Bayesian phylogenetic methods

Ziheng Yang (orcid: 0000-0003-3351-7981)[1]

[1]Department of Genetics, Evolution, and Environment, University College London, Gower Street, London WC1E 6BT, UK (Phone: +44 20 7679 4379; email: z.yang@ucl.ac.uk) .

Bayesian inference was introduced into phylogenetics in the 1990s when Markov chain Monte Carlo (MCMC) was revolutionizing Bayesian statistics. It is now the most widely used methodology for implementing advanced models of data analysis in the field. Its ability to average over variations and uncertainties (via MCMC) makes it particularly suitable for implementing models that deal with heterogeneous evolutionary processes for both molecular and morphological data. It provides a natural framework for combining information from different sources, a prime example being Bayesian relaxed clock dating, which integrates information in molecules and fossils to date the tree of life. Molecular phylogenetics has also emerged as a rich training ground for evaluating new Bayesian computational methods. Nowadays the Bayesian method has been applied to address virtually all major questions in evolutionary biology, such as inferring phylogenetic relationships and divergence times among species, detecting molecular adaptation, estimating species trees despite conflicting gene trees, inferring viral pandemic dynamics, inferring gene flow between species, delimiting species boundaries using genomic data, and reconstructing genes and genomes in extinct ancestral species. This article provides a concise introduction to Bayesian phylogenetics, including Bayesian computation through MCMC.

Keywords: Bayes theorem, Bayesian inference, Likelihood, Markov chain Monte Carlo, Phylogenetics, Prior, Posterior

Introduction

Overview of Bayesian statistics

There are two principal philosophies in statistical data analysis: the classical (Frequentist) and the Bayesian, and they are based on different concepts of probability. The Frequentist defines the probability of an event as the expected frequency of occurrence of that event in repeated random draws from a real or imaginary population. When we say that the probability of heads for a coin toss is $\frac{1}{2}$, we mean that the frequency of heads will approach $\frac{1}{2}$ when the number of tosses increases. The performance of an inference procedure is judged by its properties in repeated sampling from the data-generating model (i.e. the likelihood model), with the parameters fixed. For example, a parameter estimate (such as maximum likelihood estimate or MLE) is judged by its bias and variance, while a hypothesis test (such as the likelihood ratio test or LRT) is judged by its false-positive rate and power (or type-I and type-II errors).

In Bayesian statistics, probability measures one's degree of belief. When we say that a hypothesis (e.g., that the extinction of the dinosaurs was caused by a meteorite hitting the Earth) has probability 0.9, we mean that the hypothesis is very likely to be true, judged by currently available evidence or by one's subjective opinion. A key feature of Bayesian statistics is the use of statistical distributions to describe uncertainties in all unknowns (such as the unknown parameters in a model or the competing hypotheses for explaining the same data). In classical statistics, parameters are fixed (although unknown) constants and cannot be assigned distributions.

Suppose one wants to analyze the data (x) to estimate the unknown parameter θ under a model. The probability of the data given the parameter θ, $p(x|\theta)$, is the *likelihood function*, and is known to contain all information about the parameters θ in the data x given the model. In a Bayesian analysis, one assigns a distribution on θ before the analysis of the data. This is called the *prior distribution* and reflects one's knowledge or belief about the likely values of θ. Bayesian analysis of the data then produces the distribution of θ given the data, $f(\theta|x)$, called the *posterior distribution*. The two are related through the Bayes theorem

$$p(\theta|x) = \frac{p(\theta)p(x|\theta)}{p(x)} \propto p(\theta)p(x|\theta). \qquad (1)$$

Here the marginal probability of the data, $p(x) = \int p(\theta)p(x|\theta)\,\mathrm{d}\theta$, is a normalizing constant, and its role is to ensure that $p(\theta|x)$ is a proper statistical distribution and integrates to 1. Eq. 1 thus says that the posterior is proportional to the prior times the likelihood, or equivalently, the posterior combines information in the prior and in the data sample. Bayesian inference is then based on the posterior distribution of parameters and models. For example, for a continuous parameter, the posterior mean provides a point estimate, and the 95% credibility interval (CI) provides an interval estimate.

A major (Bayesian) criticism of classical statistics is that it does not answer relevant questions, and the

methodology comprises disconnected and sometimes contradictory set of ideas and techniques. Take the LRT as an example. A healthy test has the type-I error under control and rejects the null hypothesis H_0 in no more than 5% of datasets when the test is applied to many datasets generated under H_0. If we are not interested in the many imaginary datasets generated under H_0 and ask instead how probable H_0 is true given the particular dataset collected from our experiment, we get no answer. The probability for H_0 given the data, $\mathbb{P}(H_0|x)$, is not a meaningful concept in classical statistics.

In contrast, Bayesian inference answers such questions head-on. Inference is conditioned on the data observed, and disposes of the imaginary datasets under the fixed model and parameters. Given the data collected, the prior probability for H_0 is updated to become the posterior probability, $\mathbb{P}(H_0|x)$. The prior or the need to use a prior is thus the dividing line between Bayesian and classical statistics. To the Bayesian the prior is the premium one pays to get straight answers to relevant questions in statistical inference. For non-Bayesians, the cost of this premium is too much to bear.

There exist two flavours of Bayesians: the objective and the subjective. Objective Bayesians consider probability to be a representation of objective or rational degree of belief. This school runs into trouble when no prior information is available about the parameter and the prior is supposed to represent total ignorance. For a continuous parameter, Laplace's 'principle of insufficient reason' is used to argue for a uniform prior distribution over the range of the parameter, but this leads to contradictions. For example, $\theta \sim U(0, 1)$ may appear to be noninformative for a probability parameter θ, but it implies a nonuniform distribution for a one-to-one transform such as $\psi = \log \frac{\theta}{1-\theta}$. Surely if one has no information concerning θ one must have no information concerning ψ (since they are one-to-one and knowing one means knowing the other), but it is impossible for both to have uniform distributions. The prior is not invariant to a nonlinear transform. This contradiction is fatal, and it is now generally accepted that noninformative priors do not exist and that no prior can represent total ignorance.

To a subjective Bayesian, probability represents one's personal degree of belief. This appears to be the more popular version of Bayesian statistics. While this school is coherent, a criticism of it is that it is, uh, subjective. Biologists often find the suggestion surprising that a Bayesian analysis of their data calls for a psychoanalytical assessment of their personal beliefs, and that the conclusion drawn from their experiment may be influenced by subjective beliefs.

It should be mentioned that even though the philosophical interpretations differ in classical statistics and in objective and subjective Bayesian statistics, the laws of probability are the same.

Given the different philosophies on which classical and Bayesian statistics are based, an important question to biologists is whether the two approaches produce similar (numerical) answers. This depends on the nature of the problem. In the so-called *stable estimation problems*, a well-formulated model $f(x|\theta)$ is available, and we want to estimate parameters θ from a large dataset. The prior will have little effect, and both likelihood and Bayesian estimates will be close to the true parameter value. Furthermore, classical confidence intervals in general match posterior CIs under vague priors. In many other problems, both the prior and the likelihood may exert substantial influence on the posterior. The posterior may be sensitive to the prior, because the model is ill-formulated and parameter-rich and there are strong correlations among the parameters, or because the data lack information. In difficult problems, Bayesian inference can be very sensitive to the prior, and Baysian analysis and classic hypothesis testing can produce opposite conclusions for the same dataset. We discuss some examples later.

Overview of Bayesian phylogenetics

The Bayesian approach was introduced to molecular phylogenetics in the 1990s by three groups (Rannala and Yang, 1996; Yang and Rannala, 1997; Mau and Newton, 1997; Li *et al.*, 2000). This was a time when Bayesian MCMC algorithms were developed and implemented in various branches of sciences, revolutionizing Bayesian computation. The early studies in phylogenetics assumed simple models of sequence evolution and a constant rate of evolution (the molecular clock). Nowadays, several Bayesian phylogenetic programs exist that implement a wide range of complex models that account for various features of sequence data. General Bayesian programs for phylogenetic analysis include MrBayes (Ronquist *et al.*, 2012), RevBayes (Hohna *et al.*, 2016), BEAST (Drummond and Rambaut, 2007), and PhyloBayes (Lartillot *et al.*, 2009). Several Bayesian programs are available for estimating species divergence times incorporating information in fossil and molecular data, such as MCMCTREE (Yang, 2007) and BEAST (Drummond and Rambaut, 2007).

In molecular phylogenetics, the data x is an alignment (or alignments) of sequences of nucleotides, codons, or amino acids from several species. Here, we assume that the sequences are already aligned and ignore possible alignment errors. Our focus is the phylogenetic tree, which consists of the tree topology (τ) and the branch lengths (denoted $\boldsymbol{b}$). The branch length is measured by the expected number of substitutions per site, and quantifies the amount of evolution along the branch. Given the tree, the sequence data at the tips of the tree (for extant species) are the product of the process of sequence evolution along the branches. The process of nucleotide substitution is typically described by a continuous time Markov chain model such as the JC (Jukes and Cantor, 1969). This may include additional parameters (denoted $\boldsymbol{\phi}$), such as the relative rates between nucleotides and the equilibrium nucleotide frequencies. More complex models may include parameters that describe variable rates across sites in the sequence or the nonsynonymous/synonymous substitution rate ratio in

protein-coding genes (Yang, 1993; Goldman and Yang, 1994; Rodrigue *et al.*, 2010).

The vector of parameters is then $\theta = (\tau, \boldsymbol{b}, \boldsymbol{\phi})$. The posterior distribution of the tree topology, branch lengths, and substitution parameters is then given by eq. 1 as

$$p(\tau, \boldsymbol{b}, \boldsymbol{\phi}|x) \propto p(\tau, \boldsymbol{b})p(\boldsymbol{\phi})p(x|\tau, \boldsymbol{b}, \boldsymbol{\phi}), \qquad (2)$$

where $p(\tau, \boldsymbol{b})$ is the prior on tree topology and branch lengths, $p(\boldsymbol{\phi})$ is the prior on substitution parameters, while $p(x|\tau, \boldsymbol{b}, \boldsymbol{\phi})$ is the probability of the sequence data given the tree topology and branch lengths, also known as the phylogenetic likelihood, calculated using the pruning algorithm of Felsenstein (1981).

See Felsenstein (2004), Yang (2014), and Yang (2018) for historical accounts of Bayesian phylogenetics. Early work in the field mostly involved Bayesian re-implementation of likelihood models developed previously. Nowadays many sophisticated parameter-rich models may be available in the Bayesian framework only. *Bayesian marginalization*, which averages over uncertainties in latent variables, is an attractive way of accommodating variation in the data that we are not interested in but cannot ignore, and the ease with which one can deal with high-dimensional multi-parameter models is a great advantage of the Bayesian framework. Maximum likelihood under such models is often unfeasible computationally, as the likelihood function, which averages over such random variables and involves huge sums or high-dimensional integrals, may be too expensive to calculate.

Priors in phylogenetic models

Whether we adopt the objective or subjective views, Bayesian inference requires the specification of a prior. The prior is expected to summarize one's objective information (according to objective Bayesians) or personal beliefs (according to subjective Bayesians) about the likely values of model parameters. Thus we can construct the prior by making use of information gained in past experiments under similar conditions or from other independent evidence. The prior can also be specified by modeling the physical/biological process.

In Bayesian phylogenetics, the tree topologies (τ) represent discrete statistical models, the branch lengths ($\boldsymbol{b}$) are continuous parameters that are defined only on specific trees, while the substitution parameters ($\boldsymbol{\phi}$) are often defined for all possible trees. The parameter space of the inference problem is high-dimensional and also complex. Specification of the prior is thus a nontrivial task. Here we discuss prior specification in a simple phylogenetic analysis to reconstruct the phylogeny.

First we consider the prior on the tree topology. Most phylogenetic analyses are conducted without assuming the molecular clock and use unrooted trees. It is common to assign a uniform prior on all possible trees: each tree is assigned the probability $1/T_n$, where T_n is the number of unrooted trees for n species.

If the species are closely related, the evolutionary rate may be roughly constant among species. One can then use the molecular clock (the assumption of rate constancy) to infer rooted trees. Rooted trees are also used to infer species divergence times in the so-called molecular clock or relaxed-clock dating analysis (Thorne *et al.*, 1998; Drummond *et al.*, 2006; dos Reis *et al.*, 2016). A prior distribution over the rooted trees and node ages (branching times) can be generated using the birth-death process (Rannala and Yang, 1996; Stadler, 2010). The birth rate (speciation rate) and death rate (extinction rate), and the sampling intensity can be fixed or assigned further priors (called hyper-priors).

The birth-death process has been generalized to allow the speciation and extinction rates to vary over time. Such generalized birth-death process models have been used to analyze data of estimated divergence times on a species phylogeny (the so-called lineage-through-time or LTT data) to infer macroevolutionary processes, including changes in the speciation rate and extinction rate. Evolutionary biologists have found such inference irresistibly interesting, as the inferred rate changes may be correlated with geological, paleo-climatic and extraterrestrial events that may have impacted biodiversity on the planet. However, the birth-death process (in particular with varying speciation and extinction rates) is highly stochastic, able to generate wild fluctuations in the LTT data, and it is debatable whether typical datasets contain useful information for such inference (Louca and Pennell, 2020; Legried and Terhorst, 2022; Rannala and Yang, 2025). See Rosenberg *et al.* (2025) and related articles for recent discussions of the topic.

Next, we consider the prior for branch lengths. A binary unrooted tree for n species has $2n - 3$ branches. Given each unrooted tree topology, the $2n - 3$ branch lengths can be assigned independent and identical distributions (i.i.d.) such as the uniform or exponential. However, when applied to many variables, i.i.d. priors may imply a very strong prior statement on the mean or sum of those variables. Indeed i.i.d. priors on branch lengths are found to be problematic, as they may be informative and unreasonable about the tree length; see Rannala *et al.* (2012) for alternative priors to deal with the issue.

The following recommendations may be made concerning priors in Bayesian phylogenetic analysis.

- There is no such a thing as an uninformative prior or a prior that represents total ignorance; all priors carry information.
- Even if the user of a Bayesian program has not explicitly specified a prior, a prior has been used in the analysis. The user need consider the possibility that the 'default' priors used by the software may not be appropriate for the data and problem.
- It is important to assess the impact of the prior. The prior may have considerable influence on the posterior distribution of key parameters int he model, in particular, if the model is complex involving many parameters that are strongly correlated and the data

are not very informative. Bayesian hypothesis testing may be sensitive to the priors on parameters in the models, as discussed below.

Markov Chain Monte Carlo

A Metropolis algorithm

Note that the normalizing constant $p(x)$ in eq. 1 involves an integral. When there are many parameters in the model, this integral will be multidimensional and may be very hard to compute. Modern Bayesian inference is often achieved through a computational algorithm called MCMC. This is an iterative simulation algorithm that generates a sample from the posterior distribution $p(\theta|x)$.

Here we illustrate the main features of the MCMC algorithm by applying it to the simple phylogenetic problem of estimating the distance θ between two sequences under the JC model (Jukes and Cantor, 1969). The data consist of the human and orangutan mitochondrial 12S rRNA genes, with $x = 90$ differences at $n = 948$ sites. Parameter θ is the expected number of nucleotide substitutions per site between the two sequences. Given θ, the probability of observing the data or the likelihood is given by the binomial probability

$$\begin{aligned} p(x|\theta) &= p^x(1-p)^{n-x} \\ &= \left(\tfrac{3}{4} - \tfrac{3}{4}\,\mathrm{e}^{-4\theta/3}\right)^x \cdot \left(\tfrac{1}{4} + \tfrac{3}{4}\,\mathrm{e}^{-4\theta/3}\right)^{n-x}, \end{aligned} \tag{3}$$

where $p = \frac{3}{4} - \frac{3}{4}\,\mathrm{e}^{-4\theta/3}$ is the probability that a site is occupied by two different nucleotides in the two sequences separated by a distance θ.

The probability of the data viewed as a function of the parameter, $L(\theta) = p(x|\theta)$, is the likelihood function. By maximizing the likelihood function or its logarithm, $\ell(\theta) = \log L(\theta)$, one gets the MLE as

$$\hat{\theta} = -\tfrac{3}{4}\log\left(1 - \tfrac{4}{3} \times \tfrac{x}{n}\right). \tag{4}$$

This is the well-known JC distance formula, and in our example, gives $\hat{\theta} = 0.1015$, close to the observed proportion of different sites ($x/n = 0.095$).

To estimate θ using a Bayesian approach, we have to assign a prior on θ. Here we assign an exponential prior with mean $\mu = 0.1$, with the prior density

$$p(\theta) = \tfrac{1}{\mu}\,\mathrm{e}^{-\frac{1}{\mu}\theta}, \quad \theta > 0. \tag{5}$$

The posterior is then given by eq. 1 as

$$\begin{aligned} p(\theta|x) &= \tfrac{1}{p(x)}p(\theta)p(x|\theta) \\ &= \tfrac{1}{p(x)}\tfrac{1}{\mu}\,\mathrm{e}^{-\frac{1}{\mu}\theta}\left(\tfrac{3}{4} - \tfrac{3}{4}\,\mathrm{e}^{-4\theta/3}\right)^x \cdot \left(\tfrac{1}{4} + \tfrac{3}{4}\,\mathrm{e}^{-4\theta/3}\right)^{n-x}, \end{aligned} \tag{6}$$

where

$$p(x) = \int_0^\infty p(\theta)p(x|\theta)\,\mathrm{d}\theta \tag{7}$$

is the normalizing constant. Note that in a Bayesian analysis, the data are known and treated as constants, while our focus is on the unknown parameter θ. The normalizing constant $p(x)$ is an integral which is nontrivial to calculate even in one dimension.

The following algorithm uses a sliding window (of width w) to propose new parameter values, and generates a sample from the posterior distribution, bypassing computation of the normalizing constant, $p(x)$.

1. Initialize: $n = 948$, $x = 90$, $w = 0.25$. Set initial state: $\theta = 0.1$, say.
2. Loop through the following steps
   (a) (Propose a new value θ'.) Generate $u \sim U(0,1)$ and set $\theta' = \theta + (\frac{1}{2} - u)w$. Note that θ is a uniform random variable over the interval $\left(\theta - \frac{1}{2}w, \theta + \frac{1}{2}w\right)$.
   (b) (Accept or reject the proposed value.) Compute the posterior density ratio

$$\alpha = \frac{p(\theta'|x)}{p(\theta|x)} = \frac{\frac{p(\theta')p(x|\theta')}{p(x)}}{\frac{p(\theta)p(x|\theta)}{p(x)}} = \frac{p(\theta')p(x|\theta')}{p(\theta)p(x|\theta)}. \tag{8}$$

   If $\alpha \geq 1$, accept θ. Otherwise accept θ with probability α. This can be achieved by drawing another random number $v \sim U(0,1)$, and accepting θ' if and only if $v < \alpha$. If θ' is accepted set $\theta = \theta'$. Otherwise keep the current θ.
   (c) Print out θ.

The simple algorithm above is an instance of the Metropolis algorithm (Metropolis *et al.*, 1953). Here we note its major features.

- The algorithm simulates a Markov chain; the next θ value the algorithm will visit depends on the current θ only, but not the θ values visited in the past. This explains why the algorithm is also known as Markov chain Monte Carlo (MCMC). Monte Carlo method, also known as computer simulation, uses repeated random sampling to generate numerical results, such as the value of an integral. MCMC is a simulation algorithm that generates a Markov chain.
- The algorithm is a hill-climbing algorithm although it may go downhill. It visits θ values of high posterior density more often than those of low density. In fact, the probability that the visited θ value is in the small interval $(\theta, \theta + \Delta\theta)$ is $p(x|\theta)\Delta\theta$. Recall that if a random variable Y has the probability density function (PDF) $p(y)$, then $\mathbb{P}(y < Y < y + \Delta y) = p(y)\Delta y$. In other words, the θ values generated by the algorithm are a sample from the posterior distribution $p(x|\theta)$.
- The sliding-window proposal used here is symmetrical: the probability density of proposing θ' from θ is the same as that of proposing θ from θ'. This is known as the Metropolis algorithm. The symmetry in proposal density is relaxed in Hastings's algorithm, to be described below.
- The algorithm does not require computation of the normalizing constant $p(x)$. Note that $p(x)$ cancels out in the calculation of the posterior ratio α in eq. 8. The

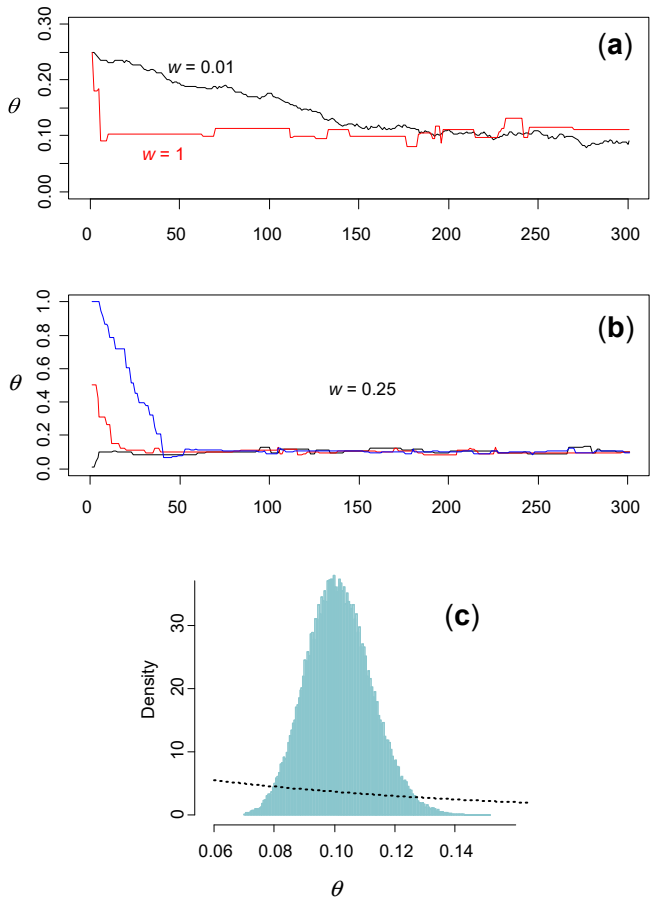

Fig. 1: MCMC for estimating sequence distance under the JC model (θ). The data consists of $x = 90$ differences between two sequences of $n = 948$ sites. (**a**) Two chains, both starting from $\theta = 0.25$, with the window size either too small ($w = 0.01$) or too large ($w = 1$). The chain with $w = 0.01$ has an acceptance rate of $P_{\text{jump}} = 91\%$, so that almost every proposal is accepted. However, this chain takes baby steps and mixes slowly. The chain with $w = 1$ has $P_{\text{jump}} = 6.9\%$, so that most proposals are rejected. The chain often stays at the same state for many iterations without a move. (**b**) Three chains started from $\theta = 0.01, 0.5$, and 1, with window size $w = 0.25$ and $P_{\text{jump}} = 13.7\%$. The optimal chain should achieve $P_{\text{jump}} \approx 40\%$ (Yang and Rodríguez, 2013), at $w = 0.1$. After about 50 iterations, the three chains become indistinguishable and have reached stationarity, so that a burn-in of 50 iterations appears sufficient for those chains. (**c**) Histogram constructed from a sample taken over 10^6 iterations. The prior (dashed line) is shown as well for comparison.

algorithm thus requires calculation of the ratio of the posteriors (at θ and θ'), but not the posterior itself. This explains how the MCMC algorithm can avoid the calculation of high-dimensional integrals.

Figure 1(a) shows two runs of the algorithm using different window sizes and different starting positions. With the small window size, $w = 0.01$, the proposed values are very close to the current value, and most proposals are accepted. However, the chain baby-walks and is very slow in exploring the parameter space, with the sampled values being high correlated. On the other hand, the window size $w = 1$ is too big, so that most proposals are from unreasonable regions of the parameter space and are rejected. The chain then stays in the same state for a long time before it jumps, causing high correlation between sample values as well. In either case the chain mixes slowly. The optimal window size is intermediate, achieved when $\sim 44\%$ of proposals are accepted (Yang and Rodríguez, 2013, table 1).

Samples taken before the chain has reached stationarity are usually discarded as *burn-in*. In the runs of figure 1b, the first 50 or 100 iterations may be so discarded. Figure 1c shows a histogram (an approximation to the posterior probability density) generated from a long chain of 10^6 iterations. Further summaries of the posterior distribution can be easily constructed as well. The posterior mean is 0.102, close to the MLE. The posterior standard deviation is 0.011, and the equal-tail 95% CI is (0.081, 0.124). This is a stable estimation problem: the prior does not have much influence and the Bayesian estimate and MLE are very similar.

Metropolis-Hastings algorithm

In general, let x be the data and θ be the parameters in the model, and $p(\theta|x)$ be the posterior. We propose the new parameter value from a proposal density $q(\theta'|\theta)$, which may not be symmetrical. The algorithm, known as the Metropolis-Hastings or MH algorithm (Metropolis *et al.*, 1953; Hastings, 1970), works as follows.

1. Set initial state to $\theta = \theta_0$.
2. Loop through the following steps
   (a) (Propose a new value θ'.) Generate $\theta' \sim q(\theta'|\theta)$.
   (b) (Accept or reject the proposed value.) Accept the new value θ' with probability

$$\min(1, \alpha) = \min\Big(1, \frac{q(\theta|\theta')}{q(\theta'|\theta)} \times \frac{p(\theta'|x)}{p(\theta|x)}\Big). \quad (9)$$

   In other words, if $\alpha \geq 1$, accept θ. Otherwise accept θ with probability α. This can be achieved by drawing a random number $v \sim U(0, 1)$, and accepting θ' if and only if $v < \alpha$. If θ' is accepted set $\theta = \theta'$. Otherwise keep the current θ.
   (c) Print out θ.

Running the algorithm through many iterations produces a sequence of values, $(\theta_0, \theta_1, \theta_2, \cdots)$, which is a sample from the posterior distribution $p(\theta|x)$.

The algorithm is nearly identical to the Metropolis algorithm of the previous section. As in eq. 8, the normalizing constant $p(x)$ cancels out in the calculation of the posterior ratio in eq. 9. In other words, the acceptance ratio α in eq. 9 can be written as

$$\alpha = \frac{q(\theta|\theta')}{q(\theta'|\theta)} \times \frac{p(\theta')}{p(\theta)} \times \frac{p(x|\theta')}{p(x|\theta)}$$
$$= \text{proposal ratio} \times \text{prior ratio} \times \text{likelihood ratio}. \quad (10)$$

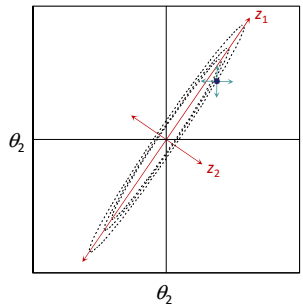

Fig. 2: A target (posterior) distribution $\pi(\theta_1, \theta_2)$ represented by the three contour lines with strong correlation between parameters θ_1 and θ_2. The MCMC algorithm will be very inefficient if it moves in the directions of θ_1 and θ_2 as the chain zigzags in very small steps. One can use the burn-in to estimate a bivariate normal approximation to the target and then move in the directions of the axes of the ellipses for contour lines: z_1 and z_2.

The major difference of the MH algorithm from the Metropolis algorithm of last section is the use of a proposal ratio, $\frac{q(\theta|\theta')}{q(\theta'|\theta)}$, also known as the Hastings ratio (Hastings, 1970). This is a correction factor applied because of the asymmetry of the proposal. If the proposal favours certain θ' values (with large $q(\theta'|\theta)$), such values will be penalized at the time of deciding whether or not to accept the proposal. As a result the chain converges to the correct target distribution despite the 'bias' in the proposal. This explains why the density for the new value θ' is in the denominator in the proposal ratio, $\frac{q(\theta|\theta')}{q(\theta'|\theta)}$, while it is in the numerator in the prior ratio, $\frac{p(\theta')}{p(\theta)}$, and likelihood ratio, $\frac{p(x|\theta')}{p(x|\theta)}$.

The proposal density $q(\theta'|\theta)$ has to satisfy certain requirements: it must specify an aperiodic recurrent Markov chain. In other words, it must be possible to move from one θ value to another and the implied Markov chain must not have a period. It is typically easy to construct such a chain and to verify that it satisfies those requirements. Note that the proposal kernel is separate from the prior and likelihood, so that the same proposal $q(\theta'|\theta)$ can be used in a variety of Bayesian inference problems.

When there are many parameters in the model, it is often unfeasible or computationally too complicated to update all parameters simultaneously. Changing many parameters simultaneously also causes poor mixing, because high-dimensional moves tend to have high rejection rates forcing one to use small steps. Instead, one can divide parameters into blocks, of possibly different dimensions, and then update the blocks one by one. This strategy often leads to computational efficiency. Also when one block is updated, one can treat the other blocks as fixed constants to design efficient proposals. Parameters that are strongly correlated may be grouped into the same block and updated simultaneously taking into account the correlation (fig. 2).

In phylogenetics, the parameter space may consist of several components: the tree topology τ, the branch lengths $\boldsymbol{b}$, and the substitution parameters $\boldsymbol{\phi}$. In each iteration, the different components may be updated in turn. For example, variants of tree search algorithms such as nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) (Swofford *et al.*, 1996) can be used to update the tree topology. The branch lengths and substitution parameters can be updated using sliding windows. The phylogenetic MCMC algorithm generates a sample from the joint posterior distribution of the tree topologies (τ), the branch lengths ($\boldsymbol{b}$), and the substitution parameters ($\boldsymbol{\phi}$).

For an extensive discussion of Markov chain Monte Carlo (MCMC) algorithms used in Bayesian phylogenetics, see Chapters 7 and 8 of Yang (2014). The edited book by Chen *et al.* (2014) summarizes recent developments in Bayesian selection of phylogenetic models.

Summaries of the posterior sample

The MCMC sample from the posterior distribution can be summarized in different ways.

For continuous parameters such as branch lengths (**b**) and substitution parameters ($\boldsymbol{\phi}$), the posterior means or medians are often used, together with the 95% posterior CIs. Two types of intervals are commonly used. The 95% equal-tail CI lies between the 2.5% and 97.5% quantiles of the posterior sample. The highest posterior density (HPD) CI includes values that make up 95% of the posterior probability and that have the highest posterior density. When the data are informative so that the posterior of the parameter is nearly symmetrical, the two intervals will be nearly identical. Otherwise they can be very different. The HPD interval is generally preferred over the equal-tail interval since it has the shortest length and includes only the most likely parameter values.

For the tree topology, a simple summary is the maximum *a posteriori* (MAP) tree, which is the tree topology with the highest posterior probability (or the tree most visited during the MCMC) (Rannala and Yang, 1996). This may be considered a point estimate of the true tree. However, when the data are not very informative, the MAP tree may have a very low posterior probability, and is a poor summary. One may also construct the 95% credible set of trees, which includes the highest-posterior trees with the total probability exceeding 95%. However, if this set contains many trees, it will not be very useful.

Even though the whole tree has weak support, some branches or splits may be well supported. Note that each internal branch on the tree defines a split (a bipartition) of the species. The posterior probability for a split is the sum of posterior probabilities for trees containing that split and can thus be estimated by the proportion of sampled trees that have that split. The majority-rule consensus tree includes splits that appear in at least half of the trees sampled, with splits receiving less than 50% posterior collapsed into a polytomy. The maximum clade probability (MCC) tree, a heuristic summary used in the program BEAST (Drummond and Rambaut, 2007), is a binary tree with the maximum sum (or product) of clade probabilities. Neither the sum nor the product of

clade probabilities makes sense, but the MCC tree is expected to be similar to the MAP tree when the data are informative.

Window size, proposal kernel and mixing efficiency

In the JC distance example (fig. 1a), we discussed that both too small and too large windows in the sliding-window proposal lead to poor mixing. Here we explain the theory for a more formal analysis of MCMC mixing efficiency.

Our focus is on the distribution of the unknown parameters θ while the data x are fixed, so we rewrite the posterior distribution as $\pi(\theta) \equiv p(\theta|x)$. We construct an MCMC algorithm to generate a sample $(\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(N)})$ from the target distribution $\pi(\theta)$ to estimate the expectation of a function of the parameters θ,

$$I = \mathbb{E}_\pi\{h(\theta)\} = \int h(\theta)\pi(\theta)\,\mathrm{d}\theta, \tag{11}$$

where the expectation or averaging is over the posterior distribution $\pi(\theta)$.

Using the MCMC sample, we calculate the sample mean as an estimate of I,

$$\hat{I} = \frac{1}{N}\sum_{i=1}^{N} h(\theta^{(i)}), \tag{12}$$

where $h(\theta^{(i)})$ is the value of the function at $\theta^{(i)}$. $\hat{I}$ is an unbiased estimate of I. Because the sampled values of θ and thus of $h(\theta)$ are correlated, the variance of the estimate may be larger than if the estimate is based on an independent sample. Thus we contrast those two variances to measure the efficiency of the MCMC algorithm, noting that a smaller variance means higher precision and a better algorithm. The efficiency (E) of the MCMC algorithm for estimating the posterior mean I is defined as the ratio of the variance of the estimate based on an independent sample to the variance of the estimate based on the MCMC sample of the same size, assuming that the sample size N is large (e.g., Peskun, 1973; Green and Han, 1992; Gelman *et al.*, 1996). This is given as

$$E = \frac{1}{1 + 2(\rho_1 + \rho_2 + \cdots)}, \tag{13}$$

where $\rho_k = \mathrm{corr}(h(\theta^{(i)}), h(\theta^{(i+k)}))$ is the lag-k autocorrelation. The independent sampler has efficiency 100%. If $E = \frac{1}{4}$, then an MCMC sample of $N = 1000$ is only as good as an independent sample of $NE = 250$ (in terms of the variance of the estimate). Here NE is known as the effective sample size (ESS).

Note that this approach of defining a function $h(\theta)$ and estimating its expectation I is general and covers almost all cases of interest. For example, suppose θ_1 is the first component in the parameter vector and define the function $h(\theta) = \theta_1$. Then the expectation I will be the posterior mean of θ_1. In a phylogenetic analysis let $\theta = (\tau, \boldsymbol{b})$, and suppose we are interested in whether the sequence data support a particular tree τ_1, say, the organismal tree. Define the function

$$h(\theta) = \begin{cases} 1, & \text{if } \tau = \tau_1, \\ 0, & \text{if } \tau \neq \tau_1. \end{cases} \tag{14}$$

Then the expectation I will be the posterior probability of tree τ_1, and the estimate $\hat{I}$ is simply the proportion of MCMC samples in which the tree is τ_1. The reader is invited to define a function $h(\theta)$ to estimate the posterior probability for a particular clade.

In the JC distance example discussed early, we observed that too small and too large windows lead to strong positive correlation between the sampled values. According to eq. 13, large ρs mean low efficiency. Instead of the uniform kernel, one can also use the Gaussian proposal kernel in the sliding window, $\theta'|\theta \sim N(\theta, \sigma^2)$, in which case the proposal standard deviation σ acts like a window size or step length. Numerical calculations suggest that when the Gaussian proposal is applied to the Gaussian target $N(0, 1)$, optimal efficiency (of $E = 0.23$) is achieved when $\sigma = 2.5$ with the acceptance rate of $P_{\text{jump}} = 0.44$ (Gelman *et al.*, 1996).

The acceptance rate is easy to monitor and can be used to adjust the step length, as the acceptance rate typically has a monotonic relationship with the step length (lower acceptance for larger steps). Indeed one can adjust the step lengths automatically during the burn-in (Yang and Rodríguez, 2013). Let σ be the current step length (the proposal standard deviation) and P_{jump} the observed acceptance rate. When the Gaussian proposal is applied to the Gaussian target $N(0, 1)$, Gelman *et al.* (1996) noted that P_{jump} has a simple relationship with σ^2:

$$P_{\text{jump}} = \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\sigma}\right). \tag{15}$$

Then

$$\sigma^* = \sigma \times \frac{\tan\left(\frac{\pi}{2}P_{\text{jump}}\right)}{\tan\left(\frac{\pi}{2}P^*_{\text{jump}}\right)}, \tag{16}$$

with the optimal acceptance rate $P^*_{\text{jump}} = 0.44$, gives the optimal step length σ^* (Yang and Rodríguez, 2013). A few rounds of automatic step-length adjustments may be used during the burn-in to account for the fact that the posterior may not be exactly Gaussian.

Besides the optimal window size or step length for a given sliding-window proposal, alternative proposal kernels may provide improved mixing. Several algorithms make use of local information of the target to guide the proposal. For example, in Hamiltonian Monte Carlo (HMC) (Neal, 2011) the gradient is used to guide the proposal toward high-probability regions; recall that positive gradient or slope indicates the direction where the density increases. Yang and Rodríguez (2013) explored a Bactrian proposal which resembles a two-humped camel and suppresses values close to the current value, reducing positive autocorrelation in the MCMC

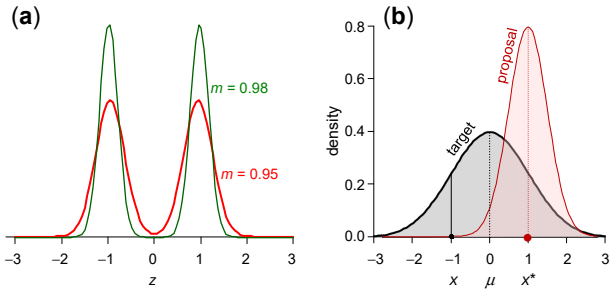

Fig. 3: (**a**) The Bactrian proposal is a 1:1 mixture of two normal distributions. When used as a proposal in an MCMC algorithm, it favours values that are different from the current value. Redrawn based on Yang and Rodríguez (2013, fig. 1a). (**b**) The mirror move applied to the Gaussian target $\pi(x) = \phi(x; \mu, \sigma^2)$ with location μ and scale σ. Given the current state x, the mirror move samples around $x^* = 2\mu - x$, the mirror image of the current state on the other side of the target, using a proposal standard deviation that is $\frac{1}{2}$ the posterior standard deviation (Thawornwattana *et al.*, 2018). In real applications, μ and σ in the target are estimated during the burn-in.

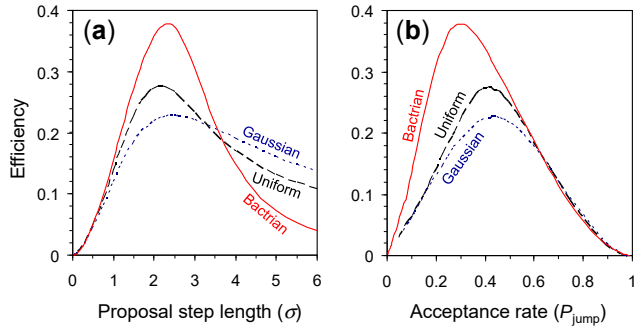

Fig. 4: MCMC mixing efficiency (defined as the variance ratio) plotted (**a**) against the proposal step length (σ) and (**b**) against the acceptance rate (P_{jump}) for three proposals applied to the $N(0, 1)$ target. The three proposals are all sliding-windows moves based on the uniform, Gaussian and Bactrian distributions (fig. 3**a**). Redrawn based on Yang and Rodríguez (2013, fig. 2).

sample (fig. 3a). For the $N(0, 1)$ target, the sliding-window proposal based on the Gaussian, the uniform, and the Bactrian kernels achieves the optimal efficiency of $E^* = 0.23$, 0.28, and 0.30, respectively, with the optimal acceptance rate at 44%, 41%, and 30%, and with the optimal proposal standard deviation to be 2.5×, 2.2×, and 2.3× the target standard deviation, (fig. 4) (Yang and Rodríguez, 2013; Thawornwattana *et al.*, 2018). The uniform sliding window is more efficient (and computationally less expensive) than the Gaussian sliding window, while the Bactrian sliding window is even most efficient. The mirror proposal (fig. 3b) samples the new value around the mirror image of the current value on the 'other side' of the target, and may introduce negative correlation in the MCMC sample, achieving super efficiency (with $E > 1$) (Thawornwattana *et al.*, 2018).

It may be noted that current MCMC algorithms in Bayesian phylogenetics achieve very low mixing efficiency. Suppose we run a Bayesian program over $\sim 10^8$ MCMC iterations and achieve ESS ~100. Then efficiency is only 10^{-6}. There is a need for developing smart MCMC proposals to improve the mixing efficiency and reduce running time and energy consumption (Douglas *et al.*, 2022).

Multiple local peaks and MC^3

Multiple modes in the posterior can cause serious convergence and mixing problems, and are a challenging problem in Bayesian computation. Multiple modes may arise due to conflict between the prior and the likelihood, or because the model is parameter rich with a complex correlation structure. A method for dealing with multiple modes in the posterior is Metropolis-coupled MCMC or MC^3 algorithm, also known as parallel tempering (Geyer, 1991; Marinari and Parisi, 1992). This is popular in phylogenetics due to its implementation in MrBayes (Altekar *et al.*, 2004).

The MC^3 algorithm involves running multiple chains in parallel. The jth chain has the stationary distribution

$$\pi_j(\theta) \propto \pi(\theta)^{1/T_j}, \tag{17}$$

where $T_j = 1 + \lambda(j-1)$, with $\lambda > 0$, is the temperature for chain j. The first chain has $\pi_1(\cdot) = \pi(\cdot)$, so it samples from the target posterior density and is called the *cold chain*. The other chains are designed to improve mixing and are called hot chains. Note that raising the density $\pi(\cdot)$ to the power $1/T$ with $T > 1$ has the effect of flattening out the surface, making it easier for the Markov chain to cross valleys and to move from one peak to another.

The hot chains will visit the local peaks easily, and swapping states between chains will let the cold chain occasionally jump across valleys, leading to better mixing. To increase the acceptance rate, one may build a temperature ladder with multiple hot chains and attempt to swap only adjacent chains in the ladder. At the end of the run, output from only the cold chain is used, while outputs from the hot chains are discarded. An obvious disadvantage of the algorithm is that m chains are run but only one chain is used for inference. MC^3 is well suited to implementation on multi-processor multi-core servers or computer clusters, since each chain will require about the same amount of computation per iteration, and there is very little communication between chains.

In molecular phylogenetics, local peaks are known to exist in the space of trees, and the problem is more serious for tree-perturbation algorithms (such as NNI) that do not induce many neighbouring trees for the current tree. Unfortunately the MC^3 algorithm is designed to help the chain move from one peak to another on the same posterior surface in problems of parameter estimation under a well-specified likelihood model. It may be ineffective for moving from one model to another in cross-model algorithms. Note that different trees

represent different likelihood models. Cross-model algorithms (Green, 1995) often have more serious mixing problems. In particular, in large datasets, the within-model posterior may be very sharp, and a move from one tree to another is like a jump from one tower in one world to another tower in another world. The proposal tends to be rejected even if the new tree has a higher posterior probability than the current, because the proposed branch lengths and other parameters are away from the posterior mode.

MCMC convergence and mixing

We may distinguish two issues with an MCMC run: slow convergence and poor mixing. Convergence means that after the MCMC algorithm runs over many iterations, the chain will approach the stationary distribution, independent of the initial state. Slow convergence means that the chain has not reached stationarity even after many iterations. Poor mixing means that the chain mixes slowly after it has reached stationarity. As mentioned above, poor mixing may be due to too small or too large step lengths, strong correlations in the parameters, etc. Both convergence and mixing problems may occur in large applications, and it may be challenging to assess and overcome such problems. Here we mention a few commonly used strategies for validating and diagnosing MCMC programs. Note that those diagnostics are able to reveal problems but may not be able to prove the correctness of the algorithm or implementation.

1. Plotting visited states over the MCMC iterations in so-called trace plots (see, e.g., fig. 1) is very useful for detecting convergence and mixing problems. It is important to monitor all parameters in the model.
2. For most proposals, the acceptance rate should be neither too high nor too low. Acceptance rates of 10-80% may be healthy but caution should be applied if acceptance is near 0% or 100%.
3. Multiple runs of the algorithm from different starting points should converge to the same posterior distribution. Running the same analysis multiple times is often a very effective approach to identifying problems with the MCMC run.
4. We can run the MCMC with no data and to compare the MCMC sample against the prior. Note that if the likelihood is set to $f(x|\theta) = 1$ in eq. 1, the posterior will become the prior. If the Bayesian program does not provide an option for 'turning off' the data, one can easily construct a dummy empty dataset (e.g., an alignment consisting of one site with missing data for every species). Often the prior means and variances etc. are analytically available for comparison. Also one can simulate larger and larger datasets under a fixed set of parameter values and analyze the simulated data under the correct model, to confirm that the Bayesian estimate becomes closer and closer to the true value. This test relies on the fact that Bayesian estimation is consistent.
5. Many Bayesian programs also provide facilities for simulating data under the likelihood model. The so-called Bayesian simulation may be very effective for identifying problems with a Bayesian MCMC implementation. In Bayesian simulation, one samples parameter values from the prior and then use them to simulate a replicate dataset under the likelihood model. One can generate and analyze many replicate datasets, with each dataset simulated using different parameter values. Then the combined MCMC sample across the datasets should be from the prior. The data size should be intermediate so that both the prior and the likelihood influence the posterior. See Flouri *et al.* (2023) for an application of this strategy.

Bayesian model comparison and hypothesis testing

Marginal likelihood and Bayes factor

When multiple competing models are available to explain the data, cross-model MCMC algorithms may be used to move between models in addition to within-model MCMC that moves in the parameter space for each model (Green, 1995). Indeed, MCMC algorithms in phylogenetics may move between different tree topologies and are thus cross-model algorithms. In general cross-model algorithms (in particular those that move between models of different dimensions) run into convergence and mixing problems more easily than within-model algorithms.

Here we focus on Bayesian comparison of two models. The commonly used device for Bayesian model comparison is the Bayes factor, which is the ratio of the marginal likelihood values under the two compared models. Let θ_0 be the parameter vector for H_0, and θ_1 be the parameter vector for H_1. The Bayes factor in support of H_1 against H_0 is

$$B_{10} = \frac{M_1}{M_0} = \frac{\mathbb{P}(x|H_1)}{\mathbb{P}(x|H_0)} = \frac{\int p(\theta_1)p(x|\theta_1, H_1)\,\mathrm{d}\theta_1}{\int p(\theta_0)p(x|\theta_0, H_0)\,\mathrm{d}\theta_0}, \tag{18}$$

where M_0 and M_1 are the marginal likelihoods for the two models, respectively (Jeffreys, 1961). Note that the marginal likelihood M for each model is the normalizing constant $p(x)$ in eq. 1.

The posterior model probabilities are given by

$$\frac{\mathbb{P}(H_1|x)}{\mathbb{P}(H_0|x)} = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \times \frac{M_1}{M_0} = \frac{\mathbb{P}(H_1)}{\mathbb{P}(H_0)} \times B_{10}. \tag{19}$$

If the prior probabilities are uniform, $\mathbb{P}(H_0) = \mathbb{P}(H_1)$, the Bayes factor will be equal to the ratio of posterior model probabilities: $B_{10} = \frac{\mathbb{P}(H_1|x)}{\mathbb{P}(H_0|x)}$ or $\mathbb{P}(H_1|x) = \frac{B_{10}}{1+B_{10}}$. This provides a 'calibration' for the Bayes factor: $B_{10} = 19$ (corresponding to $\mathbb{P}(H_1|x) = 0.95$) may be considered strong evidence for H_1 while $B_{10} = 99$ (or $\mathbb{P}(H_1|x) = 0.99$) means extremely strong evidence.

Note that the Bayes factor may strongly support H_0 as well (when $B_{10} < 0.05$ or 0.01).

The Bayes factor is a likelihood ratio, but it has important differences from the likelihood ratio in hypothesis testing.

1. In the LRT, the likelihood is optimized over the model parameters, while in calculation of the Bayes factor or in Bayesian model comparison, the marginal likelihood is an average over the model parameters. As a result, the marginal likelihood is influenced by the prior on model parameters as well as the data-generating model. For example, a conflict between the prior and the likelihood may make the model appear poor in Bayesian model selection.
2. When we add a new parameter to the model, the optimized likelihood may increase and never decrease, and the LRT is used to decide whether the improvement in likelihood is large enough to justify the inclusion of the new parameter. In contrast, the marginal likelihood may either increase or decrease when a new parameter is added to the model, and parameter-richness is penalized automatically in the calculation of marginal likelihood with no need for a test.
3. Bayes factors can be applied to non-nested models and to comparison of more than two models. In contrast hypothesis testing applies to comparison of two nested hypotheses, with the null hypothesis H_0 corresponding to the alternative hypothesis H_1 with the parameter of interest fixed at the null value. Comparison of nonnested models using the LRT is very challenging (so that information criteria such as the AIC is used instead). Even when the two hypotheses are nested, there can be considerable technical difficulties to apply the LRT if the testing problem is 'nonstandard' (Self and Liang, 1987; Brazzale and Mameli, 2024). For example, the null parameter value may be at the boundary of the parameter space for H_1 (Self and Liang, 1987), or some parameters in H_1 may become unidentifiable when the parameters of interest are fixed at the null value. In such cases, the distribution of the LRT statistic may be unknown or may not exist. The Bayes factor applies in a straightforward manner under such conditions, with no need for any special treatment.
4. Hypothesis testing and posterior model probabilities (or Bayes factor) may produce different numerical results or even contradictory conclusions when applied to the same data. In particular, the Bayes factor may lead to strong support for the null hypothesis H_0 (or strong rejection of H_1). This is not possible in classical testing, in which one may fail to reject the null but never support the null with great force. We discuss an example of this below.

Computation of the marginal likelihood or Bayes factor can be very demanding. Many methods exist (see for review Fourment *et al.*, 2020), but those that are likely to produce reliable results, such as path-sampling or thermodynamic integration (Ogata, 1989; Gelman and Meng, 1998; Lartillot and Philippe, 2006), stepping-stones (Xie *et al.*, 2011), and nested sampling (Skilling, 2006), all require many expensive MCMC runs.

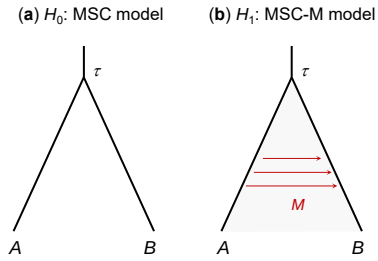

Fig. 5: (**a**) The multispecies coalescent (MSC) model is a special case (i.e., H_0) of (**b**) the MSC-with-migration model (MSC-M, H_1), with the migration rate fixed at the null value: $M = M_0 = 0$ (Flouri *et al.*, 2023). Both models involve the species split time (τ) and the population size parameter θ. The two models can be tested using the Bayes factor, which is given by the Savage-Dickey density ratio.

Test of nested hypotheses

When the two models are nested, Bayes factor is given by the Savage-Dickey density ratio (Dickey, 1971). We illustrate the approach using the example of figure 5, in which genomic data are used to test for presence of gene flow between two species. The null model is the multispecies coalescent model with no gene flow (H_0: MSC) while the alternative model is MSC-with-migration (H_1: MSC-M). When the migration rate takes the null value $M = M_0 = 0$, H_1 reduces to H_0, so the two models are nested.

When the priors on nuisance parameters (parameters that exist in both models; here these include species split times and population sizes) match between the two models, the Bayes factor B_{10} can be expressed as

$$B_{10} = \frac{p(M_0)}{p(M_0|x)}, \tag{20}$$

where $p(M_0)$ is the prior density $p(M)$ and $p(M_0|\boldsymbol{x})$ is the posterior density $p(M|\boldsymbol{x})$, both evaluated at the null value $M_0 = 0$ (Dickey, 1971).

Ji *et al.* (2023) discussed how to process an MCMC sample under H_1 to calculate eq. 20. We define a *null region* around the null value, with $M < \epsilon$, within which gene flow is negligible. Then

$$B_{10} = \frac{p(M_0) \cdot \epsilon}{p(M_0 \mid \boldsymbol{x}) \cdot \epsilon} \approx \frac{\mathbb{P}(M < \epsilon)}{\mathbb{P}(M < \epsilon \mid \boldsymbol{x})}, \tag{21}$$

Intuitively we contrast the prior and posterior probabilities for negligible gene flow. If the posterior probability is much larger than the prior probability, B_{10} will be large and the data will strongly support gene flow. This occurs when the rate of gene flow (M) is large and outside the null region in the posterior.

Calculation of B_{10} using eq. 21 involves running the MCMC under H_1 and processing the MCMC sample to calculate the posterior probability $\mathbb{P}(M < \epsilon|\boldsymbol{x})$. The prior probability is typically analytically available. This approach thus involves ~ 100 times less computation than marginal likelihood calculations based on MCMC (which may require 50 MCMC runs for each of the two models; Fourment *et al.*, 2020, table 1).

Jeffreys's paradox

Classical hypothesis testing and Bayesian model selection applied to the same data may produce strongly opposed conclusions, a situation known as Jeffreys's paradox (Jeffreys, 1935; Lindley, 1957).

Consider testing the null hypothesis $H_0 : \mu = 0$ against the alternative $H_1 : \mu \neq 0$, using a data sample, $x = (x_1, x_2, \cdots, x_n)$, from the Gaussian distribution $N(\mu, 1)$. The data can be summarized as the sample mean $\bar{x}$, with the likelihood given by $\bar{x} \sim N(0, 1/n)$ under H_0 and $\bar{x} \sim N(\mu, 1/n)$ under H_1. The likelihood under H_0 is $L_0 = \phi(\bar{x}; 0, \frac{1}{n})$, where $\phi(x; \mu, \sigma^2)$ is the PDF for $N(\mu, \sigma^2)$. The likelihood under H_1 is $L_1(\mu) = \phi(\bar{x}; \mu, \frac{1}{n})$. This is maximized at $\hat{\mu} = \bar{x}$, with $L_1(\hat{\mu}) = \phi(\bar{x}; \bar{x}, \frac{1}{n})$.

In hypothesis testing, the LRT statistic is

$$2\Delta\ell = 2\log\frac{L_1}{L_0} = 2\log\frac{\phi(\bar{x}; \bar{x}, \frac{1}{n})}{\phi(\bar{x}; 0, \frac{1}{n})} = n|\bar{x}|^2, \qquad (22)$$

in comparison with the critical value $\chi^2_{1,5\%} = 3.84$. Note that this is equivalent to the z test, which compares $\sqrt{n}\bar{x}$ with $N(0, 1)$, with the critical value 1.96 at the 5% level. (If $z \sim N(0, 1)$, then $z^2 \sim \chi^2_1$.)

In a Bayesian analysis, H_0 does not involve any parameters so that the marginal likelihood is the likelihood: $M_0 = L_0$. For H_1, we assign the prior $\mu \sim N(\mu_0, \sigma_0^2)$, and the marginal likelihood becomes

$$M_1 = \int \phi(\mu; \mu_0, \sigma_0^2)\phi\big(\bar{x}; \mu, \tfrac{1}{n}\big)\,\mathrm{d}\mu = \phi(\bar{x}; \mu_0, \sigma_0^2 + \tfrac{1}{n}). \qquad (23)$$

The Bayes factor is then

$$B_{10} = \frac{M_1}{M_0} = \frac{\phi(\bar{x}; \mu_0, \sigma_0^2 + \frac{1}{n})}{\phi(\bar{x}; 0, \frac{1}{n})}. \qquad (24)$$

While the LRT (eq. 22) depends on the data ($n|\bar{x}|^2$) only, the Bayes factor (eq. 24) depends on the prior (μ_0, σ_0^2) as well.

Consider a dataset that gives $\sqrt{n}|\bar{x}| = 1.96$ or $2\Delta\ell = 3.84$ (or $L_1/L_0 = \mathrm{e}^{1.92}$). The LRT is significant, with the p-value 0.05.

In H_1, the marginal or average likelihood M_1 must be smaller than the maximized likelihood, $M_1 \leq L_1$. Suppose we assign a prior for H_1 to support H_1 as much as possible. The best one could do is to choose the prior to match the observed data, with $\mu_0 = \bar{x}$ and $\sigma_0 = 0$. With this extreme prior, $M_1 = L_1$, and the highest posterior probability for H_1 achieved is then $\mathbb{P}(H_1|x) = L_1/(L_0 + L_1) = 0.872$ or $\mathbb{P}(H_0|x) = 1 - \mathbb{P}(H_1|x) = 0.128$. Even with such an extreme prior, 12.8% is not small enough to reject H_0. With other choices of prior on μ, $\mathbb{P}(H_0|x)$ can be larger, and can indeed reach ~100%. When that happens, the LRT and the Bayes factor will reach opposing conclusions from the same data, with the LRT rejecting H_0 and the Bayes factor strongly supporting H_0.

Jeffereys's paradox raises philosophical issues concerning the principles of statistical inference and is the topic of many discussions. Here we venture two comments that appear to be non-controversial. First, it is generally perceived that the LRT tends to reject the null model and favour parameter-rich models too often, especially in large datasets. Note that if H_0 is true, the false positive rate of the LRT stays at 5% when the sample size $n \to \infty$. In contrast in Bayesian analysis, the true model H_0 will dominate, so that $B_{10} \to 0$ and $\mathbb{P}(H_0|x) \to 1$ when $n \to \infty$. Second, Bayesian model comparison, using either Bayes factor or posterior model probabilities, may be sensitive to the prior on model parameters, in particular, priors on parameters that are in one model but not the other.

High posterior probabilities for trees and clades

Bayesian model selection is known to be consistent. When the data size $n \to \infty$, the true model 'dominates', with its posterior probability approaching 1. If several models are equally right, the model with fewer parameters dominates (Dawid, 2011). However, this theory applies only if the true model is included in the comparison.

When the competing models are equally wrong, Bayesian model selection may exhibit polarized behaviors in large datasets, supporting one model with full force while rejecting the others. If one model is slightly less wrong than the other, the less wrong model will eventually win when the amount of data increases, but the method may become overconfident before it becomes reliable (Yang and Zhu, 2018).

In molecular phylogenetics, the Bayesian method has been noted to produce very high posterior probabilities for trees or clades in analyses of large datasets. In some analyses the trees are decidedly incorrect (e.g., different trees for the three codon positions or for DNA and protein data, or conflicting trees depending on taxon sampling) and the high support values are spurious. The extreme behavior of Bayesian model selection may be a contributing factor. There have been attempts to develop methods that are less sensitive to model violation, including BayesBag, which averages posterior model probabilities over bootstrap replicate datasets (Huggins and Miller, 2023). The effectiveness of such ideas to phylogeny reconstruction is yet to be tested.

Bayesian phylogenetic software

- Beast is a Bayesian MCMC program for phylogenetic analysis of molecular and morphological data (Drummond and Rambaut, 2007). It implements strict or relaxed molecular clock models and works on rooted trees.
- Bpp is a MCMC program for Bayesian analysis of genomic sequence data from multiple species under the multispecies coalescent models, incorporating species divergences and interspecific gene flow (Yang, 2015; Flouri *et al.*, 2018).
- mcmctree in paml (for Phylogenetic Analysis by Maximum Likelihood) is a Bayesian MCMC program for dating species divergences (Yang, 2007).
- MrBayes and its update RevBayes are Bayesian MCMC programs for phylogenetic analysis using nucleotide, amino acid, and codon sequences, as well as morphological characters from extant and fossil species (Ronquist *et al.*, 2012; Hohna *et al.*, 2016).
- PhyloBayes is a Bayesian MCMC program for phylogenetic reconstruction (Lartillot *et al.*, 2009). It includes sophisticated models of amino acid substitution that account for heterogeneity in the substitution process among genes and sites, which may be important for inferring deep phylogenies.

Key points/Objectives

- Bayesian inference (BI) is widely used in phylogenetics to implement advanced high-dimensional multi-parameter models for analysis of molecular and morphological data from extant and extinct species.
- MCMC algorithms make Bayesian computation possible, by bypassing computation of multidimensional integrals.
- Bayesian marginalization is an attractive approach to accounting for heterogeneity in the data-generating process.
- BI provides a natural framework for integrating information from different sources (such as molecules and fossils in Bayesian relaxed clock dating analysis).
- In Bayesian model selection, the marginal likelihood automatically penalizes parameter-rich models and can support the null model with great force, which is impossible with hypothesis testing.
- With model misspecification, Bayesian model comparison may exhibit extreme polarized behavior.
- BI is used in nearly every aspect of phylogenetic analysis.

Conclusion

Since its introduction into phylogenetics in the 1990s, Bayesian inference has become the dominating methodology for implementing advanced models of data analysis in the field. The Bayesian framework is particularly natural for combining information from different sources, and for implementing heterogeneous models of sequence or trait evolution. Bayesian inference has been applied to address virtually all major questions in evolutionary biology, such as inferring phylogenetic relationships and divergence times among species, detecting molecular adaptation, estimating species trees despite conflicting gene trees, inferring viral pandemic dynamics, inferring gene flow between species, delimiting species boundaries using genomic data, and reconstructing genes and genomes in extinct ancestral species. The field is also a rich ground for testing novel statistical computational algorithms. We expect continual improvements in MCMC algorithms used in phylogenetics will make the methodology ever more widely used in analysis of the ever-increasing genomic data.

Acknowledgement

This work has been supported by Biotechnology and Biological Sciences Research Council (BBSRC) grants (BB/X007553/1, BB/X018571/1, BB/Y004132/1) to Z.Y.

Relevant Websites

- bpp: `https://github.com/bpp/bpp`
- FigTree: `https://beast.community/figtree`
- MrBayes: `http://mrbayes.sourceforge.net/`
- paml/mcmctree: `https://github.com/abacus-gene/paml`
- PhyloBayes: `https://github.com/bayesiancook/phylobayes`
- Tracer: `https://beast.community/tracer`

Further Reading

- Yang (2014), Chapters 7 and 8.
- Chen *et al.* (2014).
- Jiao *et al.* (2021).
- dos Reis *et al.* (2016).
- Nascimento *et al.* (2017).

See also: Consensus Methods, Phylogenetics. Directed Evolution, History of. Maximum Likelihood Phylogenetic Inference. Molecular Evolution, Models of. Searching Tree Space, Methods for

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20: 407–415.

Brazzale, A. R. and Mameli, V. 2024. Likelihood asymptotics in nonregular settings: A review with emphasis on the likelihood ratio. *Statist. Sci.*, 39: 322–345.

Chen, M.-H., Kuo, L., and Lewis, P. 2014. *Bayesian Phylogenetics: Methods, Algorithms, and Applications*. Chapman & Hall/CRC, London.

Dawid, A. 2011. Posterior model probabilities. In P. S. Bandyopadhyay and M. Forster, editors, *Philosophy of Statistics*, pages 607–630. Elsevier, New York.

Dickey, J. M. 1971. The weighted likelihood ratio, linear hypotheses on normal location parameters. *Ann. Math. Statist.*, 42(1): 204–223.

dos Reis, M., Donoghue, P. C. J., and Yang, Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.*, 17: 71–80.

Douglas, J., Jimenez-Silva, C. L., and Bouckaert, R. 2022. StarBeast3: adaptive parallelised Bayesian inference under the multispecies coalescent. *Syst. Biol.*, 71(4): 901–916.

Drummond, A., Ho, S., Phillips, M., and Rambaut, A. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol.*, 4: e88.

Drummond, A. J. and Rambaut, A. 2007. Beast: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.*, 7: 214.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17: 368–376.

Felsenstein, J. 2004. *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Flouri, T., Jiao, X., Rannala, B., and Yang, Z. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.*, 35(10): 2585–2593.

Flouri, T., Jiao, X., Huang, J., Rannala, B., and Yang, Z. 2023. Efficient Bayesian inference under the multispecies coalescent with migration. *Proc. Nat. Acad. Sci. U.S.A.*, 120(44): e2310708120.

Fourment, M., Magee, A. F., Whidden, C., Bilge, A., Matsen, F. A., and Minin, V. N. 2020. 19 dubious ways to compute the marginal likelihood of a phylogenetic tree topology. *Syst. Biol.*, 69(2): 209–220.

Gelman, A. and Meng, X. 1998. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Stat. Sci.*, 13: 163–185.

Gelman, A., Roberts, G., and Gilks, W. 1996. *Efficient Metropolis jumping rules*, volume 5, pages 599–607. Oxford University Press, Oxford.

Geyer, C. 1991. Markov chain Monte Carlo maximum likelihood. In E. Keramidas, editor, *Computing Science and Statistics: Proc. 23rd Symp. Interface*, pages 156–163. Interface Foundation, Fairfax Station.

Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11: 725–736.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82: 711–732.

Green, P. J. and Han, X. L. 1992. Metropolis methods, Gaussian proposals and antithetic variables.

Hastings, W. 1970. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, 57(1): 97–109.

Hohna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P., and Ronquist, F. 2016. Revbayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.*, 65(4): 726–736.

Huggins, J. H. and Miller, J. W. 2023. Reproducible model selection using bagged posteriors. *Bayesian Anal*, 18: 79–104.

Jeffreys, H. 1935. Some tests of significance, treated by the theory of probability. *Proc. Cam. Phil. Soc.*, 31: 203–222.

Jeffreys, H. 1961. *Theory of Probability*. Oxford University Press, Oxford, England, 3rd edition.

Ji, J., Jackson, D. J., Leache, A. D., and Yang, Z. 2023. Power of Bayesian and heuristic tests to detect cross-species introgression with reference to gene flow in the *Tamias quadrivittatus* group of North American chipmunks. *Syst. Biol.*, 72(2): 446–465.

Jiao, X., Flouri, T., and Yang, Z. 2021. Multispecies coalescent and its applications to infer species phylogenies and cross-species gene flow. *Nat. Sci. Rev.*, 8(12): DOI: 10.1093/nsr/nwab127.

Jukes, T. H. and Cantor, C. R. 1969. *Evolution of protein molecules*, pages 21–123. Academic Press, New York.

Lartillot, N. and Philippe, H. 2006. Computing Bayes factors using thermodynamic integration. *Syst. Biol.*, 55: 195–207.

Lartillot, N., Lepage, T., and Blanquart, S. 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, 25(17): 2286–2288.

Legried, B. and Terhorst, J. 2022. A class of identifiable phylogenetic birth–death models. *Proc. Natl. Acad. Sci.*, 119(35): e2119513119.

Li, S., Pearl, D., and Doss, H. 2000. Phylogenetic tree reconstruction using Markov chain Monte Carlo. *J. Amer. Statist. Assoc.*, 95: 493–508.

Lindley, D. 1957. A statistical paradox. *Biometrika*, 44: 187–192.

Louca, S. and Pennell, M. W. 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature*, 580(7804): 502–505.

Marinari, E. and Parisi, G. 1992. Simulated tempering: a new Monte Carlo scheme. *Europhysics Lett.*, 19: 451–458.

Mau, B. and Newton, M. 1997. Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *J. Computat. Graph. Stat.*, 6: 122–131.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. 1953. Equations of state calculations by fast computing machines. *J. Chem. Physi.*, 21(6): 1087–1092.

Nascimento, F. F., dos Reis, M., and Yang, Z. 2017. A biologist's guide to Bayesian phylogenetic analysis. *Nature Ecol. Evol.*, 1: 1446–1454.

Neal, R. M. 2011. Mcmc using hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, and X.-L. Meng, editors, *Handbook of Markov Chain Monte Carlo*, pages 113—-162. Chapman and Hall/CRC, London.

Ogata, Y. 1989. A Monte Carlo method for high dimensional integration. *Numer. Math.*, 55: 137–157.

Peskun, P. 1973. Optimum Monte-Carlo sampling using Markov chains. *Biometrika*, 60(3): 607–612.

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

Rannala, B. and Yang, Z. 2025. Reading tree leaves: Inferring speciation and extinction processes using phylogenies. *Phil. Trans. R. Soc. Lond.*, 380: 20230309.

Rannala, B., Zhu, T., and Yang, Z. 2012. Tail paradox, partial identifiability and influential priors in Bayesian branch length inference. *Mol. Biol. Evol.*, 29: 325–335.

Rodrigue, N., Philippe, H., and Lartillot, N. 2010. Mutation-selection models of coding sequence evolution with site-heterogeneous amino acid fitness profiles. *Proc. Natl. Acad. Sci. U.S.A.*, 107: 4629–4634.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Hohna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. 2012. Mrbayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61: 539–542.

Rosenberg, N. A., Stadler, T., and Steel, M. 2025. A mathematical theory of evolution: phylogenetic models dating back 100 years. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, 380(1919): 20230297.

Self, S. and Liang, K.-Y. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.*, 82: 605–610.

Skilling, J. 2006. Nested sampling for general Bayesian computation. *Bayesian Anal*, 1: 833–859.

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

Swofford, D., Olsen, G., Waddell, P., and Hillis, D. 1996. Phylogeny inference. In D. M. Hillis, C. Moritz, and B. K. Mable, editors, *Molecular Systematics*, pages 407–514. Sinauer Associates, Sunderland, Massachusetts, 2 edition.

Thawornwattana, Y., Dalquen, D., and Yang, Z. 2018. Designing simple and efficient Markov chain Monte Carlo proposal kernels. *Bayesian Analysis*, 13(4): 1033–1059.

Thorne, J., Kishino, H., and Painter, I. 1998. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.*, 15: 1647–1657.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., and Chen, M.-H. 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.*, 60: 150–160.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10: 1396–1401.

Yang, Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24: 1586–1591.

Yang, Z. 2014. *Molecular Evolution: A Statistical Approach*. Oxford University Press, Oxford, England.

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, 61(5): 854–865.

Yang, Z. 2018. AWF Edwards and the origin of Bayesian phylogenetics. In R. G. Winther, editor, *AWF Edwards*. Cambridge University Press, Cambridge, England.

Yang, Z. and Rannala, B. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.*, 14: 717–724.

Yang, Z. and Rodríguez, C. E. 2013. Searching for efficient markov chain monte carlo proposal kernels. *Proc. Natl. Acad. Sci. U.S.A.*, 110(48): 19307–19312.

Yang, Z. and Zhu, T. 2018. Bayesian selection of misspecified models is overconfident and may cause spurious posterior probabilities for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.*, 115(8): 1854–1859.