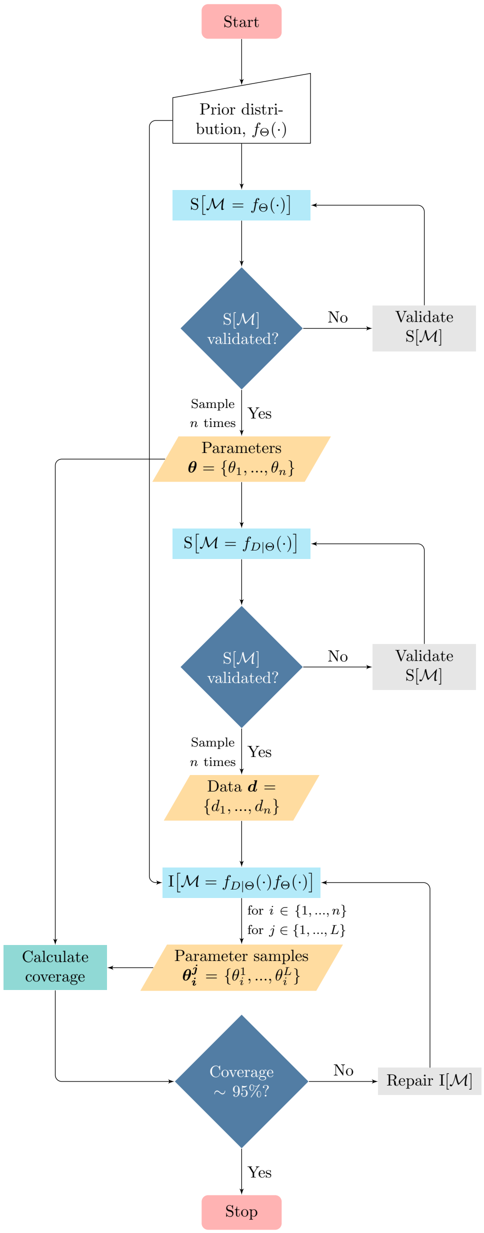
Start
Prior distribution, $f_\Theta(\cdot)$
S[$\mathcal{M} = f_\Theta(\cdot)$]
S[$\mathcal{M}$] validated?
No
Validate S[$\mathcal{M}$]
Sample n times
Yes
Parameters $\boldsymbol{\theta} = \{\theta_1, ..., \theta_n\}$
S[$\mathcal{M} = f_{D|\Theta}(\cdot)$]
S[$\mathcal{M}$] validated?
No
Validate S[$\mathcal{M}$]
Sample n times
Yes
Data $\boldsymbol{d} = \{d_1, ..., d_n\}$
I[$\mathcal{M} = f_{D|\Theta}(\cdot) f_\Theta(\cdot)$]
for $i \in \{1, ..., n\}$
for $j \in \{1, ..., L\}$
Parameter samples $\boldsymbol{\theta}_i^j = \{\theta_i^1, ..., \theta_i^L\}$
Calculate coverage
Coverage $\sim 95\%$?
No
Repair I[$\mathcal{M}$]
Yes
Stop

Figure 4: The several steps involved in determining a Bayesian model is well-calibrated. Note that even though we present simulation in two , one for prior $f_\Theta(\cdot)$ and one for the likelihood $f_D(\cdot|\theta)$, a Bayesian model $\mathcal{M}$ is defined by the density functions of both prior and likelihood.

When a new model $\mathcal{M}$ is initially proposed, however, authors must ensure that their methods can at the very least robustly recover generating parameters. In this section, we discuss a few techniques that can be employed to assess the correctness of a parameter-estimation routine. These techniques assume that one can simulate from a probabilistic data-generating process (see previous section).

Coverage-based validation of Bayesian inference machinery

Our discussion on how to ensure a Bayesian model is well-calibrated and thus correct will mostly follow the ideas in Cook et al. (2006) and Talts et al. (2018). The basic idea is presented in the flowchart in figure 4, and consists of three stages: simulation, inference, and coverage calculation. Once we have a validated simulator for model $\mathcal{M}$, we start by sampling n parameter sets from $\boldsymbol{\theta} = \{\theta_1, \theta_2, ..., \theta_n\}$ from its prior, $f(\theta)$. For each parameter set θ_i, we then sample a data set d_i from $f(d|\theta)$. These two steps conclude the "simulation" stage of this validation protocol. With $\boldsymbol{d} = \{d_1, d_2, ..., d_n\}$, we use the inferential machinery I[$\mathcal{M}$] under evaluation to compute $f(\theta|d_i)$ for each d_i. Recall that we assume the posterior distribution defined by $f(\theta|d)$ over θ will be approximated with MCMC, an algorithm that generates a large sample of size L of parameter values given a data set (d_i), $\boldsymbol{\theta}_i^j = \{\theta_i^1, \theta_i^2, ..., \theta_i^L\}$. At this point, we have concluded the

k	$\Pr(x = k)$
90	0.0167
91	0.0349
92	0.0649
93	0.1060
94	0.1500
95	0.1800
96	0.1781
97	0.1396
98	0.0812
99	0.0312
100	0.0059

Table 2: Under a correctly implemented model, coverage x (the number of true simulated values that fall within their corresponding 95%-HPDs) is binomially distributed with n trials ($n = 100$ in this case), and probability of success $p = 0.95$. @Fabio, is this table really necessary?

inference stage of this validation pipeline.

The third stage and final stage consists of investigating coverage properties of uncertainty intervals. The critical expectation here is that if the inferential engine is correct, we will be able to obtain interval estimates with precise coverage properties. More concretely, let us first define the highest posterior density (HPD) interval. For a credibility level $\alpha \in (0, 1)$, we define $I_\alpha(y) := (a(y, \alpha), b(y, \alpha))$ such that

$$\frac{1}{f(d)} \int_{a(y,\alpha)}^{b(y,\alpha)} f(d|\theta) f(\theta)\, d\theta = \alpha,$$

where $f(d) = \int_\Theta f(d|\theta) f(\theta) d\theta$ is a constant that can be ignored. As there are potentially multiple intervals which attain the required credibility, we find the shortest such interval.

Now taking a set of parameter values θ_1, and data points d_1, sampled from $f_\Theta(\cdot)$ and $f_D(\cdot|\theta)$, respectively:

$$\theta_1 \sim f_\Theta(\cdot),$$
$$y_1 \sim f_D(\cdot|\theta_1),$$

it can be shown that $\Pr(\theta_1 \in I_\alpha(d)) = \alpha$, i.e., that $100 \times \alpha\%$ HPDs have nominal coverage under the true generative model. A proof is provided in the supplementary material. @Fabio There's a notation change here. I do prefer this notation to what has come before, but we need to be consistent. So I added a proviso at the top of the paper to say that we drop the subscripts **when convenient**.

We define a method's *coverage* as the number of intervals that trap the data-generating parameter value under repeated sampling. The inference engine $\mathrm{I}[\mathcal{M}]$ of a Bayesian model is said to be well-calibrated and correct if we

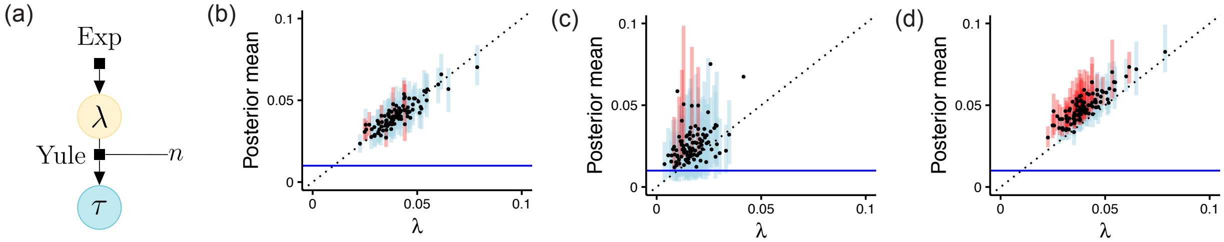

Figure 5: [This figure is being updated] Coverage-validation analyses of a simple Bayesian hierarchical model (Fig. 1). Panels show the true (i.e., simulated or generating) parameter values plotted against their mean posteriors (the dashed line gives $f(x) = x$). Dots and lines represent true values and their 95%-HPDs, respectively. Simulations for which 95%-HPDs contained the true value are highlighted in blue, otherwise are presented in red. (a) Model is correctly specified. (b) Model is misspecified in inference (log-normal on λ is misspecified). (c) Model is correctly specified, but only small trees (with up to fifteen taxa) are generated.

ascertain that its coverage lies within the expected bounds. More specifically, the coverage of n intervals obtained as above will be distributed as binomial random variable with n trials and success probability α. When $n = 100$ and $\alpha = 0.95$, the confidence interval for the number of simulations containing the correct data-generating parameter is between 90 and 99 (Table 2).

We provide an example in figure 5, which shows coverage graphical summaries for the model represented in figure 1. This model is deliberately simple for the sake of brevity and clarity in the discussion below. The parameters in this model are the phylogenetic tree Φ, the species birth rate λ, and the continuous-trait evolutionary rate r. We assume the continuous trait-root value, $\boldsymbol{y_0}$, is known and set it to **0.0** for all simulated data sets. When the model is correctly specified (Fig. 5a), coverage is close to 95% and is adequate. In figure 5b, however, we misspecify the model during inference, setting the prior distribution on λ to be a log-normal with a mean of -2.25 (rather than -3.25, as specified in the simulation procedure; Fig. 1). As can be seen, coverage is much lower (X%), which suggests that the model is not behaving as expected. Of course, in a real-world validation experiment the model should be correctly specified, and such a result would suggest a problem with the inferential machinery – provided that the simulator has been previously validated.

Finally, we can further capitalize on this validation setup and gauge how accurate our inferential tool can be for different parameters. The more identifiable a parameter is, the higher should the correlation between its posterior mean and its generating “true” value be.‘ In our first example (Fig. 5a), the species birth-rate λ is largely identifiable for trees of the simulated size (Fig. 5a). In a separate analysis, we analyzed the same model save for one difference: phylogenetic trees could only have up to fifteen terminal taxa. The results are shown in figure 5c. Here, it can be seen that the inference machinery is less accurate and that the λ parameter is less identifiable. We note that unidentifiability should not be taken as a sign that a model is incorrect – in figure 5c, coverage is still appropriate. @Fabio: I’m not sure I understand the claim here: what would be a more ’identifiable’ parameter? Do you mean a parameter for which inferences converge faster to the true, data-generating value? This is a separate issue from identifiability; it has to do with the rate of strong consistency. I advise against using ’identifiability’ in such a way.

Simulation-based calibration (SBC)

We now discuss another technique that might be combined with coverage-based checking to further scrutinize the inference machinery using the same samples generated in the coverage step.

Talts et al. (2018) show that one can devise other tests that might be more powerful to detect problems than just looking at the coverage of Bayesian HPD intervals. In particular, they show (Theorem 1 therein) that if the inference machinery $\mathrm{I}[\mathcal{M}]$ works as intended, the distribution of the rank r^j of the j-th sample from an MCMC chain ($\boldsymbol{\theta}^j = \{\theta^1, \theta^2, ..., \theta^L\}$) (@Fabio, this notation does not make sense: if j is a superscript, it does not make sense for $1, 2, \ldots, L$ to also be superscripts) will follow a uniform distribution on $[1, L+1]$. In other words, if the sampling algorithm is correctly implemented, we expect the "true" data-generating parameter to fall uniformly within the posterior samples. Adherence to this distribution can be investigated by constructing histograms (Talts et al., 2018) as well as by looking at the empirical cumulative distribution function (ECDF) and their confidence bands (Säilynoja et al., 2021).

We illustrate the SBC procedure on the model represented in figure 1... [@Remco or @Luiz: to be done]

Items to discuss/consider after example:

- Would it help to have a flowchart for the SBC procedure?
- Investigate whether SBC flags problems that coverage doesn't when one restricts the tree sizes;
- Show how a misspecified model can cause this test to fail (if we take the Yule + phyloBM model as an example, as I suggest, there are only a few parameters to be misspecified, so maybe it's easy to get this test to fail and know exactly why. One way to do that with that model, I'd suspect, is to throw away all trees with fewer than, say, 15 leaves – I would imagine that would distort our λ rank distribution, even if the model still pass the well-calibrated test);

Importantly, the patterns of deviation from uniformity displayed by the rank histograms resulting from an incorrectly implemented inference machinery are also informative. Here, we will give a very brief overview of how to interpret the shape of the rank histograms; for more details, the reader is referred to Section 4.2 of Talts et al. (2018). If the histogram display a ∪ shape, with "horns", one of two things can be happening: either the posterior samples being used have high autocorrelation or that the posterior being sampled from is narrower than the true posterior. Determining which of these two possibilities is likely the cause can be achieved by thinning the posterior samples in order to reduce autocorrelation and see whether that changes the resulting histogram. Another possible shape for the histogram is the ∩ shape, where the ranks clump in the middle; this usually indicates that the posterior being targeted is wider (over-dispersed) than the true posterior. Finally, the histogram might be asymmetric, with a spike on either end. In this situation, there likely is a location mismatch between the posterior being sampled from and and the true posterior.

Box 1: Validating a phylogenetic model with respect to its phylogenetic tree parameter

Given the centrality of the phylogenetic tree (Φ) in comparative analyses, we must pay close attention to how we investigate this parameter when validating a phylogenetic model. Analyzing phylogenetic trees is made challenging by tree space being a complex mix of a discrete and continuous component, due to trees being comprised by both a topology and set of node times (Semple and Steel, 2003; Gavryushkin and Drummond, 2016). Additionally, there is no canonical total-ordering structure for trees, which complicates a validation procedure such as SBC.

To get around this difficulty, we propose computing one or more phylogenetic metrics, or functionals, for which total-ordering holds and thus for which ranks can be obtained. One such metric is the well-known Robinson-Foulds distance (RF; Robinson and Foulds, 1981), which counts the number of different clades (splits) relative to another phylogenetic tree. In order to compute a relational metric like the RF distance during validation, we must have a reference phylogeny Φ_0 to which we can compare our focal generating phylogeny Φ and its posterior MCMC samples. The simulation-based calibration protocol remains the same, with an additional step in which we generate Φ_0 (see Algorithm 1 in the supplementary material). Figure X [@Luiz: see my description of new Fig. X below] shows validation results for the RF metric for trees simulated under the model illustrated in Fig 1. Figure Xa and Xb show the coverage of 95% HPD intervals, and the rank distribution of the RF metric, respectively. The coverage of the RF statistic is very close to the nominal, and the rank distribution is approximately uniform on $[1, L+1]$; together, both panels indicate this implementation is correct. We consider other phylogenetic tree metrics that could be used as an alternative to RF in the supplementary material (Supplementary Figs. 2 and 3).

[@Luiz: include Fig Xa and FigXb here; this figure would consist of just the RF panels; all other metrics have would be placed in the supp. material]

Concluding remarks

As more data is generated and made publicly available, the more will researchers in the life sciences require computational methods with which to analyze it. If such methods are not correctly implemented, conclusions drawn from the data will be of reduced or void of any significance. Here we discussed guidelines for the validation of computational methods implementing Bayesian probabilistic models. This manuscript is also followed by a supplementary text further illustrating these guidelines that is linked with a live document available on https://github.com/rbouckaert/DeveloperManual. We hope our guidelines can help raise the standards for software package releases required by users, developers and reviewers alike, and consequently lead to computational tools that are more efficient, better documented, and most importantly, correctly implemented.

Funding

F.K.M. and A.J.D. were supported by Marsden grant 16-UOA-277. R.B. was supported by Marsden grant .

References

Aldous, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Stat. Sci.*, 16:23–24.

Bates, K. A., Clare, F. C., O'Hanlon, S., Bosch, J., Brookes, L., Hopkins, K., McLaughlin, E. J., Daniel, O., Garner, T. W. J., Fisher, M. C., and Harrison, X. A. (2018). Amphibian chytridiomycosis outbreak dynamics are linked with host skin bacterial community structure. *Nature Comm.*, 9:1–11.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kühnert, D., Maio, N. D., Matschiner, N., Müller, N. F., Ogilvie, H., du Plessis, L., Popinga, A., Mendes, F. K., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comp. Biol.*, 15:e1006650.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. In Launer, R. L. and Wilkinson, G. N., editors, *Robustness in statistics*, pages 201–236. Academic Press.

Brawand, D. et al. (2014). The genomic substrate for adaptive radiation in African cichlid fish. *Nature*, 513(7518):375–381.

Brown, J. M. and Thomson, R. C. (2018). Evaluating model performance in evolutionary biology. *Annu. Rev. Ecol. Evol. Syst.*, 49:95–114.

Clavel, J., Escarguel, G., and Merceron, G. (2015). mvMORPH: an R package for fitting multivariate evolutionary models to morphometric data. *Methods Ecol. Evol.*, 6:1311–19.

Cook, S. R., Gelman, A., and Rubin, D. B. (2006). Validation of software for bayesian models using posterior quantiles. *J. Comput. Graph. Stat.*, 15(3):675–692.

Darriba, D., Flouri, T., and Stamatakis, A. (2018). The state of software for evolutionary biology. *Mol. Biol. Evol.*, 35:1037–46.

de Manuel, M. et al. (2016). Chimpanzee genomic diversity reveals ancient admixture with bonobos. *Science*, 354:477–81.

Dobzhansky, T. (1936). Studies on hybrid sterility. II. Localization of sterility factors in *Drosophila pseudoobscura* hybrids. *Genetics*, 21:113–35.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–22.

Felsenstein, J. (1973). Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.*, 25:471–92.

Gavryushkin, A. and Drummond, A. J. (2016). The space of ultrametric phylogenetic trees. *J. Theor. Biol.*, 403:197–208.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis.* CRC Press, Boca Raton, Florida.

Gibson, A. K. and Fuentes-González, J. A. (2015). A phylogenetic test of the Red Queen Hypothesis: outcrossing and parasitism in the Nematode phylum. *Evolution*, 69:530–40.

Haller, B. C. and Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–Fisher model. *Mol. Biol. Evol.*, 36.

Hasegawa, M., Kishino, H., and Yano, T. A. (1985). Dating of the human age splitting by a molecular clock of mitochondrial-DNA. *J. Mol. Evol.*, 22:160–74.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57.

Hopkins, R. and Rausher, M. D. (2012). Pollinator-mediated selection on flower color allele drives reinforcement. *Science*, 335:1090–92.

Hudson, R. R. (1990). Gene genealogies and the coalescent process. *Oxf. Surv. Evol. Biol.*, 11:1–44.

Hudson, R. R. (2002). Generating samples under a Wright-Fisher neutral model. *Bioinformatics*, 18:337–8.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature*, 486:215–221.

Jarvis, E. D. et al. (2014). Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, 346(6215):1320–1331.

Kawahara, A. Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E. F. A., Donath, A., Gimnich, F., Frandsen, P. B., Zwick, A., dos Reis, M., Barber, J. R., Peters, R. S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J. E., Misof, B., and Breinholdt, J. W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. *Proc. Natl. Acad. Sci. USA.*, 116:22657–63.

Lack, J. B., Lange, J. D., Tang, A. D., Corbett-Detig, R. B., and Pool, J. E. (2016). A thousand fly genomes: an expanded *Drosophila* genome nexus. *Mol. Biol. Evol.*, 33:3308–13.

Lartillot, N. and Poujol, R. (2011). A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.*, 28.

Lively, C. M. (1987). Evidence from a New Zealand snail for the maintenance of sex by parasitism. *Nature*, 328:519–21.

Luo, A., Duchêne, D. A., Zhang, C., Zhu, C.-D., and Ho, S. Y. W. (2020). A simulation-based evaluation of tip-dating under the fossilized birth-death process. *Syst. Biol.*, 69:325–344.

Lynch, M. (2007). Population genomics of *Daphnia pulex*. *Genetics*, 206:315–32.

Maddison, W. P., Midford, P. E., and Otto, S. P. (2007). Estimating a binary character's effect on speciation and extinction. *Syst. Biol.*, 56:701–710.

Martin, S. H., Dasmahapatra, K. K., Nadeau, N. J., Salazar, C., Walters, J. R., Simpson, F., Blaxter, M., Manica, A., Mallet, J., and Jiggins, C. D. (2013). Genome-wide evidence for speciation with gene flow in Heliconius butterflies. *Genome Res.*, 23(11):1817–28.

Maynard Smith, J. (1978). *The evolution of sex.* Cambridge University Press.

Mendes, F. K., Fuentes-González, J. A., Schraiber, J., and Hahn, M. W. (2018). A multispecies coalescent model for quantitative traits. *eLife*, 7:e36482.

Mendes, F. K. and Hahn, M. W. (2018). Why concatenation fails near the anomaly zone. *Systematic Biology*, 67(1):158–169.

Mendes, F. K., Livera, A. P., and Hahn, M. W. (2019). The perils of intralocus recombination for inferences of molecular convergence. *Philosophical Transactions of the Royal Society B*, 374:20180244.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., and Teller, A. H. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087–1092.

Moore, B. R., Höhna, S., May, M. R., Rannala, B., and Huelsenbeck, J. P. (2016). Critically evaluating the theory and performance of bayesian analysis of macroevolutionary mixtures. *Proc. Natl. Acad. Sci. USA.*, 113:9569–9574.

Morran, L. T., Schmidt, O. G., Gelarden, I. A., II, R. C. P., and Lively, C. M. (2011). Running with the Red Queen: host-parasite coevolution selects for biparental sex. *Science*, 333:216–18.

Muller, H. J. (1940). Bearing of the *Drosophila* work on systematics. In Huxley, J. S., editor, *The new systematics*, pages 185–268. Clarendon Press, Oxford.

Neafsey, D. E. et al. (2015). Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science*, 347(6217):1258522.

Novikova, P. Y. et al. (2016). Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.*, 48(9):1077–1082.

Ogilvie, H. A., Mendes, F. K., Matzke, N. J., Stadler, T., Welch, D., and Drummond, A. J. (22). Novel integrative modeling of molecules and morphology across evolutionary timescales. *Systematic Biology*, 71:208–220.

Pease, J. B., Haak, D. C., Hahn, M. W., and Moyle, L. C. (2016). Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.*, 14(2):e1002379.

Rabosky, D. L. and Goldberg, E. E. (2015). Model inadequacy and mistaken inferences of trait-dependent speciation. *Syst. Biol.*, 64:340–355.

Rabosky, D. L., Santini, F., Eastman, J., Smith, S. A., Sidlauskas, B., Chang, J., and Alfaro, M. E. (2013). Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.*, 4(1958):1–8.

Rambaut, A. and Grass, N. C. (1997). Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13:235–38.

Reinhold, K., Engqvist, L., Misof, B., and Kurtz, J. (1999). Meiotic drive and evolution of female choice. *Proc. R. Soc. Lond. B*, 266:1341–45.

Revell, L. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.*, 3:217–23.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Math. Biosci.*, 53(1-2):131–147.

Roda, F., Mendes, F. K., Hahn, M. W., and Hopkins, R. (2017). Genomic evidence of gene flow during reinforcement in Texas *Phlox*. *Mol. Ecol.*, 26:2317–30.

Rosenblum, E. B. et al. (2013). Complex history of the amphibian-killing chytrid fungus revealed with genome resequencing data. *Proc. Natl. Acad. Sci. USA.*, 110:9385–90.

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2021). Graphical test for discrete uniformity and its applications in goodness of fit evaluation and multiple sample comparison. *arXiv preprint arXiv:2103.10522*.

Semple, C. and Steel, M. (2003). Phylogenetics. Oxford University Press.

Servedio, M. R., Brandvain, Y., Dhole, S., Fitzpatrick, C. L., Goldberg, E. E., Stern, C. A., Cleve, J. V., and Yeh, D. J. (2014). Not just a theory – the utility of mathematical models in evolutionary biology. *PLoS Biology*, 12:e1002017.

Shepherd, D. A. and Klaere, S. (2018). How well does your phylogenetic model fit your data? *Syst. Biol.*, 68:157–167.

Siepel, A. (2019). Challenges in funding and developing genomic software: roots and remedies. *Genome Biol.*, 20(147).

Smith, A. F. M. and Roberts, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Stat. Soc. B*, 55:3–23.

Stadler, T. (2010). Sampling-through-time in birth–death trees. *J. Theor. Biol.*, 267(3):396–404.

Stadler, T. (2011). Simulating trees with a fixed number of extant species. *Syst. Biol.*, 60:676–84.

Sullivan, J. and Swofford, D. L. (1997). Are guinea pigs rodents? the importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.*, 4:77–86.

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2018). Validating Bayesian inference algorithms with simulation-based calibration. *arXiv preprint arXiv:1804.06788*.

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature*, 526:68–74.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Ann. Stat.*, 22:1701–62.

Upham, N. S., Esselstyn, J. A., and Jetz, W. (2019). Inferring the mammal tree: Species-level sets of phylogenies for questions in ecology, evolution, and conservation. *PLoS Biol.*, 17:e3000494.

Van Valen, L. (1973). A new evolutionary law. *Evolutionary Theory*, 1:1–30.

Vaughan, T. G. and Drummond, A. J. (2013). A stochastic simulator of birth–death master equations with application to phylodynamics. *Mol. Biol. Evol.*, 30:1480.

Yule, G. U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, FRS. *Philos. Trans. R. Soc. London Ser. B*, 213:21–87.

Zhang, C. and Matsen, F. A. (2019). Variational bayesian phylogenetic inference. In *International Conference on Learning Representations*.