[49] Swofford, D., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogenetic inference. In *Molecular Systematics* (2nd edn) (ed. D. Hillis, C. Moritz, and B. Mable), pp. 438–514. Sinauer, Sutherland, MA.

[50] Thorne, J.L., Goldman, N., and Jones, D.T. (1996). Combining protein evolution and secondary structure. *Molecular Biology and Evolution*, **13**(5), 666–673.

[51] Tillier, E.R.M. and Collins, R.A. (1998). High apparent rate of simultaneous compensatory base-pair substitutions in ribosomal RNA. *Genetics*, **148**, 1993–2002.

[52] Tuffley, C. and Steel, M.A. (1998). Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, **147**, 63–91.

[53] Van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[54] Vinh, L.S. and von Haeseler, A. (2004). IQPNNI: Moving fast through tree space and stopping in time. *Molecular Biology and Evolution*, **21**, 1565–1571.

[55] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Molecular Biology and Evolution*, **10**(6), 1396–1401.

[56] Yang, Z. (1994). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *Journal of Molecular Evolution*, **39**(3), 306–314.

[57] Yang, Z. (1995). A space-time process model for the evolution of DNA sequences. *Genetics*, **139**, 993–1005.

[58] Yang, Z. (2000). Phylogenetic analysis by maximum likelihood (PAML), version 3.0.

[59] Yang, Z. and Roberts, D. (1995). On the use of nucleic acid sequences to infer early branchings in the tree of life. *Molecular Biology and Evolution*, **12**(3), 451–458.

[60] Yang, Z., Swanson, W.J., and Vacquier, V.D. (2000). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution*, **17**(10), 1446–1455.

[61] Yang, Z. and Wang, T. (1995). Mixed model analysis of DNA sequence evolution. *Biometrics*, **51**(2), 552–561.

Yang, Z. 2005. In "Bayesian Inference in Molecular phylogenetics". In "Mathematics of Evolution and Phylogeny" (Ed. O. Gascuel). Oxford Univ. Press, Oxford, pp. 63–90.

3

BAYESIAN INFERENCE IN MOLECULAR PHYLOGENETICS

Ziheng Yang

The Bayesian method of statistical inference combines the prior for parameters with the data to generate the posterior distribution of parameters, upon which all inferences about the parameters are based. The method has become very popular due to recent advances in computational algorithms. In molecular evolution and phylogenetics, Bayesian inference has been applied to address fundamental biological problems under sophisticated models of sequence evolution. This chapter introduces Bayesian statistics through comparison with the likelihood method. I will discuss Markov chain Monte Carlo algorithms, the major modern computational methods for Bayesian inference, as well as two applications of Bayesian inference in molecular phylogenetics: estimation of species phylogenies and estimation of species divergence times.

3.1 The likelihood function and maximum likelihood estimates

The probability of observing the data D, when viewed as a function of the unknown parameters θ with the data given, is called the likelihood function: $L(\theta; D) = f(D \mid \theta)$. According to the likelihood principle, the likelihood function contains all information in the data about the parameters. The best point estimate of θ is given by the θ that maximizes the likelihood L or the log likelihood $\ell(\theta; D) = \log\{L(\theta; D)\}$. Furthermore, the likelihood curve provides information about the uncertainty in the point estimate. In this chapter, I use estimation of the distance between two sequences under the Jukes and Cantor model [23] as an example to contrast the likelihood and Bayesian methodologies (see Chapter 2, this volume for more about likelihood methods in phylogenetics).

Suppose x of the n sites are different between the two sequences, with the proportion of different sites to be x/n. The distance is the expected number of nucleotide substitutions per site, $\theta = \lambda t$, where λ is the substitution rate and t is the time that separates the two sequences—since rate and time are confounded, we estimate one single parameter θ using the data x. The probability that a site is different between two sequences separated by distance θ is

$$p = \frac{3}{4}(1 - e^{(-4/3)\theta}). \tag{3.1}$$

Thus the likelihood, or the probability of observing x differences out of n sites, is given by the binomial probability

$$L(\theta; x) = f(x \mid \theta) = Cp^x(1-p)^{n-x}, \tag{3.2}$$

where $C = n!/[x!(n-x)!]$ is constant (independent of parameter θ) and can be ignored. By setting $\mathrm{d}L/\mathrm{d}\theta = 0$ or $\mathrm{d}\ell/\mathrm{d}\theta = 0$, one can determine that the likelihood is maximized at

$$\hat{\theta} = -\frac{3}{4}\log\left(1 - \frac{4}{3} \times \frac{x}{n}\right). \tag{3.3}$$

Thus $\hat{\theta}$ is the maximum likelihood estimate (MLE) of θ. This is the familiar Jukes–Cantor distance formula [23]. In most problems in molecular phylogenetics to which maximum likelihood is applied, the solution is not analytical and numerical algorithms are needed to find the MLEs.

The MLEs are invariant to transformations or re-parametrizations. The MLE of a function of parameters is the same function of the MLEs of the parameters: $\hat{h}(\theta) = h(\hat{\theta})$. For example, we can use the expected proportion of different sites p as the parameter; this is still a measure of distance although it is non-linear with time. Its MLE is $\hat{p} = x/n$ from the binomial likelihood (equation (3.2)). We can then view θ as a function of p through equation (3.1), and obtain its MLE $\hat{\theta}$, as given in equation (3.3). Whether we use p or θ as the parameter, the same inference is made, and the same log likelihood is achieved: $\ell(\hat{p}) = \ell(\hat{\theta}) = x\log(x/n) + (n-x)\log((n-x)/n)$.

As an example, suppose $x = 10$ differences are observed out of $n = 100$ sites. The log-likelihood curves are shown in Fig. 3.1(a) and (b) for parameters θ and p, respectively. The log likelihood is maximized at $\hat{\theta} = 0.107326$ and $\hat{p} = x/n = 0.1$, with $\ell(\hat{\theta}) = \ell(\hat{p}) = -32.508$.

Two approaches can be used to calculate a confidence interval for the MLE. The first relies on the theory that $\hat{\theta}$ is asymptotically normally distributed around the true θ when the sample size $n \to \infty$. This is equivalent to using a quadratic function to approximate the log likelihood around the MLE. The variance of the asymptotic normal distribution can be calculated using the curvature of the log-likelihood surface around the MLE:

$$\mathrm{var}(\hat{\theta}) = -\left[\frac{\mathrm{d}^2\ell}{\mathrm{d}\theta^2}\right]^{-1} = \frac{9\hat{p}(1-\hat{p})}{(3-4\hat{p})^2 n}. \tag{3.4}$$

Thus an approximate 95% confidence interval for θ can be constructed as $\hat{\theta} \pm 1.96\sqrt{\mathrm{var}(\hat{\theta})}$. For our example of $x = 10$ differences in $n = 100$ sites, we have $\mathrm{var}(\hat{\theta}) = 0.001198$, and the 95% confidence interval is $0.10733 \pm 1.96 \times 0.06784$ or $(0.03948, 0.17517)$. Similarly, $\mathrm{var}(\hat{p}) = \hat{p}(1-\hat{p})/n = 0.0009$, so that the 95% confidence interval for p is $(0.04120, 0.15880)$. Note that those two intervals do not match each other.

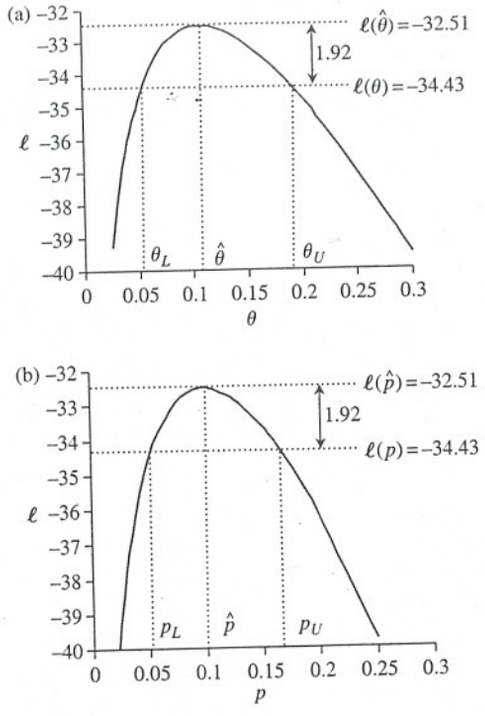

FIG. 3.1. Log-likelihood curves for estimation of sequence distance θ or p under the JC69 model [23]. Log-likelihood curves as a function of the sequence distance θ (a) or p (b). The data are two sequences, each of length $n = 100$ with $x = 10$ different sites. The likelihood interval is constructed by lowering the log likelihood ℓ from the optimum value by 1.92.

A second approach is based on the result that the likelihood ratio test statistic, $2[\ell(\hat{\theta}) - \ell(\theta)]$, where θ is the true parameter and $\hat{\theta}$ is the MLE, has a χ^2_1 distribution in large samples. Thus, we can lower the log likelihood by, say, $\frac{1}{2}\chi^2_{1,5\%} = 3.84/2 = 1.92$ from $\ell(\hat{\theta})$, to construct a 95% likelihood interval (θ_L, θ_U) (Fig. 3.1(a)). Thus at $\ell = \ell(\hat{\theta}) - 1.92 = -34.43$, the likelihood interval is found to be $(0.05327, 0.19119)$ for θ. Note that this interval is asymmetrical and is shifted to the right compared with the interval based on the normal approximation, due to the steeper drop of log likelihood and thus more information on the left side of $\hat{\theta}$ than on the right side. The corresponding likelihood interval for p is $(0.05142, 0.16876)$. This approach in general gives more reliable intervals than the normal approximation to MLEs. The normal approximation works well for some parameterizations but not for others; the use of the likelihood interval is equivalent to using the best parametrization.

The likelihood method may run into problems when the model involves too many parameters. If the number of parameters increases without bound with

the increase of the sample size, the MLEs may not even be consistent. Dealing with the so-called nuisance parameters is also a difficult area for likelihood. For example, if we are interested in the sequence distance under the substitution model of Kimura [24], we might consider distance θ as the parameter of interest, while the transition/transversion rate ratio κ is a nuisance parameter. Similarly, if our interest is in the phylogeny for a group of species, branch lengths as well as all parameters in the substitution model are nuisance parameters. Perhaps the biggest problem for the application of likelihood to molecular phylogeny reconstruction is the unconventional nature of the tree topology parameter, and the resulting difficulties in attaching a confidence interval for the maximum likelihood tree [51] (see Chapter 4, this volume).

3.2 The Bayesian paradigm

The central idea of Bayesian inference is that parameters θ have distributions. Before the data are observed, θ have a prior distribution $f(\theta)$. This is combined with the likelihood or the probability of the data given the parameters, $f(D \mid \theta)$, to give the posterior distribution, $f(\theta \mid D)$, through the Bayes theorem

$$f(\theta \mid D) = \frac{f(\theta)f(D \mid \theta)}{f(D)} = \frac{f(\theta)f(D \mid \theta)}{\int f(\theta)f(D \mid \theta)\,\mathrm{d}\theta}. \tag{3.5}$$

The marginal probability of the data, $f(D)$, is a normalizing constant, to make $f(\theta \mid D)$ integrate to one. Equation (3.5) thus says that the posterior $f(\theta \mid D)$ is proportional to the prior $f(\theta)$ times the likelihood $f(D \mid \theta)$. Or equivalently, the posterior information is the sum of the prior information and the sample information.

The posterior distribution is the basis for all Bayesian inference concerning θ. For example, the mean, median, or mode of the distribution can be used as the point estimate. For interval estimation, one can use the interval encompassing the highest 95% of the density mass as the 95% highest posterior density (HPD) interval. This works even if there are multiple peaks in the distribution; the interval may include disconnected regions. For a single-moded posterior density, the 2.5% and 97.5% quantiles can be used to construct the 95% equal-tail credibility interval (CI). In general, the posterior expectation of any function of the parameters, $h(\theta)$, is constructed as $E[h(\theta) \mid D)] = \int h(\theta)f(\theta \mid D)\,\mathrm{d}\theta$.

Consider estimation of sequence distance θ under the JC69 model [23] using the data of $x = 10$ differences out of $n = 100$ sites. Suppose we use an exponential prior $f(\theta) = \mu^{-1}e^{(-\theta/\mu)}$, with mean $\mu = 0.1$. The posterior distribution of θ is

$$f(\theta \mid x) = \frac{f(\theta)f(x \mid \theta)}{f(x)} = \frac{f(\theta)f(x \mid \theta)}{\int f(\theta)f(x \mid \theta)\,\mathrm{d}\theta}, \tag{3.6}$$

where the likelihood $f(x \mid \theta)$ is given in equation (3.2). It seems awkward, although possible, to calculate the integral for $f(x)$ in equation (3.6) analytically. Instead I use Mathematica to evaluate it numerically. Figure 3.2 shows the resulting posterior density, plotted together with the prior and scaled likelihood. In this

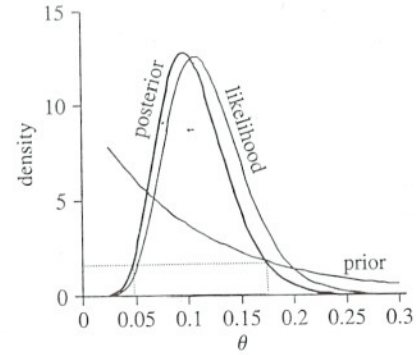

FIG. 3.2. Prior and posterior densities for sequence distance θ under the JC69 model. The likelihood is also shown, rescaled to match up with the posterior density. The data are two sequences, each of length $n = 100$ with $x = 10$ different sites. The 95% highest posterior density interval is (0.04758, 0.17260), indicated on the graph.

case the posterior is dominated by the likelihood. The posterior mean is found to be 0.10697, with standard deviation 0.03290. The 95% equal-tail credibility interval is (0.05284, 0.18077), while the 95% HPD interval is (0.04758, 0.17260).

The Bayesian paradigm also provides a natural way of dealing with nuisance parameters. Let $\theta = \{\lambda, \eta\}$, with λ to be the parameters of interest and η the nuisance parameters. The joint conditional distribution of λ and η given the data is

$$f(\lambda, \eta \mid D) = \frac{f(\lambda, \eta)f(D \mid \lambda, \eta)}{f(D)} = \frac{f(\lambda, \eta)f(D \mid \lambda, \eta)}{\iint f(\lambda, \eta)f(D \mid \lambda, \eta)\,\mathrm{d}\lambda\,\mathrm{d}\eta} \tag{3.7}$$

from which the (marginal) posterior density of λ can be obtained as

$$f(\lambda \mid D) = \int f(\lambda, \eta \mid D)\,\mathrm{d}\eta. \tag{3.8}$$

3.3 Prior

Specification of the prior distribution for parameters, and indeed the need for such specification is where all controversies surrounding Bayesian inference lies. If the physical process can be used to model uncertainties in the quantities of interest, it is standard in the likelihood framework to treat such quantities as random variables, and derive their conditional probability distribution given the data. An example relevant to this chapter is the use of the Yule branching process [5] and the birth–death process [34] to specify the probability distributions of phylogenies. The parameters in the models are the birth and death rates, estimated from the marginal likelihood, which averages over the tree topologies and branch lengths, while the phylogeny is estimated from the conditional

probability distribution of phylogenies given the data. The controversy arises when no physical model is available to specify the distribution of parameters, and when subjective beliefs or diffuse distributions are used as "vague" priors. Modern terminology does not distinguish whether or not the prior is based on a model of the physical process; in either case the quantities of interest are considered parameters, the approach considered Bayesian, and the conditional probability is known as the posterior probability.

Approaches for specifying the prior include (1) use of a physical model, as mentioned above, (2) use of past observations of the parameters in similar situations, and (3) subjective beliefs of the researcher. To avoid undue influence of the prior on the posterior, uniform distributions are often used as vague priors. For a discrete parameter that can take m possible values, this means assigning probability $1/m$ to each element. For a continuous parameter, this means a uniform distribution over the range of the parameters. However, saying that distance θ is equally likely to be any value between 0 and 10 is not the same as saying that nothing is known about θ, so one should not consider any prior as entirely non-informative. Another criticism is that unlike the MLEs, the prior is not invariant to reparametrizations. For example, a uniform prior for parameter p is very different from a uniform prior for θ (see below).

Another class of priors is the conjugate priors. Here the prior and the posterior have the same distributional form, and the role of the data or likelihood is to update the parameters in that distribution. Well-known examples include (1) the binomial (n, p) distribution of data with a beta prior for the probability parameter p; (2) poisson(λ) distribution of data with a gamma prior for the rate parameter λ; and (3) normal distribution of data $N(\mu, \sigma^2)$ with a normal prior for the mean μ. In our example of estimating sequence distance under the JC69 model, if we use the probability of different sites p as the distance, we can assign a beta prior beta(α, β). When the data have x differences out of n sites, the posterior distribution of p is beta$(\alpha + x, \beta + n - x)$. This result also illustrates the information contained in the beta prior: beta(α, β) is equivalent to observing α differences out of $\alpha + \beta$ sites. Conjugate priors are possible only for special combinations of the prior and likelihood. They are theoretically convenient as the integrals are tractable analytically, but they may not be realistic models for the problem at hand. Conjugate priors have not found a use in molecular phylogenetics (except for the trivial one above), as the problem is typically too complex.

When the prior distribution involves unknown parameters, one can assign priors for them, called hyper-priors. Unknown parameters in the hyper-prior can have their own priors. This is known as the hierarchical or full Bayesian approach. Typically one does not go beyond two or three levels, as the effect will become unimportant. For example, the mean μ in the exponential prior in our example of distance calculation under JC69 in equation (3.6) can be assigned a hyper-prior. An alternative is to estimate the hyper-parameters from the marginal likelihood, and use them in posterior probability calculation for parameters of interest. This is known as the empirical Bayesian approach. For example, μ can be estimated by maximizing $f(x \mid \mu) = \int f(\theta \mid \mu) f(x \mid \theta)\, \mathrm{d}\theta$, and the estimate can be used to calculate $f(\theta \mid x)$ in equation (3.6). Empirical Bayesian approach has been used widely in molecular phylogenetics, for example, to estimate evolutionary rates at sites [55], to reconstruct ancestral DNA or protein sequences on a phylogeny [52], to identify amino acid residues under positive Darwinian selection [31], to infer secondary structure categories of a protein sequence [13], and to construct sequence alignments under models of insertions and deletions [46, 47].

An important question in real data analysis is whether the posterior is sensitive to the prior. It is always prudent to assess the influence of the prior. If the posterior is dominated by the data, the choice of the prior is inconsequential. When this is not the case, the effect of the prior has to be assessed carefully and reported. Due to advances in computational algorithms (see below), the Bayesian methodology is now very powerful and allows the researcher to fit sophisticated parameter-rich models. As a result, the researcher might be tempted to add parameters that are barely identifiable [33], and the posterior may be unduly influenced by some aspects of the prior even without the knowledge of the researcher. In our example of distance estimation under the JC69 model, identifiability problems will arise if we attempt to estimate both the substitution rate λ and time t instead of one parameter θ. It is thus important for the researcher to understand which aspects of the data provide information about the parameters, what parameters are knowable and what are not, to avoid overloading the model with parameters.

3.4 Markov chain Monte Carlo

Until recently, computational difficulties had prevented the use of the Bayesian method as a general inference methodology. For most problems, the prior and the likelihood are easy to calculate, but the marginal probability of the data $f(D)$, that is, the normalizing constant, is hard to calculate. Except for trivial problems such as cases involving conjugate priors, analytical results are unavailable. We have noted above the difficulty of calculating the marginal likelihood $f(D)$ (in equation (3.6)) in our extremely simple problem of distance estimation. More complex Bayesian models can involve hundreds or thousands of parameters and high-dimensional integrals have to be evaluated (see equations (3.7) and (3.8)). For example, to calculate posterior probabilities for phylogenetic trees, one has to evaluate the marginal probability of data $f(D)$, which is a sum over all possible tree topologies and integration over all branch lengths in those trees and over all parameters in the substitution model. The breakthrough is the development of Markov chain Monte Carlo (MCMC) algorithms, which provide a powerful method for achieving Bayesian computation.

3.4.1 *Metropolis–Hastings algorithm*

Here we describe the algorithm of Metropolis *et al.* [30]. The goal is to generate a Markov chain, whose states are the parameters θ, and whose steady-state

(stationary) distribution is $\pi(\theta) = f(\theta \mid D)$, the posterior distribution of θ. Suppose the current state of the Markov chain is θ. The algorithm proposes a new state θ^* through a proposal density or jumping kernel $q(\theta^* \mid \theta)$, which is symmetrical: $q(\theta^* \mid \theta) = q(\theta \mid \theta^*)$. For example, one can use a uniform distribution around θ, so that $\theta^* = U(\theta - w/2, \theta + w/2)$, with w controlling the size of steps taken. This is a sliding window with window size w. The candidate state θ^* is accepted with probability

$$\alpha = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta)}\right). \tag{3.9}$$

If the new state θ^* is accepted, the chain moves to θ^*. If it is rejected, the chain stays at the current state θ. Both acceptance and rejection are counted as an iteration, and the procedure is repeated for many iterations. The values of θ over iterations generated this way form a Markov chain, as they satisfy the Markovian property that "given the present, the future is independent of the past." This Markov chain has $\pi(\theta)$ as the stationary distribution as long as the proposal density $q(. \mid .)$ specifies an irreducible and aperiodic chain. In other words, $q(. \mid .)$ should allow the chain to reach any state from any other state, and that the chain should not have a period.

Intuitively, one may think of the algorithm as describing a wanderer climbing a hill, the height at location θ being the target density $\pi(\theta)$. A random step in a random direction is chosen from the current location. If the step is uphill, that is, if $\pi(\theta^*) > \pi(\theta)$, it is always taken. However, if the step is downhill, it is not rejected straightaway but instead accepted with probability $\pi(\theta^*)/\pi(\theta) < 1$. If the wanderer is allowed to wander around for a very long time, he will explore the hill extensively and spend time in each location θ in proportion to the height of that location $\pi(\theta)$. Thus a sample of his visits can be used to estimate the target distribution $\pi(\theta)$.

Hastings [18] extended the Metropolis algorithm to allow the use of asymmetrical proposal densities, that is, if $q(\theta^* \mid \theta) \neq q(\theta \mid \theta^*)$. This involves a simple correction in calculation of the acceptance probability

$$\alpha = \min\left(1, \frac{\pi(\theta^*)q(\theta \mid \theta^*)}{\pi(\theta)q(\theta^* \mid \theta)}\right). \tag{3.10}$$

We might suppose that the wanderer has a tendency to move north, and takes a northward step three times as likely as a southward step. Then by accepting northward moves only $\frac{1}{3}$ times as often as southward moves, the Markov chain will still recover the correct target distribution $\pi(\theta)$ even if the proposal density is biased. The correction term, $q(\theta \mid \theta^*)/q(\theta^* \mid \theta)$, is called the proposal ratio or the Hastings ratio.

When the MCMC algorithm is used to approximate the posterior distribution of parameters θ, we have $\pi(\theta) = f(\theta \mid D) = f(\theta)f(D \mid \theta)/f(D)$, so that

$$\frac{\pi(\theta^*)}{\pi(\theta)} = \frac{f(\theta^*)f(D \mid \theta^*)}{f(\theta)f(D \mid \theta)}.$$

Importantly note that the normalizing constant $f(D)$ in equation (3.5) cancels. The acceptance probability is thus

$$\begin{aligned}\alpha &= \min\left(1, \frac{f(\theta^*)}{f(\theta)} \times \frac{f(D \mid \theta^*)}{f(D \mid \theta)} \times \frac{q(\theta \mid \theta^*)}{q(\theta^* \mid \theta)}\right) \\ &= \min(1, \text{ prior ratio} \times \text{likelihood ratio} \times \text{proposal ratio}).\end{aligned} \tag{3.11}$$

In typical applications of MCMC algorithms to molecular phylogenetics, the prior ratio $f(\theta^*)/f(\theta)$ is easy to calculate. The likelihood ratio $f(D \mid \theta^*)/f(D \mid \theta)$ is often easy to calculate as well even though computationally expensive. The proposal ratio $q(\theta \mid \theta^*)/q(\theta^* \mid \theta)$ affects greatly the efficiency of the MCMC algorithm. So much of practical effort is spent on developing good proposal algorithms.

Here we use the example of distance estimation under the JC69 model to explain MCMC algorithms. Those who have not written any Bayesian MCMC program are invited to implement the algorithm below, using any programming language such as C/C++, Java, Basic, or Mathematica. The data are $x = 10$ differences out of $n = 100$ sites. We use an exponential prior

$$f(\theta \mid \mu) = \frac{1}{\mu}e^{-(1/\mu)\theta}$$

with $\mu = 0.1$. The proposal algorithm uses a sliding window of size w.

1. Initialize: $n = 100$, $x = 10$, $w = 0.01$.
2. Initial state $\theta = 0.5$.
3. Propose a new state as $\theta^* \sim U(\theta - w/2, \theta + w/2)$. That is, generate a $U(0, 1)$ random number r, and set $\theta^* = \theta - w/2 + wr$. If $\theta^* < 0$, set $\theta^* = -\theta^*$.
4. Calculate the acceptance probability, using equations (3.1) and (3.2) to calculate the likelihood $f(x \mid \theta)$.

$$\alpha = \min\left(1, \frac{f(\theta^* \mid \mu)}{f(\theta \mid \mu)} \times \frac{f(x \mid \theta^*)}{f(x \mid \theta)}\right).$$

5. Accept or reject the proposal θ^*. Draw $r \sim U(0, 1)$. If $r < \alpha$ set $\theta = \theta^*$. Otherwise set $\theta = \theta$.
6. Go to step 3.

Figures 3.3(a) and (b) show the first 500 iterations of five independent chains, starting from different initial values and using different window sizes. Figure 3.3(c) shows the posterior probability density estimated from a long chain with 10 million iterations. This is indistinguishable from the distribution calculated using numerical integration (Fig. 3.2).

A number of variations to the general Metropolis–Hastings algorithm exist. Below we mention three commonly used ones: the single-component Metropolis–Hastings algorithm, the Gibbs sampler, and Metropolis-coupled MCMC or MC^3.

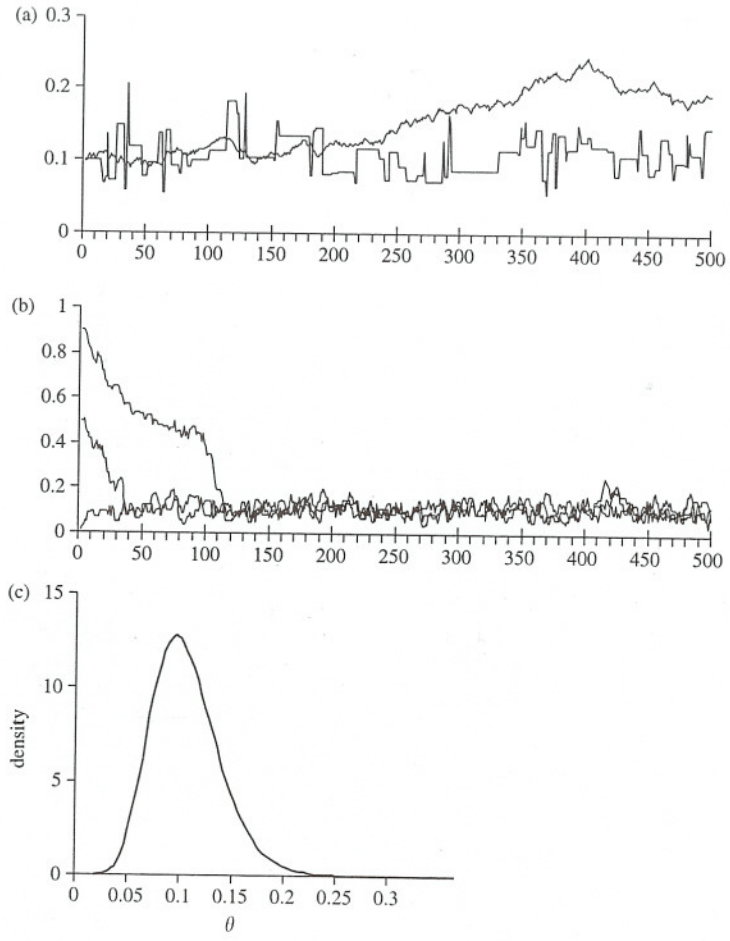

FIG. 3.3. MCMC runs for estimating sequence distance θ under the JC69 substitution model. The data consists of $x = 10$ differences between two sequences of $n = 100$ sites. (a) Two chains with the window size either too small ($w = 0.01$) or too large ($w = 1$). Both chains started at $\theta = 0.1$. The chain with $w = 0.01$ has an acceptance rate of 97%, so that almost every proposal is accepted. However, this chain takes tiny baby steps and mixes very poorly. The other chain, with $w = 1$, has an acceptance rate of 20%, so that 80% of proposals are rejected. The chain often stays at the same state for many iterations without a move. This window size is slightly too large. Further experiment shows that the window size $w = 0.2$ leads to an acceptance rate of 48%, and is near optimum (see text). (b) Three chains started from $\theta = 0.01, 0.5$, and 1. The window size is 0.1, with an acceptance rate of 70%. It appears that after about 120 iterations, the three chains become indistinguishable and have reached stationarity, so that a burn-in of 200 iterations should be sufficient for those chains. (c) Posterior density estimated from a long chain (with 10,000,000 iterations) with window size $w = 0.1$, estimated by kernel density smoothing [40].

3.4.2 *Single-component Metropolis–Hastings algorithm*

Simple single-parameter problems are straightforward to deal with using the likelihood methodology. The advantage of Bayesian inference mostly lies in the ease with which it can deal with sophisticated multi-parameter models. In particular, Bayesian "marginalization" of nuisance parameters (equation (3.8)) provides an attractive way of accommodating variation in the data that we are not really interested in. In MCMC algorithms for such multi-parameter models, it is often unfeasible or computationally too complicated to update all parameters in θ simultaneously. Instead, it is more convenient to divide θ into components or blocks, of possibly different dimensions, and then update those components one by one. Different proposals are often used to update different components. This is known as "blocking." Many models have a structure of conditional independence, and blocking often leads to computational efficiency.

A variety of strategies are possible concerning the order of updating the components. One can use a fixed order, or a random permutation of the components. There is no need to update every component in every iteration. One can also select components for updating with fixed probabilities. However, the probabilities should be fixed and not dependent on the current state of the Markov chain, as otherwise the stationary distribution may no longer be the target distribution $\pi(\cdot)$. It is advisable to update highly correlated components more frequently. It is also advantageous to group into one block components that are highly correlated in the posterior density, and update them simultaneously using a proposal density that accounts for the correlation (see below).

3.4.3 *Gibbs sampler*

The *Gibbs sampler* [11] is a special case of the single-component Metropolis–Hastings algorithm. The proposal distribution for updating the ith component is the conditional distribution of the ith component given all the other components. This proposal leads to an acceptance probability of 1; that is, all proposals are accepted. The Gibbs sampler has been widely used, especially in linear models involving normal prior and posterior densities. However, it has not been used in molecular phylogenetics as it is in general impossible to obtain the conditional distributions analytically.

3.4.4 *Metropolis-coupled MCMC*

If the target distribution has multiple peaks, separated by low valleys, the Markov chain may have difficulties in moving from one peak to another. As a result, the chain may get stuck on one peak and the resulting samples will not approximate the posterior density correctly. This is a serious practical concern for phylogeny reconstruction, as multiple local peaks are known to exist in the tree space during heuristic tree search under the maximum parsimony (MP), maximum likelihood (ML), and minimum evolution (ME) criteria, and the same can be expected for stochastic tree search using MCMC. Some strategies have been proposed to improve mixing of Markov chains in presence of multiple local

peaks in the posterior density. One such algorithm is the Metropolis-coupled MCMC or MCMCMC (MC^3) algorithm suggested by Geyer [12].

In this algorithm, m chains are run in parallel, with different stationary distributions $\pi_j(\cdot), j = 1, 2, \ldots, m$, where $\pi_1(\cdot) = \pi(\cdot)$ is the target density, while $\pi_j(\cdot)$, $j = 2, 3, \ldots, m$ are chosen to improve mixing. For example, one can use incremental heating of the form

$$\pi_j(\theta) = \pi(\theta)^{1/[1+\lambda(j-1)]}, \quad \lambda > 0, \tag{3.12}$$

so that the first chain is the cold chain with the correct target density, while chains $2, 3, \ldots, m$ are heated chains. Note that raising the density $\pi(\cdot)$ to the power $1/T$ with $T > 1$ has the effect of flattening out the distribution, similar to heating a metal. In such a distribution, it is easier to traverse between peaks across the valleys than in the original distribution. After each iteration, a swap of states between two randomly chosen chains is proposed through a Metropolis–Hastings step. Let $\theta^{(j)}$ be the current state in chain j, $j = 1, 2, \ldots, m$. A swap between the states of chains i and j is accepted with probability

$$\alpha = \min\left(1, \frac{\pi_i(\theta_j)\pi_j(\theta_i)}{\pi_i(\theta_i)\pi_j(\theta_j)}\right). \tag{3.13}$$

At the end of the run, output from only the cold chain is used, while those from the hot chains are discarded. Heuristically, the hot chains will visit the local peaks rather easily, and swapping states between chains will let the cold chain occasionally jump valleys, leading to better mixing. However, if $\pi_i(\theta)/\pi_j(\theta)$ is very unstable, proposed swaps will seldom be accepted; this is the reason for using several chains which differ only incrementally. An obvious disadvantage of the algorithm is that m chains are run but only one chain is used for inference. MC^3 is ideally suited to implementation on parallel machines or network workstations, since each chain will in general require about the same amount of computation per iteration, and interactions between chains are minimal.

3.5 Simple moves and their proposal ratios

The proposal ratio is separate from the likelihood or the prior and is solely dependent on the proposal algorithm. Thus simple proposals can be used in a variety of Bayesian inference problems. As mentioned earlier, the proposal density has only to specify an aperiodic recurrent Markov chain to guarantee convergence of the MCMC algorithm. One can easily construct such chains and it is also typically easy to verify that the proposal density satisfies those conditions. For a discrete parameter that takes a set of values, calculation of the proposal ratio often amounts to counting the number of candidate elements in the source and target states, which is easy. Calculation for continuous parameters requires more care. In this section, I list a few commonly used proposals and their proposal ratios. I may use x instead of θ to represent the state of the chain.

Two results are particularly useful in deriving proposal ratios. So I mention them in the form of two theorems, before describing the proposals. The first result concerns the distribution of functions of random variables (see, for example, [15]: pp. 107–112).

Theorem 3.1 *(a) If x is a random variable with density $f(x)$, and $y = y(x)$ and $x = x(y)$ is a one-to-one mapping between x and y, then the random variable y has the density*

$$f(y) = f(x(y)) \times \left|\frac{\mathrm{d}x}{\mathrm{d}y}\right|. \tag{3.14}$$

(b) The multivariate version is very similar. Suppose random variables $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$ and $\mathbf{y} = \{y_1, y_2, \ldots, y_m\}$ constitute a one-to-one mapping through $y_i = y_i(\mathbf{x})$, and $x_i = x_i(\mathbf{y})$, $i = 1, 2, \ldots, m$, and that $\mathbf{x}$ has probability density $f(\mathbf{x})$. Then $\mathbf{y}$ has density

$$f(\mathbf{y}) = f(\mathbf{x}(\mathbf{y})) \times |J(\mathbf{y})|, \tag{3.15}$$

where $|J(\mathbf{y})|$ is the absolute value of the Jacobian determinant of the transform

$$J(\mathbf{y}) = \frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_m} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_m}{\partial y_1} & \frac{\partial x_m}{\partial y_2} & \cdots & \frac{\partial x_m}{\partial y_m} \end{vmatrix}. \tag{3.16}$$

As an example, suppose that the probability of different sites p has a uniform prior distribution $f(p) = 4/3$, $0 \le p < 3/4$. What is the distribution of the sequence distance θ? From equation (3.1), we have $\mathrm{d}p/\mathrm{d}\theta = e^{(-4/3)\theta}$. Thus the distribution of θ is $f(\theta) = 4/3 \times e^{(-4/3)\theta}$, $0 \le \theta < \infty$. This is the exponential distribution with mean $3/4$.

The second useful result gives the proposal ratio when the proposal is made though transformed variables.

Theorem 3.2 *Suppose the Markov chain is run using the original variables $x_1, x_2, \ldots, x_m$, but the proposal is through transformed variables $y_1, y_2, \ldots, y_m$. Then*

$$\frac{q(\mathbf{x} \mid \mathbf{x}^*)}{q(\mathbf{x}^* \mid \mathbf{x})} = \frac{q(\mathbf{y} \mid \mathbf{y}^*)}{q(\mathbf{y}^* \mid \mathbf{y})} \times \frac{|J(\mathbf{y}^*)|}{|J(\mathbf{y})|}. \tag{3.17}$$

The proposal ratio in the original variables is the proposal ratio in the transformed variables times the ratio of the Jacobian.

The statement can be proved by noting that

$$q(\mathbf{y}^* \mid \mathbf{y}) = q(\mathbf{y}^* \mid \mathbf{x}) = q(\mathbf{x}^* \mid \mathbf{x}) \times J(\mathbf{y}^*). \tag{3.18}$$

The first equation is because conditioning on $\mathbf{y}$ is equivalent to conditioning on $\mathbf{x}$ due to the one-to-one mapping. The second equation applies Theorem 3.1(b) to derive the density of $\mathbf{y}^*$ as functions of $\mathbf{x}^*$.

3.5.1 *Sliding window using uniform proposal*

This proposal chooses the new state x^* as a random variable from a uniform distribution around the current state x:

$$x^* \sim U\left(x - \frac{w}{2}, x + \frac{w}{2}\right). \tag{3.19}$$

The window size w is a fixed constant, chosen to achieve a reasonable acceptance rate. The proposal ratio is 1 since $q(x^* \mid x) = q(x \mid x^*)$. If x is constrained in the interval (a, b) and x^* is outside the range, the excess is reflected back into the interval; that is, if $x^* < a$, x^* is reset to $a + (a - x^*) = 2a - x^*$, and if $x^* > b$, x^* is reset to $b - (b - x^*) = 2b - x^*$. The proposal ratio is 1 even with reflection, because if x can reach x^* through reflection, x^* can reach x through reflection as well. The window size w should be smaller than the range $b - a$. Note that it is incorrect to simply set the unfeasible proposed values to a or b.

3.5.2 *Sliding window using normally distributed proposal*

This algorithm uses a normal proposal density centred around the current state; that is, x^* has a normal distribution with mean x and variance σ^2, with σ controlling the step size

$$x^* \sim N(x, \sigma^2). \tag{3.20}$$

As $q(x^* \mid x) = (1/(\sigma\sqrt{2\pi}))\exp\{-(x^* - x)^2/(2\sigma^2)\} = q(x \mid x^*)$, the proposal ratio is 1. This proposal works also if x is constrained in the interval (a, b). If x^* is outside the range, the excess is reflected back into the interval, and the proposal ratio remains one. Both with and without reflection, the number of routes from x to x^* is the same as from x^* to x, and the densities are the same in the opposite directions, even if not between the routes. Note that sliding window algorithms using either uniform or normal jumping kernels are Metropolis algorithms with symmetrical proposals.

How do we choose σ? Suppose the target density is the standard normal $N(0, 1)$, and the proposal is $x^* \sim N(x, \sigma^2)$. A large σ will cause most proposals to be in unreasonable regions of the parameter space and be rejected. The chain then stays at the same state for a long time, causing high correlation. A σ too small means that the proposed states are very close to the current state, and most proposals will be accepted. However, the chain baby-walks in the same region of the parameter space for a long time, leading again to high correlation. Proposals that minimize the auto correlations are thus optimal.

More formally, consider the sample mean $\hat{\theta} = (1/N)\sum x^{(t)}$, where $x^{(t)}$ is the state in iteration t, with $t = 1, 2, \ldots, N$. With independent sampling, $\text{var}(\hat{\theta}) = 1/N$. The large-sample variance of a dependent sample is

$$\text{var}(\hat{\theta}) = \frac{1}{N}[1 + 2(\rho_1 + \rho_2 + \rho_3 + \cdots)], \tag{3.21}$$

where ρ_k is the autocorrelation of the Markov chain at lag k. In effect, a dependent sample of size N is equivalent to an independent sample of size $N/[1 + 2(\rho_1 + \rho_2 + \rho_3 + \cdots)]$. By minimizing $\text{var}(\hat{\theta})$ in equation (3.21), Gelman *et al.* [9] found the optimum σ to be about 2.4. Thus if the target density is a general normal density $N(\mu, \tau^2)$, the optimum proposal density should be $N(x, \tau^2\sigma^2)$ with $\sigma = 2.4$. As τ is unknown, one can monitor the acceptance rate or jumping probability, which is slightly below 0.5 at the optimum σ.

3.5.3 *Sliding window using normal proposal in multidimensions*

If the target density is a m-dimensional standard normal with density $N_m(0, I)$ where I is a $m \times m$ identity matrix, one can use the proposal density $q(\mathbf{x}^* \mid \mathbf{x}) = N_m(\mathbf{x}, \mathbf{I}\sigma^2)$. The proposal ratio is one. The Gelman *et al.* [9] analysis suggests that the optimum scale factor σ is 2.4, 1.7, 1.4, 1.2, 1, 0.9, 0.7 for m =1, 2, 3, 4, 6, 8, 10, respectively, with an optimal acceptance rate of about 0.26 for $m > 6$. It is interesting to note that at low dimensions, the optimal proposal density is over-dispersed relative to the target density, suggesting that one should take big steps, while at high dimensions, one should use under-dispersed proposal densities and take small steps. In general one should try to achieve an acceptance rate of about 20–70% for 1-D proposals, and 15–40% for multi-dimensional proposals.

Those results are more useful than for just standard normal densities. When the target density is $\mathbf{x} \sim N_m(\mu, \mathbf{S})$, with variance–covariance matrix $\mathbf{S}$, several strategies can be used. One is to reparametrize the model using $\mathbf{y} = \mathbf{S}^{-1/2}\mathbf{x}$ as parameters, where $\mathbf{S}^{-1/2}$ is the square root of $\mathbf{S}^{-1}$. Note that $\mathbf{y}$ has unit variance, and the above proposal can be used. The second strategy is to propose new states using the transformed variables $\mathbf{y}$, that is, $q(\mathbf{y}^* \mid \mathbf{y}) = N_m(\mathbf{y}, \mathbf{I}\sigma^2)$, and then derive the proposal ratio in the original variables $\mathbf{x}$. The proposal ratio is one according to Theorem 3.2. A third approach is to simply use the proposal $x^* \sim N_m(\mathbf{x}, \sigma^2\mathbf{S})$, where σ^2 is chosen according to the above discussion. The three approaches are equivalent and all of them take care of possible differences in the scales and possible correlations among the variables. In real data analysis, $\mathbf{S}$ is unknown. One can perform short runs of the Markov chain to obtain an estimate $\hat{\mathbf{S}}$ of the variance–covariance matrix in the posterior density, and then use it in the proposal. If $\mathbf{S}$ is estimated in the same run, samples taken to estimate $\mathbf{S}$ should be discarded. If the normal distribution is a good approximation to the posterior density, those guidelines should work well.

3.5.4 *Proportional shrinking and expanding*

For a variable that is always positive or always negative, this proposal multiplies the current value by a random number that is around 1. Let

$$\begin{aligned} c &= e^{\epsilon(r - 1/2)}, \\ x^* &= cx, \end{aligned} \tag{3.22}$$

where $r \sim U(0, 1)$ and $\epsilon > 0$ is a small finetuning parameter. Note that x is shrunk or expanded depending on whether r is $<$ or $> 1/2$. To calculate the proposal ratio, derive the proposal density $q(x^* \mid x)$ through variable transform, noting that r and x^* are random variables while ϵ and x are constants. Since

$r = 1/2 + \log(x^*/x)/\epsilon$, and $\mathrm{d}r/\mathrm{d}x^* = 1/(\epsilon x^*)$, we have from Theorem 3.1(a)

$$q(x^* \mid x) = f(r(x^*)) \times \left|\frac{\mathrm{d}r}{\mathrm{d}x^*}\right| = \frac{1}{\varepsilon|x^*|}. \tag{3.23}$$

Similarly $q(x \mid x^*) = 1/\epsilon|x|$, so the proposal ratio is $q(x \mid x^*)/q(x^* \mid x) = c$.

This proposal can be used to shrink or expand many variables by the same factor c: $x_i^* = cx_i$, $i = 1, 2, \ldots, m$. This is useful for variables with a fixed order, such as the ages of nodes in a phylogenetic tree [48]. It is also effective in bringing all variables, such as branch lengths on a phylogeny, into the right scale if all of them are either too large or too small. Although all m variables are altered, the proposal is really in one dimension (along a line in the m-D space). We can derive the proposal ratio using the transform: $y_1 = x_1, y_i = x_i/x_1$, $i = 2, 3, \ldots, m$. The proposal changes y_1, but $y_2, \ldots, y_m$ remain unchanged. The proposal ratio in the transformed variables is c. The Jacobian is $J(y_1, y_2, \ldots, y_m) = |\partial \mathbf{x}/\partial \mathbf{y}| = y_1^{m-1}$. The proposal ratio in the original variables is thus $c \times (y_1^*/y_1)^{m-1} = c^m$, according to Theorem 3.2. Similarly, if the proposal multiplies m variables by c and divides n variables by c, the proposal ratio is $c^{m/n}$.

3.6 Monitoring Markov chains and processing output

3.6.1 *Diagnosing and validating MCMC algorithms*

An MCMC algorithm can suffer from two problems: slow convergence and poor mixing. The former means that it takes very long for the chain to reach stationarity. The latter means that the sampled states are highly correlated and the chain is very inefficient in exploring the parameter space. While it is often obvious that the proposal density $q(. \mid .)$ satisfies the required regularity conditions so that the MCMC is in theory guaranteed to converge to the target distribution, it is much harder to determine in real data problems whether the chain has reached stationarity. A number of heuristic methods have been suggested to diagnose the Markov chain. However, those diagnostics are able to reveal problems but unable to prove the correctness of the algorithm or implementation. Model misspecification, programming errors, and slow convergence all pose difficulties to program validation. A Bayesian MCMC program is notably harder to debug than a maximum likelihood program implementing a similar model. In a likelihood iteration, the convergence is to a point while in Bayesian MCMC, it is to a statistical distribution. In likelihood iteration, the log likelihood should always go up (at least if the optimizer is non-decreasing), and the gradient converges to zero. In a Bayesian MCMC algorithm, no statistics have a fixed direction of change. It is usually hard to independently calculate the posterior probability distribution. The temptation to use sophisticated models with excessive parameters in Bayesian modelling adds further difficulty. Often when the algorithm converges slowly or mixes poorly, it is difficult to decide whether this is due to faulty theory, buggy program, or inefficient but correct algorithm.

The following are some of the commonly used strategies for diagnosing and validating an MCMC program. (1) One can plot parameters of interest or their functions against the iterations. Such time-series plots can often reveal lack of convergence and/or poor mixing (see, for example, Figs. 3.3(a) and (b)). Often the chain appears to have converged with respect to some parameters but not to others. (2) The acceptance rate for each proposal should be neither too high nor too low. (3) It is advisable to run multiple chains from different starting points and make sure that the chains all converge to the same distribution. Gelman and Rubin's [10] statistic can be used to analyse multiple chains; see the next section. (4) Another technique is to run the chain without data, that is, to fix $f(D \mid \theta) = 1$ in equation (3.11). The posterior should then be the prior, which might be analytically available for comparison. (5) Simulation is also commonly used to validate MCMC algorithms. For example, Wilson *et al.* [49] simulated data under the prior to calculate the "hit probability" and "coverage probability" to validate their BATWING program. The former is the probability that the $100\alpha\%$ posterior credibility interval of a parameter includes the correct value. This should equal α. The latter is the average, across data replicates, of posterior coverage probability of a fixed interval. If this fixed interval has $100\alpha\%$ coverage probability in the prior, the average posterior coverage probability should also equal α [37, 49]. This is a more precise criterion for assessing interval coverage than the hit probability.

3.6.2 *Gelman and Rubin's potential scale reduction statistic*

Gelman and Rubin [10] suggested a diagnostic statistic called estimated "potential scale reduction," based on variance-components analysis of samples taken from several chains run using "over-dispersed" starting points. The idea is that after convergence, the within-chain variance should be indistinguishable from the between-chain variation while before convergence, the within-chain variance should be too small and the between-chain variance should be too large. The statistic can be used to monitor any or every parameter of interest. Let this be x, and its variance in the target distribution be τ^2. Suppose there are m chains, each run for n iterations, after the burn-in is discarded. Let x_{ij} be the parameter sampled at the jth iteration from the ith chain. Gelman and Rubin [10] defined the between-chain variance

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{x}_{i.} - \bar{x}_{..})^2, \tag{3.24}$$

and the within-chain variance

$$W = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2, \tag{3.25}$$

where $\bar{x}_{i.} = (1/n)\sum_{j=1}^{n} x_{ij}$ is the mean within the ith chain, and $\bar{x}_{..} = (1/m)\sum_{i=1}^{m} \bar{x}_{i.}$ is the overall mean. If all the m chains have reached stationarity and x_{ij} are samples from the same target density, both B and W are

unbiased estimates of τ^2, and so is their weighted mean

$$\hat{\tau}^2 = \frac{n-1}{n}W + \frac{1}{n}B. \tag{3.26}$$

If the m chains have not reached stationarity, W will be an *underestimate* of τ^2 since each chain has not traversed the whole parameter space and does not contain enough variation, while B will be an overestimate as the chains are from overdispersed starting points. Gelman and Rubin [10] showed that in this case $\hat{\tau}^2$ is also an overestimate of τ^2. The estimated "potential scale reduction" is defined as

$$\hat{R} = \frac{\hat{\tau}^2}{W}. \tag{3.27}$$

This should get smaller and approach one when the parallel chains reach the same target distribution. In real data problems, values of $\hat{R} < 1.1$ or 1.2 indicate convergence.

3.6.3 *Processing output*

Before we process the output, the beginning part of the chain before it has converged to the stationary distribution is discarded as "burn-in." Some programs do not sample every iteration but instead only takes a sample for every certain number of iterations. This is known as "thinning" the chain, as the thinned samples have reduced autocorrelations across iterations. While in theory sampling every iteration is more efficient (with smaller variances) than thinned samples, MCMC algorithms easily produce huge output files and it is often necessary to thin the chain to reduce the disk requirement.

After the burn-in, the samples taken from the MCMC can be summarized in a straightforward way. The sample mean, median, or mode can be used as a point estimate of the parameter, while the HPD or equal-probability credibility intervals can be constructed from the sample as well. For example, a 95% CI can be constructed by sorting the MCMC output for the variable and then using the 2.5% and 97.5% percentiles. The whole posterior distribution can be estimated by using a histogram, perhaps with further smoothing [40].

3.7 Applications to molecular phylogenetics

MCMC algorithms have been widely used in population genetics to analyse genetic data (DNA sequences, micro-satellites, etc.) under the coalescent models of variable complexity. Such applications include estimation of mutation rates (e.g. [4]), inference of population demographic processes or gene flow between subdivided populations (e.g. [3, 49]), and estimation of ancestral population sizes [35, 50], to name a few. See recent reviews by Griffiths and Tavaré [14] and Stephens and Donnelly [42]. Here I will discuss two major applications of Bayesian inference to molecular phylogenetics: estimation of phylogenetic trees and estimation of species divergence times under stochastic models of evolutionary rate change.

3.7.1 *Estimation of phylogenies*

Brief history. The Bayesian method was introduced to molecular phylogenetics by Rannala and Yang [34, 53], Mau and Newton [29], and Li *et al.* [28]. Those early studies assumed a constant rate of evolution (the molecular clock) as well as equal-probability prior for rooted trees either with or without ordered node ages (rooted trees or labelled histories). Since then, much more efficient MCMC algorithms have been implemented in the computer programs BAMBE [27] and MrBayes [21, 36]. The clock constraint is also relaxed, enabling phylogenetic inference under more realistic evolutionary models. A number of innovations have been introduced in those programs, adapting tree perturbation algorithms used in heuristic tree search (such as nearest-neighbour interchange, NNI, and subtree pruning and regrafting, SPR [44]), into flexible and efficient MCMC proposal algorithms for moving around in the tree space. In particular, MrBayes 3 has essentially incorporated all evolutionary models developed for likelihood inference, and can accommodate heterogeneous data sets from multiple gene loci in a combined analysis. A Metropolis-coupled MCMC algorithm (MC^3) is implemented in MrBayes to overcome multiple local peaks in the tree space. The parallel algorithm is efficient on network workstations that are becoming accessible to empirical biologists [2, 36]. MrBayes is now widely used in phylogeny reconstruction and is the top-cited paper in August 2002 in the whole field of computer science!

General framework. To formulate the problem of phylogeny reconstruction in the general framework of Bayesian inference described requires no more than definition of symbols. Let D be the sequence data. Let θ include all parameters in the model, with a prior distribution $f(\theta)$. Let τ_i be the ith tree topology, $i = 1, 2, \ldots, N(s)$, where $N(s)$ is the total number of tree topologies for s species. Usually a uniform prior $f(\tau_i) = 1/N(s)$ is assumed. Let b_i be branch lengths on tree τ_i, with prior probability $f(b_i)$. MrBayes 3 assumes that branch lengths have independent uniform or exponential priors with the parameter (upper bound for the uniform or mean for the exponential) set by the user. The posterior probability of tree τ_i is then

$$P(\tau_i \mid D) = \frac{\int\int f(\theta)f(b_i \mid \theta)f(\tau_i \mid \theta)f(D \mid \tau_i, b_i, \theta)\,\mathrm{d}b_i\,\mathrm{d}\theta}{\sum_{j=1}^{N(s)} \int\int f(\theta)f(b_j \mid \theta)f(\tau_j \mid \theta)f(D \mid \tau_j, b_j, \theta)\,\mathrm{d}b_j\,\mathrm{d}\theta}. \tag{3.28}$$

Note that calculating the denominator, the marginal probability of the data $f(D)$, would involve summing over all possible tree topologies and, for each tree topology τ_j, integrating over all branch lengths b_i and parameters θ, a virtually impossible task except for very small trees. The MCMC algorithm avoids direct calculation of $f(D)$, but integrates over branch lengths b_i and parameters θ through MCMC.

Summarizing output. It is straightforward to summarize the posterior probability distribution of trees, and several summaries are provided by MrBayes. One can take the tree with the maximum posterior probability (MAP) as a point

estimate, the so-called MAP tree [34]. This should be identical or very similar to the maximum likelihood tree under the same model. An approximate 95% credibility set of trees can be constructed by including trees with the highest posterior probabilities until the total probability exceeds 95%. Similarly to summarizing bootstrap support values for clades (subtrees) [8], posterior clade probabilities can also be collected and shown on a majority-rule consensus tree [27]. It may be noted that the branch lengths on the consensus tree produced by MrBayes 3 should be ignored as those are averages over different tree topologies; branch lengths are meaningful only on a fixed topology and their posterior probabilities should be calculated by running the MCMC on the fixed tree topology.

Comparison with likelihood. In terms of computational efficiency, stochastic tree search by MrBayes appears to be more efficient than heuristic tree search under likelihood using David Swofford's PAUP program [45]. Nevertheless, running time of the MCMC algorithm is proportional to the number of iterations the algorithm is run for. In general, longer chains are needed to achieve convergence in larger data sets due to the increased number of parameters to be averaged over. However, many users ran shorter chains for larger data sets because larger trees require more computation per iteration. As a result, it is not always certain that the MCMC algorithm has converged in Bayesian analyses of very large data sets. Furthermore, dramatic improvements to heuristic tree search under likelihood are still being made [16]. So it seems possible that for the purpose of obtaining a point estimate, likelihood heuristic search using numerical optimization can be faster than Bayesian stochastic search using MCMC. However, no one knows how to use the information in the likelihood tree search to attach a confidence interval or some other measure of sampling errors in the maximum likelihood tree—as one can use the local curvature or Hessian matrix calculated in a non-linear programming algorithm to construct a confidence interval for a conventional parameter. As a result, one currently resorts to bootstrapping. Bootstrapping under likelihood is an expensive procedure, and appears slower than Bayesian MCMC.

To many, Bayesian inference of molecular phylogenies enjoys a theoretical advantage over maximum likelihood with bootstrapping. Posterior probabilities have an easy interpretation: the posterior probability of a tree or clade is the probability that the tree or clade is correct given the data and the model [27, 34]. In contrast, the interpretation of bootstrap in phylogenetics has been controversial (e.g. [6, 19], Chapter 4, this volume). As a result, posterior probabilities of trees can be used in a straightforward manner in a variety of phylogeny-based evolutionary analyses to accommodate phylogenetic uncertainty; for example, they were used in comparative analysis to average the results over phylogenies [20, 22].

It has been noted that Bayesian posterior probabilities calculated from real data sets using MrBayes are often extremely high. One may observe that while bootstrap clade proportions are shown on published trees only if they are >50% (as otherwise the relationships may not be considered trustable), posterior clade probabilities are reported only if they are <100% (as most of them are 100%!). Recently a number of simulation studies suggested that the posterior probabilities are often misleadingly high (e.g. [1, 7, 43]). Some of the high posterior probabilities from real data sets may be genuine and indicate high but correct confidence in the phylogenetic relationship. Some may be due to lack of convergence of the MCMC algorithm or inadequate evolutionary model, which could be resolved by running longer chains or implementing more realistic substitution models. However, the problem seems more serious. Extremely high probabilities were observed by Rannala and Yang [34], who studied only small trees and used numerical integration, in which case algorithm performance is not an issue. Yang and Rannala [54] note that the posterior probabilities of trees vary widely over simulated replicate data sets and that they can be unduly influenced by the prior on the internal branch lengths. It is easy to see that high posterior probabilities will decrease when the internal branch lengths assumed in the prior get smaller; in the extreme when internal branch lengths are assumed to be 0, all trees will have the same probability. It is not clear to what extent the high posterior probabilities observed in real data sets can be attributed to this sensitivity. The problem raises serious practical concern about the methodology and further investigation is urgently needed.

3.7.2 *Estimation of species divergence times*

Bayesian inference has also been successfully applied by Thorne and co-workers [26, 48] to estimate species divergence times under models of rate change, that is, when the evolutionary rate itself evolves. Traditionally the molecular clock has been assumed for divergence time estimation. However, in many data sets, especially when the species are not closely related, the clock assumption is seriously violated. Because the sequence data contain information only about the branch length, which is the product of time and rate, but not about time and rate individually, incorrectly assuming that the clock can lead to seriously biased time estimates.

The likelihood approach to this problem has been to classify the branches on the tree into a few rate classes and then to estimate the divergence times as well as those few branch rates by maximum likelihood [25, 32, 57]. The methods have the drawback of requiring the researcher to assign branches to rate groups, although ideas of heuristic rate smoothing [38, 39] can be used to automate that process. The likelihood method has also been extended to incorporate fossil calibration information at multiple nodes on the phylogeny and to account for the heterogeneity in evolutionary process of multiple gene loci in combined analysis [56]. Yang and Yoder [56] emphasized the importance of such combined analysis as a way of circumventing the serious confounding effect between time and rate; the rates vary over lineages in different ways among gene loci, but the divergence times are shared, so that the internal constraints in the model might lead to reliable estimation of divergence times even when the clock is violated in every gene.

The Bayesian method specifies a prior distribution $f(t)$ of divergence times (t) and a prior distribution $f(r)$ of evolutionary rates (r). Let θ be all parameters in the model, with prior $f(\theta)$. The joint posterior distribution of times and rates are then

$$f(t,r \mid D) = \frac{\int f(\theta)f(t \mid \theta)f(r \mid t,\theta)f(D \mid t,r,\theta)\,\mathrm{d}\theta}{\iiint f(\theta)f(t \mid \theta)f(r \mid t,\theta)f(D \mid t,r,\theta)\,\mathrm{d}r\,\mathrm{d}t\,\mathrm{d}\theta}. \quad (3.29)$$

This is approximated by the MCMC algorithm. The marginal posterior of divergence times

$$f(t \mid D) = \int f(t,r \mid D)\,\mathrm{d}r \quad (3.30)$$

can be constructed from the samples taken from the MCMC.

Thorne *et al.* [48] and Kishino *et al.* [26] used a recursive procedure to specify the prior for the rates, proceeding from the root of the tree towards the tips. The rate at the root is assumed to have a gamma prior. Then the rate at each node is specified conditioning on the rate at the ancestral node. Specifically, given the log rate, $\log(r_A)$, of the ancestral node, the log rate of the current node, $\log(r)$, follows a normal distribution with mean $\log(r_A) - c$ and variance νt, where t is the time duration separating the two nodes. The correction term c in the mean is to remove any trend in the rate but is unimportant to the present description. Parameter ν controls how quickly the rate drifts and determines how clock-like the tree is *a priori*. This is a geometric Brownian motion model.

The prior for divergence times is specified using another recursive procedure [26], starting from the root and moving towards the tips. The age of the root has a gamma prior. Then each path from a tip to the root or an ancestral node is broken into random segments, corresponding to branches on the path, with the segment lengths having a Dirichlet density with equal probabilities (see [48]). Fossil calibration information is incorporated in the prior for times as constraints on node ages.

Thorne's program implements an efficient algorithm for divergence time estimation under the models of Thorne *et al.* [48] and Kishino *et al.* [26]. It incorporates fossil information at multiple nodes as lower and upper bounds. The likelihood is calculated using a normal approximation to the branch lengths estimated without the clock assumption, to achieve computational efficiency. Recent extensions made the method suitable for combined analysis of multiple data sets. The method and program has been used extensively to date divergences of major species groups, such as the radiation of mammals [17, 41].

While many factors including the substitution model can potentially affect divergence time estimation in the Bayesian method, the most difficult and important of those appear to be the priors for rates and times. An infinite amount of sequence data combined with a perfectly correct substitution model will reduce the errors in branch lengths to zero, but the errors in time estimates will persist as long as there is uncertainty in the fossil calibrations, or mismatch between the model and prior on one hand and reality on the other. Yoder and Yang [58] described a case where species sampling had a major effect on Bayesian divergence time estimation. The authors estimated divergence times on a tree of mammals, when either two or nine mouse lemur species are included in the data. The estimated age of the mouse lemur clade in the bigger data set was 25% older than in the small data set. The reason appears to be the assumed prior model of times. As discussed above, the method assumes similar branch lengths on the tree. However, branches within the mouse lemurs are very short, and inclusion of more mouse lemur species in the large data set made the prior rather unrealistic and pushed back the age of the mouse lemur clades.

In sum, recent developments in Bayesian and likelihood frameworks make it possible to estimate divergence times without the molecular clock through integrated analysis of heterogeneous genetic data sets incorporating multiple fossil calibrations. However, one has to bear in mind that estimation of divergence times without a clock is an extremely difficult problem whatever method is used, and should critically assess the effects of assumptions about rates and times on time estimates. The quality of fossils is critically important.

3.8 Conclusions and perspectives

The Baysian method, especially combined with MCMC algorithms, provides exciting opportunities to model-based analysis in molecular phylogenetics. Use of the likelihood function makes it straightforward to conduct integrated analysis of heterogeneous data sets from multiple loci while accommodating differences in their evolutionary characteristics, obliterating the need for ad hoc approaches such as supermatrix and supertree analyses. However, a number of computational and theoretical problems remain, which will no doubt prompt active research in the future. Computational problems include development of ingenious and efficient proposal mechanisms that will lead to improved mixing of the MCMC algorithms. While likelihood and Bayesian algorithms will probably never be fast enough to scale up with the ever-increasing sizes of real data sets analysed by molecular systematicists, any gain in performance is highly beneficial. Theoretical problems include understanding the power and limitations of the Bayesian methods and its robustness to assumptions in the prior and in the substitution model. The complexity of likelihood estimation of phylogeny has been extensively discussed (Chapter 2, this volume). That complexity appears to apply also in the Bayesian framework, and it remains an open question whether Bayesian posterior probabilities will be the ultimate answer to molecular phylogeny reconstruction.

Program availability

The programs mentioned in this chapter are available at the following web sites:
MrBayes: http://morphbank.ebc.uu.se/mrbayes/;
Divergence time estimation by Bayesian methods (T^3: Thornian Time Traveller):
ftp://abacus.gene.ucl.ac.uk/pub/T3/ and
http://statgen.ncsu.edu/thorne/multidivtime.html;

Tree reconstruction by likelihood:
PAUP: http://paup.csit.fsu.edu/;
Time estimation by likelihood:
PAML: http://abacus.gene.ucl.ac.uk/software/paml.html.

Acknowledgments

I thank Olivier Gascuel, Bret Larget, and an anonymous referee for comments. This work is supported by a grant from the Biotechnology and Biological Sciences Research Council (UK) to Z.Y.

References

[1] Alfaro, M.E., Zoller, S., and Lutzoni, F. (2003). Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, **20**, 255–266.

[2] Altekar, G., Dwarkadas, S., Huelsenbeck, J.P., and Ronquist, F. (2004). Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, **20**, 407–415.

[3] Beerli, P. and Felsenstein, J. (2001). Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proceedings of National Academy of Sciences USA*, **98**, 4563–4568.

[4] Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., and Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. *Genetics*, **161**, 1307–1320.

[5] Edwards, A.W.F. (1970). Estimation of the branch points of a branching diffusion process (with discussion). *Journal of the Royal Statistics Society, Series B*, **32**, 155–174.

[6] Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees [corrected and republished article originally printed in *Proceedings of National Academy of Sciences USA*, 1996, **93**, 7085–7090]. *Proceedings of National Academy of Sciences USA*, **93**, 13429–13434.

[7] Erixon, P., Svennblad, B., Britton, T., and Oxelman, B. (2003). Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic Biology*, **52**, 665–673.

[8] Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution*, **39**, 783–791.

[9] Gelman, A., Roberts, G.O., and Gilks, W.R. (1996). Efficient metropolis jumping rules. In *Bayesian Statistics*, Volume 5 (ed. J. Bernardo, J. Berger, A. Dawid, and A. Smith), pp. 599–607. Oxford University Press, Oxford.

[10] Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, **7**, 457–511.

[11] Gelman, S. and Gelman, G.D. (1984). Stochastic relaxation, Gibbs distributions and the Bayes restoration of images. *IEEE Transactions of Pattern Analysis and Machine Intelligence*, **6**, 721–741.

[12] Geyer, C.J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium of the Interface* (ed. E.M. Keramidas), pp. 156–163. Interface Foundation, Fairfax Station, VA.

[13] Goldman, N., Thorne, J.L., and Jones, D.T. (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics*, **149**, 445–458.

[14] Griffiths, R.C. and Tavaré, S. (1997). Computational methods for the coalescent. In *Progress in Population Genetics and Human Evolution: IMA Volumes in Mathematics and its Applications*, Volume 87 (ed. P. Donnelly and S. Tavaré), pp. 165–182. Springer-Verlag, Berlin.

[15] Grimmett, G.R. and Stirzaker, D.R. (1992). *Probability and Random Processes* (2 edn). Clarendon Press, Oxford.

[16] Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696–704.

[17] Hasegawa, M., Thorne, J.L., and Kishino, H. (2003). Time scale of Eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes and Genetic Systems*, **78**, 267–283.

[18] Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their application. *Biometrika*, **57**, 97–109.

[19] Hillis, D.M. and Bull, J.J. (1993). An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Systematic Biology*, **42**, 182–192.

[20] Huelsenbeck, J.P., Rannala, B., and Masly, J.P. (2000). Accommodating phylogenetic uncertainty in evolutionary studies. *Science*, **288**, 2349–2350.

[21] Huelsenbeck, J.P. and Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

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

[23] Jukes, T.H. and Cantor, C.R. (1969). Evolution of Protein Molecules. In *Mammalian Protein Metabolism* (ed. H. Munro), pp. 21–123. Academic Press, New York.

[24] Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, **16**, 111–120.

[25] Kishino, H. and Hasegawa, M. (1990). Converting distance to time: Application to human evolution. *Methods in Enzymology*, **183**, 550–570.

[26] Kishino, H., Thorne, J.L., and Bruno, W.J. (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Molecular Biology and Evolution*, **18**, 352–361.

[27] Larget, B. and Simon, D.L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, **16**, 750–759.

[28] Li, S., Pearl, D., and Doss, H. (2000). Phylogenetic tree reconstruction using Markov chain Monte Carlo. *Journal of American Statistics Association*, **95**, 493–508.

[29] Mau, B. and Newton, M.A. (1997). Phylogenetic inference for binary data on dendrograms using Markov chain Monte Carlo. *Journal of Computational Graphics and Statistics*, **6**, 122–131.

[30] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

[31] Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*, **148**, 929–936.

[32] Rambaut, A. and Bromham, L. (1998). Estimating divergence dates from molecular sequences. *Molecular Biology and Evolution*, **15**, 442–448.

[33] Rannala, B. (2002). Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Systematic Biology*, **51**, 754–760.

[34] Rannala, B. and Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, **43**, 304–311.

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

[36] Ronquist, F. and Huelsenbeck, J.P. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

[37] Rubin, D.B. and Schenker, N. (1986). Efficiently simulating the coverage properties of interval estimates. *Applied Statistics*, **35**, 159–167.

[38] Sanderson, M.J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, **14**, 1218–1232.

[39] Sanderson, M.J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, **19**, 101–109.

[40] Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[41] Springer, M.S., Murphy, W.J., Eizirik, E., and O'Brien, S.J. (2003). Placental mammal diversification and the cretaceous–tertiary boundary. *Proceedings of National Academy of Sciences USA*, **100**, 1056–1061.

[42] Stephens, M. and Donnelly, P. (2000). Inference in molecular population genetics (with discussions). *Journal of Royal Statistics Society, Series B*, **62**, 605–655.

[43] Suzuki, Y., Glazko, G.V., and Nei, M. (2002). Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proceedings of National Academy of Sciences USA*, **99**, 16138–16143.

[44] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogeny inference. In *Molecular Systematics* (2 edn) (ed. D.M. Hillis, C. Moritz, and B.K. Mable), pp. 411–501. Sinauer Associates, Sunderland, MA.

[45] Swofford, D.L. (1999). PAUP*: Phylogenetic analysis by parsimony, version 4.

[46] Thorne, J.L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences [published erratum appears in *Journal of Molecular Evolution* 1992, **34**, 91]. *Journal of Molecular Evolution*, **33**, 114–124.

[47] Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: An improved likelihood model of sequence evolution. *Journal of Molecular Evolution*, **34**, 3–16.

[48] Thorne, J.L., Kishino, H., and Painter, I.S. (1998). Estimating the rate of evolution of the rate of molecular evolution. *Molecular Biology and Evolution*, **15**, 1647–1657.

[49] Wilson, I.J., Weal, M.E., and Balding, D.J. (2003). Inference from DNA data: Population histories, evolutionary processes and forensic match probabilities. *Journal of Royal Statistics Society, Series A*, **166**, 155–201.

[50] Yang, Z. (2002). Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. *Genetics*, **162**, 1811–1823.

[51] Yang, Z., Goldman, N., and Friday, A.E. (1995). Maximum likelihood trees from DNA sequences: A peculiar statistical estimation problem. *Systematic Biology*, **44**, 384–399.

[52] Yang, Z., Kumar, S., and Nei, M. (1995). A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics*, **141**, 1641–1650.

[53] Yang, Z. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution*, **14**, 717–724.

[54] Yang, Z. and Rannala, B. (2004). Branch-length models bias Bayesian probability of phylogeny. *Systematic Biology*, in press.

[55] Yang, Z. and Wang, T. (1995). Mixed model analysis of DNA sequence evolution. *Biometrics*, **51**, 552–561.

[56] Yang, Z. and Yoder, A.D. (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Systematic Biology*, **52**, 705–716.

[57] Yoder, A.D. and Yang, Z. (2000). Estimation of primate speciation dates using local molecular clocks. *Molecular Biology and Evolution*, **17**, 1081–1090.

[58] Yoder, A.D. and Yang, Z. (2004). Divergence dates for malagasy lemurs estimated from multiple gene loci: Geological and evolutionary context. *Molecular Ecology*, **13**, 757–773.

4

STATISTICAL APPROACH TO TESTS INVOLVING PHYLOGENIES

Susan Holmes

This chapter reviews statistical testing involving phylogenies. We present both the classical framework with the use of sampling distributions involving the bootstrap and permutation tests and the Bayesian approach using posterior distributions.

We give some examples of direct tests for deciding whether the data support a given tree or trees that share a particular property, comparative analyses using tests that condition on the phylogeny being known are also discussed.

We introduce a continuous parameter space that enables one to avoid the delicate problem of comparing exponentially many possible models with a finite amount of data. This chapter contains a review of the literature on parametric tests in phylogenetics and some suggestions of non-parametric tests. We also present some open questions that have to be solved by mathematical statisticians to provide the theoretical justification of both current testing strategies and as yet underdeveloped areas of statistical testing in non-standard frameworks.

4.1 The statistical approach to phylogenetic inference

From our point of view, as statisticians, we see the phylogenetic inference as both estimation and testing problems that are set in an unusual space. In most standard statistical theory, the parameter space is either the real line $\mathbb{R}$ or an Euclidean space of higher dimension, $\mathbb{R}^d$ for instance. One notable exception for which there are a number of available statistical models and tests are ranked data. These sit in the symmetric group $\mathfrak{S}_n$ of permutations of n elements. See [58] for a book long treatment on statistics in such spaces, see [15] for some examples of data and relevant statistical analyses based on decompositions of the space, and [27] on the use of distances and their applications in that context. Of course other relevant high dimensional parameters that statisticians use are probability distributions themselves (non-parametric statistics). The authors of [16] use them to show conditions on consistency for Bayes estimates. Thus, as opposed to some authors in systematics, statisticians actually do believe that both distributions and trees can be true parameters. Although some references [4, 76, 80] do not agree with this approach, we will confer the status of true parameters to both the