How to validate a Bayesian model

Fábio K. Mendes[1*], Remco Bouckaert[2*],

Christiaan Swanepoel[2], Alexei J. Drummond[1,2]

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

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

Corresponding authors*: f.mendes@auckland.ac.nz, remco@cs.auckland.ac.nz

October 25, 2021

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. This manuscript is also presented in a reproducible and executable version [LINK TO STENCILA VERSION], and is accompanied by code examples for these practices (with the BEAST 2 platform as a reference; Bouckaert et al., 2019).

Probabilistic models

Probabilistic models mathematically formalize natural phenomena having an element of randomness. 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, $\mathrm{P}(D|\theta)$, describes the probability distribution over the data given the model parameters. This probability mass function (pmf; or its continuous countepart, the probability density function $f(D|\theta)$) is sometimes referred to as the likelihood function. 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, $\mathrm{P}(D|\theta)$ is the sole component of a probabilistic model, and is maximized across parameter (θ) 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, $\mathrm{P}(\theta|D) = \frac{\mathrm{P}(D|\theta)\mathrm{P}(\theta)}{\mathrm{P}(D)}$. Here, our prior inferences or beliefs about the natural world – represented by the prior distribution $\mathrm{P}(\theta)$ – are confronted and updated by the data through the likelihood function (or multiple likelihood functions). $\mathrm{P}(D) = \int_\theta \mathrm{P}(D|\theta)\mathrm{P}(\theta)d\theta$, the probability of the data, is also known as the marginal likelihood or the model evidence. Crucially, a Bayesian model includes a prior, $\mathrm{P}(\theta)$: when models are compared, for example, $\mathrm{P}(\theta)$ needs to be taken into account when computing the model evidence $\mathrm{P}(D)$.

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 sample the posterior distribution. This is because the MCMC algorithm that generates the Markov chain, called the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970) algorithm, only requires 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 will spend time considering MCMC in particular as it is the commonly chosen technique for obtaining $f(\theta|D)$ under an implementation of model $\mathcal{M}$. A thorough validation effort thus entails verifying the correctness of (i) the model (i.e., $f(D|\theta)f(\theta)$), and (ii) the components involved in the MCMC transition mechanism. We note that the latter are not part of the model, however, and it is possible 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).

Finally, we stress that we are interested in practices for verifying model *correctness*. There are other tests employed

to ensure that a particular MCMC analysis is converging as anticipated. Determining that one or more independent Markov chains converged on very similar posterior distributions is not a correctness test, as those distributions might be very different from the target distribution.

Validating a Bayesian model

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 $I[\mathcal{M}]$. It is $I[\mathcal{M}]$ that will be 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 $\boldsymbol{\theta}$.

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 $\boldsymbol{\theta} = \{\lambda, r, r_m, y_0\}$), which jointly simulates all parameters in $\boldsymbol{\theta}$, (ii) $S[f(\Phi|\lambda)]$, which simulates a Yule tree Φ, (iii) $S[f(\mathbf{A}|\Phi, r)]$, which simulates a multiple sequence alignment $\mathbf{A}$, and (iv) $S[f(\mathbf{M}|\Phi, r_m, y_0])$, which simulates a continuous character vector $\mathbf{M}$ for all species in Φ. 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 of validating a probabilistic model simulator is by sampling a large number of points, calculating summary statistics from the sample, and comparing those statistics to either their true value counterparts (i.e., the values used as input in the simulation), or to their analytical expectations. In Box 1, we illustrate this procedure for a known parametric distribution, the multivariate normal distribution characterizing the phylogenetic Brownian motion model (Felsenstein, 1973).

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:

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

where $\mathbf{y}$ corresponds to the observed trait values representing n species, $\boldsymbol{y_0}$ is the trait value at the root of the tree, r is the variance of the process (also known as 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 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 $D = \{\mathbf{y}\}$ (but note that sometimes researchers treat Φ and consequently $\boldsymbol{T}$ as data).

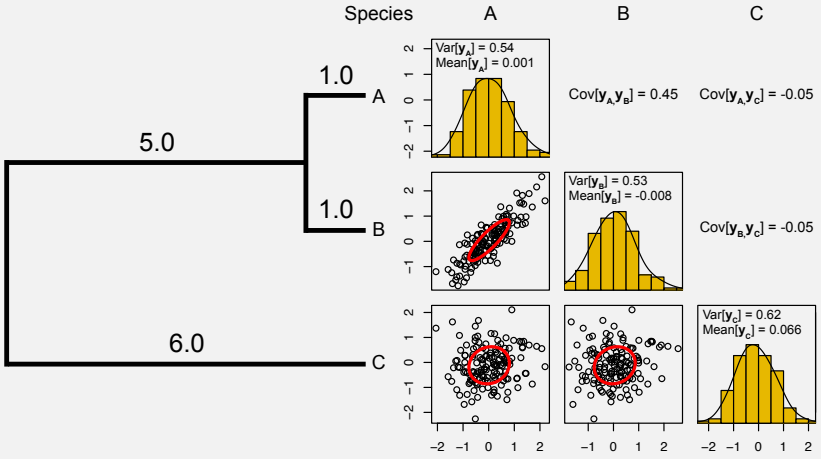

Figure 1: A sample of 200 draws from a MVN distribution, each representing the evolutionary trajectory of one continuous trait along the species tree on the left. The root trait value, $\boldsymbol{y_0}$, and the evolutionary rate of the process, r, were set to 0.0 and 0.1, respectively. The panel on the right shows histograms of trait values sampled from the MVN for each species, as well as their covariation.

Validating a phylogenetic BM simulator

Because the MVN is a **known parametric distribution**, it is trivial to verify the correctness of a phylogenetic BM model simulator, $S[f(\mathbf{y}|\boldsymbol{y_0}, r, \boldsymbol{T})]$. Figure 1 shows the result of 200 simulations under the MVN with $\boldsymbol{y_0} = [0.0, 0.0, 0.0]$, $r = 0.1$, along the tree shown on the left, which determines the variance-covariance matrix:

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

One can then verify that the simulated mean trait values and their variances in each species fall within their 95% confidence intervals 95% of the time. More specifically (i) $\boldsymbol{y}_\mathrm{A}$, $\boldsymbol{y}_\mathrm{B}$ and $\boldsymbol{y}_\mathrm{C}$ fall within [] X, Y and Z times out of 200, and (ii) Var$[\boldsymbol{y}_\mathrm{A}]$,

(a)

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

(b)

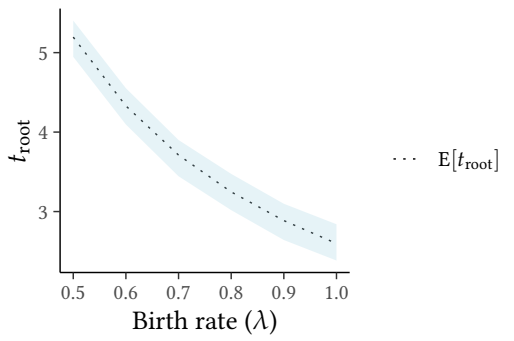

Figure 2: Validation of Yule tree simulator. (a) Number of simulated data sets (out of 100) for which the expected tree height (t_{root}) was inside the 95% CI about its sample average. Each data set consisted of 50 twenty-taxon simulated Yule trees. (b) The area shaded in light blue represents the 95% 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).

$\mathrm{Var}[\boldsymbol{y}_{\mathrm{B}}]$ and $\mathrm{Var}[\boldsymbol{y}_{\mathrm{C}}]$ fall within [] X, Y and Z times out of 200.

Another 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, $\mathrm{E}[t_{\mathrm{root}}]$, for a tree with n tips is:

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

One can then verify if $\mathrm{E}[t_{\mathrm{root}}]$ is 95% of the time within ± 1.96 standard errors of the average Yule-simulated tree height. Confirming that this is the case indicates $S[f(\Phi|\lambda)]$ is correctly implemented (Fig. 2).

We note that $S[\mathcal{M}]$ represents a *direct* simulator under model $\mathcal{M}$ (Table 1), meaning each and every sample generated by $S[\mathcal{M}]$ 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 $S[\mathcal{M}]$ has not been yet implemented, and it is predicated upon the existence of correct implementations of both an inferential engine $I[\mathcal{M}]'$ and of proposal functions. We distinguish $I[\mathcal{M}]'$ from $I[\mathcal{M}]$ because simulations are being carried out precisely to validate $I[\mathcal{M}]$. Unless MCMC simulations are done with $I[\mathcal{M}]'$ – an independent 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, $I(M)$

In order to achieve correct inferences, it is of utmost importance to guarantee that inferential machinery returns quantifiably correct answers. In practice, however, this is impossible to achieve due to the models under consideration being almost always badly misspecified for the data at hand. Simulating data from a probabilistic generative model can thus be helpful to understand when the inferential machinery under development can at the very least recover generating parameters in a robust fashion. In this section we discuss a few techniques that can be employed to assess the correctness of a parameter-estimation routine when one can simulate from a probabilistic data-generating process.

The discussion will mostly follow the ideas in Cook et al. (2006) and Talts et al. (2018). The basic idea is to draw $\theta^{(i)}, i = 1, \ldots, M$ from its prior, $\pi(\theta)$, simulate some data $y^{(i)}$ from $f(y^{(i)} \mid \theta^{(i)})$, use the computational machinery under evaluation to compute $p(\theta \mid y^{(i)})$ and then produce a sample $\boldsymbol{\theta}_s^{(i)} = \{\theta_{s1}^{(i)}, \ldots, \theta_{sL}^{(i)}\}$ of size L.

We begin our investigation with the coverage properties of uncertainty intervals; if the inferential engine is correct, we will be able to obtain interval estimates that have precise coverage properties. More concretely, let us first define the highest posterior density (HPD) interval. For a credibility level $\alpha \in (0, 1)$ define $I_\alpha(y) := (a(y, \alpha), b(y, \alpha))$ such that

$$\frac{1}{m(y)} \int_{a(y,\alpha)}^{b(y,\alpha)} f(y \mid t)\pi(t)\, dt = \alpha,$$

where $m(y) = \int_\Theta f(y \mid t)\pi(t)\, dt$. We say $\mathrm{Cred}(I_\alpha(y)) = \alpha$. Taking

$$\inf_{b(y,\alpha)-a(y,\alpha)} \{I_\alpha(y) : \mathrm{Cred}(I_\alpha(y)) = \alpha\},$$

yields the shortest interval with the required credibility. We are now prepared to discuss how to validate a (Bayesian)

k	$P(x = k)$
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 (the number k of true simulated values that fall within their corresponding 95%-HPDs) is binomially distributed, with $p = 0.95$.

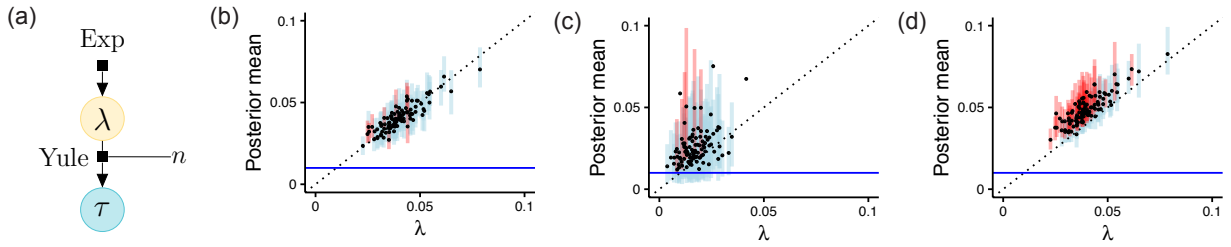

Figure 3: Calibrated validation analyses of a simple Bayesian hierarchical model. (a) The graphical representation of the model used in the analyses, where λ is the Yule birth-rate, n is the number of species, and τ is the set of speciation times. The prior is an exponential density with arbitrary mean of 0.01. Panels (b-d) show true λ values plotted against their mean posteriors (the dashed line gives $f(x) = x$). (b) Trees with $50 < n < 200$, (c) Trees with $5 < n < 10$, (d) Model is misspecified (log-normal prior is used in inference, rather than exponential).

inference engine. Consider the following generative model:

$$\theta_0 \sim \Pi,$$
$$\tilde{y} \sim F(\cdot \mid \theta_0).$$

It can be shown that in this setup we have $\Pr\left(\theta_0 \in I_\alpha(Y)\right) = \alpha$, i.e. that $100 \times \alpha\%$ HPDs have nominal coverage under the true generative model. A proof is provided in Appendix A.

We can thus validate the inferential engine by ascertaining that the coverage probabilities are within their expected bounds under the null hypothesis that the engine is working as designed. In particular, the coverage of M intervals obtained as above will be distributed as binomial random variable with M trials and success probability α. When $M = 100$ and $\alpha = 0.95$, the confidence interval for the number of simulations containing the correct data-generating parameter is between 90 and 99.

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 intervals. In particular, they show (Theorem 1 therein) that if the algorithm

works as intended the distribution of the rank $r^{(i)}$ of $\theta^{(i)}$ in $\boldsymbol{\theta}_s^{(i)}$ will follow a uniform distribution on $[1, L+1]$. Adherence to this distribution can be investigated through histograms (Talts et al., 2018) and the empirical cumulative distribution function (ECDF) and its confidence bands (Säilynoja et al., 2021). In phylogenetics, however, we are usually interested in the phylogeny, for which there is no canonical total-ordering structure. To get around this difficulty, we propose computing phylogenetic distances, for which total-ordering holds and thus for which ranks make sense. The general algorithm for simulation-based calibration for phylogenetics can be described briefly as

0. Generate a reference tree from the prior $\bar{\tau}_0 \sim \pi_T(\tau|\boldsymbol{\gamma})$;

   for each iteration in 1:N, **do**:

1. Generate $\bar{\tau} \sim \pi_T(\tau|\boldsymbol{\gamma})$;
2. Compute the distance $\bar{\delta} = d_\sigma(\bar{\tau}, \bar{\tau}_0)$ according to the metric of choice;
3. Generate some (aligment) data $\tilde{y} \sim f(y|\bar{\tau}, \boldsymbol{\alpha})$;
4. Draw (approximately) $\boldsymbol{\tau}_s = \{\tau_s^{(1)}, \tau_s^{(2)}, \ldots, \tau_s^{(L)}\}$ from the posterior $\pi(\tau|\tilde{y})$;
5. Compute distances $\boldsymbol{\delta}_s = \{\delta_1, \delta_2, \ldots, \delta_L\}$ with $\delta_i = d_\sigma(\tau_s^{(i)}, \bar{\tau}_0)$;
6. Compute the rank $r(\boldsymbol{\delta}_s, \bar{\delta}) = \sum_{i=1}^{L} \mathbb{I}(\delta_i < \bar{\delta})$.

Functionals There are a plethora of functionals one might want to compute in order to use SBC to study correctness and the key to effective SBC is choosing functionals that reflect relevant estimators of the quantities of interest. As with any other high-dimensional model, we explore different functionals of τ in order to study different aspects of the estimation procedure. Unlike many models, however, phylogenies are very hard to summarise using univariate measures. We exploit the metric nature of the space of phylogenies and compute distances with respect to a reference phylogeny τ_0 under different tree distances. In this study I have used the following functionals:

- The largest branch length in τ, $M(\tau)$;
- The length of the phylogeny, i.e., the sum of branch lengths $S(\tau)$;
- The length of the external branch leading to taxon s_1, $T_1(\tau)$;
- The Robinson-Foulds (Robinson and Foulds, 1981) distance between τ and τ_0, $\mathrm{RF}_0(\tau)$;
- The Kendall-Coljin (Kendall and Colijn, 2016) distance between τ and τ_0, $\mathrm{KC}_0(\tau; \lambda)$;
- The Billera-Holmes-Vogtman (BHV) (Billera et al., 2001) distance between τ and τ_0, $\mathrm{BHV}_0(\tau)$;

An example In order to illustrate how a typical phylogenetic SBC analysis might work, we discuss a simple example where [REMCO will describe this in detail].

Figure 4 shows the results for the functionals discussed in the previous section, i.e., tree length, maximum branch length and the length of the first external branch as purely continuous parameters and the Billera-Holmes-Vogtman (BHV), Robinson-Foulds (RF) and Kendall-Coljin phylogenetic distances. As can be seen, the sampler yielded samples that do not point to any anomaly; the ECDFs lie well inside their confidence ellipses and the observed ranks also lie inside their confidence bands.

The plots in Figure 5 also show that the HPDs do cover the generating parameters with probability compatible with what is expected theoretically. In this instance, the tests would fail to detect problems with the algorithm. These plots can be supplemented with a table showing attained coverage and binomial confidence intervals for this quantity (see Table 2. A further advantage of graphically investigating coverage is that one can identify consistent areas for which estimation fails – e.g. high/low simulated values of a given variable.

Validating the MCMC transition mechanism (operators)

At the heart of MCMC as an approach to sample a target distribution (conventionally, $\mathrm{P}(\theta|D)$) is the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970). This algorithm uses a transition mechanism (i.e., a set of "operators") characterized by a proposal density $q(\theta'|\theta)$ that perturbs parameter values θ toward new values θ'. The Markov chain progresses as these perturbations are accepted, which occurs with probability α:

$$\alpha = \min\left(1, \frac{\mathrm{P}(\theta'|D)}{\mathrm{P}(\theta|D)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right), \tag{4}$$

where $\frac{q(\theta|\theta')}{q(\theta'|\theta)}$ is also referred to as the Hastings ratio (Smith and Roberts, 1993; Tierney, 1994; Gelman et al., 2013). For simple proposal densities, the Hastings ratio will often be unity and only the prior and likelihood ratios must be computed. It is sometimes the case, however, that more complex proposals will increase or reduce the dimensionality of parameter space, in which case the derivation of the Hastings ratio will be less straightforward. For the sake of brevity, we point the reader to the relevant theory and examples in (Green, 1995; Huelsenbeck et al., 2004; Drummond and Suchard, 2010).

Although operators are not strictly part of the model (as mentioned above, MCMC is not the only way to sample or approximate a target distribution), it is absolutely vital to validate them prior or together with the model per se, should MCMC be the approach of choice. Only correctly implemented operators will lead to ergodic (i.e., irreducible, positive recurrent, and aperiodic) Markov chains with a stationary distribution that will hopefully match the target distribution, should the chain be long enough.

In the context of MCMC, unless a direct simulator $S(M)$ is available to be used as a proposal mechanism (see above), model and operator validation can be seen as two sides of the same coin. Outside of unit-testing, validating models requires carrying out MCMC (see items (ii) and (iv) in the previous section, for example), which in turn can

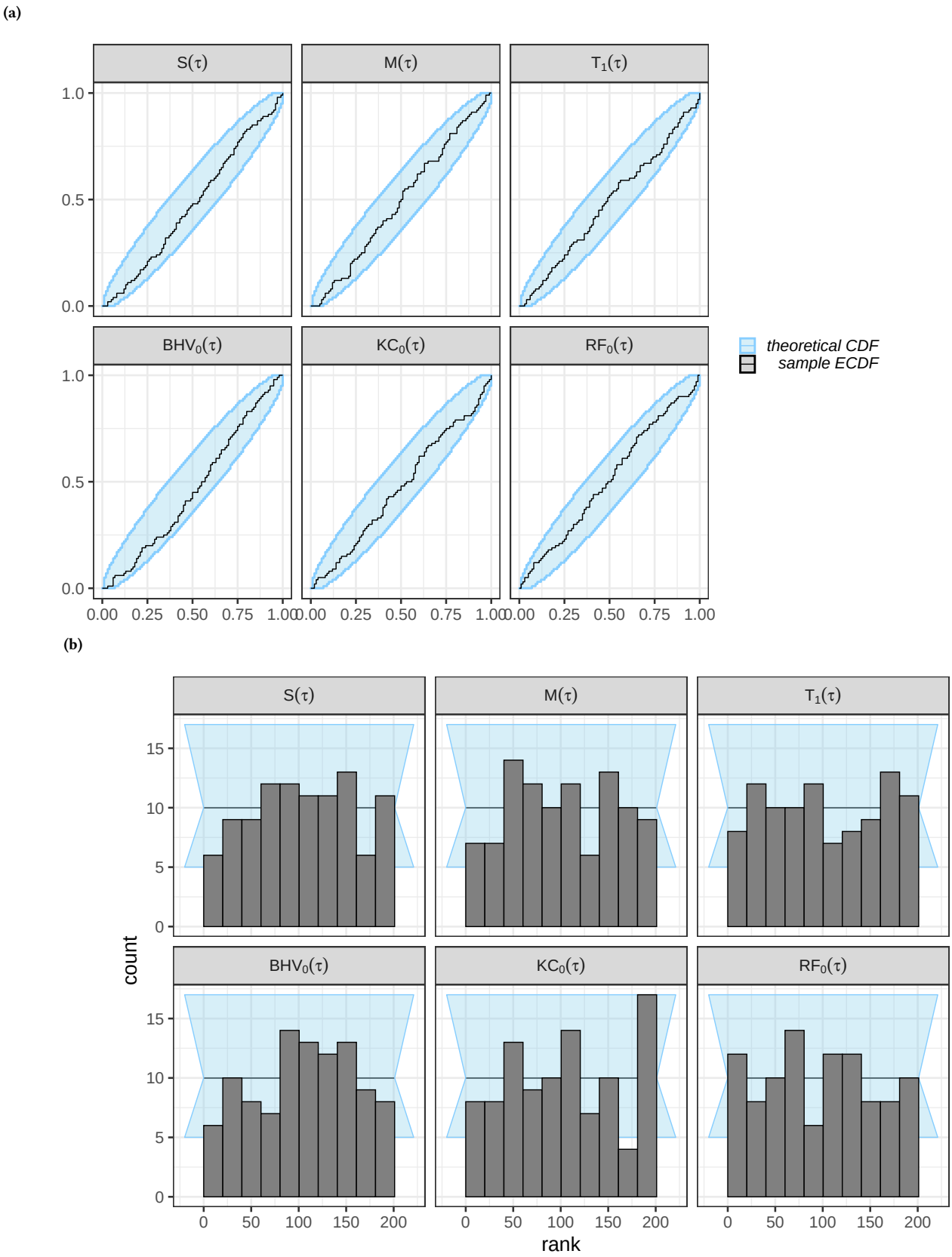

Figure 4: Simulation-based calibration of a phylogenetic model. We show the empirical cumulative distribution function (ECDF) for each functional described in Section in panel (a), whereas the distribution of the attained ranks is shown in panel (b).

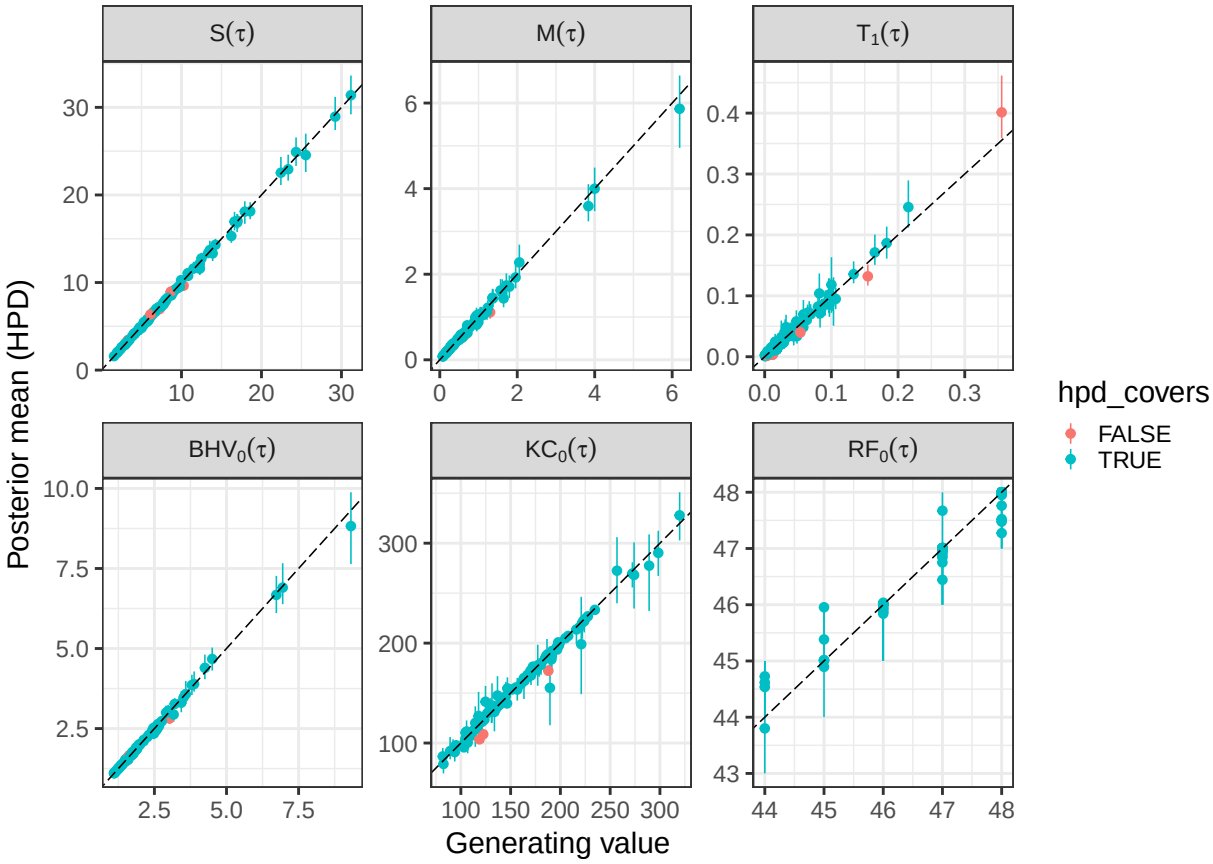

Figure 5: Coverage in simulation-based calibration. We show posterior mean (points) and 95% HPDs (bars) for all six functionals of interest. Red dots and lines show intervals that failed to include the generating value, whilst blue ones show those that did conver the "true" value.

only be a meaningful procedure if correctly implemented operators are available. Conversely, if the intention is to validate new operators, then a model under which to evaluate proposals must have been correctly implemented prior to MCMC.

In the latter case, the principle behind validation is the same: one compares the stationary distribution (with respect to one or more summary statistics) obtained using the operator(s) being tested, $P_I(\theta)$, with an expected distribution, $P_S(\theta)$ (see listing 5 in supplementary material). This second, expected distribution can be obtained through MCMC with a mutually exclusive set of correctly implemented operators capable of generating an ergodic chain (see item (a) in supplementary table S1).

Finally, a well-calibrated validation study also serves the purpose of validating the transition mechanism. Low coverage might indicate not only that the model was incorrectly implemented, but also that operators are not functioning as expected. Determining which of the two possibilities – or whether both are happening – is not a trivial task, and careful diagnostics are necessary on the part of method developers.

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 discuss 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.

Billera, L. J., Holmes, S. P., and Vogtmann, K. (2001). Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767.

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 Computational Biology*, 15:e1006650.

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

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. *Journal of Computational and Graphical Statistics*, 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.

Drummond, A. J. and Suchard, M. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biol.*, 8:114.

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.

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.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–32.

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.

Huelsenbeck, J. P., Larget, B., and Alfaro, M. E. (2004). Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.*, 21(6):1123–33.

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.

Kendall, M. and Colijn, C. (2016). Mapping phylogenetic trees to reveal distinct patterns of evolution. *Molecular biology and evolution*, 33(10):2735–2743.

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.

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

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

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.

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.

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. *Nature Genetics*, 48(9):1077–1082.

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.

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. *Mathematical biosciences*, 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*.

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.

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

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*.

A Proofs

For a number M of simulations, simulate

$$\theta^{(i)} \sim \pi(\cdot)$$
$$y^{(i)} \mid \theta^{(i)} \sim f(\cdot \mid \theta^{(i)})$$

Now for brevity, define $a^{(i)} := a(y^{(i)}, \alpha)$ and $b^{(i)} := b(y^{(i)}, \alpha)$ and recall $I_\alpha\left(y^{(i)}\right)$ is such that

$$Q\left(b^{(i)} \mid y^{(i)}\right) - Q\left(a^{(i)} \mid y^{(i)}\right) = p_1 - p_2 = \alpha,$$

where $Q_y(x)$ is the posterior CDF and $p_1, p_2 \in (0,1)$, $p_1 < p_2$. A natural quantity to compute is

$$S_M = M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(\theta^{(i)} \in I_\alpha\left(y^{(i)}\right)\right),$$

i.e. the attained coverage of the Bayesian intervals. Let $F_U(x) = x$ be the CDF of a $\text{Uniform}(0,1)$ random variable. We may re-write the limit as

$$\begin{aligned}
\lim_{M\to\infty} S_M &= \lim_{M\to\infty} M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(\theta^{(i)} \in I_\alpha\left(y^{(i)}\right)\right), \\
&= \lim_{M\to\infty} M^{-1} \sum_{i=1}^{M} \left\{\mathbb{I}\left(\theta^{(i)} \le b^{(i)}\right) - \mathbb{I}\left(\theta^{(i)} \le a^{(i)}\right)\right\}, \\
&= \lim_{M\to\infty} M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(\theta^{(i)} \le b^{(i)}\right) - M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(\theta^{(i)} \le a^{(i)}\right), \\
&= \lim_{M\to\infty} M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(Q^{-1}_{y^{(i)}}\left(\theta^{(i)}\right) \le p_1\right) - \lim_{M\to\infty} M^{-1} \sum_{i=1}^{M} \mathbb{I}\left(Q^{-1}_{y^{(i)}}\left(\theta^{(i)}\right) \le p_2\right), \\
&= F_U(p_1) - F_U(p_2) = \alpha,
\end{aligned}$$

where the last line follows from the fact that the CDF of $\theta^{(i)}$ is uniformly distributed on $(0,1)$ (Theorem 1 in Cook et al. (2006)) and almost sure convergence of the ECDF to the true CDF (Glivenko-Cantelli theorem).