7

Bayesian phylogenetic analysis using MrBayes

THEORY

Fredrik Ronquist, Paul van der Mark, and John P. Huelsenbeck

7.1 Introduction

What is the probability that Sweden will win next year's world championships in ice hockey? If you're a hockey fan, you probably already have a good idea, but even if you couldn't care less about the game, a quick perusal of the world championship medalists for the last 15 years (Table 7.1) would allow you to make an educated guess. Clearly, Sweden is one of only a small number of teams that compete successfully for the medals. Let's assume that all seven medalists the last 15 years have the same chance of winning, and that the probability of an outsider winning is negligible. Then the odds of Sweden winning would be 1:7 or 0.14. We can also calculate the frequency of Swedish victories in the past. Two gold medals in 15 years would give us the number 2:15 or 0.13, very close to the previous estimate. The exact probability is difficult to determine but most people would probably agree that it is likely to be in the vicinity of these estimates.

You can use this information to make sensible decisions. If somebody offered you to bet on Sweden winning the world championships at the odds 1:10, for instance, you might not be interested because the return on the bet would be close to your estimate of the probability. However, if you were offered the odds 1:100, you might be tempted to go for it, wouldn't you?

As the available information changes, you are likely to change your assessment of the probabilities. Let's assume, for instance, that the Swedish team made it to

The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing, Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme (eds.). Published by Cambridge University Press.

Table 7.1 Medalists in the ice hockey world championships 1993–2007

Year	Gold	Silver	Bronze
1993	Russia	Sweden	Czech Republic
1994	Canada	Finland	Sweden
1995	Finland	Sweden	Canada
1996	Czech Republic	Canada	United States
1997	Canada	Sweden	Czech Republic
1998	Sweden	Finland	Czech Republic
1999	Czech Republic	Finland	Sweden
2000	Czech Republic	Slovakia	Finland
2001	Czech Republic	Finland	Sweden
2002	Slovakia	Russia	Sweden
2003	Canada	Sweden	Slovakia
2004	Canada	Sweden	United States
2005	Czech Republic	Canada	Russia
2006	Sweden	Czech Republic	Finland
2007	Canada	Finland	Russia

the finals. Now you would probably consider the chance of a Swedish victory to be much higher than your initial guess, perhaps close to 0.5. If Sweden lost in the semifinals, however, the chance of a Swedish victory would be gone; the probability would be 0.

This way of reasoning about probabilities and updating them as new information becomes available is intuitively appealing to most people and it is clearly related to rational behavior. It also happens to exemplify the Bayesian approach to science. Bayesian inference is just a mathematical formalization of a decision process that most of us use without reflecting on it; it is nothing more than a probability analysis. In that sense, Bayesian inference is much simpler than classical statistical methods, which rely on sampling theory, asymptotic behavior, statistical significance, and other esoteric concepts.

The first mathematical formulation of the Bayesian approach is attributed to Thomas Bayes (*c.* 1702–1761), a British mathematician and Presbyterian minister. He studied logic and theology at the University of Edinburgh; as a Non-Conformist, Oxford and Cambridge were closed to him. The only scientific work he published during his lifetime was a defense of Isaac Newton's calculus against a contemporaneous critic (*Introduction to the Doctrine of Fluxions*, published anonymously in 1736), which apparently got him elected as a Fellow of the Royal Society in 1742. However, it is his solution to a problem in so-called inverse probability that made him famous. It was published posthumously in 1764 by his friend Richard Price in the *Essay Towards Solving a Problem in the Doctrine of Chances.*

Assume we have an urn with a large number of balls, some of which are white and some of which are black. Given that we know the proportion of white balls, what is the probability of drawing, say, five white and five black balls in ten draws? This is a problem in forward probability. Thomas Bayes solved an example of the converse of such problems. Given a particular sample of white and black balls, what can we say about the proportion of white balls in the urn? This is the type of question we need to answer in Bayesian inference.

Let's assume that the proportion of white balls in the urn is p. The probability of drawing a white ball is then p and the probability of drawing a black ball is $1 - p$. The probability of obtaining, say, two white balls and one black ball in three draws would be

$$\Pr(2\text{white}, 1\text{black}|p) = p \times p \times (1 - p) \times \binom{3}{2} \tag{7.1}$$

The vertical bar indicates a condition; in this case we are interested in the probability of a particular outcome given (or conditional) on a particular value of p. It is easy to forget the last factor (3 choose 2), which is the number of ways in which we can obtain the given outcome. Two white balls and one black ball can be the result of drawing the black ball in the first, second or third draw. That is, there are three ways of obtaining the outcome of interest, 3 choose 2 (or 3 choose 1 if we focus on the choice of the black ball; the result is the same). Generally, the probability of obtaining a white balls and b black balls is determined by the function

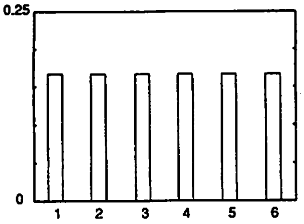
$$f(a, b|p) = p^a(1 - p)^b\binom{a+b}{a} \tag{7.2}$$

which is the ***probability mass function*** (Box 7.1) of the so-called binomial distribution. This is the solution to the problem in forward probability, when we know the value of p. Bayesians often, somewhat inappropriately, refer to the forward probability function as the ***likelihood function***.

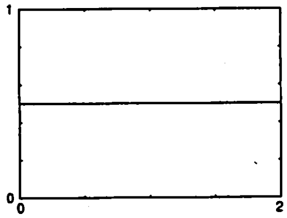
But given that we have a sample of a white balls and b black balls, what is the probability of a particular value of p? This is the reverse probability problem, where we are trying to find the function $f(p|a, b)$ instead of the function $f(a, b|p)$. It turns out that it is impossible to derive this function without specifying our prior beliefs about the value of p. This is done in the form of a probability distribution on the possible values of p (Box 7.1), the ***prior probability distribution*** or just ***prior*** in everyday Bayesian jargon. If there is no previous information about the value of p, we might associate all possible values with the same probability, a so-called uniform probability distribution (Box 7.1).

Box 7.1 Probability distributions

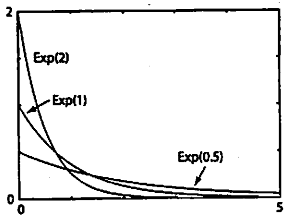
A function describing the probability of a discrete random variable is called a ***probability mass function.*** For instance, this is the probability mass function for throwing a dice, an example of a *discrete uniform distribution*:



For a continuous variable, the equivalent function is a ***probability density function.*** The value of this function is not a probability, so it can sometimes be larger than one. Probabilities are obtained by integrating the density function over a specified interval, giving the probability of obtaining a value in that interval. For instance, a *continuous uniform distribution* on the interval (0,2) has this probability density function:



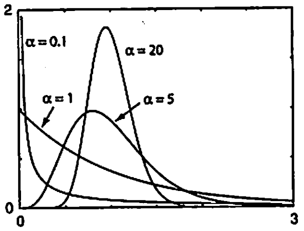
Most prior probability distributions used in Bayesian phylogenetics are uniform, exponential, gamma, beta or Dirichlet distributions. Uniform distributions are often used to express the lack of prior information for parameters that have a uniform effect on the likelihood in the absence of data. For instance, the discrete uniform distribution is typically used for the topology parameter. In contrast, the likelihood is a negative exponential function of the branch lengths, and therefore the *exponential distribution* is a better choice for a *vague* prior on branch lengths. The exponential distribution has the density function $f(x) = \lambda e^{-\lambda x}$, where λ is known as the *rate* parameter. The expectation (mean) of the exponential distribution is $1/\lambda$.



The *gamma* distribution has two parameters, the shape parameter α and the scale parameter β. At small values of α, the distribution is L-shaped and the variance is large;

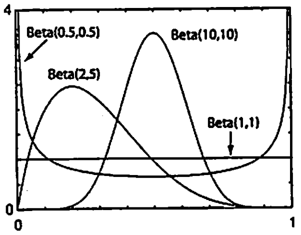
Box 7.1 *(cont.)*

at high values it is similar to a normal distribution and the variance is low. If there is considerable uncertainty concerning the shape of the prior probability distribution, the gamma may be a good choice; an example is the rate variation across sites. In these cases, the value of α can be associated with a uniform or an exponential prior (also known as a ***hyperprior*** since it is a prior on a parameter of a prior), so that the MCMC procedure can explore different shapes of the gamma distribution and weight each according to its posterior probability. The sum of exponentially distributed variables is also a gamma distribution. Therefore, the gamma is an appropriate choice for the prior on the tree height of clock trees, which is the sum of several presumably exponentially distributed branch lengths.



The *beta* and *Dirichlet* distributions are used for parameters describing proportions of a whole, so called simplex parameters. Examples include the stationary state frequencies that appear in the instantaneous rate matrix of the substitution model. The exchangeability or rate parameters of the substitution model can also be understood as proportions of the total exchange rate (given the stationary state frequencies). Another example is the proportion of invariable and variable sites in the invariable sites model. The beta distribution, denoted Beta(α_1, α_2), describes the probability on two proportions, which are associated with the weight parameters $\alpha_1 > 0$ and $\alpha_2 > 0$. The Dirichlet distribution is equivalent except that there are more than two proportions and associated weight parameters.

A Beta(1, 1) distribution, also known as a flat beta, is equivalent to a uniform distribution on the interval (0,1). When $\alpha_1 = \alpha_2 > 1$, the distribution is symmetric and emphasizes equal proportions, the more so the higher the weights. When $\alpha_1 = \alpha_2 < 1$, the distribution puts more probability on extreme proportions than on equal proportions. Finally, if the weights are different, the beta is skewed towards the proportion defined by the weights; the expectation of the beta is $\alpha/(\alpha + \beta)$ and the mode is $(\alpha - 1)/(\alpha + \beta - 2)$ for $\alpha > 1$ and $\beta > 1$.



Box 7.1 *(cont.)*

Assume that we toss a coin to determine the probability p of obtaining heads. If we associate p and $1 - p$ with a flat beta prior, we can show that the posterior is a beta distribution where $\alpha_1 - 1$ is the number of heads and $\alpha_2 - 1$ is the number of tails. Thus, the weights roughly correspond to counts. If we started with a flat Dirichlet distribution and analyzed a set of DNA sequences with the composition 40 A, 50 C, 30 G, and 60 T, we might expect a posterior for the stationary state frequencies around Dirichlet(41, 51, 31, 61) if it were not for the other parameters in the model and the blurring effect resulting from looking back in time. Wikipedia (http://www.wikipedia.org) is an excellent source for additional information on common statistical distributions.

Thomas Bayes realized that the probability of a particular value of p, given some sample (a, b) of white and black balls, can be obtained using the probability function

$$f(p|a, b) = \frac{f(p)f(a, b|p)}{f(a, b)} \tag{7.3}$$

This is known as Bayes' theorem or Bayes' rule. The function $f(p|a, b)$ is called the ***posterior probability distribution***, or simply the ***posterior***, because it specifies the probability of all values of p after the prior has been updated with the available data.

We saw above how we can calculate $f(a, b|p)$, and how we can specify $f(p)$. How do we calculate the probability $f(a, b)$? This is the unconditional probability of obtaining the outcome (a, b) so it must take all possible values of p into account. The solution is to integrate over all possible values of p, weighting each value according to its prior probability:

$$f(a, b) = \int_0^1 f(p)f(a, b|p)\,\mathrm{d}p \tag{7.4}$$

We can now see that the denominator is a normalizing constant. It simply ensures that the posterior probability distribution integrates to 1, the basic requirement of a proper probability distribution.

A Bayesian problem that occupied several early workers was an analog to the following. Given a particular sample of balls, what is the probability that p is larger than a specified value? To solve it analytically, they needed to deal with complex integrals. Bayes made some progress in his *Essay*; more important contributions were made later by Laplace, who, among other things, used Bayesian reasoning and novel integration methods to show beyond any reasonable doubt that the probability of a newborn being a boy is higher than 0.5. However, the analytical complexity of most Bayesian problems remained a serious problem for a long time and it is only in the last few decades that the approach has become popular due to the combination of efficient numerical methods and the widespread availability of fast computers.

7.2 Bayesian phylogenetic inference

How does Bayesian reasoning apply to phylogenetic inference? Assume we are interested in the relationships between man, gorilla, and chimpanzee. In the standard case, we need an additional species to root the tree, and the orangutan would be appropriate here. There are three possible ways of arranging these species in a phylogenetic tree: the chimpanzee is our closest relative, the gorilla is our closest relative, or the chimpanzee and the gorilla are each other's closest relatives (Fig. 7.1).

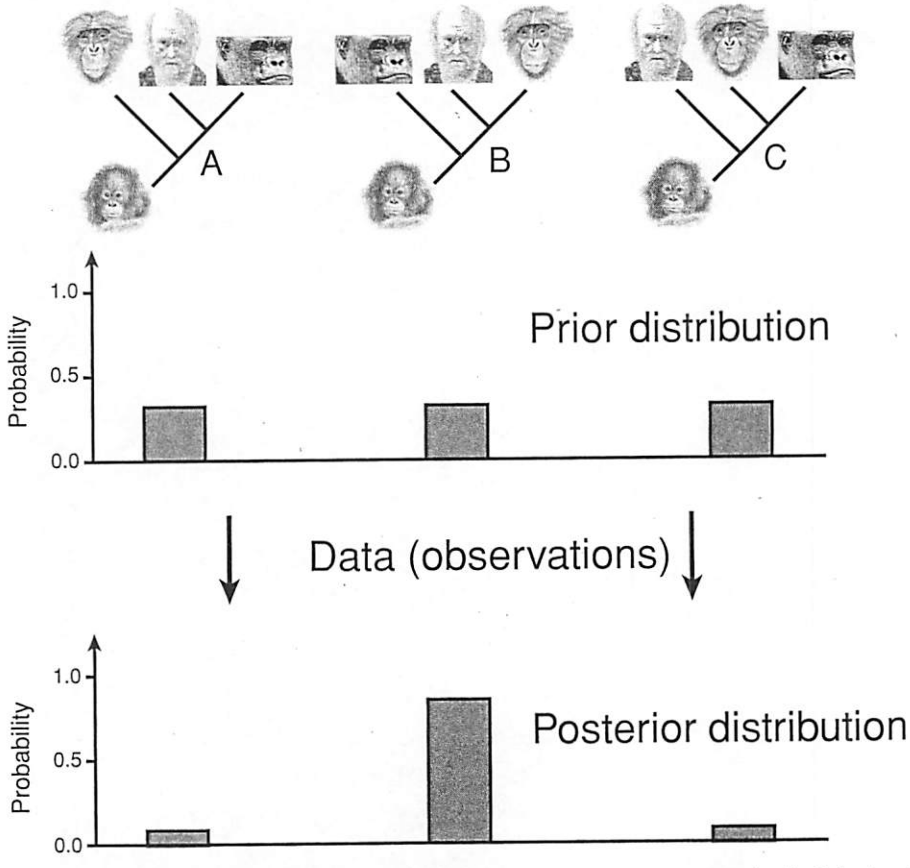


Fig. 7.1 A Bayesian phylogenetic analysis. We start the analysis by specifying our prior beliefs about the tree. In the absence of background knowledge, we might associate the same probability to each tree topology. We then collect data and use a stochastic evolutionary model and Bayes' theorem to update the prior to a posterior probability distribution. If the data are informative, most of the posterior probability will be focused on one tree (or a small subset of trees in a large tree space).

Before the analysis starts, we need to specify our prior beliefs about the relationships. In the absence of background data, a simple solution would be to assign equal probability to the possible trees. Since there are three trees, the probability of each would be one-third. Such a prior probability distribution is known as a *vague* or *uninformative prior* because it is appropriate for the situation when we do not have any prior knowledge or do not want to build our analysis on any previous results.

To update the prior we need some data, typically in the form of a molecular sequence alignment, and a stochastic model of the process generating the data on the tree. In principle, Bayes' rule is then used to obtain the posterior probability distribution (Fig. 7.1), which is the result of the analysis. The posterior specifies the probability of each tree given the model, the prior, and the data. When the data are informative, most of the posterior probability is typically concentrated on one tree (or a small subset of trees in a large tree space).

If the analysis is performed correctly, there is nothing controversial about the posterior probabilities. Nevertheless, the interpretation of them is often subject to considerable discussion, particularly in the light of alternative models and priors.

To describe the analysis mathematically, designate the matrix of aligned sequences X. The vector of model parameters is contained in θ (we do not distinguish in our notation between vector parameters and scalar parameters). In the ideal case, this vector would only include a topology parameter τ, which could take on the three possible values discussed above. However, this is not sufficient to calculate the probability of the data. Minimally, we also need branch lengths on the tree; collect these in the vector v. Typically, there are also some ***substitution model*** parameters to be considered but, for now, let us use the Jukes Cantor substitution model (see below), which does not have any free parameters. Thus, in our case, $\theta = (\tau, v)$.

Bayes' theorem allows us to derive the posterior distribution as

$$f(\theta|X) = \frac{f(\theta)f(X|\theta)}{f(X)} \tag{7.5}$$

The denominator is an integral over the parameter values, which evaluates to a summation over discrete topologies and a multidimensional integration over possible branch length values:

$$f(X) = \int f(\theta)f(X|\theta)\,\mathrm{d}\theta \tag{7.6}$$

$$= \sum_{\tau}\int_{v} f(v)f(X|\tau, v)\,\mathrm{d}v \tag{7.7}$$

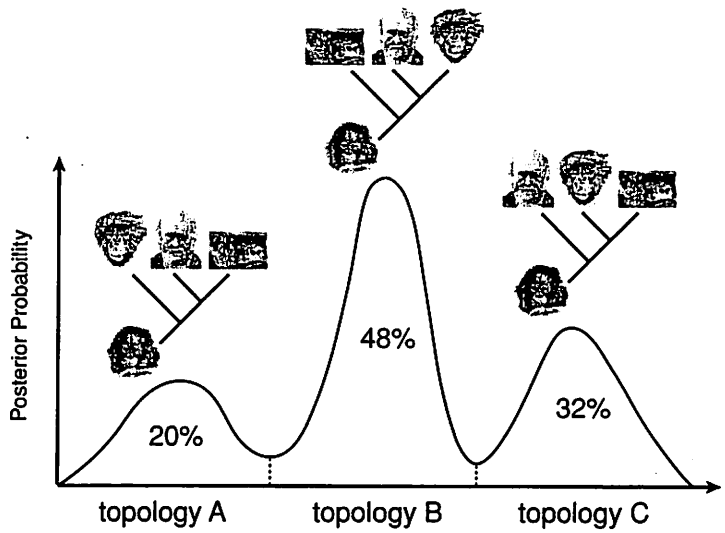


Fig. 7.2 Posterior probability distribution for our phylogenetic analysis. The *x*-axis is an imaginary one-dimensional representation of the parameter space. It falls into three different regions corresponding to the three different topologies. Within each region, a point along the axis corresponds to a particular set of branch lengths on that topology. It is difficult to arrange the space such that optimal branch length combinations for different topologies are close to each other. Therefore, the posterior distribution is multimodal. The area under the curve falling in each tree topology region is the posterior probability of that tree topology.

Even though our model is as simple as phylogenetic models come, it is impossible to portray its parameter space accurately in one dimension. However, imagine for a while that we could do just that. Then the parameter axis might have three distinct regions corresponding to the three different tree topologies (Fig. 7.2). Within each region, the different points on the axis would represent different branch length values. The one-dimensional parameter axis allows us to obtain a picture of the posterior probability function or surface. It would presumably have three distinct peaks, each corresponding to an optimal combination of topology and branch lengths.

To calculate the posterior probability of the topologies, we integrate out the model parameters that are not of interest, the branch lengths in our case. This corresponds to determining the area under the curve in each of the three topology regions. A Bayesian would say that we are marginalizing or deriving the ***marginal probability distribution*** on topologies.

Why is it called marginalizing? Imagine that we represent the parameter space in a two-dimensional table instead of along a single axis (Fig. 7.3). The columns in this table might represent different topologies and the rows different branch length values. Since the branch lengths are continuous parameters, there would actually

Topologies

Joint probabilities

Branch length vectors	τ_A	τ_B	τ_C	
V^A	0.10	0.07	0.12	0.29
V^B	0.05	0.22	0.06	0.33
V^C	0.05	0.19	0.14	0.38
	0.20	0.48	0.32	

Marginal probabilities

Fig. 7.3 A two-dimensional table representation of parameter space. The columns represent different tree topologies, the rows represent different branch length bins. Each cell in the table represents the joint probability of a particular combination of branch lengths and topology. If we summarize the probabilities along the margins of the table, we get the marginal probabilities for the topologies (bottom row) and for the branch length bins (last column).

be an infinite number of rows, but imagine that we sorted the possible branch length values into discrete bins, so that we get a finite number of rows. For instance, if we considered only short and long branches, one bin would have all branches long, another would have the terminal branches long and the interior branch short, etc.

Now, assume that we can derive the posterior probability that falls in each of the cells in the table. These are ***joint probabilities*** because they represent the joint probability of a particular topology and a particular set of branch lengths. If we summarized all joint probabilities along one axis of the table, we would obtain the marginal probabilities for the corresponding parameter. To obtain the marginal probabilities for the topologies, for instance, we would summarize the entries in each column. It is traditional to write the sums in the margin of the table, hence the term marginal probability (Fig. 7.3).

It would also be possible to summarize the probabilities in each row of the table. This would give us the marginal probabilities for the branch length combinations (Fig. 7.3). Typically, this distribution is of no particular interest but the possibility of calculating it illustrates an important property of Bayesian inference: there is no sharp distinction between different types of model parameters. Once the posterior probability distribution is obtained, we can derive any marginal distribution of interest. There is no need to decide on the parameters of interest before performing the analysis.

7.3 Markov chain Monte Carlo sampling

In most cases, including virtually all phylogenetic problems, it is impossible to derive the posterior probability distribution analytically. Even worse, we can't even estimate it by drawing random samples from it. The reason is that most of the posterior probability is likely to be concentrated in a small part of a vast parameter space. Even with a massive sampling effort, it is highly unlikely that we would obtain enough samples from the interesting region(s) of the posterior. This argument is particularly easy to appreciate in the phylogenetic context because of the large number of tree topologies that are possible even for small numbers of taxa. Already at nine taxa, you are more likely to be hit by lightning (odds 3:100 000) than to find the best tree by picking one randomly (odds 1:135, 135). At slightly more than 50 taxa, the number of topologies outnumber the number of atoms in the known universe – and this is still considered a small phylogenetic problem.

The solution is to estimate the posterior probability distribution using ***Markov chain Monte Carlo sampling***, or ***MCMC*** for short. ***Markov chains*** have the property that they converge towards an equilibrium state regardless of starting point. We just need to set up a Markov chain that converges onto our posterior probability distribution, which turns out to be surprisingly easy. It can be achieved using several different methods, the most flexible of which is known as the *Metropolis algorithm*, originally described by a group of famous physicists involved in the Manhattan project (Metropolis *et al.*, 1953). Hastings (1970) later introduced a simple but important extension, and the sampler is often referred to as the ***Metropolis–Hastings*** method.

The central idea is to make small random changes to some current parameter values, and then accept or reject those changes according to the appropriate probabilities. We start the chain at an arbitrary point θ in the landscape (Fig. 7.4). In the next generation of the chain, we consider a new point θ^* drawn from a proposal distribution $f(\theta^*|\theta)$. We then calculate the ratio of the posterior probabilities at the two points. There are two possibilities. Either the new point is uphill, in which case we always accept it as the starting point for the next cycle in the chain, or it is downhill, in which case we accept it with a probability that is proportional to the height ratio. In reality, it is slightly more complicated because we need to take asymmetries in the proposal distribution into account as well. Formally, we accept

Markov chain Monte Carlo steps

1. Start at an arbitrary point (θ)
2. Make a small random move (to θ^*)
3. Calculate height ratio (r) of new state (to θ^*) to old state (θ)
   (a) $r > 1$: new state accepted
   (b) $r < 1$: new state accepted with probability r
   if new state rejected, stay in old state
4. Go to step 2

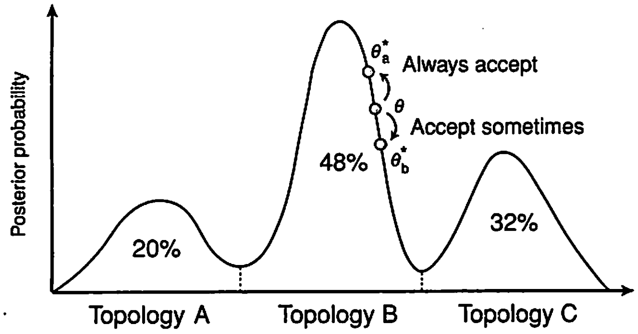


Fig. 7.4 The Markov chain Monte Carlo (MCMC) procedure is used to generate a valid sample from the posterior. One first sets up a Markov chain that has the posterior as its stationary distribution. The chain is then started at a random point and run until it converges onto this distribution. In each step (generation) of the chain, a small change is made to the current values of the model parameters (step 2). The ratio r of the posterior probability of the new and current states is then calculated. If $r > 1$, we are moving uphill and the move is always accepted (3a). If $r < 1$, we are moving downhill and accept the new state with probability r (3b).

or reject the proposed value with the probability

$$r = \min\left(1, \frac{f(\theta^*|X)}{f(\theta|X)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)}\right) \tag{7.8}$$

$$= \min\left(1, \frac{f(\theta^*)f(X|\theta^*)/f(X)}{f(\theta)f(X|\theta)/f(X)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)}\right) \tag{7.9}$$

$$= \min\left(1, \frac{f(\theta^*}{f(\theta)} \times \frac{f(X|\theta^*)}{f(X|\theta)} \times \frac{f(\theta|\theta^*)}{f(\theta^*|\theta)}\right) \tag{7.10}$$

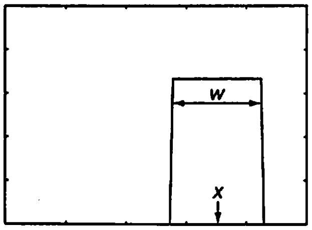
The three ratios in the last equation are referred to as the *prior ratio*, the *likelihood ratio*, and the *proposal ratio* (or *Hastings ratio*), respectively. The first two ratios correspond to the ratio of the numerators in Bayes' theorem; note that the complex integral in the denominator of Bayes' theorem, $f(X)$, cancels out in the second step because it is the same for both the current and the proposed states. Because of this, r is easy to compute.

The Metropolis sampler works because the relative equilibrium frequencies of the two states θ and θ^* is determined by the ratio of the rates at which the chain moves back and forth between them. Equation (7.10) ensures that this ratio is the same as the ratio of their posterior probabilities. This means that, if the Markov chain is allowed to run for a sufficient number of generations, the amount of time it spends sampling a particular parameter value or parameter interval is proportional to the posterior probability of that value or interval. For instance, if the posterior probability of a topology is 0.68, then the chain should spend 68% of its time sampling that topology at ***stationarity***. Similarly, if the posterior probability of a branch length being in the interval (0.02, 0.04) is 0.11, then 11% of the chain samples at stationarity should be in that interval.

For a large and parameter-rich model, a mixture of different Metropolis samplers is typically used. Each sampler targets one parameter or a set of related parameters (Box 7.2). One can either cycle through the samplers systematically or choose among them randomly according to some proposal probabilities (MrBayes does the latter).

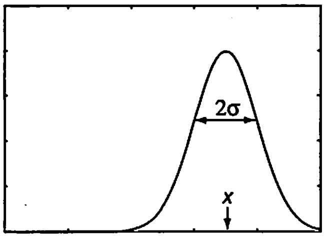
Box 7.2 Proposal mechanisms

Four types of proposal mechanisms are commonly used to change continuous variables. The simplest is the *sliding window* proposal. A continuous uniform distribution of width w is centered on the current value x, and the new value x^* is drawn from this distribution. The "window" width w is a tuning parameter. A larger value of w results in more radical proposals and lower acceptance rates, while a smaller value leads to more modest changes and higher acceptance rates.



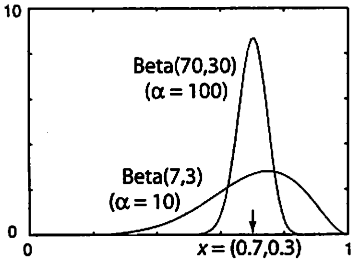
The *normal* proposal is similar to the sliding window except that it uses a normal distribution centered on the current value x. The variance σ^2 of the normal distribution determines how drastic the new proposals are and how often they will be accepted.

Box 7.2 *(cont.)*



Both the sliding window and normal proposals can be problematic when the effect on the likelihood varies over the parameter range. For instance, changing a branch length from 0.01 to 0.03 is likely to have a dramatic effect on the posterior but changing it from 0.51 to 0.53 will hardly be noticeable. In such situations, the *multiplier* proposal is appropriate. It is equivalent to a sliding window with width λ on the log scale of the parameter. A random number u is drawn from a uniform distribution on the interval $(-0.5, 0.5)$ and the proposed value is $x^* = mx$, where $m = e^{\lambda u}$. If the value of λ takes the form $2 \ln a$, one will pick multipliers m in the interval $(1/a, a)$.

The *beta* and *Dirichlet* proposals are used for simplex parameters. They pick new values from a beta or Dirichlet distribution centered on the current values of the simplex. Assume that the current values are (x_1, x_2). We then multiply them with a value α, which is a tuning parameter, and pick new values from the distribution Beta$(\alpha x_1, \alpha x_2)$. The higher the value of α, the closer the proposed values will be to the current values.



More complex moves are needed to change topology. A common type uses stochastic branch rearrangements (see Chapter 8). For instance, the extending ***subtree pruning and regrafting*** (extending SPR) move chooses a subtree at random and then moves its attachment point, one branch at a time, until a random number u drawn from a uniform on $(0, 1)$ becomes higher than a specified extension probability p. The extension probability p is a tuning parameter; the higher the value, the more drastic rearrangements will be proposed.

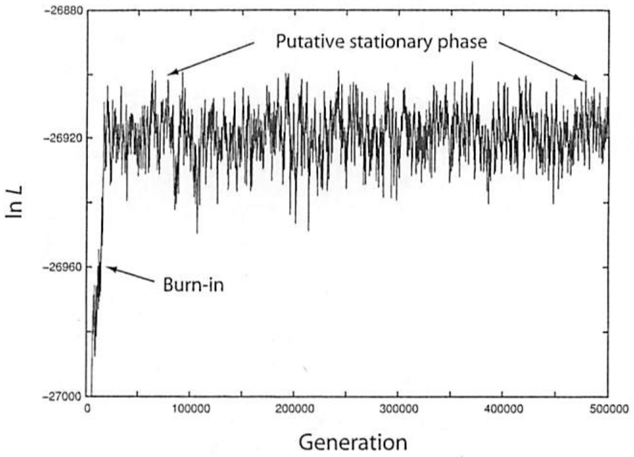


Fig. 7.5 The likelihood values typically increase very rapidly during the initial phase of the run because the starting point is far away from the regions in parameter space with high posterior probability. This initial phase of the Markov chain is known as the burn in. The burn-in samples are typically discarded because they are so heavily influenced by the starting point. As the chain converges onto the target distribution, the likelihood values tend to reach a plateau. This phase of the chain is sampled with some thinning, primarily to save disk space.

7.4 Burn-in, mixing and convergence

If the chain is started from a random tree and arbitrarily chosen branch lengths, chances are that the initial likelihood is low. As the chain moves towards the regions in the posterior with high probability mass, the likelihood typically increases very rapidly; in fact, it almost always changes so rapidly that it is necessary to measure it on a log scale (Fig. 7.5). This early phase of the run is known as the ***burn-in***, and the burn-in samples are often discarded because they are so heavily influenced by the starting point.

As the chain approaches its stationary distribution, the likelihood values tend to reach a plateau. This is the first sign that the chain may have converged onto the target distribution. Therefore, the plot of the likelihood values against the generation of the chain, known as the ***trace plot*** (Fig. 7.5), is important in monitoring the performance of an MCMC run. However, it is extremely important to confirm convergence using other diagnostic tools because it is not sufficient for the chain to reach the region of high probability in the posterior, it must also cover this region adequately. The speed with which the chain covers the interesting regions of the posterior is known as its ***mixing behavior***. The better the mixing, the faster the chain will generate an adequate sample of the posterior.

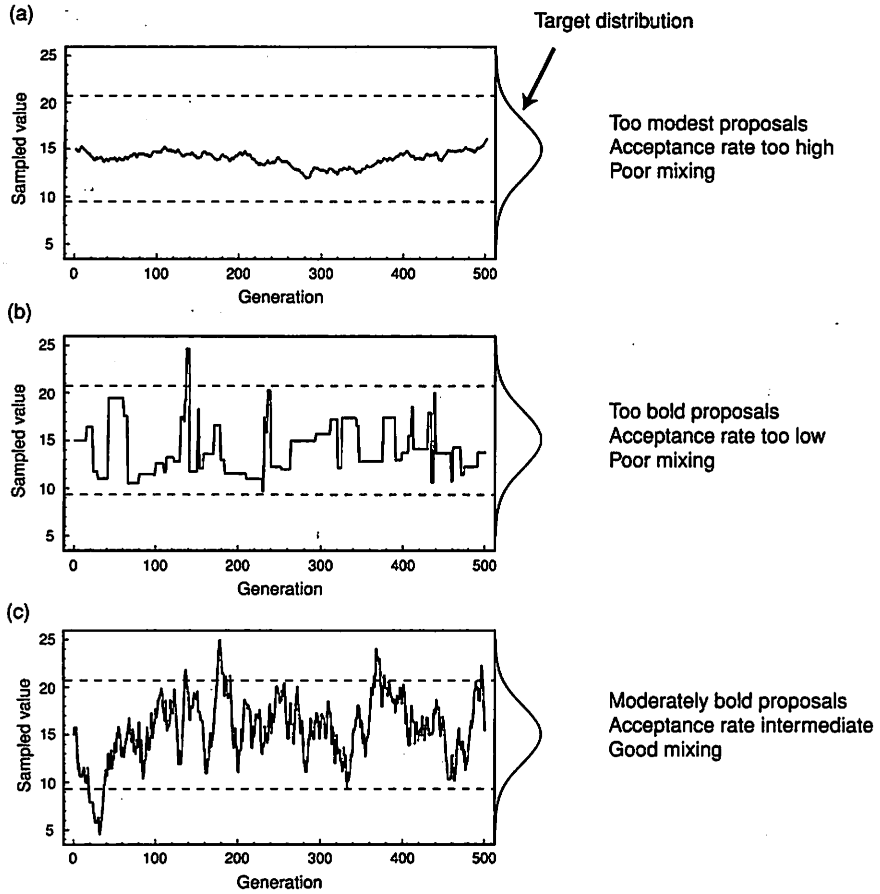


Fig. 7.6 The time it takes for a Markov chain to obtain an adequate sample of the posterior depends critically on its mixing behavior, which can be controlled to some extent by the proposal tuning parameters. If the proposed values are very close to the current ones, all proposed changes are accepted but it takes a long time for the chain to cover the posterior; mixing is poor. If the proposed values tend to be dramatically different from the current ones, most proposals are rejected and the chain will remain on the same value for a long time, again leading to poor mixing. The best mixing is obtained at intermediate values of the tuning parameters, associated with moderate acceptance rates.

The mixing behavior of a Metropolis sampler can be adjusted using its tuning parameter(s). Assume, for instance, that we are sampling from a normal distribution using a sliding window proposal (Fig. 7.6). The sliding window proposal has one tuning parameter, the width of the window. If the width is too small, then the proposed value will be very similar to the current one (Fig. 7.6a). The posterior probabilities will also be very similar, so the proposal will tend to be accepted. But each proposal will only move the chain a tiny distance in parameter space, so it will take the chain a long time to cover the entire region of interest; mixing is poor.

A window that is too wide also results in poor mixing. Under these conditions, the proposed state is almost always very different from the current state. If we have reached a region of high posterior probability density, then the proposed state is also likely to have much lower probability than the current state. The new state will therefore often be rejected, and the chain remains in the same spot for a long time (Fig. 7.6b), resulting in poor mixing. The most efficient sampling of the target distribution is obtained at intermediate acceptance rates, associated with intermediate values of the tuning parameter (Fig. 7.6c).

Extreme acceptance rates thus indicate that sampling efficiency can be improved by adjusting proposal tuning parameters. Studies of several types of complex but unimodal posterior distributions indicate that the optimal acceptance rate is 0.44 for one-dimensional and 0.23 for multi-dimensional proposals (Roberts *et al.*, 1997; Roberts & Rosenthal, 1998, 2001). However, multimodal posteriors are likely to have even lower optimal acceptance rates. Adjusting the tuning parameter values to reach a target acceptance rate can be done manually or automatically using adaptive tuning methods (Roberts & Rosenthal, 2006). Bear in mind, however, that some samplers used in Bayesian MCMC phylogenetics have acceptance rates that will remain low, no matter how much you tweak the tuning parameters. In particular, this is true for many tree topology update mechanisms.

Convergence diagnostics help determine the quality of a sample from the posterior. There are essentially three different types of diagnostics that are currently in use: (1) examining autocorrelation times, ***effective sample sizes***, and other measures of the behavior of single chains; (2) comparing samples from successive time segments of a single chain; and (3) comparing samples from different runs. The last approach is arguably the most powerful way of detecting convergence problems. The drawback is that it wastes computational power by generating several independent sets of burn-in samples that must be discarded.

In Bayesian MCMC sampling of phylogenetic problems, the tree topology is typically the most difficult parameter to sample from. Therefore, it makes sense to focus our attention on this parameter when monitoring convergence. If we start several parallel MCMC runs from different, randomly chosen trees, they will initially sample from very different regions of tree space. As they approach stationarity, however, the tree samples will become more and more similar. Thus, an intuitively appealing convergence diagnostic is to compare the variance among and within tree samples from different runs.

Perhaps the most obvious way of achieving this is to compare the frequencies of the sampled trees. However, this is not practical unless most of the posterior probability falls on a small number of trees. In large phylogenetic problems, there is often an inordinate number of trees with similar probabilities and it may be extremely difficult to estimate the probability of each accurately.

The approach that we and others have taken to solve this problem is to focus on ***split*** (clade) *frequencies* instead. A split is a partition of the tips of the tree into two non-overlapping sets; each branch in a tree corresponds to exactly one such split. For instance, the split ((human, chimp),(gorilla, orangutan)) corresponds to the branch uniting the human and the chimp in a tree rooted on the orangutan. Typically, a fair number of splits are present in high frequency among the sampled trees. In a way, the dominant splits (present in, say, more than 10% of the trees) represent an efficient diagnostic summary of the tree sample as a whole. If two tree samples are similar, the split frequencies should be similar as well. To arrive at an overall measure of the similarity of two or more tree samples, we simply calculate the average standard deviation of the split frequencies. As the tree samples become more similar, this value should approach zero.

Most other parameters in phylogenetic models are continuous scalar parameters. An appropriate convergence diagnostic for these is the ***Potential Scale Reduction Factor*** (***PSRF***) originally proposed by Gelman and Rubin (1992). The PSRF compares the variance among runs with the variance within runs. If chains are started from over-dispersed starting points, the variance among runs will initially be higher than the variance within runs. As the chains converge, however, the variances will become more similar and the PSRF will approach 1.0.

7.5 Metropolis coupling

For some phylogenetic problems, it may be difficult or impossible to achieve convergence within a reasonable number of generations using the standard approach. Often, this seems to be due to the existence of isolated peaks in tree space (also known as tree islands) with deep valleys in-between. In these situations, individual chains may get stuck on different peaks and have difficulties moving to other peaks of similar probability mass. As a consequence, tree samples from independent runs tend to be different. A topology convergence diagnostic, such as the standard deviation of split frequencies, will indicate that there is a problem. But are there methods that can help us circumvent it?

A general technique that can improve mixing, and hence convergence, in these cases is ***Metropolis Coupling***, also known as MCMCMC or $(MC)^3$ (Geyer, 1991). The idea is to introduce a series of Markov chains that sample from a *heated* posterior probability distribution (Fig. 7.7). The heating is achieved by raising the posterior probability to a power smaller than 1. The effect is to flatten out the posterior probability surface, very much like melting a landscape of wax.

Because the surface is flattened, a Markov chain will move more readily between the peaks. Of course, the heated chains have a target distribution that is different from the one we are interested in, sampled by the ***cold chain***, but we can use them

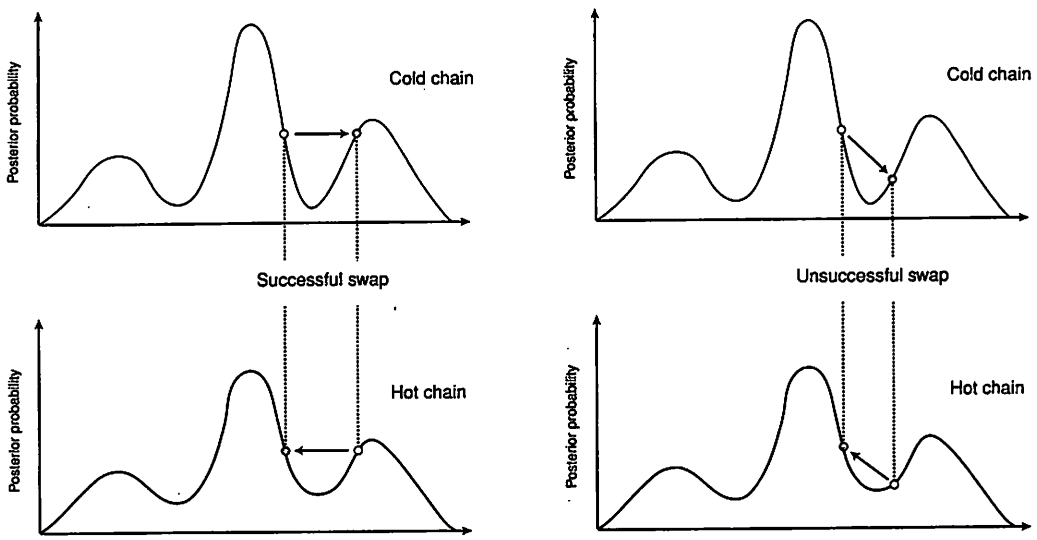


Fig. 7.7 Metropolis Coupling uses one or more *heated* chains to accelerate mixing in the so-called *cold* chain sampling from the posterior distribution. The heated chains are flattened out versions of the posterior, obtained by raising the posterior probability to a power smaller than one. The heated chains can move more readily between peaks in the landscape because the valleys between peaks are shallower. At regular intervals, one attempts to swap the states between chains. If a swap is accepted, the cold chain can jump between isolated peaks in the posterior in a single step, accelerating its mixing over complex posterior distributions.

to generate proposals for the cold chain. With regular intervals, we attempt to swap the states between two randomly picked chains. If the cold chain is one of them, and the swap is accepted, the cold chain can jump considerable distances in parameter space in a single step. In the ideal case, the swap takes the cold chain from one tree island to another. At the end of the run, we simply discard all of the samples from the heated chains and keep only the samples from the cold chain.

In practice, an incremental heating scheme is often used where chain i has its posterior probability raised by the temperature factor

$$T = \frac{1}{1 + \lambda i} \tag{7.11}$$

where $i \in \{0, 1, \ldots, k\}$ for k heated chains, with $i = 0$ for the cold chain, and λ is the temperature factor. The higher the value of λ, the larger the temperature difference between adjacent chains in the incrementally heated sequence.

If we apply too much heat, then the chains moving in the heated landscapes will walk all over the place and are less likely to be on an interesting peak when we try to swap states with the cold chain. Most of the swaps will therefore be rejected and the heating does not accelerate mixing in the cold chain. On the other hand, if we do not heat enough, then the chains will be very similar, and the heated chain will not mix more rapidly than the cold chain. As with the proposal tuning parameters, an intermediate value of the heating parameter λ works best.

7.6 Summarizing the results

The stationary phase of the chain is typically sampled with some thinning, for instance every 50th or 100th generation. This is done primarily to save disk space, since an MCMC run can easily generate millions of samples. Once an adequate sample is obtained, it is usually trivial to compute an estimate of the marginal posterior distribution for the parameter(s) of interest. For instance, this can take the form of a frequency histogram of the sampled values. When it is difficult to visualize this distribution or when space does not permit it, various summary statistics are used instead.

Most phylogenetic model parameters are continuous variables and their estimated posterior distribution is summarized using statistics such as the mean, the median, and the variance. Bayesian statisticians typically also give the 95% ***credibility interval***, which is obtained by simply removing the lowest 2.5% and the highest 2.5% of the sampled values. The credibility interval is somewhat similar to a confidence interval but the interpretation is different. A 95% credibility interval actually contains the true value with probability 0.95 (given the model, prior, and data) unlike the confidence interval, which has a more complex interpretation.

The posterior distribution on topologies and branch lengths is more difficult to summarize efficiently. If there are few topologies with high posterior probability, one can produce a list of the best topologies and their probabilities, or simply give the topology with the maximum posterior probability. However, most posteriors contain too many topologies with reasonably high probabilities, and one is forced to use other methods.

One way to illustrate the topological variance in the posterior is to list the topologies in order of decreasing probabilities and then calculate the cumulative probabilities so that we can give the estimated number of topologies in various ***credible sets***. Assume, for instance, that the five best topologies have the estimated probabilities (0.35, 0.25, 0.20, 0.15, 0.03), giving the cumulative probabilities (0.35, 0.60, 0.80, 0.95, 0.98). Then the 50% credible set has two topologies in it, the 90% and the 95% credible sets both have four trees in them, etc. We simply pass down the list and count the number of topologies we need to include before the target probability is met or superseded. When these credible sets are large, however, it is difficult to estimate their sizes precisely.

The most common approach to summarizing topology posteriors is to give the frequencies of the most common splits, since there are much fewer splits than topologies. Furthermore, all splits occurring in at least 50% of the sampled trees are guaranteed to be compatible and can be visualized in the same tree, a ***majority rule consensus tree***. However, although the split frequencies are convenient, they do have limitations. For instance, assume that the splits ((A,B),(C,D,E)) and ((A,B,C),(D,E)) were both encountered in 70% of the sampled trees. This could mean that 30% of the sampled trees contained neither split or, at the other extreme, that all sampled trees contained at least one of them. The split frequencies themselves only allow us to approximately reconstruct the underlying set of topologies.

The sampled branch lengths are even more difficult to summarize adequately. Perhaps the best way would be to display the distribution of sampled branch length values separately for each topology. However, if there are many sampled topologies, there may not be enough branch length samples for each. A reasonable approach, taken by MRBAYES, is then to pool the branch length samples that correspond to the same split. These pooled branch lengths can also be displayed on the consensus tree. However, one should bear in mind that the pooled distributions may be multimodal since the sampled values in most cases come from different topologies, and a simple summary statistic like the mean might be misleading.

A special difficulty appears with branch lengths in clock trees. Clock trees are rooted trees in which branch lengths are proportional to time units (see Chapter 11). Even if computed from a sample of clock trees, a majority rule consensus tree with mean pooled branch lengths is not necessarily itself a clock tree. This problem is easily circumvented by instead using mean pooled node depths instead of branch lengths (for Bayesian inference of clock trees, see also Chapter 18).

7.7 An introduction to phylogenetic models

A phylogenetic model can be divided into two distinct parts: a tree model and a substitution model. The tree model we have discussed so far is the one most commonly used in phylogenetic inference today (sometimes referred to as the different-rates or ***unrooted model***, see Chapter 11). Branch lengths are measured in amounts of expected evolutionary change per site, and we do not assume any correlation between branch lengths and time units. Under time-reversible substitution models, the likelihood is unaffected by the position of the root, that is, the tree is unrooted. For presentation purposes, unrooted trees are typically rooted between a specially designated reference sequence or group of reference sequences, the ***outgroup***, and the rest of the sequences.

Alternatives to the standard tree model include the strict and ***relaxed clock*** tree models. Both of these are based on trees, whose branch lengths are strictly proportional to time. In strict clock models, the evolutionary rate is assumed to be constant so that the amount of evolutionary change on a branch is directly proportional to its time duration, whereas relaxed clock models include a model component that accommodates some variation in the rate of evolution across the tree. Various prior probability models can be attached to clock trees. Common examples include the uniform model, the birth-death process, and the coalescent process (for the latter two, see Chapter 18).

The substitution process is typically modeled using Markov chains of the same type used in MCMC sampling. For instance, they have the same tendency towards an equilibrium state. The different substitution models are most easily described in terms of their instantaneous rate matrices, or *Q* ***matrices.*** For instance, the general time-reversible model (GTR) is described by the rate matrix

$$Q = \begin{bmatrix} - & \pi_C r_{AC} & \pi_G r_{AG} & \pi_T r_{AT} \\ \pi_A r_{AC} & - & \pi_G r_{CG} & \pi_T r_{CT} \\ \pi_A r_{AG} & \pi_C r_{CG} & - & \pi_T r_{GT} \\ \pi_A r_{AT} & \pi_C r_{CT} & \pi_G r_{GT} & - \end{bmatrix}$$

Each row in this matrix gives the instantaneous rate of going from a particular state, and each column represents the rate of going to a particular state; the states are listed in alphabetical sequence A, C, G, T. For instance, the second entry in the first row represents the rate of going from A to C. Each rate is composed of two factors; for instance, the rate of going from A to C is a product of π_C and r_{AC}. The rates along the diagonal are commonly omitted since their expressions are slightly more complicated. However, they are easily calculated since the rates in each row always sum to zero. For instance, the instantaneous rate of going from A to A (first entry in the first row) is $-\pi_C r_{AC} - \pi_G r_{AG} - \pi_T r_{AT}$.

It turns out that, if we run this particular Markov chain for a long time, it will move towards an equilibrium, where the frequency of a state i is determined exactly by the factor π_i given that $\sum \pi_i = 1$. Thus, the first rate factor corresponds to the stationary state frequency of the receiving state. The second factor, r_{ij}, is a parameter that determines the intensity of the exchange between pairs of states, controlling for the stationary state frequencies. For instance, at equilibrium we will have π_A sites in state A and π_C sites in state C. The total instantaneous rate of going from A to C over the sequence is then π_A times the instantaneous rate of the transition from A to C, which is $\pi_C r_{AC}$, resulting in a total rate of A to C changes over the sequence of $\pi_A \pi_C r_{AC}$. This is the same as the total rate of the reverse changes over the sequence, which is $\pi_C \pi_A r_{AC}$. Thus, there is no net change of the state proportions, which is the definition of an equilibrium, and the factor r_{AC} determines how intense the exchange between A and C is compared with the exchange between other pairs of states.

Many of the commonly used substitution models are special cases or extensions of the GTR model. For instance, the Jukes Cantor model has all rates equal, and the Felsenstein 81 (F81) model has all exchangeability parameters (r_{ij}) equal. The *covarion* and *covariotide* models have an independent on–off switch for each site, leading to a composite instantaneous rate matrix including four smaller rate matrices: two matrices describing the switching process, one being a zero-rate matrix, and the last describing the normal substitution process in the on state.

In addition to modeling the substitution process at each site, phylogenetic models typically also accommodate rate variation across sites. The standard approach is to assume that rates vary according to a gamma distribution (Box 7.1) with mean 1. This results in a distribution with a single parameter, typically designated α, describing the shape of the rate variation (see Fig. 4.8 in Chapter 4). Small values of α correspond to large amounts of rate variation; as α approaches infinity, the model approaches rate constancy across sites. It is computationally expensive to let the MCMC chain integrate over a continuous gamma distribution of site rates, or to numerically integrate out the gamma distribution in each step of the chain. The standard solution is to integrate out the gamma using a discrete approximation with a small number of rate categories, typically four to eight, which is a reasonable compromise. An alternative is to use MCMC sampling over discrete rate categories.

Many other models of rate variation are also possible. A commonly considered model assumes that there is a proportion of invariable sites, which do not change at all over the course of evolution. This is often combined with an assumption of gamma-distributed rate variation in the variable sites.

It is beyond the scope of this chapter to give a more detailed discussion of phylogenetic models but we present an overview of the models implemented in MrBayes 3.2, with the command options needed to invoke them (Fig. 7.8). The MrBayes manual provides more details and references to the different models. A simulation-based presentation of Markov substitution models is given in (Huelsenbeck & Ronquist, 2005) and further details can be found in Chapter 4 and Chapter 10.

7.8 Bayesian model choice and model averaging

So far, our notation has implicitly assumed that Bayes' theorem is conditioned on a particular model. To make it explicit, we could write Bayes' theorem:

$$f(\theta|X, M) = \frac{f(\theta|M)f(X|\theta, M)}{f(X|M)} \tag{7.12}$$

It is now clear that the normalizing constant, $f(X|M)$, is the probability of the data given the chosen model after we have integrated out all parameters. This

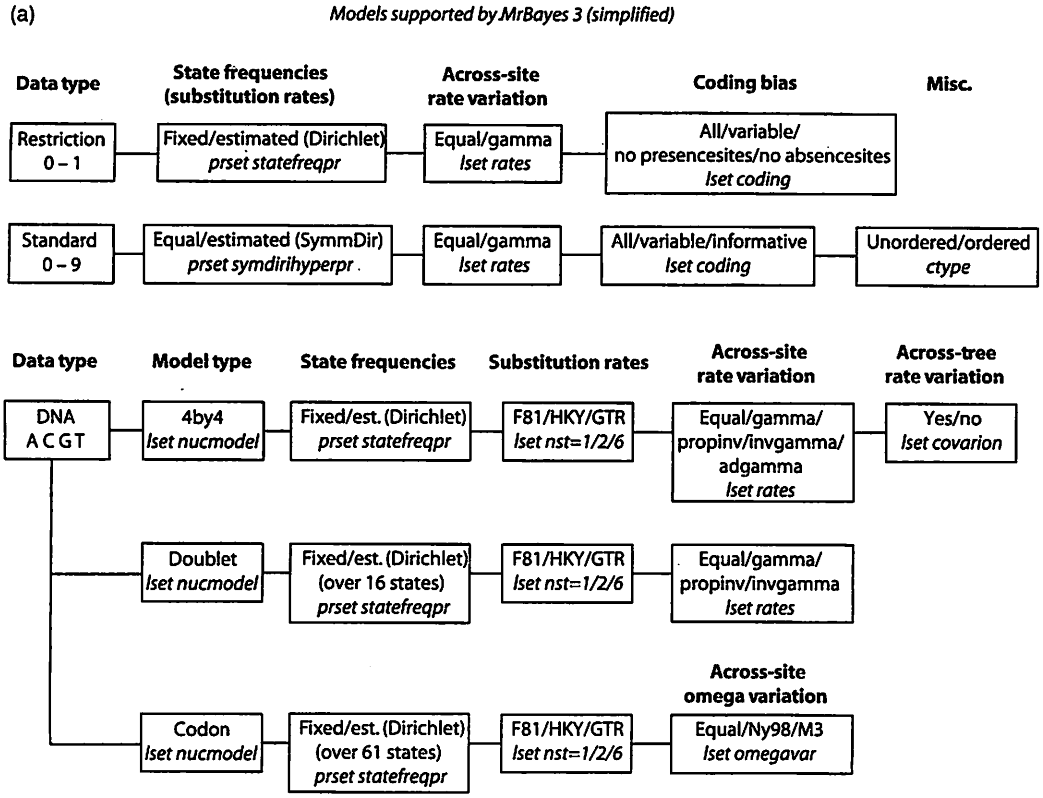

(a)
Models supported by MrBayes 3 (simplified)
Data type
State frequencies (substitution rates)
Across-site rate variation
Coding bias
Misc.
Restriction 0 – 1
Fixed/estimated (Dirichlet) prset statefreqpr
Equal/gamma lset rates
All/variable/ no presencesites/no absencesites lset coding
Standard 0 – 9
Equal/estimated (SymmDir) prset symdirihyperpr.
Equal/gamma lset rates
All/variable/informative lset coding
Unordered/ordered ctype
Data type
Model type
State frequencies
Substitution rates
Across-site rate variation
Across-tree rate variation
DNA ACGT
4by4 lset nucmodel
Fixed/est. (Dirichlet) prset statefreqpr
F81/HKY/GTR lset nst=1/2/6
Equal/gamma/ propinv/invgamma/ adgamma lset rates
Yes/no lset covarion
Doublet lset nucmodel
Fixed/est. (Dirichlet) (over 16 states) prset statefreqpr
F81/HKY/GTR lset nst=1/2/6
Equal/gamma/ propinv/invgamma lset rates
Across-site omega variation
Codon lset nucmodel
Fixed/est. (Dirichlet) (over 61 states) prset statefreqpr
F81/HKY/GTR lset nst=1/2/6
Equal/Ny98/M3 lset omegavar

Fig. 7.8 Schematic overview of the models implemented in MrBayes 3. Each box gives the available settings in normal font and then the program commands and command options needed to invoke those settings in italics.

quantity, known as the *model likelihood*, is used for Bayesian model comparison. Assume we are choosing between two models, M_0 and M_1, and that we assign them the prior probabilities $f(M_0)$ and $f(M_1)$. We could then calculate the ratio of their posterior probabilities (the posterior odds) as

$$\frac{f(M_0|X)}{f(M_1|X)} = \frac{f(M_0)f(X|M_0)}{f(M_1)f(X|M_1)} = \frac{f(M_0)}{f(M_1)} \times \frac{f(X|M_0)}{f(X|M_1)} \tag{7.13}$$

Thus, the posterior odds is obtained as the prior odds, $f(M_0)/f(M_1)$, times a factor known as the ***Bayes factor***, $B_{01} = f(X|M_0)/f(X|M_1)$, which is the ratio of the model likelihoods. Rather than trying to specify the prior model odds, it is common to focus entirely on the Bayes factor. One way to understand the Bayes factor is that it determines how much the prior model odds are changed by the data when calculating the posterior odds. The Bayes factor is also the same as

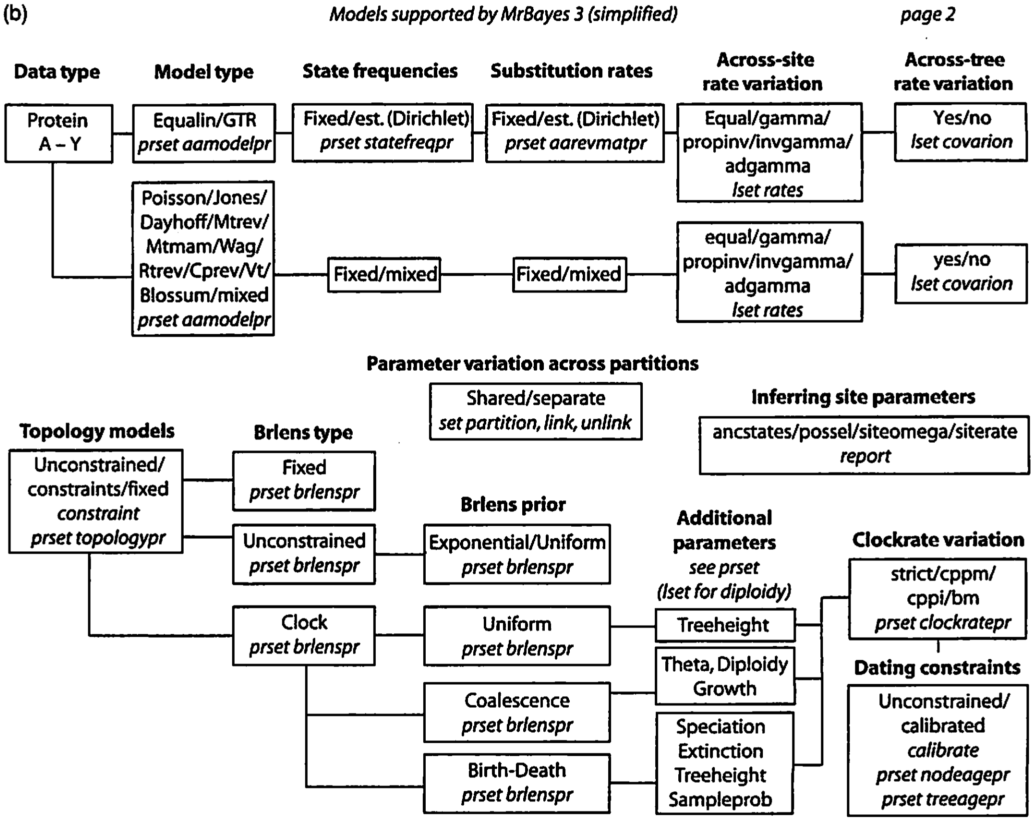

(b)
Models supported by MrBayes 3 (simplified)
page 2
Data type
Model type
State frequencies
Substitution rates
Across-site rate variation
Across-tree rate variation
Protein A – Y
Equalin/GTR prset aamodelpr
Fixed/est. (Dirichlet) prset statefreqpr
Fixed/est. (Dirichlet) prset aarevmatpr
Equal/gamma/ propinv/invgamma/ adgamma lset rates
Yes/no lset covarion
Poisson/Jones/ Dayhoff/Mtrev/ Mtmam/Wag/ Rtrev/Cprev/Vt/ Blossum/mixed prset aamodelpr
Fixed/mixed
Fixed/mixed
equal/gamma/ propinv/invgamma/ adgamma lset rates
yes/no lset covarion
Parameter variation across partitions
Shared/separate set partition, link, unlink
Inferring site parameters
ancstates/possel/siteomega/siterate report
Topology models
Unconstrained/ constraints/fixed constraint prset topologypr
Brlens type
Fixed prset brlenspr
Unconstrained prset brlenspr
Clock prset brlenspr
Brlens prior
Exponential/Uniform prset brlenspr
Uniform prset brlenspr
Coalescence prset brlenspr
Birth-Death prset brlenspr
Additional parameters see prset (lset for diploidy)
Treeheight
Theta, Diploidy Growth
Speciation Extinction Treeheight Sampleprob
Clockrate variation
strict/cppm/ cppi/bm prset clockratepr
Dating constraints
Unconstrained/ calibrated calibrate prset nodeagepr prset treeagepr

Fig. 7.8 (cont.)

the posterior odds when the prior odds are 1, that is, when we assign equal prior probabilities to the compared models.

Bayes factor comparisons are truly flexible. Unlike ***likelihood ratio tests***, there is no requirement for the models to be nested. Furthermore, there is no need to correct for the number of parameters in the model, in contrast to comparisons based on the ***Akaike Information Criterion*** (Akaike, 1974) or the confusingly named ***Bayesian Information Criterion*** (Schwarz, 1978). Although it is true that a more parameter-rich model always has a higher ***maximum likelihood*** than a nested submodel, its model likelihood need not be higher. The reason is that a more parameter-rich model also has a larger parameter space and therefore a lower prior probability density. This can lead to a lower model likelihood unless it is compensated for by a sufficiently large increase in the likelihood values in the peak region.

The interpretation of a Bayes factor comparison is up to the investigator but some guidelines were suggested by Kass and Raftery (1995) (Table 7.2).

Table 7.2 Critical values for Bayes factor comparisons

$2 \ln B_{01}$	B_{01}	Evidence against M_1
0 to 2	1 to 3	Not worth more than a bare mention
2 to 6	3 to 20	Positive
6 to 10	20 to 150	Strong
>10	>150	Very strong

From Kass & Raftery (1995).

The easiest way of estimating the model likelihoods needed in the calculation of Bayes factors is to use the *harmonic mean* of the likelihood values from the stationary phase of an MCMC run (Newton & Raftery, 1994). Unfortunately, this estimator is unstable because it is occasionally influenced by samples with very small likelihood and therefore a large effect on the final result. A stable estimator can be obtained by mixing in a small proportion of samples from the prior (Newton & Raftery, 1994). Even better accuracy, at the expense of computational complexity, can be obtained by using thermodynamic integration methods (Lartillot & Philippe, 2006). Because of the instability of the harmonic mean estimator, it is good practice to compare several independent runs and only rely on this estimator when the runs give consistent results.

An alternative to running a full analysis on each model and then choosing among them using the estimated model likelihoods and Bayes' factors is to let a single Bayesian analysis explore the models in a predefined model space (using ***reversible-jump MCMC***). In this case, all parameter estimates will be based on an average across models, each model weighted according to its posterior probability. For instance, MrBayes 3 uses this approach to explore a range of common fixed-rate matrices for amino acid data (see practice in Chapter 9 for an exercise).

Different topologies can also be considered different models and, in that sense, all Markov chains that integrate over the topology parameter also average across models. Thus, we can use the posterior sample of topologies from a single run to compare posterior probabilities of topology hypotheses.

For instance, assume that we want to test the hypothesis that group A is monophyletic against the hypothesis that it is not, and that 80% of the sampled trees have A monophyletic. Then the posterior model odds for A being monophyletic would be $0.80/0.20 = 4.0$. To obtain the Bayes factor, one would have to multiply this with the inverse of the prior model odds (see 7.13). If the prior assigned equal prior probability to all possible topologies, then the prior model odds would be determined by the number of trees consistent with each of the two hypotheses, a ratio that is easy to calculate. If one class of trees is empty, a conservative estimate of the Bayes factor would be obtained by adding one tree of this class to the sample.

7.9 Prior probability distributions

We will end with a few cautionary notes about priors. Beginners often worry excessively about the influence of the priors on their results and the subjectivity that this introduces into the inference procedure. In most cases however, the exact form of the priors (within rather generous bounds) has negligible influence on the posterior distribution. If this is a concern, it can always be confirmed by varying the prior assumptions.

The default priors used in MRBAYES are designed to be vague or uninformative probability distributions on the model parameters. When the data contain little information about some parameters, one would therefore expect the corresponding posterior probability distributions to be diffuse. As long as we can sample adequately from these distributions, which can be a problem if there are many of them (Nylander *et al.*, 2004), the results for other parameters should not suffer. We also know from simulations that the Bayesian approach does well even when the model is moderately overparameterized (Huelsenbeck & Rannala, 2004). Thus, the Bayesian approach typically handles weak data quite well.

However, the parameter space of phylogenetic models is vast and occasionally there are large regions with inferior but not extremely low likelihoods that attract the chain when the data are weak. The characteristic symptom is that the sample from the posterior is concentrated on parameter values that the investigator considers unlikely or unreasonable, for instance in comparison with the maximum likelihood estimates. We have seen a few examples involving models of rate variation applied to very small numbers of variable sites. In these cases, one can either choose to analyze the data under a simpler model (probably the best option in most cases) or include background information into the priors to emphasize the likely regions of parameter space.