A biologist's guide to Bayesian phylogenetic analysis

Fabrícia F. Nascimento[1*], Mario dos Reis[2] and Ziheng Yang[3*]

1. Department of Zoology, University of Oxford, OX1 3PS, UK.
2. School of Biological and Chemical Sciences, Queen Mary University of London, E1 4NS, UK
3. Department of Genetics, Evolution and Environment, University College London, WC1E 6BT, UK

*Correspondent author
FFN: nascimentoff@yahoo.com.br
ZY: z.yang@ucl.ac.uk

Abstract

Bayesian methods have become very popular in molecular phylogenetics due to the availability of user-friendly software implementing sophisticated models of evolution. However, Bayesian phylogenetic models are complex, and analyses are often carried out using default settings, which may not be appropriate. Here, we summarize the major features of Bayesian phylogenetic inference and discuss Bayesian computation using Markov chain Monte Carlo (MCMC), the diagnosis of an MCMC run, and ways of summarising the MCMC sample. We discuss the specification of the prior, the choice of the substitution model, and partitioning of the data. Finally, we provide a list of common Bayesian phylogenetic software and provide recommendations as to their use.

Introduction

Bayesian phylogenetic methods were introduced in the 1990s[1,2] and have since revolutionised the way we analyse genomic sequence data[3]. Examples of such analyses include phylogeographic analysis of virus spread in humans[4-7], inference of phylogeographic history and migration between species[8-10], analysis of species diversification rates[11,12], divergence time estimation[13-15], and inference of phylogenetic relationships among species or populations[13,16-20]. The popularity of Bayesian methods appears to be due to two factors: (1) the development of powerful models of data analysis; and (2) the availability of user-friendly computer programs implementing the models (Table 1).

Models implemented in Bayesian software programs are becoming increasingly complicated, and the priors and model assumptions made in those programs are not always clear to the user. Analyses are often conducted using default priors, which may not be appropriate and may lead to biased or incorrect results. Likewise, over-simplified likelihood models may produce biased results, while over-complicated models may lead to loss of power as well as inefficient computation.

The workhorse underlying all modern Bayesian phylogenetic programs is the Markov chain Monte Carlo (MCMC) or Metropolis-Hastings algorithm[21,22]. However, MCMC is both art and science, and a basic understanding of its workings is essential for the correct use of those programs. In this review, we explain the basic concepts of Bayesian statistics and discuss the major features of MCMC algorithms, such as the prior and the likelihood, MCMC proposals, diagnosis of MCMC convergence and mixing, and summary of the posterior sample. Our intended reader is the empirical biologist who needs to use Bayesian phylogenetic programs to analyse their data. We lay out and answer a set of questions important for setting up a Bayesian analysis. We focus on Bayesian estimation of phylogenetic trees. However, the basic concepts discussed here apply to other phylogenetic problems as well, such as divergence time estimation or species tree estimation under the multi-species coalescent model. Extensive reviews of these are available elsewhere[23-25].

What is the Bayesian method?

The Bayesian method is a statistical inference methodology. Its main feature is the use of probability distributions to describe the uncertainty of all unknowns including the model parameter(s). Let D be the observed data and θ the unknown parameter. We assign a distribution $f(\theta)$, called the *prior distribution*, based on our knowledge about θ before analysis of the data. After the data are observed, we use Bayes's theorem to calculate the *posterior distribution* of θ given the data:

$$f(\theta \mid D) = \frac{1}{z} f(\theta) f(D \mid \theta), \qquad (1)$$

where the probability of the data given the parameter, $f(D|\theta)$, is called the *likelihood*. This summarises the information about θ in the data. The *normalising constant* $z = \int f(\theta) f(D \mid \theta) \mathrm{d}\theta$ ensures that $f(\theta|D)$ integrates to 1 and is a proper statistical distribution.

Equation (1) indicates that the posterior is proportional to the prior times the likelihood, or the posterior combines the information in the prior and in the data. An example of the prior, likelihood and posterior for a two-parameter phylogenetic example is given in Figure 1.

In the above we assume that the model for generating the data is known. In the so-called trans-model inference, we have several competing models, with each model m having its own parameters θ_m. Then a prior, $f(m, \theta_m) = f(m)\, f(\theta_m|m)$, is assigned to both the model (m) and its parameters (θ_m), and the posterior of the model and parameter is similarly given by Bayes's theorem: $f(m, \theta_m|D) \propto f(m, \theta_m)\, f(D|m, \theta_m)$.

In phylogenetics, the tree topology and the substitution model together specify the statistical model for the data. Different tree topologies thus correspond to different models, while the branch lengths or divergence times as well as the substitution parameters (such as the transition/transversion rate ratio) are parameters in the model. The data are usually a molecular sequence alignment or an alignment of morphological characters (or a combination of both).

An appealing property of Bayesian inference is that it makes direct probabilistic statements about the model or unknown parameter. The posterior probability of a model, $f(m|D)$, is the probability that the model is correct, given the data. The 95% *credibility interval* (CI) of a parameter covers the true parameter with probability 0.95, given the data. Such statements are impossible using confidence intervals and p-values in classical statistics, which treat parameters as unknown constants[26].

What type of data can I use?

The most common type of data used in phylogenetic analyses is DNA and amino acid sequence alignments. Morphological characters can also be used[27]. Here, we focus on DNA sequences. The sequences must be aligned before they are used as input data in phylogenetic programs, and alignment accuracy is important in phylogenetic analysis. Much effort has been made to develop models of insertions and deletions[28-30]. For species phylogeny estimation, the sequences must be orthologs, as incorrect use of paralogs may lead to incorrect phylogenies. Several methods are now available to infer paralogy/orthology[31,32].

How do I select a substitution model for my data?

A number of models have been developed to describe nucleotide or amino acid substitutions[26,33,34]. For nucleotide sequences, these range from the simple JC69 (for Jukes and Cantor)[35] to the complex GTR (for General Time Reversible)[36-38], and the unrestricted model (UNREST)[37]. In JC69 all nucleotide changes occur at the same rate, while in GTR or UNREST substitutions occur at different rates depending on the source and target nucleotides. It is also common to assume a gamma model of variable rates across sites, in particular, in analysis of coding DNA or protein sequences[39-41].

Programs such as jModelTest[42], Modelgenerator[43] or PartitionFinder[44] are commonly used to choose a substitution model. Those programs examine the goodness of fit of the model to the data but never consider the robustness of the analysis to model assumptions. For example, it is well known that the transition/tranversion bias typically has a greater impact on the fit of the model to data (judged by the improvement in likelihood), but less effect on estimation of the tree topology and branch lengths than rate variation among sites[41]. Although there does not seem to be serious harm in mechanical use of those programs, it may be unnecessary to do so in many cases. As a rule of thumb, different substitution models tend to give very similar sequence distance estimates when sequence divergence is less than 10%, so that a simple model can be used even though it may not fit the data. Complex models are necessary in reconstruction of deep phylogenies. Two of the most complex nucleotide substitution models, HKY+Γ and GTR+Γ, often produce similar estimates of phylogenetic trees and branch lengths[37,45]. When in doubt, note that it is more problematic to under-specify than to over-specify the model in Bayesian phylogenetics[46].

For discrete morphological data, the Mk model, an extension of the JC69 model to *k* morphological character states, can be used[27]. An extension that allows for unequal rates of substitution is available in MrBayes[47]. A correction for assertion bias is applied in calculation of the likelihood function because only variable characters are used[27]. For continuous characters, diffusion process models (such as the Wiener or the Ornstein-Uhlenbeck process) can be used[48]. Definitions and detailed review of these models are given elsewhere[49]. There has been much interest in the joint analysis of morphological and molecular data to estimate divergence times for extant and fossil species[50-52].

What is over- and under-parameterisation?

A model is non-identifiable if different values of parameters make the same predictions about the data, so that such data can never be used to estimate those parameters; in other words, the model is non-identifiable if $f(D|\theta_1) = f(D|\theta_2)$ for certain $\theta_1 \neq \theta_2$ and for all possible data D [53]. A simple phylogenetic example is estimation of the geological time of divergence between two species (t) and the molecular evolutionary rate (r) using data of a pair of aligned sequences. The likelihood depends only on the molecular distance, $d = rt$, and not on t and r separately, and is the same for, say, $t = 1$ and $r = 0.1$, or $t = 0.1$ and $r = 1$, or any other combination of t and r such that $rt = d = 0.1$. In theory, non-identifiability (or over-parameterisation) is not a serious problem for Bayesian analysis, especially if informative priors are assigned on the parameters. In practice, over-parameterisation can cause both inference difficulties (such as loss of power, strong correlations between parameters, large variance in the posterior, and extreme sensitivity to the prior and model assumptions) and computational problems (such as poor mixing of the MCMC). Sometimes, a model is identifiable, but the data contain only weak information about the parameters with the likelihood surface being nearly flat. Then similar symptoms will show up in the data analysis.

An example is the popular I+G model of rate variation among sites, which assumes a proportion of sites p_0 in the alignment are invariable with rate 0, while the other sites $(1 - p_0)$ evolve according to a discrete gamma distribution[54]. Because the gamma distribution allows for extremely conserved sites with rates close to 0, p_0 and the gamma shape parameter α are strongly correlated[55]. The MCMC algorithm may have to spend a long time exploring a ridge on the posterior surface.

A similar case applies to the use of parameter-rich GTR+Γ model in analysis of highly similar sequences from closely related species as in Bayesian species delimitation or species tree estimation under the multi-species coalescent model[24,56]. The GTR model has eight parameters that describe the exchangeabilities between nucleotides. If there are only a few variable sites in the alignment, there will be little information about those parameters. Simple models, such as JC69 and K80, may be adequate in such analysis.

On the other hand, the use of overly simplistic model or under-parametrisation can cause systematically incorrect phylogenetic trees and seriously biased estimates of branch lengths and substitution parameters, and over-confident assessment of uncertainties such as spuriously high posterior probabilities for trees or clades[46]. For example, ignoring variable substitution rates among sites leads to underestimated branch lengths[41]. Systematic errors tend to be greater when sequences are more divergent. In short, the substitution model is a trade-off between bias on one hand and variance and computation expense on the other, and should ideally be chosen by a careful consideration of its role on the analysis rather than mechanistic use of a model selection procedure.

How do I decide to concatenate or partition my data?

The rationale for partitioned analysis is that sites in the same partition have similar evolutionary characteristics while those in different partitions have different characteristics[40,44,57]. The characteristics here may be substitution rates, base composition, branch lengths, or even the tree topology. The Bayesian program will estimate different parameter values or even different gene tree topologies for the different partitions, thus accounting for their heterogeneity in the evolutionary process.

For example, genes with different G+C compositions or evolutionary rates may be analysed as separate partitions in phylogeny reconstruction. Vertebrate mitochondrial genes coded on the same strand of the genome have similar G+C content and may be concatenated and analysed as a single partition, although the three codon positions may be treated as different partitions to account for their large differences in rate and in base compositions[58]. Non-coding mitochondrial genes (rRNAs and tRNAs) may be analysed as another partition. Likewise, mitochondrial and nuclear sequences should also be analysed as different partitions[59]. For nuclear sequences, exons and introns should be analysed as different partitions, and the three codon positions should be placed in their own partitions. Some partitioning software may suggest the use of different substitution models for partitions[44] (e.g., HKY for one partition and GTR+G for another). This is unnecessary because with the same model for all partitions, different parameter values will accommodate the heterogeneity among partitions.

An important issue is whether partitions should share the same tree topology. In traditional phylogenetic inference, topology is assumed to be the same across partitions. However, a number of biological processes, such as gene duplication, horizontal gene transfer, and incomplete lineage sorting can cause different genes to have different trees[60,61]. Recently, a number of methods for species tree estimation have been developed under the multi-species coalescent (MSC) model[24,62,63], which account for the process of incomplete lineage sorting (the so-called deep coalescent, due to polymorphism in ancestral species, where coalescence may occur in ancient ancestors leading to gene trees that differ from the species tree). Under the MSC different genomic regions (or exons) are placed into different partitions and allowed to have their own gene-trees, which are embedded into the species tree. The mitochondrial genome does not recombine and mitochondrial genes should be treated as one partition within the MSC. In some viruses, such as influenza, different genome segments can re-assort (i.e. be horizontally transferred) among related strains[64], and thus different segments can have different topologies and should be treated as different partitions.

How do I choose the prior for my Bayesian analysis?

In theory the prior should summarize the biologist's best knowledge about the model or parameters before the data are analysed[26,65]. In practice, specification of the prior is often a thorny issue, especially if there are multiple parameters with complex correlations or if little is known about the parameters. While we are supposed to specify a joint prior distribution for all parameters, the common practice is to ignore the correlation, and assign independent priors for the parameters. When there are many parameters of the same kind, such independent and identically distributed (i.i.d.) prior can sometimes cause problems because they may make a strong statement about the mean or sum of those parameters. For example, it is common to assign independent exponential or uniform priors for branch lengths in the unrooted tree, but this i.i.d. prior can cause very long trees in analysis of highly similar sequence data[66,67]. In relaxed-clock dating analysis, the i.i.d prior for substitution rates among different partitions makes a strong statement about the average rate over loci, leading to biased but over-confident divergence time estimates[68], in particular as the number of partitions increases. Such i.i.d. priors should be avoided.

Default priors in many Bayesian software packages may not be appropriate for the data being analysed and should be used with caution. Specification of the prior is the biologist's responsibility even though it may not be an easy task. Robustness analysis should also be an important component of any Bayesian analysis. By evaluating the posteriors generated under different priors, the biologist can evaluate whether the posterior is robust to the prior.

In Bayesian estimation of phylogenetic trees without the assumption of a molecular clock, it is common to assign a uniform prior on the unrooted tree topologies. When phylogenetic analysis is conducted on rooted trees under the clock or relaxed clock models[69], rooted trees are commonly assigned a prior using a model of cladogenesis such as the Yule process and the birth-death-sampling process[70]. Note that all those models favour balanced trees, and the impact of the prior on the posterior probabilities of the rooted trees can be substantial if the tree is large. For coalescent-based species tree estimation, the MSC model specifies a probability distribution for the rooted gene trees (topologies and node ages)[71]. This is part of the model rather than a prior on gene trees to be specified. In molecular clock dating analysis, fossils may be used to specify minimum and maximum bounds on clade age, which are used to construct a so-called calibration density to calibrate the age of the clade, it is also advisable to include a prior on the age of the root of the tree. For an overview on calibration densities for use in divergence dating, see[72]. It is also necessary to specify a prior on the evolutionary rates for the different loci or partitions. A gamma-Dirichlet prior can be used instead of the i.i.d. prior mentioned above[68]. In relaxed-clock models, the rates not only vary among partitions, but also drift along branches on the tree. Current Bayesian implementations assume that rates drift independently among partitions so that different partitions are independent realizations of the rate-drift process[73,74]. A discussion of the different rate-drift models is given in[68].

What is Markov chain Monte Carlo (MCMC)?

Once the biologist has decided on the data, model and prior, the next step is to obtain a sample from the posterior. This is done by using MCMC, a simulation technique for sampling from a probability distribution that is known up to a normalising constant[21,22]. Note that all terms on the right hand side of equation (1) are straightforward to calculate except the normalizing constant z, which involves multidimensional integrals and may be too expensive to compute. Thus, MCMC is particularly suitable for Bayesian computation. Instead of calculating the posterior distribution $f(\theta|D)$, the algorithm generates a sample from the posterior, which can be used to estimate the mean, the standard deviation of the posterior, or even the whole posterior distribution.

Here we illustrate the major features of MCMC by applying it to the problem of estimating the sequence distance d and the transition/transversion rate ratio κ under the K80 model[75] using a pair of DNA sequences. The data (D) are an alignment of the human and orangutan mitochondrial 12S rRNA genes, summarized as $n_S = 84$ transitional differences and $n_V = 6$ transversional differences at $n = 948$ sites[26, p.7]. We assign independent gamma priors, $d \sim \mathrm{G}(2, 20)$ and $\kappa \sim \mathrm{G}(2, 0.1)$, with densities (Fig. 1a):

$$\begin{aligned} f(d) &= \tfrac{\beta^\alpha}{\Gamma(\alpha)} \times d^{\alpha-1} \mathrm{e}^{-\beta d}, \quad \text{with } \alpha = 2, \beta = 20, \\ f(\kappa) &= \tfrac{\beta^\alpha}{\Gamma(\alpha)} \times \kappa^{\alpha-1} \mathrm{e}^{-\beta \kappa}, \quad \text{with } \alpha = 2, \beta = 0.1. \end{aligned} \tag{2}$$

The likelihood (Fig. 1b) is given by the K80 model[26,75] as

$$f(D|d, \kappa) = (\tfrac{p_0}{4})^{n-n_S-n_V} (\tfrac{p_1}{4})^{n_S} (\tfrac{p_2}{4})^{n_V}, \tag{3}$$

where

$$\begin{aligned} p_0(t) &= \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4\beta t} + \tfrac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4d/(\kappa+2)} + \tfrac{1}{2}\mathrm{e}^{-2d(\kappa+1)/(\kappa+2)}, \\ p_1(t) &= \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4\beta t} - \tfrac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t} = \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4d/(\kappa+2)} - \tfrac{1}{2}\mathrm{e}^{-2d(\kappa+1)/(\kappa+2)}, \\ p_2(t) &= \tfrac{1}{4} - \tfrac{1}{4}\mathrm{e}^{-4\beta t} = \tfrac{1}{4} - \tfrac{1}{4}\mathrm{e}^{-4d/(\kappa+2)}. \end{aligned} \tag{4}$$

Thus, the unnormalized posterior (Fig. 1c) is

$$f(d, \kappa | D) \propto f(d) f(\kappa) f(D|d, \kappa). \tag{5}$$

We give a sketch of an MCMC algorithm in Box 1, and then discuss its main features. We use two sliding windows (uniform distributions centred around the current parameter value) to update parameters d and κ. The sliding window (even with reflection) is a symmetrical proposal, in the sense that the probability density of proposing d^* from d is equal to that of proposing d from d^*. If the proposal is asymmetrical, a correction term, called the Hastings ratio[22] needs to be applied.

Note that the parameter values (d and κ) visited in the next iteration depend on the current values but not values visited in the past. The algorithm has no memory. This memoryless property is called the *Markovian* property. As a result, the sequence of visited parameter values form a Markov chain, and the algorithm is called Markov chain Monte Carlo. An important feature of the algorithm is that it requires the calculation of the ratio of posterior densities, but not the posterior density itself. The normalizing constant z of equation (1) cancels in the calculation of the acceptance ratio α in steps 2a & 2b, and algorithm thus avoids its calculation. It is easy to see that the algorithm visits parameter values with high posterior more often than those with low posterior. Indeed, it visits the parameter values exactly in proportion to their posterior. One runs the algorithm over many iterations, and then uses the visited values of d and κ to construct a histogram to estimate the posterior distribution or to calculate the mean and standard deviation of the posterior (Fig. 2).

The window size (or step length) in the sliding window proposal (w_d and w_κ) can affect the mixing efficiency of the chain (Box 2). If the window is too large, most of the proposals will fall in the tails of the posterior and be rejected. The chain then stays at the current value and does not move (Fig 2a’). If the window is too small, the chain takes tiny baby steps, almost all of which are accepted but the chain is ineffective in exploring the posterior surface (Fig 2b’). Thus, both small steps (with high acceptance proportion) and large steps (with very low acceptance proportion) lead to inefficient algorithms. The step lengths should be adjusted to achieve a near optimal acceptance proportion, at about 30-40%. Fine-tuning a phylogenetic MCMC chain to be efficient is important because MCMC runs may take weeks or months. It is easy to monitor the acceptance proportion and use it to adjust the step length automatically[76]. Most current MCMC phylogenetic programs have automatic fine-tuning algorithms and this is in most cases not a concern for the user.

In *trans-model MCMC* algorithms, both the model index m and the model parameters θ_m change over the chain. The algorithm will involve both within-model proposals, which change parameters of the current model, and trans-model proposals, which move from the current model to another new model[77]. In the long run, the frequency at which the MCMC visits each model is an estimate of the posterior probability of that model. There are a number of differences between within-model and trans-model algorithms[26], and here we note a few concerning mixing efficiency and acceptance proportion. First, for a within-model move (such as a sliding window changing the sequence distance or branch length), we can make the window size small enough so that the acceptance proportion is arbitrarily close to 100%. However, in trans-model moves, the acceptance proportion is constrained by the posterior model probabilities. If the maximum *a posteriori* (MAP) model (the model with the highest posterior probability) has the posterior P_1, then the acceptance proportion cannot exceed $2(1 - P_1)$ [26]. Thus, if the MAP tree has posterior 99%, the highest acceptance proportion for cross-tree moves is 2%. Second, while an acceptance proportion of near 0 indicates a poor proposal (e.g., the window size is too large) for a within-model move, this may and may not indicate a mixing problem in cross-model moves because it may be caused by the MAP model having posterior near 100%. Third, for a within-model move, the optimal acceptance proportion is intermediate at 30-40%, but for a trans-model move, a mobile chain is in general more efficient than a lazy chain, so that we should strive to achieve as high an acceptance proportion as possible.

All those comments apply to Bayesian phylogenetic MCMC algorithms, which include both within-tree moves that change the branch lengths and substitution parameters without changing the tree topology and cross-tree moves that change the tree topology. The cross-tree moves are typically constructed using tree-perturbation (branch-swapping) algorithms such as nearest-neighbour interchange (NNI), subtree pruning and re-grafting (SPR) and tree bisection and reconnection (TBR)[26,78]. About a dozen MCMC phylogenetic programs are now available (Table 1).

What are convergence, burn-in and mixing of the MCMC?

An MCMC algorithm may suffer from two problems: slow convergence and poor mixing. In the long run, the Markov chain should be spending most of the time visiting high-probability regions of the posterior. The *convergence rate* is the rate at which a chain starting from any initial position (which may be in the tails of the posterior) moves to the high-posterior region of the parameter space[79]. Parameter values sampled before reaching this stationary phase are usually discarded as the *burn-in*. Thus, if convergence is slow, a long burn-in will be necessary. Convergence rate is affected by the proposals used and by the shape of the posterior in the tails[67]. If the posterior is nearly flat in the tail, it will be difficult for the chain to get out of the tail and move to the high-posterior region.

Mixing efficiency refers to how efficiently the chain traverses the posterior after it has reached the stationary distribution. If the chain is more efficient, the estimate based on the MCMC sample will have a smaller variance, and the results will show less variation among independent runs (Box 2) and a relatively short chain will provide acceptable estimate. The proposal (such as the uniform sliding window *vs*. the normal-distribution sliding window) as well as the step length for the same proposal (such as the width of the sliding window) can have a great effect on mixing efficiency[76].

Both convergence and mixing problems can be diagnosed by using a trace plot, in which we plot the log likelihood or sampled parameter values against the MCMC iteration, for example, using R[80] or Tracer[81]. It is also very important to run the same algorithm multiple times to check consistency between runs. With fast convergence, different chains that started from very different positions become indistinguishable very quickly. Efficient mixing is indicated by different runs generated nearly identical means, standard deviations, and histograms. If the runs are healthy, samples from different runs can be combined to produce posterior summaries.

The trace plots of Figures 2a and 2b are from an efficient chain with good mixing, while those of Figures 2a' and 2b' have poor mixing and low efficiency. The histograms from the efficient algorithm match each other much better than those from the inefficient algorithm (Fig. 2c and 2c'). In theory, the consistency among multiple runs could be because all runs got stuck in a region of the parameter space, giving the false impression that convergence was reached. This may happen when there are multiple peaks in the posterior. Thus, it is important to initiate the runs from widely dispersed starting points.

How many iterations should I run my chain for? How many samples should I take?

Ideally one would like to run the MCMC long enough to obtain a reliable estimation of the posterior distribution, but not overly too long as to waste computational resources. However, currently reliable automatic stopping rules do not exist. As a result, the user has to specify the number of iterations, and then decide whether the chain is long enough or additional iterations are necessary using certain diagnosis tools. MCMC algorithms tend to generate huge output files. To save disk space, one takes a sample only for every certain number of iterations. For example, running an MCMC chain for 10^7 iterations and using a sample frequency of 10^3 iterations will produce 10^4 samples.

Note that in some programs (such as MCMCtree and BPP), each MCMC iteration consists of a fixed sequence of MCMC proposals, while in some others (such as MrBayes and BEAST), it consists of one proposal, chosen at random from a collection of proposals. Thus, if there are 1,000 parameters in the model and if each proposal changes one parameter, each MCMC iteration in the former programs is worth about 1,000 iterations in the latter programs. Thus, MCMC iterations from different programs are not comparable. The biologist should instead aim to accumulate a reasonable (as large as practically possible) effective sample size (ESS) for each parameter (Box 2).

Why should an MCMC analysis be run with an "empty alignment"? Is the data informative?

It is useful to run the MCMC algorithm sampling from the prior. This is achieved by setting the likelihood to 1 in equation (1). Some programs generate a dummy "empty" alignment that can be used to achieve the same effect. Runs should also be assessed for good convergence and mixing. Running the chain without data is a good way of checking the correctness of the software, because the mean, variance, etc. of the prior are often analytically available and can be checked against the MCMC sample. In molecular clock dating using fossil calibrations, the prior on divergence times incorporates the calibration information and is typically intractable. Running the program without using the sequences allows one to generate the prior used by the program.

The sample from the prior can also be compared with the sample from the posterior (which is generated by using the data) to assess how informative the data are, and whether there are serious conflicts between the prior and the data. High similarity between the prior and the posterior suggests that the data contain little information about the parameters. Considerable overlap between the prior and posterior but with the posterior being much more concentrated than the prior means that the data are informative and the prior is reasonable. In contrast, if the prior and posterior do not overlap well, there may be a conflict between the prior and the data, possibly caused by misspecified priors. One can also modify the prior to assess the impact of the prior on the posterior. Note, however, that it is incorrect to specify the prior by trying to match the posterior, since the prior is supposed to reflect our knowledge before the analysis of the data.

Conclusions

Bayesian phylogenetics has undergone explosive growth during the past decade. The implementation of sophisticated models in easy-to-use software programs has made the method extremely appealing to biologists. The method is especially powerful in combining different sources of information in an integrated data analysis. As a result, Bayesian MCMC methods are the most commonly used framework for development of new models of data analysis, especially in the areas of divergence time estimation integrating molecular, morphological and fossil information[82], species tree estimation using multi-locus genomic sequence data[24], and species delimitation incorporating genetic and morphological/ecological information[83]. The potential of the Bayesian method to deal with these and future questions has never been greater. For further reading on the Bayesian method and Bayesian phylogenetics the reader may consult[26,84,85]

A tutorial that helps the user to write a simple R program to conduct phylogenetic MCMC to reproduce the figures of this paper is available at: http://github.com/thednainus/Bayesian_tutorial

Author contributions

FFN conceived the idea, FFN, MdR and ZY wrote the paper.

Correspondence should be addressed to FFN or ZY

Acknowledgements

This work was supported by Biotechnology and Biological Sciences Research Council (UK) grant BB/N000609/1. FFN was supported by a Royal Society and British Academy Newton International Fellowship (UK) grant number NF140338.

Table 1. List of Bayesian programs

Program	Brief description	Refs
BEAST	Implements a vast number of models. Examples are simultaneous estimation of the tree topology and divergence times, phylodynamics, phylogeography, and species tree estimation under the multispecies coalescent model.	86
MrBayes	Implements a large number of models for analysis of nucleotide, amino acid, and morphological data. Estimates species phylogenies and species divergence times.	87
RevBayes	Similar to MrBayes, but with its own programming language to set up complex hierarchical Bayesian models.	88
MCMCTree	Estimates divergence times on a fixed phylogenetic tree.	89
Phycas	Estimates phylogenetic trees based on nucleotide data. This allows for multifurcating trees, helping to reduce spuriously high posterior probabilities for phylogenies.	90,91
PhyloBayes	Reconstructs phylogenetic trees using infinite mixture models to account for among-site and among-lineage heterogeneity in nucleotide or amino acid compositions, which may be important for inferring deep phylogenies.	92
BPP	Implements species tree estimation and species delimitation under the multi-species coalescent model using multi-loci genomic sequence data.	56
Migrate	Estimates population sizes and migration rates under the population-subdivision model based on molecular data.	93
IMa2	Estimates divergence times, population sizes and migration rates under the isolation-with-migration model using multi-loci DNA sequence data and a fixed phylogenetic tree for populations.	94
Structure	Estimates population structure from multi-locus genotype data.	95
BAMM	Estimates clade diversification rates on phylogenies.	96
Tracer	A program for MCMC diagnostics and summaries.	81
AWTY	A package for MCMC diagnostics for Bayesian phylogenetic inference.	97

Box 1. MCMC algorithm to estimate *d* and κ under the K80

- 1. (Initialization): Initialize window seizes w_d and w_k. Choose random starting values (d, κ).

- 2. (Main loop)

  - 2a (Proposal to change distance d): Propose a new value d^* by sampling from a uniform sliding window (with reflection) around the current value: $d^* = \mathrm{U}(d - w_d/2, d + w_d/2)$, where w_d is the width of the window. If $d^* < 0$, set $d^* = -d^*$ (reflection). If the unnormalised posterior is higher at the new value, accept the proposal. Otherwise accept with probability equal to the ratio of the posteriors:

$$\alpha = \frac{f(d^*, \kappa \mid D)}{f(d, \kappa \mid D)} = \frac{f(d^*) f(\kappa) f(D \mid d^*, \kappa)}{f(d) f(\kappa) f(D \mid d, \kappa)} \tag{6}$$

    If the proposal is accepted, set $d = d^*$. If it is rejected, stay where it is ($d = d$).

    2b (Proposal to change κ): Use a similar sliding window of width w_k to propose a new value $\kappa^* = \mathrm{U}(\kappa - w_k/2, \kappa + w_k/2)$. If $\kappa^* < 0$, reflect by setting $\kappa^* = -\kappa^*$. Accept the proposal with probability $\min\{1, \alpha\}$, where

$$\alpha = \frac{f(d, \kappa^* \mid D)}{f(d, \kappa \mid D)} = \frac{f(d) f(\kappa^*) f(D \mid d, \kappa^*)}{f(d) f(\kappa) f(D \mid d, \kappa)} \tag{7}$$

    If the proposal is accepted, set $\kappa = \kappa^*$. Otherwise stay where it is ($\kappa = \kappa$).

  - 2c (Save the state of the chain): Print out d and κ. Go back to 2a and iterate to obtain as many samples as desired.

Box 2. Efficiency of the MCMC and the effective sample size (ESS)

Parameter values sampled during the MCMC are autocorrelated because the current value is either the same as the previous value (if the proposed value is rejected) or a modification of it (e.g., a value sampled from the sliding window around the current value). Stronger autocorrelations mean that the Markov chain is less efficient in traversing the posterior space. More formally, we use the mean of the MCMC sample ($\tilde{x}$) to estimate the posterior mean of any parameter. This has the variance

$$v_{\mathrm{MCMC}} = v_{\mathrm{IND}} \times [1 + 2(\rho_1 + \rho_2 + \dots)], \tag{8}$$

where v_{IND} is the variance for an independent sample of the same size from the posterior distribution, and where $\rho_k = \mathrm{corr}(x_t, x_{t+k})$ is the correlation between the values of the parameter in the MCMC sample that are k iterations apart, known as the lag k autocorrelation. Both the independent-sample variance v_{IND} and the MCMC-sample variance v_{MCMC} are typically proportional to $1/n$, with n to be the sample size. The efficiency of an MCMC chain is defined as the variance ratio

$$\mathrm{Eff} = \frac{v_{\mathrm{IND}}}{v_{\mathrm{MCMC}}} = \frac{1}{1 + 2(\rho_1 + \rho_2 \dots)}. \tag{9}$$

For example, an Eff = 0.25 means that an MCMC sample of size n is as efficient as an independent sample of size $n/4$, so that we need to generate an MCMC sample four times as large as the independent sample to have the same variance. The effective sample size, ESS, is simply

$$\mathrm{ESS} = n \times \mathrm{Eff}.$$

As a rule of thumb, one should aim for ESS = 1,000 or 10,000 [98]. Bayesian phylogenetic algorithms are computationally intensive, so that ESS = 200 is commonly recommended, but this may be too small for calculation of the 95% or 99% credibility intervals. A good strategy may be to conduct multiple runs of the same analysis, and then combine the samples before producing the posterior summary. If ESS = 200 for each sample, 10 replicate runs will give a combined sample of ESS = 2000.

Figure 1 | Prior, likelihood and posterior distribution for a two-parameter phylogenetic example. The data of the 12s RNA mitochondrial genes from human and orang-utan are used to estimate of the evolutionary distance (*d*) and the transition/transversion ratio (κ) model[75].

Figure 2 | Trace plots and histograms for parameters *d* and *κ* sampling the posterior distribution of Figure 1c using efficient and inefficient MCMC chains. Parts **a** and **b** show the trace plots of *d* and κ for an efficient chain with good mixing. The window sizes are $w_d = 0.12$ and $w_\kappa = 180$, with acceptance proportions $P_{jump} = 30.4\%$ for *d* and 29.8% for κ, achieving efficiency Eff = 23% for *d* and 20% for κ. Parts **a'** and **b'** show the trace plots for an inefficient chain with poor mixing, with $w_d = 5$ and $w_\kappa = 1$. In **a'**, the window for *d* is too wide, and most proposals are rejected ($P_{jump} = 1.5\%$), so that the chain is often stuck at the same value for many iterations, leading to poor mixing with Eff = 1.79%. In **b'**, the window for κ is too small, so that most of the proposals are accepted (with $P_{jump} = 98.6\%$), but the chain makes small baby steps and is very slow in traversing the posterior parameter space, with Eff = 1.28%. Parts **c** and **c'** show histograms of κ for two runs of the efficient and inefficient chains (sample size $n = 10{,}000$). The posterior mean (and standard deviation) calculated using a very long run of the efficient chain is 0.104 (0.0114) for *d*, and 29.2 (10.0) for κ.

References

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

2 Mau, B. & Newton, M. A. Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *J. Comp. Graph. Stat.* **6**, 122-131 (1997).

3 Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. Bayesian inference of phylogeny and its impact on evolutionary biology. *Science* **294**, 2310-2314 (2001).

4 Wilfert, L. *et al.* Deformed wing virus is a recent global epidemic in honeybees driven by *Varroa* mites. *Science* **351**, 594-597 (2016).

5 Pybus, O. G. *et al.* Unifying the spatial epidemiology and molecular evolution of emerging epidemics. *Proc. Natl. Acad. Sci. USA* **109**, 15066-15071 (2012).

6 Faria, N. R. *et al.* HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56-61 (2014).

7 Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. Phylogeography takes a relaxed random walk in continuous space and time. *Mol. Biol. Evol.* **27**, 1877-1885 (2010).

8 Bloomquist, E. W., Lemey, P. & Suchard, M. A. Three roads diverged? Routes to phylogeographic inference. *Trends Ecol. Evol.* **25**, 626-632 (2010).

9 Nascimento, F. F. *et al.* The role of historical barriers in the diversification processes in open vegetation formations during the Miocene/Pliocene using an ancient rodent lineage as a model. *PLoS One* **8**, e61924 (2013).

10 Werneck, F. P., Leite, R. N., Geurgas, S. R. & Rodrigues, M. T. Biogeographic history and cryptic diversity of saxicolous Tropiduridae lizards endemic to the semiarid Caatinga. *BMC Evol. Biol.* **15**, 94 (2015).

11 Merckx, V. S. F. T. *et al.* Evolution of endemism on a young tropical mountain. *Nature* **524**, 347-350 (2015).

12 Hoorn, C. *et al.* Amazonia through time: Andean uplift, climate change, landscape evolution, and biodiversity. *Science* **330**, 927-931 (2010).

13 Prum, R. O. *et al.* A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**, 569-573 (2015).

14 dos Reis, M. *et al.* Uncertainty in the timing of origin of animals and the limits of precision in molecular timescales. *Curr. Biol.* **25**, 2939-2950 (2015).

15 Meredith, R. W. *et al.* Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521-524 (2011).

16 Nascimento, F. F. *et al.* Evolution of endogenous retroviruses in the Suidae: evidence for different viral subpopulations in African and Eurasian host species. *BMC Evol. Biol.* **11**, 139 (2011).

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

18 Misof, B. *et al.* Phylogenomics resolves the timing and pattern of insect evolution. *Science* **346**, 763-767 (2014).

19 Raymann, K., Brochier-Armanet, C. & Gribaldo, S. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. USA* **112**, 6670-6675 (2015).

20 Foley, N. M., Springer, M. S. & Teeling, E. C. Mammal madness: is the mammal tree of life not yet resolved? *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **371** (2016).

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

22 Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109 (1970).

23 Liu, L., Xi, Z., Wu, S., Davis, C. C. & Edwards, S. V. Estimating phylogenetic trees from genome-scale data. *Ann. NY. Acad. Sci.* **1360**, 36-53 (2015).

24 Xu, B. & Yang, Z. Challenges in species tree estimation under the multispecies coalescent model. *Genetics* **204**, 1353-1368 (2016).

25 Szöllosi, G. J., Tannier, E., Daubin, V. & Boussau, B. The inference of gene trees with species trees. *Syst. Biol.* **64**, e42-e62 (2015).

26 Yang, Z. *Molecular Evolution: A statistical Approach*. (Oxford Univ. Press, 2014).

27 Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913-925 (2001).

28 Redelings, B. D. & Suchard, M. A. Joint Bayesian estimation of alignment and phylogeny. *Syst. Biol.* **54**, 401-418 (2005).

29 Löytynoja, A. & Goldman, N. Uniting alignments and trees. *Science* **324**, 1528-1529 (2009).

30 Chatzou, M. *et al.* Multiple sequence alignment modeling: methods and applications. *Brief. Bioinform.* **17**, 1009-1023 (2016).

31 Altenhoff, A. M. & Dessimoz, C. Inferring orthology and paralogy. *Methods Mol. Biol.* **855**, 259-279 (2012).

32 Altenhoff, A. M. *et al.* The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* **43**, D240-D249 (2015).

33 Dimmic, M. in *Statistical Methods in Molecular Evolution* (ed R. Nielsen) (Springer-Verlag, 2005).

34 Liò, P. & Goldman, N. Models of molecular evolution and phylogeny. *Genome Res.* **8**, 1233-1244 (1998).

35 Jukes, T. H. & Cantor, C. R. in *Mammalian Protein Metabolism* (ed H.N. Munro) 21-132 (Academic Press, 1969).

36 Tavaré, S. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* **17**, 57-86 (1986).

37 Yang, Z. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**, 105-111 (1994).

38 Zharkikh, A. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**, 315-329 (1994).

39 Mayrose, I., Graur, D., Ben-Tal, N. & Pupko, T. Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. *Mol. Biol. Evol.* **21**, 1781-1791 (2004).

40 Yang, Z., Lauder, I. J. & Lin, H. J. Molecular evolution of the hepatitis B virus genome. *J. Mol. Evol.* **41**, 587-596 (1995).

41 Yang, Z. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol. Evol.* **11**, 367-372 (1996).

42 Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).

43 Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McLnerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol. Biol.* **6**, 29 (2006).

44 Lanfear, R., Calcott, B., Ho, S. Y. & Guindon, S. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695-1701 (2012).

45 Hoff, M., Orf, S., Riehm, B., Darriba, D. & Stamatakis, A. Does the choice of nucleotide substitution models matter topologically? *BMC Bioinformatics* **17**, 143 (2016).

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

47 Wright, A. M., Lloyd, G. T. & Hillis, D. M. Modeling character change heterogeneity in phylogenetic analyses of morphology through the use of priors. *Syst. Biol.* **65**, 602-611 (2016).

48 Felsenstein, J. Maximum-likelihood estimation of evolutionary trees from continuous characters. *Am. J. Hum. Genet.* **25**, 471-492 (1973).

49 Felsenstein, J. *Inferring Phylogenies*. (Sinauer Associates Sunderland, 2004).

50 Ronquist, F. *et al.* A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* **61**, 973-999 (2012).

51 Heath, T. A., Huelsenbeck, J. P. & Stadler, T. The fossilized birth-death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. USA* **111**, E2957-2966 (2014).

52 O'Reilly, J. E., dos Reis, M. & Donoghue, P. C. Dating tips for divergence-time estimation. *Trends Genet.* **31**, 637-650 (2015).

53 Rannala, B. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* **51**, 754-760 (2002).

54 Gu, X., Fu, Y. X. & Li, W. H. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12**, 546-557 (1995).

55 Sullivan, J., Swofford, D. L. & Naylor, G. J. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. *Mol. Biol. Evol.* **16**, 1347-1356 (1999).

56 Yang, Z. The BPP program for species tree estimation and species delimitation. *Curr. Zoo.* **61**, 854-865 (2015).

57 Shapiro, B., Rambaut, A. & Drummond, A. J. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Mol Biol Evol* **23**, 7-9 (2006).

58 Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212-226 (2006).

59 Nylander, J. A., Ronquist, F., Huelsenbeck, J. P. & Nieves-Aldrey, J. L. Bayesian phylogenetic analysis of combined data. *Syst. Biol.* **53**, 47-67 (2004).

60 Maddison, W. P. Gene trees in species trees. *Syst. Biol.* **46**, 523-536 (1997).

61 Nichols, R. Gene trees and species tree are not the same. *Trends Ecol. Evol.* **16**, 358-364 (2001).

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

63 Edwards, S. V. *et al.* Implementing and testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Mol. Phylogenet. Evol.* **94**, 447-462 (2016).

64 Vijaykrishna, D., Mukerji, R. & Smith, G. J. D. RNA virus reassortment: an evolutionary mechanism for host jumps and immune evasion. *PLoS Pathog.* **11**, e1004902 (2015).

65 Ronquist, F., van der Mark, P. & Huelsenbeck, J. P. in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (eds P. Lemey, M. Salemi, & A.M. Vandamme) (Cambridge Univ. Press, 2009).

66 Brown, J. M., Hedtke, S. M., Lemmon, A. R. & Lemmon, E. M. When trees grow too long: investigating the causes of highly inaccurate bayesian branch-length estimates. *Syst. Biol.* **59**, 145-161 (2010).

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

68 dos Reis, M., Zhu, T. & Yang, Z. The impact of the rate prior on Bayesian estimation of divergence times with multiple loci. *Syst. Biol.* **63**, 555-565 (2014).

69 Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4**, e88 (2006).

70 Yang, Z. & Rannala, B. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**, 717-724 (1997).

71 Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645-1656 (2003).

72 Ho, S. Y. & Phillips, M. J. Accounting for calibration uncertainty in phylogenetic estimation of evolutionary divergence times. *Syst. Biol.* **58**, 367-380 (2009).

73 Thorne, J. L., Kishino, H. & Painter, I. S. Estimating the rate of evolution of the rate of molecular evolution. *Mol. Biol. Evol.* **15**, 1647-1657 (1998).

74 Rannala, B. & Yang, Z. Inferring speciation times under an episodic molecular clock. *Syst. Biol.* **56**, 453-466 (2007).

75 Kimura, M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**, 111-120 (1980).

76 Yang, Z. & Rodriguez, C. E. Searching for efficient Markov chain Monte Carlo proposal kernels. *Proc. Natl. Acad. Sci. USA* **110**, 19307-19312 (2013).

77 Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711-732 (1995).

78 Lakner, C., van der Mark, P., Huelsenbeck, J. P., Larget, B. & Ronquist, F. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. Biol.* **57**, 86-103 (2008).

79 Green, P. J. & Han, X. L. in *Stochastic Models, Statistical Methods, and Algorithms in Image Analysis* (eds P. Barone, A. Frigessi, & M. Piccioni) (Springer, 1992).

80 R Core Team. *R: A language and environment for statistical computing*, <http://www.r-project.org/> (2015).

81 Rambaut, A., Suchard, M. A., Xie, D. & Drummond, A. J. *Tracer v1.6*, <http://beast.bio.ed.ac.uk/Tracer> (2014).

82 dos Reis, M., Donoghue, P. C. & Yang, Z. Bayesian molecular clock dating of species divergences in the genomics era. *Nat. Rev. Genet.* **17**, 71-80 (2016).

83 Solís-Lemus, C., Knowles, L. L. & Ané, C. Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* **69**, 492-507 (2015).

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

85 Gelman, A. *et al. Bayesian Data Analysis*. (Chapman and Hall/CRC, 2013).

86 Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

87 Ronquist, F. *et al.* MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539-542 (2012).

88 Höhna, S. *et al.* RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.* **65**, 726-736 (2016).

89 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586-1591 (2007).

90 Lewis, P. O., Holder, M. T. & Swofford, D. L. Phycas: software for Bayesian phylogenetic analysis. *Syst. Biol.* **64**, 525-531 (2015).

91 Lewis, P. O., Holder, M. T. & Holsinger, K. E. Polytomies and Bayesian phylogenetic inference. *Syst. Biol.* **54**, 241-253 (2005).

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

93 Beerli, P. Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. *Bioinformatics* **22**, 341-345 (2006).

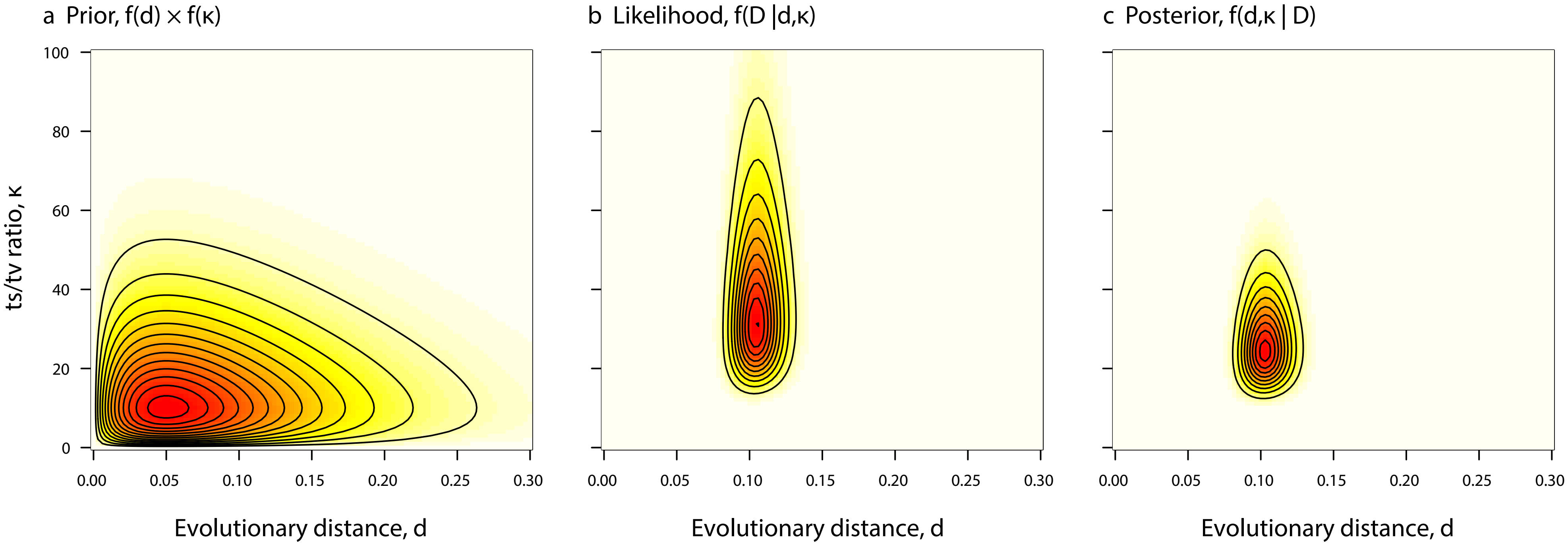
94 Hey, J. & Nielsen, R. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proc. Natl. Acad. Sci. USA* **104**, 2785-2790 (2007).

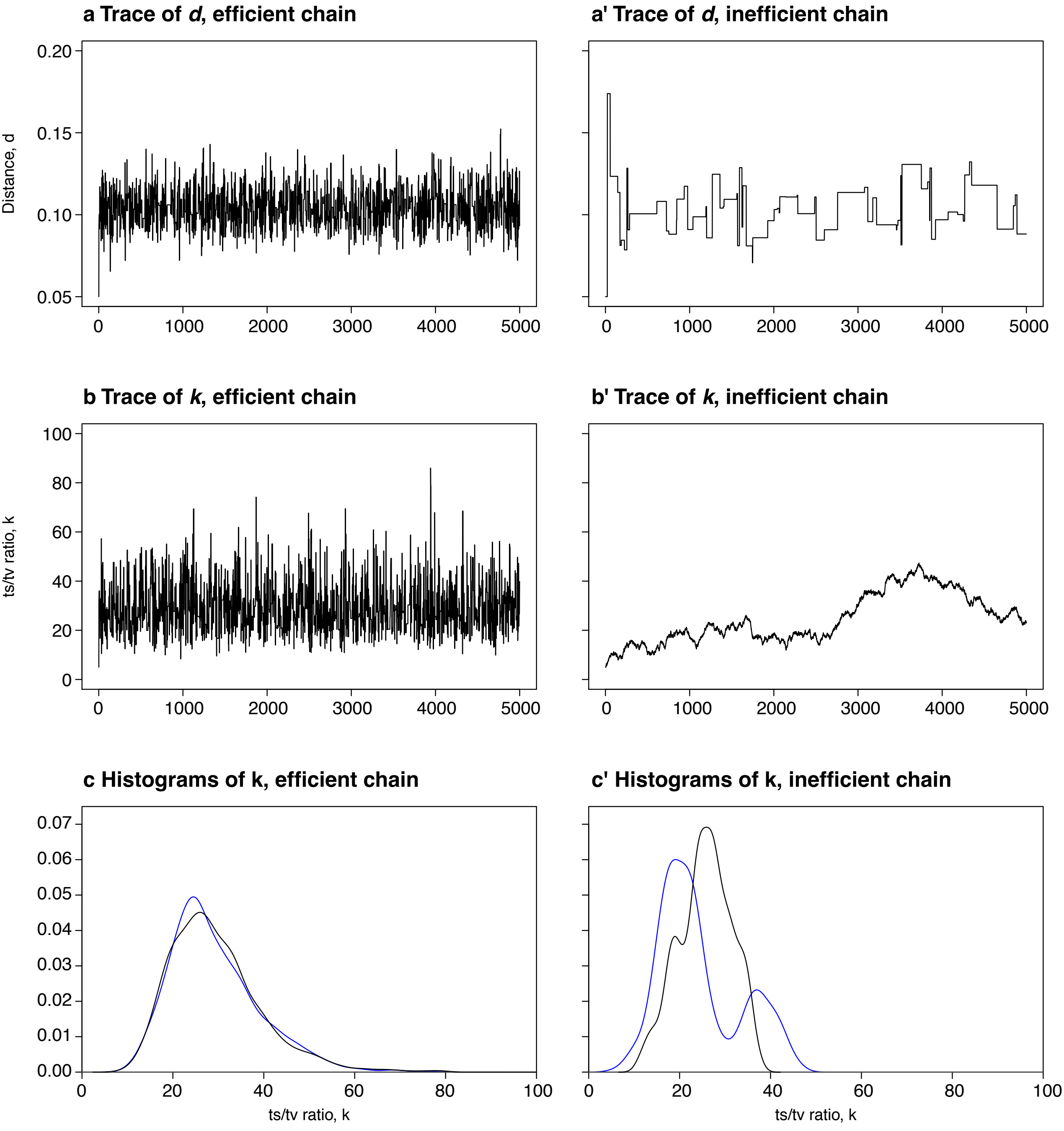
95 Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945-959 (2000).

96 Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS One* **9**, e89543 (2014).

97 Nylander, J. A., Wilgenbusch, J. C., Warren, D. L. & Swofford, D. L. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. *Bioinformatics* **24**, 581-583 (2008).

98 Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap*. (Chapman & Hall/CRC, 1994).


a Prior, f(d) × f(κ)
b Likelihood, f(D |d,κ)
c Posterior, f(d,κ | D)
ts/tv ratio, κ
100
80
60
40
20
0
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Evolutionary distance, d
Evolutionary distance, d
Evolutionary distance, d



a Trace of *d*, efficient chain

a' Trace of *d*, inefficient chain

b Trace of *k*, efficient chain

b' Trace of *k*, inefficient chain

c Histograms of k, efficient chain

c' Histograms of k, inefficient chain