The State of Bayesian Phylogenetics: Bayes for the Uninitiated

Joseph W. Brown
Department of Biology
Queen's University
December 22, 2003
Current email: josephwb@umich.edu

I think that I shall never see
A poem lovely as a tree...

Alfred Joyce Kilmer

ACKNOWLEDGEMENTS

Thank you to Dr. Stephen Lougheed for the time required to research and conceptualize this paper, and Dr. Theresa Burg for a space free from teething graduate students and frenzied, underpaid babies in which to write it.

CONTENTS

1. INTRODUCTION AND DISCLAIMER

The bulk of this paper is aimed at an audience that has been exposed, however tangentially, to Bayesian inference of phylogeny; who feel the excitement and intuit its promise, but are intimidated by both the mathematics and the Bayesian/MCMC jargon. We are in a unique position in that, as Paul O. Lewis (2002) states, the field is young enough that "the literature on Bayesian phylogenetics is still small enough that you can have some hope of actually reading all the papers on the topic!" This will undoubtedly become much more difficult within the decade. The purpose of this paper in part is to summarize the research to date; to sort out and explain the jargon, to point out some of the advantages of Bayesian inference, and to elaborate on some possible pitfalls.

Mathematical equations are given, but not overly dwelled on; instead, it is hoped that the reader will get a sense of the process of navigating through tree space and assessing reliability rather than dissecting high-dimensional functions. I have written this review in a specific mindset: to write a paper that I myself would liked to have read when first being exposed to the field of Bayesian phylogenetics. As such, particular attention has been paid to making it as readable as possible, in parts relying on figures to expand on points. In this respect I would like to thank those authors that contributed (unknowingly) many of the figures contained herein.

The reader is expected to have knowledge of rudimentary classical statistics and a working understanding of model-based phylogenetic inference. Numerous papers (Kuhner and Felsenstein, 1994; Huelsenbeck and Hillis, 1993; Huelsenbeck, 1995; Gaut and Lewis, 1995; etc.) have explained in detail the superiority of maximum likelihood in most situations when compared to such inference methods as maximum parsimony, minimum evolution, and various distance methods; I will not, therefore, repeat those arguments here. Nor will I compare Bayesian inference with those proven inferior methods. Rather, what is of interest is whether Bayesian inference of phylogeny can perform at least as well as maximum likelihood.

* Readers proceeding beyond this page do so at their own risk. I do not hold myself responsible for any psychological trauma that may ensue, nor will I pay psychiatrist fees.

2. A BRIEF INTRODUCTION TO BAYESIAN STATISTICS

Despite the relative novelty of the word "Bayesian" in the molecular systematics literature, this school of statistics is anything but new. Bayesian statistics was borne of a posthumously published paper by Reverend Thomas Bayes in the year 1763; as such, Bayesian inference predates classical (Frequentist) inference by 150 years (Huelsenbeck *et al*., 2002). In his *Essay Towards Solving a Problem in the Doctrine of Chances* Bayes developed a formal method for incorporating prior evidence into the inference of the probability that an event occurs. All subsequent work in Bayesian inference stems ultimately from this 1763 paper, though many of the recent developments involve technological and programming advancements that allow evaluation of extremely complex problems. A Bayesian framework offers many advantages over its classical counterpart in terms of the questions that can be asked, the incorporation of relevant prior evidence, the speed at which conclusions are reached, and the straightforward interpretation of results.

In this chapter I will walk through the derivation of Bayes' theorem, illustrate its utility through use of a simple example, discuss the controversy surrounding prior probabilities and how they are constructed, show how posterior probabilities and credible intervals are defined and interpreted, and finally contrast the major differences between the classical and Bayesian schools of statistics.

BAYES' THEOREM

Bayesian statistics is a formal method for inferring the probability that an event occurs from consideration of both the prior probability of that event occurring and the current data. Bayes' theorem, the instrument used to perform this task, is described below.

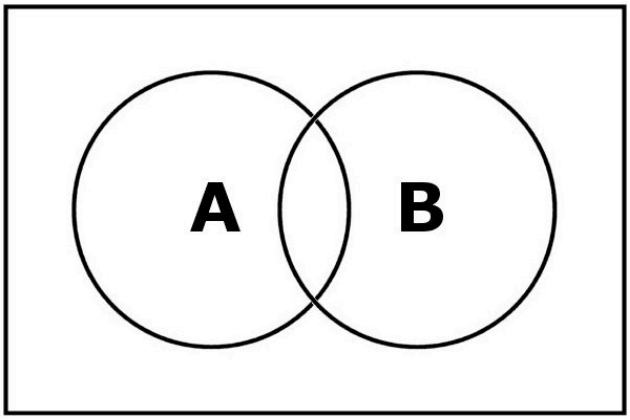

Figure 2.1: A Venn diagram. The rectangle represents the universe set while A and B represent two separate events within that universe. Areas where the circles overlap indicate both A and B have occurred.

Let A and B represent two separate events. From Figure 2.1 it is clear that:

$$\Pr(A,B) = \Pr(B,A)$$

This equation is trivial. What it says is that the probability of A and B both occurring is equal to the probability of B and A both occurring. These are both *joint* probabilities. Using simple probability theory a joint probability can be rewritten as the product of a conditional probability and the probability that the condition is true. The probability of both A and B occurring, $\Pr(A,B)$, can thus be rewritten as $\Pr(A,B) = \Pr(A)\Pr(B \mid A)$. What this equation says is that the probability of A and B both occurring is equal to the marginal (unconditional) probability of A occurring multiplied by the conditional probability of B occurring *given* that A has occurred. The above equation thus becomes:

$$\Pr(A)\Pr(B \mid A) = \Pr(B)\Pr(A \mid B)$$

which can be rearranged to the familiar form of Bayes' theorem:

$$\Pr(B \mid A) = \frac{\Pr(B) \times \Pr(A \mid B)}{\Pr(A)}$$

Biologists are not usually interested in the association of separate events; rather, we are interested in the relationship between an event and a particular hypothesis. Let us rewrite the above equation in a more useful form:

$$\Pr(hypothesis \mid data) = \frac{\Pr(hypothesis) \times \Pr(data \mid hypothesis)}{\Pr(data)}$$

The left side of the equation, $\Pr(hypothesis \mid data)$, is the conditional probability of the hypothesis given the data, and is called the *posterior probability*. This is the quantity that we are all after: how well our model agrees with the observed data. Contrast this with the quantity $\Pr(data \mid hypothesis)$ on the right hand side of the equation which represents the *likelihood*, the vehicle for data analysis in the classical framework. Clearly these two quantities measure very different things. Personally, I feel that that the posterior probability delivers something more intuitive and useful than the likelihood. We will return to this issue shortly.

$\Pr(hypothesis)$ in the above equation is the *prior probability* of the hypothesis. The prior probability, or simply "prior", represents our state of knowledge (or ignorance) about the truth of a hypothesis before we have observed the data (Sivia, 2002). Thus prior probabilities are determined *before* any data are observed. The prior is modified by the data through the likelihood function to yield the posterior probability. Priors, because they are true probabilities, obey the laws of probability and thus must sum (or integrate) to 1 across hypotheses. The final quantity, $\Pr(data)$, is the marginal probability of the data given the model, and can be expressed as:

$$\Pr(data) = \sum_{hypotheses} \Pr(hypothesis) \times \Pr(data \mid hypothesis)$$

This is simply the sum of the numerators over all possible hypotheses (Felsenstein, 2003). At first sight $\Pr(data)$ appears completely incalculable for complex (e.g. phylogenetic) problems,

not to mention incomprehensible (Lewis, 2002). However, close examination reveals that Pr(*data*) does not depend on the hypothesis of interest and so is a constant. Pr(*data*) essentially acts as a scaling factor to ensure that the posterior probability lies in the interval (0,1). Bayesians thus often write Bayes' theorem as follows:

$$\Pr(hypothesis \mid data) \propto \Pr(hypothesis) \times \Pr(data \mid hypothesis)$$

or more simply:

$$Posterior \propto Prior \times Likelihood$$

Some investigators like to think of the posterior probability as an updated form of the prior in the light of the observed data. The working out of a Bayesian posterior probability is thus a very logical procedure and effectively encapsulates the process of learning (Sivia, 2002).

A SIMPLE EXAMPLE

We will now go through a trivial application of Bayes' theorem to get a *feel* for the computations involved. Imagine two populations of *Amazona bowie*, a neotropical parrot species famed for spectacular nocturnal vocal displays and wicked guitar rifts. Individuals from these two populations are nearly identical, save for eye colour. Table 2.1 (below) describes the known eye colour frequencies for the two populations that were determined from extensive previous observations.

Table 2.1: Eye colour frequencies for the two populations of *Amazona bowie* derived from previous observation. To avoid rounding-off error these numbers are represented as fractions in the calculations.

Eye Colour	Population A	Population B
Blue	20%	60%
Green	30%	25%
Red	50%	15%

Because of these marked differences you decide that the populations warrant genetic comparison. Unfortunately both populations are deep within the Amazon and your NSERC grant wasn't large enough to allow you to travel there yourself, so you must depend on genetic samples sent to you by local authorities. You do this, but after having received two boxes of samples you notice that the population identification stickers lost in transport. The result: two separate boxes of samples but no idea of which is which. Fortunately a collector in one of the populations included two pictures of birds that contributed to the samples from that particular population. Both birds have blue eyes. What you would like to do is calculate the probability that this box of samples represents population B.

Because we are working within a Bayesian paradigm we can incorporate evidence from the previous studies into our posterior probability estimates. Recall from above that:

$$\Pr(hypothesis \mid data) = \frac{\Pr(hypothesis) \times \Pr(data \mid hypothesis)}{\sum_{hypotheses} \Pr(hypothesis) \times \Pr(data \mid hypothesis)}$$

The above equation rewritten for the *Amazona bowie* problem becomes:

$$\Pr(Pop.B \mid 2\ blue) = \frac{\Pr(Pop.B) \times \Pr(2\ blue \mid Pop.B)}{\Pr(Pop.B) \times \Pr(2\ blue \mid Pop.B) + \Pr(Pop.A) \times \Pr(2\ blue \mid Pop.A)}$$

Remember that $\Pr(data \mid hypothesis)$ is a likelihood. Assuming independence of the photographs (i.e. the birds are not close relatives), for population A the likelihood of the data given the hypothesis is:

$$\Pr(2\ blue \mid Pop.A) = \frac{1}{5} \times \frac{1}{5} = \frac{1}{25}$$

and for population B is:

$$\Pr(2\ blue \mid Pop.B) = \frac{3}{5} \times \frac{3}{5} = \frac{9}{25}$$

Also recall that $\Pr(hypothesis)$ is the prior probability of the hypothesis. Because we really have no idea which box is which, and because prior probabilities across hypotheses must sum to one, appropriate prior probabilities would seem to be:

$$\Pr(Pop.A) = \Pr(Pop.B) = 0.5$$

Substituting the above quantities into Bayes' theorem gives us:

$$\Pr(Pop.B \mid 2\ blue) = \frac{\frac{1}{2} \times \frac{9}{25}}{\frac{1}{2} \times \frac{9}{25} + \frac{1}{2} \times \frac{1}{25}} = 0.9$$

We can therefore say that we are 90% sure that the box from which the pictures originated represent population B, given the data (granted, these results are based on little data). Another way to think about this result is that your prior beliefs of equal probability have been updated to a probability of 0.9 for population B after considering the data. It follows that, again because we are dealing with true probabilities, that the posterior probability for the hypothesis "population A" is 0.1 (it may be useful for you to try this for yourself using the same steps as above). Now imagine that you find a third photograph in the same box as the others and it depicts a parrot with red eyes. How would this new piece of information change your conclusions? The likelihoods would be as follows:

$$\Pr(2\ blue, 1\ red \mid Pop.A) = \frac{1}{5} \times \frac{1}{5} \times \frac{1}{2} = \frac{1}{50} \quad \text{and} \quad \Pr(2\ blue, 1\ red \mid Pop.B) = \frac{3}{5} \times \frac{3}{5} \times \frac{3}{20} = \frac{27}{500}$$

You can see that because the frequency of the red-eye morph in population B is relatively low this new information greatly diminishes the likelihood of the box in question representing population B. Using the same prior probabilities as above the posterior probability for population B becomes:

$$\Pr(Pop.B \mid 2\ blue, 1\ red) = \frac{\frac{1}{2} \times \frac{27}{500}}{\frac{1}{2} \times \frac{27}{500} + \frac{1}{2} \times \frac{1}{50}} = 0.73$$

We can see that our beliefs have been shifted again in the light of the new evidence, and we are no longer so certain that the photographs came from population B, though it is still 2.7 times more likely $(0.73 \div 0.27 = 2.7)$. Tearing the box apart you are able to find 6 more pictures to give a total of 4 blue-eyed birds, 3 with green eyes, and 2 of the red-eyed morph. The likelihood equations are now:

$$\Pr(4\ blue, 3\ green, 2\ red \mid Pop.A) = \left(\frac{1}{5}\right)^4 \left(\frac{3}{10}\right)^3 \left(\frac{1}{2}\right)^2 = \frac{3}{181}$$

and

$$\Pr(4\ blue, 3\ green, 2\ red \mid Pop.B) = \left(\frac{3}{5}\right)^4 \left(\frac{1}{4}\right)^3 \left(\frac{1}{20}\right)^2 = \frac{9}{121}$$

and the posterior probability for the "hypothesis B" becomes:

$$\Pr(Pop.B \mid 4\ blue, 3\ green, 1\ red) = \frac{\frac{1}{2} \times \frac{9}{121}}{\frac{1}{2} \times \frac{9}{121} + \frac{1}{2} \times \frac{3}{181}} = 0.82$$

Once again the probability for "hypothesis B" has changed with increased data. Because these last results are based on more data than the first (2 photograph) calculation, they may be considered more robust. As a last twist let us imagine that we get a phone call from our collaborator in São Paulo and he says that he is about 75% sure that the box with the photographs contains samples from population A. Up until now we have been using ignorant prior probabilities simply because we had no reason to do otherwise, but from this simple phone call we are furnished with information that allows us to alter our prior beliefs. This is where Bayesian inference diverges from that of likelihood, because prior information can be naturally incorporated into the Bayesian analysis via Bayes' theorem whereas the information cannot be made use of in a likelihood framework. Using the same data as above, the posterior probability for "hypothesis B" changes to:

$$\Pr(Pop.B \mid 4\ blue, 3\ green, 1\ red) = \frac{\frac{1}{4} \cdot \frac{9}{121}}{\frac{1}{4} \cdot \frac{9}{121} + \frac{3}{4} \cdot \frac{3}{181}} = 0.60$$

As a result we are less certain that the box represents population B as it is now only moderately (1.5 times) more likely than population A. We will see in forthcoming sections that the prior probability has a large influence on the posterior probability when there is little data, but diminishes as data increases. Because we have very little data indeed in this example, it is clear that the prior probabilities are having a considerable effect on the results. The prior probabilities described above are unquestionably contrived and dubious, and were only put forth for illustrative purposes.

MARGINAL ESTIMATION

The likelihood function in complex applications of Bayes' theorem typically contains many parameters of the statistical model, only one of which is of real interest to the investigator. For example, a phylogeneticist may be interested in a tree topology, but could care less about transition rate parameters, gamma shape parameter, ancestral states, etc. These parameters that are required to evaluate a problem but are of no direct interest are commonly referred to as "nuisance parameters". In a classical approach, all parameters are jointly estimated in order to find the highest peak on the likelihood landscape (Holder and Lewis, 2003). This can be problematic when the ratio of data points to parameters is low, as parameter estimates can be quite unreliable. Bayesian statistics, alternatively, deals with nuisance parameters in a straightforward and intuitive manner. Rather than find the values of the nuisance parameters that maximize the likelihood function (as in the classical framework), Bayes' theorem allows for evaluation of the parameter of interest while "marginalizing" over the nuisance parameters.

Marginalizing over, or "integrating out", nuisance parameters is a way to "take account of all possible values of" each of these parameters (Lewis, 2002). Let Z be a nuisance parameter, required for computational purposes but otherwise superfluous. Marginalization over Z effectively integrates over all possible values of Z, or accounts for uncertainty in Z, when evaluating the parameter(s) of interest. Marginalization is of utmost importance for all Bayesian probability inference: the information about a subset of the system's variables is derived by integrating out all nuisance parameters. More generally, given parameters X, Y, and Z, marginalization is the process to derive information about X and Y, given all possible values of Z, as in the following equation:

$$\Pr(X,Y) = \int \Pr(X,Y,Z)dz$$

A Bayesian approach is hence not as affected by a low ratio of data points to parameters (though this is never a good thing) because the results do not rely on point estimates of the parameter, but instead considers all possible parameter values (Holder and Lewis, 2003).

PRIOR PROBABILITIES

The long-standing schism between Frequentist and Bayesian schools of statistics is due largely to the idea of prior probabilities. The derivation of Bayes' theorem is not in question; any statistician will tell you that it is a valid and simple equation relating conditional probabilities. Nor are prior probabilities *per se* in debate. The controversy involves the *choice* of priors. If prior probabilities are universally agreed upon then there is no debate; the issue is whether usable prior probabilities exist (Felsenstein, 2003). Bayesian prior probabilities can be based on theoretical expectations or previous experience by the investigator. For example, an investigator will have had flipped many a-coin in his or her lifetime, and hence has a "gut feeling" that the probability of heads p for a fair coin should be about 0.5. However, because different investigators *will* have had different experiences, they may assign different prior probabilities to the same problem. In the coin flipping example, one investigator may allow a fair coin to have a p value in the range of 0.4 to 0.6 for a finite number of flips, while another might constrict the range to $0.45 < p < 0.55$. The objections made by classical statisticians, then, is simply the inherent *subjectivity* of specified prior probabilities. Since there may be no single correct prior distribution, then all conclusions drawn from the posterior distribution are suspect (Bullard, 2001). Felsenstein (2003) notes that a Bayesian is defined not by using a prior probability, but by willing to use a controversial one.

Bayesian statisticians, on the other hand, view prior probabilities as a strength of their school. For one thing, Bayesian statistics provides a solid, formal framework for incorporating prior information into the statistical analysis. Information gleaned from experiments made in the past can thus be incorporated into the current analysis. Prior probabilities, then, while being subjective, need not be arbitrary (Bullard, 2001). This property alone counts for much of the attractiveness of Bayesian statistics. Why should we not make use of information obtained through extensive effort by a multitude of investigators? Secondly, the subjectivity in selecting a prior is *explicit* (as compared, for example, to cutoff values for significance, choice of null and alternative hypotheses, and choice of likelihoods in a classical setting; Bullard, 2001) and must be defensible (Lewis, 2001). Lastly, and perhaps more alluring to skeptical Frequentists than to Bayesians, the effect of the prior decreases with increasing data. This means that given enough data the prior will not overly influence the results (Figure 2.2).

Prior probabilities usually do not take specific values (as with our simplistic example above), but instead form probability distributions. [Technically, if the distribution deals with discrete data then it is a probability *mass* function, and if it deals with continuous data then it is a probability *density* function. Both distributions have the property that the area beneath the probability surface is equal to the posterior probability for the range of values of the parameter of interest. The difference between the two is that in the discrete case the area is obtained by summing across parameter values, while the continuous case requires integration. However, to avoid jumping back and forth between "probability mass function" and "probability density function" for the duration this paper I will use the vague term of "probability distribution" for both prior and posterior probabilities, the context making it clear the type of distribution involved.] The form of the prior probability distribution depends on the availability of relevant prior information and the nature of the question being asked. A primary distinction can be made between informative and uninformative priors. An informative prior can make use of information from

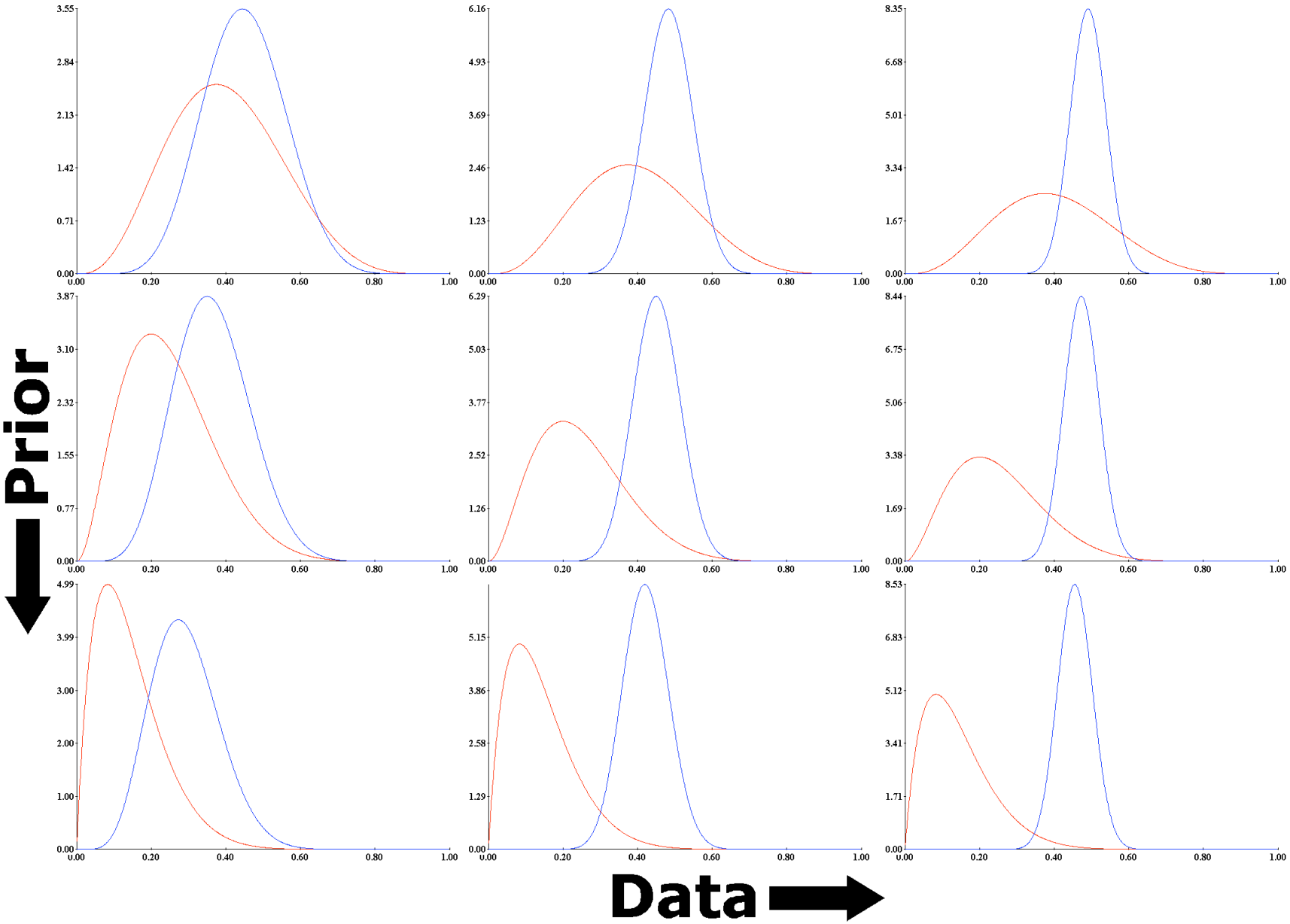

Figure 2.2: The effect of increasing data on the influence of the prior probability on the posterior probability in a coin flipping experiment. The y-axis represents the probability (blue curve = posterior, red curve = prior) and the x-axis represents the parameter being estimated (p, the probability of heads). Data increases from left to right: 5 heads in 10 flips, 25 heads in 50 flips, and 50 heads in 100 flips, respectively. "Strength" of misleading prior increases from top to bottom. As can be seen, posterior probabilities for 100 flips are nearly identical despite substantial differences in specified prior probabilities.

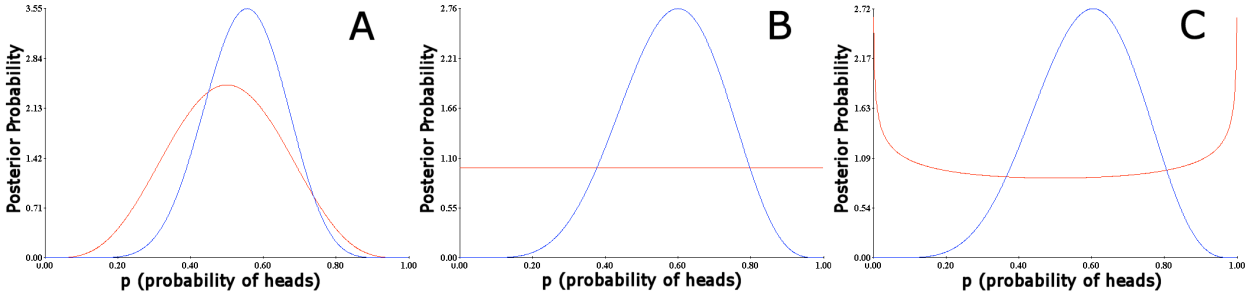

Figure 2.3: Prior probability distributions used for binomial data (posterior probability distributions, centered at approximately p=0.6, are also shown). Figure A represents an informative prior: the distribution gives added weight to the more probable values of the parameter of interest. This weighting information comes from either theoretical expectations or prior experimental results. Figures B and C represent vague priors. Figure B gives equal weight to all possible values of p *a priori*, and hence is often referred to as a "flat" prior. Figure C is an example of a "bathtub" prior: this prior has higher variance and thus has the least influence on the posterior probability distribution.

previous research (for example, from the output of a previous experiment) or directly from theoretical expectations. Such a prior will have a distribution that gives added weight to the more probable values of the parameter of interest, as in Figure 2.3A.

An uninformative (vague) prior, alternatively, will have a distribution such that the influence of the prior on the posterior probability is minimal. In this case the prior distribution illustrates the level of our ignorance about the truth of a hypothesis. Often, uniform (flat) priors are used as uninformative priors, as they attribute equal probability to all possible values of a parameter *a priori* (Figure 2.3B). When flat priors are used the posterior probability is directly proportional to the likelihood (Lewis, 2002). However, flat priors are not necessarily the best choice in all situations (Holder, 2003). This is because the posterior probability mode is tugged (the degree determined by the amount of data) towards the mean of the prior distribution (Jones and Browning, 2003). A more vague prior, then, would have a higher level of variance so that this effect is minimized. Figure 2.3C above shows examples of vague prior probability distributions used for binomial data. The bathtub-shaped distribution on the right has increased variance and hence a decreased influence of the prior on the posterior. For firsthand experience in exploring the relationship between priors, posteriors, and sample size, I strongly encourage the reader to download the Windows program "Bayesian Coin Tosser" by Paul O. Lewis, available online at the following address: http://lewis.eeb.uconn.edu/lewishome/software.html.

A second distinction in prior probability distributions has to do with the type of data being analyzed (or the complexity of question being asked). Binomial data (e.g. flipping a coin) typically have beta distributions of the form Beta(α, β) for priors. Beta distributions are a family of distributions which are non-zero only over a finite interval $0 < X < 1$ (Lee, 1997). The variables α and β are shape parameters that allow the distribution to take on variety of shapes. These distributions are useful for modeling purposes as they are extremely flexible and can be applied to a myriad of data sets. Figure 2.3 (above) illustrates the plasticity of the beta distribution.

Multinomial data cannot be expressed in terms of Beta(α, β), and so we must use another family of distributions. The Dirichlet distribution is a multivariate generalization of the beta distribution and is as flexible for multinomial data (e.g. the substitution rates of the GTR model) as the beta

is for binomial data. A Dirichlet distribution behaves in much the same way as a beta distribution (i.e. with respect to the construction of informative and vague priors) but cannot be represented in a two dimensional figure because it has *n* dimensions, where *n* is the number of parameters involved.

A final distinction involves proper and improper prior probability distributions. A proper prior is one that obeys the laws of probability, specifically that the distribution integrates to 1. A prior is improper if it does not integrate to 1 (i.e. there is no proportionality constant which will make the integral of the probability function equal to 1), and thus not a true probability distribution (Lee, 1997). Such a distribution would arise, for example, when placing uniform probability on all values of a continuous parameter. A prime example of an improper prior is Jeffery's prior (Figure 2.4), equivalent to a uniform prior distribution on the logarithm of the parameter, which represents complete ignorance about the value of a scale parameter on behalf of the investigator (Sivia, 2002). The use of improper priors has the danger that it often, but not always, leads to an improper posterior probability distribution (Huelsenbeck *et al*., 2002).

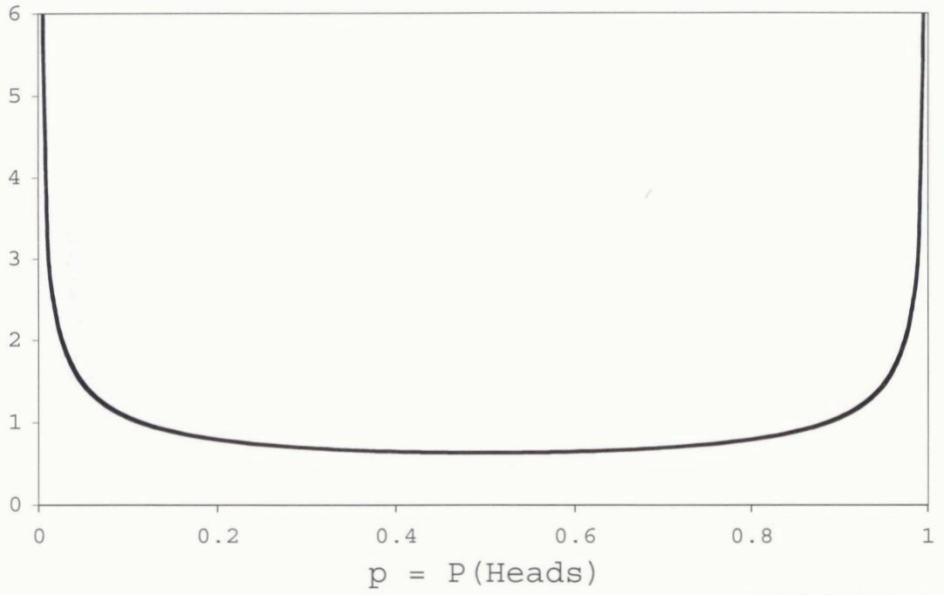

Figure 2.4: An example of Jeffrey's prior used in a coin flipping experiment (figure from Holder, 2003). The prior is an improper probability distribution because it fails to integrate to 1. Jeffrey's prior reflects the complete ignorance of the value of the scale parameter. This distribution has the property that rescaling the horizontal axis makes no difference on the distribution that is assigned.

Regardless of the degree of comfort with the idea and construction of prior probabilities, some effort should be made on the part of the investigator to examine the sensitivity of the results to the choice of prior probability distribution (Jones and Browning, 2003). Prior probability misspecification is completely analogous to model misspecification, and as such should be of utmost concern.

POSTERIOR PROBABILITIES AND CREDIBLE INTERVALS

In the simple example above we had distinct values for the posterior probability, one value for hypothesis one and another for hypothesis two. However, like prior probabilities, for more

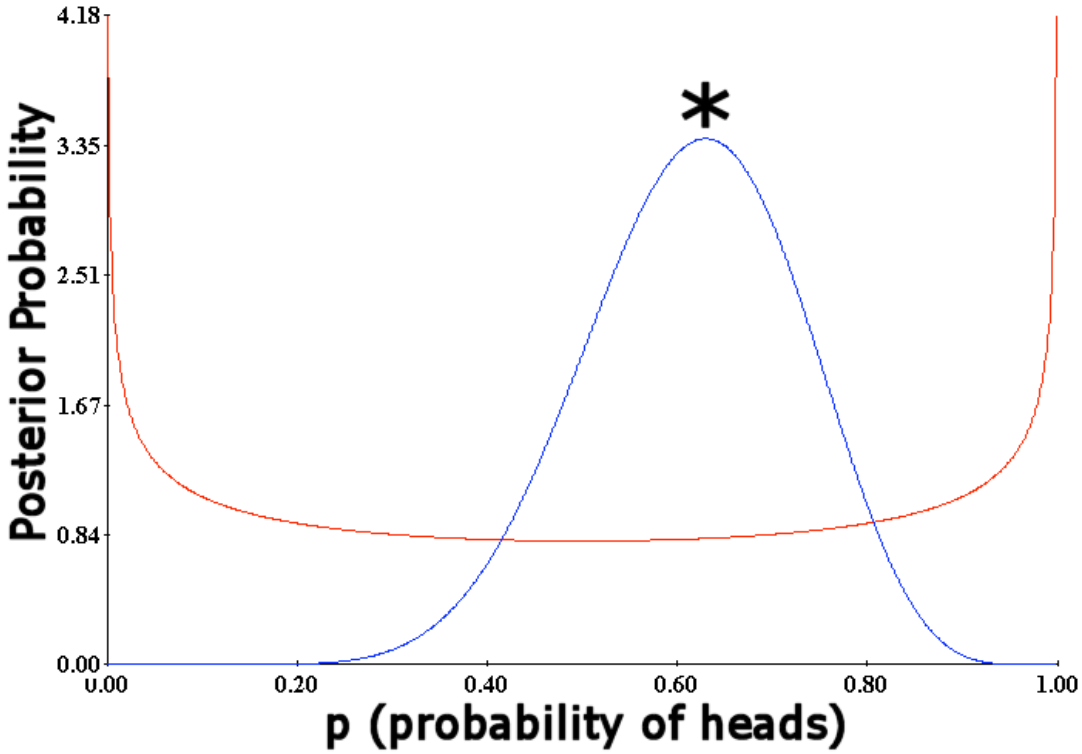

Figure 2.5: Posterior probability distribution for a simulated coin tossing experiment (the prior probability distribution, beta(0.7,0.7), is also shown). A coin was flipped 16 times, 10 flips of which resulted in heads. The mode of the distribution, marked by an asterisk, represents the most highly probable value of p.

complex problems posteriors usually take the form of a probability distribution. An example of a posterior probability distribution for a coin tossing experiment in given in Figure 2.5.

How do we summarize the information in such a figure? We might be first interested in obtaining a "best guess" for the parameter of interest, here being p, the probability of heads. In the figure above the best guess would be represented by the asterisk which corresponds to a value of p = 0.625. We could thus state "our best estimate of the probability of heads is 0.625". However, upon close examination of the posterior probability distribution above we see that a value 0.625 has a posterior probability of only 3.5! This tells us that we are only 3.5% sure that the true probability of heads is 0.625. Granted that these results are based on little data, this appears to be a very low value indeed and not very trustworthy at all.

To understand this result we must first note that Bayesian statistics treats parameters as random variables rather than fixed (but unknown) values. A Bayesian conclusion for complex problems will thus often be expressed as a range of values rather than a point estimate. [In the classical sense, a parameter is unknown but constant (not random) and hence cannot have a distribution (Bullard, 2001)]. The beauty of using a Bayesian framework is that the posterior, being a probability distribution, must integrate to 1. To be 95% certain of a conclusion we need only find the area under the curve which represents 95% of the distribution. The range of values that flank this region is called a "credible interval" and is shown in Figure 2.6.

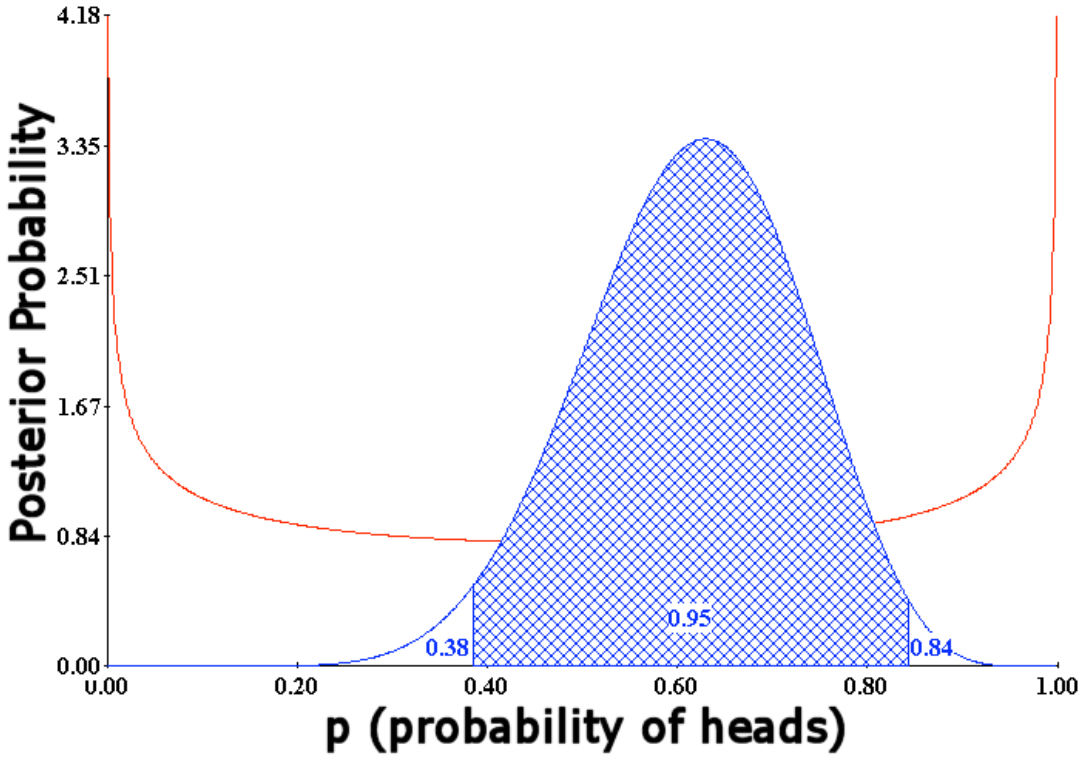

Figure 2.6: Posterior probability distribution for a simulated coin tossing experiment (the prior probability distribution, beta(0.7,0.7), is also shown). A coin was flipped 16 times, 10 flips of which resulted in heads. The 95% credible interval of the distribution is shown by the shaded region, (0.38,0.84).

From the figure above we can say "we are 95% certain that the probability of heads, p, is between 0.38 and 0.84". Posterior probabilities have the immense advantage in that probability statements are made about the parameters themselves rather than about the data (Shoemaker *et al*., 1999). As such a Bayesian uses probability as a direct measure of uncertainty on behalf of the investigator. Contrast this with the classical confidence interval, where a 95% confidence interval is interpreted as (given the model) one result of a procedure that had a 95% chance of generating an interval that would contain the parameter being estimated. In the classical sense probability is interpreted as a long-term frequency and is made specifically about the data. We will return to the idea of probability in both the Bayesian and classical senses at the end of this chapter.

BAYES FACTORS

Bayesian hypothesis testing sometimes takes the form of Bayes Factors, which is simply the odds-ratio form of Bayes' theorem. Let H_1 and H_2 be competing hypotheses for some data. From Bayes' theorem above we can write the posterior probability of each hypothesis as:

$$\Pr(H_i \mid data) = \frac{\Pr(H_i) \times \Pr(data \mid H_i)}{\Pr(data)}$$

Put in ratio form, the two hypotheses can be compared via a Bayes Factor as in the following equation:

$$BF = \frac{\Pr(H_1 \mid data)}{\Pr(H_2 \mid data)} = \frac{\Pr(H_1)}{\Pr(H_2)} \times \frac{\Pr(data \mid H_1)}{\Pr(data \mid H_2)}$$

Given a parameter of interest, θ, the likelihood ratio in the above equation, becomes:

$$\frac{\Pr(data \mid H_1)}{\Pr(data \mid H_2)} = \frac{\int_{\Theta_1} \Pr(data \mid \theta_1, H_1) \times \Pr(\theta_1 \mid H_1) d\theta_1}{\int_{\Theta_2} \Pr(data \mid \theta_1, H_2) \times \Pr(\theta_2 \mid H_2) d\theta_2}$$

Bayes Factors can then be interpreted as the odds in favour of H_1 against H_2 that are given by the data (Lee, 1997). The posterior odds ratio is simply equal to the prior odds ratio times the ratio of the likelihoods under the data (Felsenstein, 2003). A scale for interpreting Bayes Factors is given in Table 2.2.

Table 2.2: A suggested scale for interpreting Bayes Factors (from Jones and Browning, 2003).

2log(BF)	Interpretation
0 to 2	Not worth more than a bare mention
2 to 5	Positive
5 to 10	Strong
> 10	Decisive

Bayes Factors are more straightforward than classical approaches and circumvent some important issues with likelihood ratio tests. Most significantly, Bayes Factors do *not* require that the hypotheses be strictly nested. This aspect alone enables the investigator using a Bayesian framework to ask many more and complex questions than their classical counterparts. The concern with Bayes Factors, as with Bayes' theorem itself, is whether usable priors exist (Felsenstein, 2003).

CONTRASTING CLASSICAL AND BAYESIAN STATISTICS

Regarding the controversial uses of Bayesian statistics, Felsenstein (2003) states that "the arguments were old long before anyone thought of using Bayesian approached to inferring phylogenies. Nothing that biologists say is going to settle the matter." Nevertheless, it is valuable for the biologist to understand the distinctions between classical and Bayesian inference so that results from both approaches can be read critically. This section serves to contrast the major differences between the classical and Bayesian schools of statistics in terms of the interpretations of probability, inference, and reliability.

Both Frequentist and Bayesian methods use probability to assess statistical confidence, but interpret probability in very different ways. Frequentist (classical) statistical methods are named for their definition of probability as a long-term frequency (Shoemaker *et al*., 1999). [*Strictly speaking*, once an event is in the past it is no longer random, and so in a Frequentist sense it is meaningless to discuss the probability that the event occurred (Bullard, 2001)]. A Frequentist thus views probability in terms of (hypothetical) replicated experiments where most variables of the experimental environment are kept constant to ensure identical conditions across replicates. An experimental P value is then interpreted as follows: given that the null hypothesis of no difference is true, a result as extreme or more so than the observed result would be expected to

occur a proportion P of the time. As was mentioned above, this is a statement made specifically about the data (or the randomness of the sampling process) rather than about the parameter of interest (Bullard, 2001).

Bayesian methods view probability in a different sense. Rather than recognize probability as a measure of repeatability, a Bayesian sees probability as a direct measure of uncertainty (made directly about the parameter) and may or may not represent a long-term frequency (Shoemaker *et al*., 1999). What is more, only the data that are actually observed by the investigator are relevant in determining the probability that any particular model is true; data that are not observed (e.g. 'more extreme' values) are irrelevant (Bullard, 2001). On close examination it is clear that the Bayesian interpretation of probability is a much more straightforward and intuitive measure of "sureness". However, as Lee (1997) states, "the mere fact that [people] have difficulty with a theory does not prove it wrong." Still, a Bayesian probability can apply to many situations where a classical interpretation of probability does not conform naturally (Shoemaker *et al*., 1999). For example, consider the probability that I will complete my thesis within the next six months. It is hard to fathom this problem in terms of long-term frequency, but a Bayesian conclusion is easily applied: "I am 63% certain that I will finish my thesis within the next six months." A Bayesian framework thus addresses questions more directly and can be naturally applied to situations that are unrepeatable (e.g. products of evolution).

Closely related to the idea of probability is that of hypothesis testing and inference, and again Frequentists and Bayesians disagree on how to go about carrying out these tasks. In the classical sense, inference is performed by evaluating the probability of the observed data (or data more extreme) given a hypothesized model. As discussed above, this requires the viewpoint of observing the data generated from many experiments run under similar conditions. Here a null hypothesis of no difference is typically assumed between two quantities of interest, the experiment is run, and a test statistic is calculated from the data. The test statistic is in turn compared with a distribution of the test statistic under the null hypothesis. An experimental test statistic that is located in the extremes of the distribution is interpreted as evidence *against* the null hypothesis. A major problem with this approach, however, is that testing the significance of the results confounds the amount of evidence with the degree to which the null hypothesis is violated (Shoemaker *et al*., 1999). Take a coin flipping experiment from Lewis (2002) as an example. Here the parameter of interest is p, the probability of heads. Given a particular coin, the null hypothesis for this experiment is that $p = 0.5$ (or the coin is fair) and the alternative hypothesis is that it is not a fair coin ($p \neq 0.5$). However, because it is essentially impossible that p equals exactly 0.5 (and not, for example 0.51, 0.501, or 0.50001), it is only a matter of collecting enough data to prove this. This, then, is not a test of whether the coin is fair, but actually a test of whether we have flipped the coin enough times to prove that it is not perfectly fair (Lewis, 2002).

A Bayesian approach to hypothesis testing, conversely, gives the probability of the model, given the data (which is what we all really want), and provides evidence *in favour* of the model rather than against it (Shoemaker *et al*., 1999). Inference then is based on the posterior probability distribution of the parameter of interest, and so conclusions are statements about the parameter rather than the data. What is more, while in the classical framework only two hypotheses can be compared at a time, in a Bayesian setting multiple hypotheses can be compared at once with a posterior probability for each hypothesis being calculated (Shoemaker *et al*., 1999). A major distinction to note between Frequentists and Bayesians is that instead of true and false being discrete logical states, a Bayesian sees a logical continuum between zero and one, where zero

represents the statement is false, or has probability of zero of being true. A Bayesian thus sees the idea of null and alternative hypotheses as being fundamentally flawed (Lewis, 2002). Take the coin-flipping example above. While a Frequentist would say that a fair coin is one in which $p = 0.5$, a Bayesian would define a fair coin by a range of values, for example $0.45 < p < 0.55$, and find the probability that p for a particular coin falls within this range.

Lastly, Frequentists and Bayesians disagree on how reliability estimates are generated (and thus interpreted). A classical confidence interval is interpreted as follows: a 95% confidence interval is one result of a procedure that had a 95% chance of generating an interval that would contain the parameter being estimated (Bullard, 2001). This convoluted definition has driven many a young statistics student to an emotional breakdown, and on dissection of the definition it clearly is not what we are after. The probability statement involved above refers to the randomness of the sampling process rather than confidence in the estimated parameter value. Contrast this with the Bayesian interpretation of a credible interval: the probability is 95% that the parameter being estimated lies in a 95% credible interval. This statement is direct, intuitive, and gives us exactly what we want: the probabilities of different hypotheses in the light of data (Felsenstein, 2003). As we touched on above, this is accomplished through the use of the posterior probability distribution. Because (true) probability distributions integrate to 1, we need only take the area under the posterior probability surface to give direct probability statements about a range of values for a particular parameter. This is impossible in the classical framework because likelihood surfaces do not integrate to a fixed value, and even if they did they deal with the probability of the data rather than the model. The distinction, then, between reliability intervals in the Frequentist and Bayesian senses is that the former deals with errors in sampling while the latter deals with uncertainty in parameter value estimates.

A BAYESIAN FUTURE?

From the above it is clear that, despite uninformed claims to the alternative, classical inference is "not just a special case of [Bayesian inference] corresponding to a conventional choice of prior beliefs" (Lee, 1997). Classical and Bayesian statistics deal with problems in very different manners, and in many respects the Bayesian approach is more useful and direct. The first and most obvious distinction is on the use of prior information. A Bayesian framework naturally allows the incorporation of relevant prior information, be it theoretical or empirical in nature. An attractive corollary of this is that many different *types* of data (e.g. morphological, paleontological, genetic, behavioural, etc.) can be incorporated for resolution of the same problem. Such information is not admissible into a Frequentist framework as it does not allow for the incorporation of prior information.

Secondly, Frequentists and Bayesians deal with nuisance parameters differently. In a likelihood framework, the maximum likelihood estimate (MLE), or the highest peak in parameter space, requires that all parameters are jointly estimated. As we saw above, this may lead to unreliable results if the ratio of data points to parameters is low (Holder and Lewis, 2003). Within a Bayesian analysis, however, nuisance parameters are "integrated out" while the parameter of interest is focused on. Such an approach is not as adversely affected by a low ratio of data points to parameters as it explicitly incorporates uncertainty in the nuisance parameter values into the statistical problem via marginalization. This is enormously attractive as rogue nuisance parameters (i.e. those that are not well defined by the data) will not overly effect the results as might occur in a classical framework.

In addition to the benefits of incorporating prior information and marginalizing over nuisance parameters, Bayesian statistics also has the advantage of direct testing of hypotheses and straightforward, intuitive results. We discussed above how classical approaches provide evidence against a null hypothesis (this always being possible to reject) and give probability statements that deal with errors in the sampling process (i.e. the data) rather than about the parameters involved. A Bayesian approach, conversely, provides evidence in favour of certain parameter values and gives probability statements about the parameters themselves which directly expresses the level of certainty on behalf of the investigator. What is more, a Bayesian framework allows for multiple hypothesis testing while classical analyses are limited to comparing null and alternative hypotheses (Shoemaker *et al*., 1999). All of this, together with the ability to deal with unreplicable events, illustrates the immense freedom Bayesians have in designing testable questions and the ease of interpretation of the results that are returned.

Lastly, the results of a Bayesian analysis provides the investigator with an idea of the shape of the posterior probability distribution, rather than a point estimate (as in a maximum likelihood approach). In the likelihood school, many simulated data sets must be subjected to analysis rather than one (or a few) search(es) of parameter space. In contrast, the Bayesian school allows for simultaneous estimation of parameter values and support. Because of this (as we will see in the next chapter) Bayesian methods are much faster than likelihood methods with regards to generating intervals of 'sureness'. For particularly complex problems the speed of analysis becomes a limiting factor and is a large reason for the recent interest in Bayesian methods.

Despite the many (and large) advantages of working in the Bayesian paradigm (Figure 2.7), there remains the controversial use of (subjective) prior probabilities. This is a very contentious issue, and some people get very worked up about priors (Jones and Browning, 2003). As stated previously, prior specification is analogous to model specification, and as such should be treated with the same caution. We currently have two methods in our analytical toolbox to minimize the influence of dubious priors: make the prior distribution vague by increasing the variance, or increase the amount of data. Clearly the latter is preferable, but often funding is a limiting factor. It is not the aim of this paper to give suggestions on prior construction, but instead to inform the reader on the issues so that informed choices can be made.

Bayesian statistics has also been limited in use until quite recently because of the complex implementations required for statistically thorny problems. However, as we will see in chapter 4, Markov chain Monte Carlo methods are allowing previously intractable questions to be addressed in far less time than comparable likelihood methods.

Figure 2.7 (follow page): A mock debate between Sir Ronald Fisher and Reverend Thomas Bayes on the relative merits of classical and Bayesian statistical approaches (adapted from a table in Bullard, 2001).

***On Probability*:**

Probability is interpreted as the long-run relative frequency with which an event occurs in many repeated similar trials. Probability lies objectively in the world, not in the observer.

Probability is interpreted as a measure of one's degree of uncertainty about an event. Probability lies in the mind of the observer and may be different for people having different information or different past experiences.

On Inference:

Inference is performed by evaluating the probability of the *observed data*, or data more extreme, given a hypothesized model.

Inference is performed by evaluating the probability of a *hypothesized model* given the observed data.

On Intervals:

A 95% confidence interval is one result of a procedure that had a 95% chance of generating an interval that would contain the parameter being estimated.

The probability is 95% that the parameter being estimated lies in a 95% credible interval.

On Testing:

The *P*-value in a test of significance is the probability of getting a result at least as extreme as the one observed, given that the null hypothesis is true.

One may evaluate the probability of any particular model or set of models given observed data. Data not observed (e.g. 'more extreme' values) are irrelevant.

On Music:

David Bowie kicks ass.

I concur.

3. APPLYING BAYESIAN THINKING TO PHYLOGENETIC INFERENCE

This chapter serves to recount the young history of Bayesian approaches to phylogenetics, to contrast phylogenetic reconstruction in the traditional and Bayesian senses, and illustrate both the current advantages and future promises of a probabilistic phylogenetic paradigm.

A BAYESIAN HISTORY

We saw in chapter 2 that Bayesian inference predates Frequentist thinking by well over a century. Why should we get so excited about the application of such an old theorem to the phylogeny problem? Huelsenbeck *et al*. (2002) give four reasons. First, in the Bayesian paradigm the likelihood function is the vehicle that extracts information from the data matrix. Bayesian analyses can then use the same complex models of sequence evolution as conventional maximum likelihood investigations. As alluded to in the introduction, maximum likelihood is a proven winner in phylogenetics, and Bayesian methodologies inherit some of the desirable statistical properties via inclusion of the likelihood function. Second, Bayesian inference allows incorporation of relevant prior information into the phylogenetic analysis, and this is a property that maximum likelihood lacks. This is enormously attractive because it permits the inclusion of past results and also different types of data (for example, morphological, behavioural, and genetic data). Third, Bayesian inference via MCMC methods is a computationally efficient approach to approximating posterior probabilities. MCMC methodologies are the subject of chapter 4, so here we will simply state that these analyses can incorporate arbitrarily complex models of evolution and run in a relatively reasonable amount of time. Lastly, Bayesian inference is the first approach that treats the phylogeny as a random variable. This property allows direct probability statements to be made at the conclusion of the analysis. We will return to this issue later.

It is generally agreed that the introduction of Bayesian inference to the phylogenetic problem was due largely to the independent pioneering work of PhD dissertations by Bob Mau of the University of Wisconsin and Shuying Li of Ohio State University, and papers by Bruce Rannala and Ziheng Yang, all of which were published in 1996. [However, as Huelsenbeck *et al*. (2002) point out, inklings of Bayesian phylogenetics can be seen as far back as Joseph Felsenstein's 1968 thesis where he discussed both posterior probabilities and credible sets of trees, though he was unable to calculate these quantities. Felsenstein (2003) cites further hints of Bayesian leanings in the pre-1990's phylogenetics literature, but these were not developed fully and hence have had little to no influence on current work in the field]. The efforts of these three groups are recognized as the impetus for much of the subsequent work in the adolescent field of Bayesian phylogenetics, and so are ultimately responsible for the current incarnation of Bayesian phylogenetic inference.

Li (1996; further developed in Li, Pearl, and Doss, 2000) developed an MCMC strategy for approximating the posterior probability distribution of trees assuming a molecular clock, the Jukes-Cantor model of DNA sequence evolution, and that all possible rooted trees were of equal probability *a priori*. Applied to both simulated data and empirical data sets they found their implementation to recover the "true" phylogeny reliably in a reasonably short run time. The approach of Mau (1996; Mau and Newton, 1997; Mau, Newton, and Larget, 1999; Newton, Mau and Larget, 1999) allows for analysis of both DNA and restriction site data. Their implementation permitted use of a more complex model of sequence evolution (specifically the HKY85 model, but easily extended to more general models) and removal of the molecular clock

assumption. In this case all labeled trees are assumed to have equal probability *a priori*. Felsenstein (2003) argues that both of these distributions are inadmissible as you cannot apply a uniform non-zero probability distribution of node times if there is no limit on the age of the oldest node. Neither group explains how these issues are addressed. We will return to use of prior probabilities in phylogenetic inference shortly.

The method of Rannala and Yang (1996; Yang and Rannala, 1997) differ from those above in that they assumed a birth-death process as a model for speciation and extinction, and used this model to specify priors of phylogeny and branching times. The initial incarnation of this method required numerical integration to calculate posterior probabilities, and hence was limited to phylogenies of few taxa. The improved method (Yang and Rannala, 1997) used Markov Chain Monte Carlo strategies to approximate the posterior probability distribution, and consequently allowed the analysis of much more complicated data matrices.

Despite the pioneering efforts of the three groups described above, and subsequent influential work by Larget and Simon (1999) and Li, Pearl and Doss (2000), it is really due to John Huelsenbeck that the gospel of Bayesian statistics has spread and proliferated throughout the phylogenetics community (Huelsenbeck, 2000; Huelsenbeck and Ronquist, 2001a, 2001b; Huelsenbeck *et al*., 2000, 2001, 2002). The program MrBayes (Huelsenbeck and Ronquist, 2001a, 2001b; Ronquist and Huelsenbeck, in press) is currently the state-of-the-art software package for the Bayesian inference of phylogeny, though other packages exist (e.g. Larget and Simon, 1999; McGuire *et al*., 2001). In the next section we will discuss how Bayesian methods such as those utilized in MrBayes differ from conventional phylogenetic reconstructions.

TRADITIONAL AND BAYESIAN APPROACHES TO PHYLOGENETIC RECONSTRUCTION

Maximum likelihood and Bayesian approaches to inferring phylogenies differ in many respects, and these differences are intimately related to the statistical contrasts described in chapter 2. With respect to the phylogeny problem these differences can be crudely classified into two categories: what is being estimated, and how these estimates are reached.

We saw in chapter 2 that Frequentists and Bayesians view statistical problems in very different ways, and the inference of phylogeny is no different. Recall the equations:

$$\Pr(\textit{data} \mid \textit{hypothesis}) \text{ vs. } \Pr(\textit{hypothesis} \mid \textit{data})$$

The term on the left is the likelihood while that on the right is the posterior probability. Put into a phylogenetic context we can rewrite the equations as:

$$\Pr(\textit{data} \mid \textit{topology, model parameters}) \text{ vs. } \Pr(\textit{topology, model parameters} \mid \textit{data})$$

[Though topology is a parameter of the statistical model, I have separated it from the others as it is of prime interest.] Take a moment to consider the two equations. Which quantity would you rather evaluate? In my mind there is no question that the equation on the right is more useful, as it uses the data to arbitrate between different models (topologies). Also, implicit in this equation is that the Bayesian stance treats topology (and other parameters) as a random variable (Huelsenbeck *et al*., 2002; Douady *et al*., 2003). As such it is valid for topology to have a distribution of values, weighted by the posterior probability. This leads to the advantages of interpretation discussed earlier, in particular that the posterior probability for a given tree gives a

direct index of certainty for that pattern of evolutionary relationships, given the data. The likelihood equation, contrastingly, regards the data as uncertain, and tries to quantify the probability that these data could be generated by the given model (topology). Clearly these methods evaluate two very different aspects of the phylogenetic problem.

Not only does the statistical objective differ between likelihood and Bayesian approaches, but also how these objectives are accomplished. In a likelihood framework, the ML tree is found by finding the parameter values that *jointly* maximize the likelihood function (Huelsenbeck and Rannala, 1997; Huelsenbeck and Ronquist, 2001b). Clearly one would require very large (# of nucleotides) datasets to obtain accurate estimates for all of the parameters in complex models of evolution (e.g. GTR + I + G). Holder and Lewis (2003) show that the accuracy of topology estimation can be compromised when the ratio of data points to parameters becomes low.

Bayesians attack the problem in a different way. Like their Frequentist counterparts, investigators using Bayesian inference are usually interested primarily in only one parameter, the other parameters of the statistical model being so-called "nuisance parameters". As with the parameter of concern, these uninteresting but required parameters are not known with certainty, but this can be dealt with through marginalization, or "integrating out". Marginalizing thus allows the investigator to focus on one parameter while taking into account uncertainty in all other parameters. Figure 3.1 shows a simple cartoon example of marginalizing in a phylogenetic problem (from Holder, 2003). In phylogenetics, marginalization can involve any parameter at all in the statistical model, though usually topology is of direct interest and so all of the remaining parameters (transition rate parameters, gamma shape parameter, ancestral states, etc.) are integrated out. A Bayesian topological posterior probability is thus a marginalized probability, rather than a joint probability, as in a likelihood treatment.

Closely related to this joint/marginalized dichotomy is the algorithmic search itself. Maximum likelihood, as the names suggests, maximizes the objective likelihood function (i.e. finds the values of all statistical parameters that jointly gives the most positive likelihood value). This can be accomplished through either exact (exhaustive or branch-and-bound) or heuristic searches. Regardless, the end result is the same: the most likely point found in parameter space is taken to represent the final estimate. In other words, it is a straightforward point estimate. Points in parameter space visited previously by the algorithm are disregarded. As a result, though the tree found does have the best likelihood score, there is no immediate indication of whether this score is significantly better than that for any other tree. Confidence estimates are typically generated through use of the nonparametric bootstrap (Felsenstein, 1985).

Unlike maximum likelihood inference, Bayesian phylogenetics seeks to glean information about the *shape* of the posterior probability landscape rather than simply locating the global maximum (Lewis and Swofford, 2001; Figure 3.2). Put another, rather poetic, way, to a Bayesian the journey itself (through parameter space) is of greater interest than the destination (i.e. a best guess; Cummings *et al*., 2003). In a Bayesian analysis, then, not only is a best estimate obtained (the tree with the highest posterior probability), but also an idea of the relative merit of all other trees visited in the search. Confidence in a particular evolutionary relationship is thus accomplished through integration. The result is a simultaneous estimation of both phylogeny and support (Douady *et al*., 2003). We will return to the issue of phylogenetic confidence in detail in chapter 5.

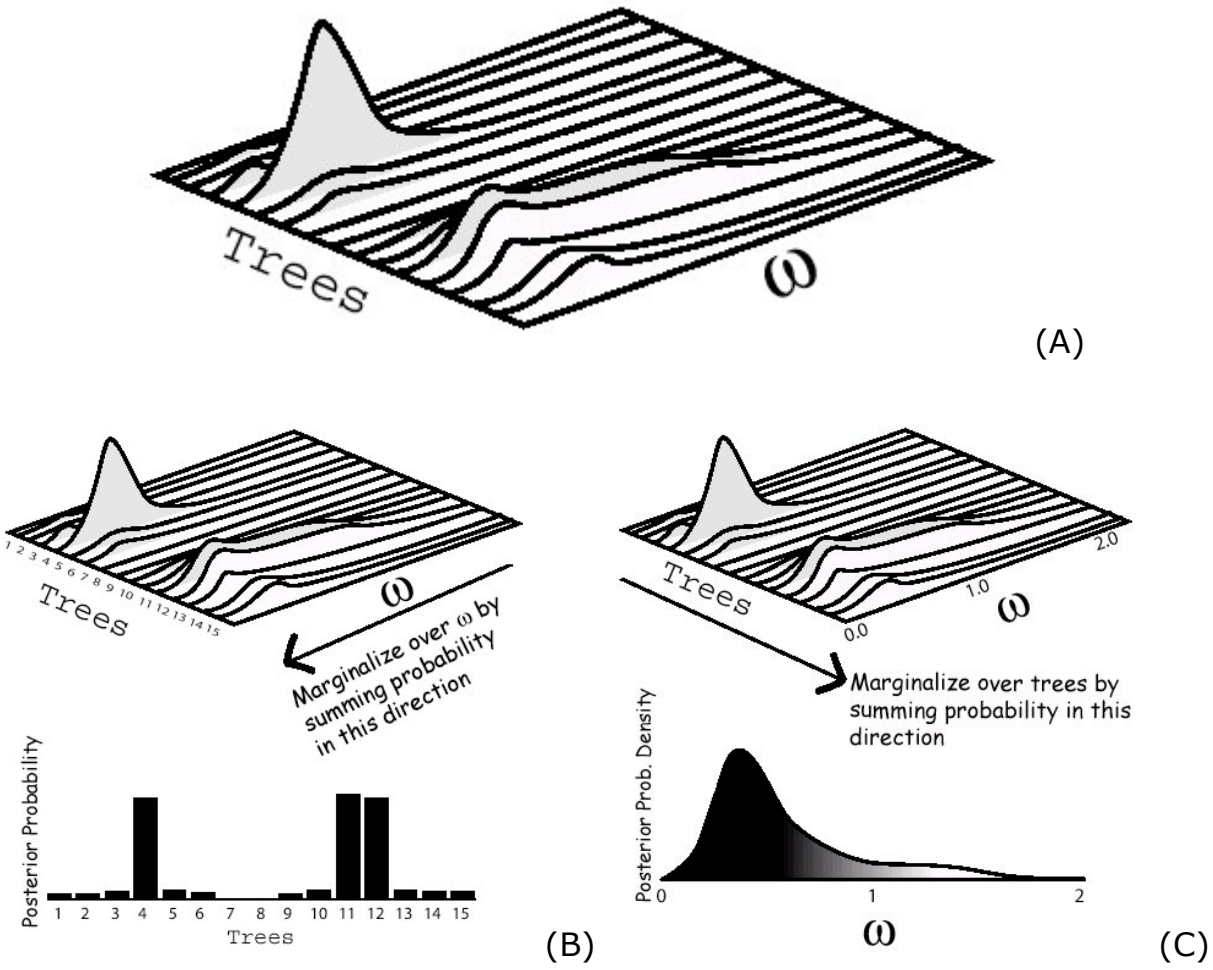

Figure 3.1: Simple marginalization in the phylogenetic inference problem (figures taken from Holder, 2003). (A) A hypothetical trivial parameter space where only two parameters are involved, the tree parameter (topology) and ω (a representative substitution model parameter). (B) Marginalization over ω: the posterior probability of each tree is calculated by integrating over ω; trees with larger volumes have higher posterior probabilities. (C) Marginalizing over topology: here ω is of primary interest, and so uncertainty in topology is accounted for by integrating over trees.

The parameters of a statistical model in the phylogenetic context are the trees (/), branch lengths (/), substitution model parameters ( ), and gamma shape parameter (/). Therefore, let X be the data and $\Theta = \{\tau, \vec{v}, \vec{r}, \alpha\}$ be a specific tree with a particular combination of branch lengths, substitution parameters and gamma shape parameter (Huelsenbeck and Ronquist, 2001a). The probability distribution of interest is then the joint posterior probability distribution of the parameters above and can be represented by Bayes' theorem as:

$$f(\Theta \mid X) = \frac{f(X \mid \Theta) f(\Theta)}{f(X)}$$

The likelihood function alone is integrated over all possible values for the branch lengths and substitution model parameters. The *i*th tree, for example, has the following likelihood function:

$$f(X \mid \theta_i) = \int_{\theta_i} \int_{\theta} \int_{\pi} f(X \mid \theta_i, \theta_i, \theta, \pi) f(\theta_i) f(\theta) f(\pi) d\theta_i \, d\theta \, d\pi$$

As can be seen, the posterior probability distribution involves many high-dimensional integrations, and as such cannot be solved analytically except in the most simplest cases, which is of course no use to the practical phylogeneticist. Instead, then, Bayesian phylogeneticists must rely on this distribution being approximated. The tool used in approximating the posterior probability distribution in Bayesian inference is the Markov chain Monte Carlo (MCMC) algorithm, and is the subject of the next chapter.

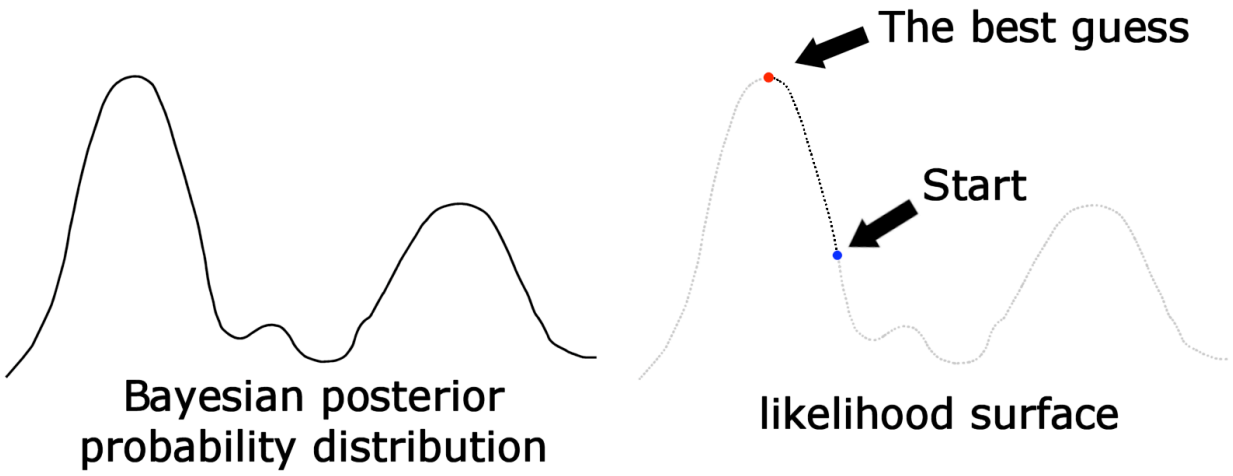

Figure 3.2: Contrasting search strategies in Bayesian and maximum likelihood inference. For ease of comparison the two underlying surfaces are depicted as identical. The goal of a Bayesian analysis is the learn about the shape of the posterior probability surface; the algorithm has no predefined termination step, and can be run arbitrarily long. Maximum likelihood inference, on the other hand, works to maximize the likelihood function; beginning from a particular point in parameter space the algorithm seeks to find parameter values that jointly maximize the function, and terminates when it can find no more likely parameter value combinations (i.e. reaches a peak).

PRIOR PROBABILITIES FOR THE PHYLOGENETIC PROBLEM

The missing distinction between maximum likelihood and Bayesian inferences of phylogeny above is the use of prior probability distributions. In the previous chapter we dealt with the concept of prior probabilities and the controversy that surrounds them. In the phylogenetics literature priors are no less controversial than in any other discipline, though they seem at present to be less well understood. This section briefly describes how priors are constructed and implemented for the phylogenetic problem.

What constitutes a prior probability in phylogenetic inference? To be a *fully* Bayesian implementation, all parameters in the statistical model must have associated prior probability distributions. This includes, then, branch lengths, nucleotide frequencies, substitution rates, etc. Because the field of Bayesian phylogenetics is still young, at present only vague priors are used (Lewis, 2002). For example, a transition-transversion ratio would have a simple flat beta distribution prior, while priors for the substitution rate parameters of the GTR model would be a flat Dirichlet distribution. The idea here is to minimize the influence of specified priors on the results (posterior probabilities) of the analysis. This is done partly as a consequence of our

ignorance with respect to the underlying distribution of a particular parameter, and partly to placate the skeptics. [Joseph Felsenstein (2003) argues that because priors are not universal, researchers should publish likelihoods instead and let the reader provide their own prior. As a reader, I would personally much rather see both the posterior *and* prior probabilities; if the priors seem sensible then I am willing to accept the posterior probabilities]. As such, prior probabilities are currently viewed in the phylogenetics literature as mere imposed requirements for using Bayes' theorem rather than powerful conduits for the incorporation of relevant prior information. In the future, as we become more comfortable with the idea of prior probabilities and we accumulate the essential empirical data, we will no doubt experience a shift from looking at priors as requirements to seeing them as compellingly valuable tools.

That said, there exists still other objections to the use of prior probabilities (which is the equivalent to saying there exists objections to using Bayesian inference in phylogenetics), most formidably by Joseph Felsenstein (2003). First he points out the dependence of a parameter estimate on the scale used in the analysis. As a hypothetical example he inferred a tree with only two species using the Jukes-Cantor model of evolution, so the only characteristic being estimated is the distance between the two species. If we place a flat prior on the net probability p of change at a site we will get a uniform distribution on the interval (0, 3/4). Figure 3.3 shows this distribution and the implied prior on the branch length. Clearly the prior on branch length is not flat, and so his point is that the property of the estimate being independent of the scale (which is true in the case of maximum likelihood) does not hold for Bayesian inference.

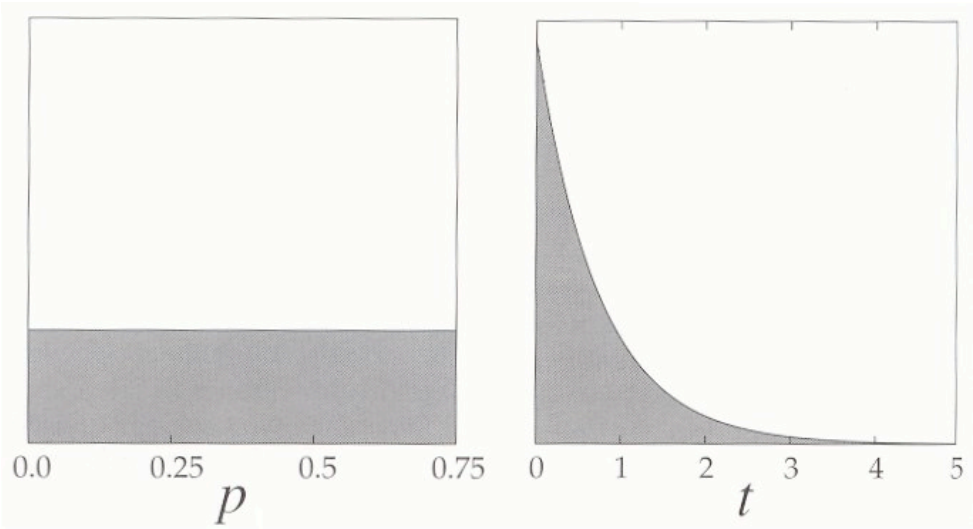

Figure 3.3: Priors probabilities for the 2 taxon hypothetical problem (figure modified from Felsenstein, 2003). On the left is the flat prior probability specified for the net probability p of change at a site, and on the right is the prior distribution for branch length t that the prior on the left implies.

A second objection deals with flat priors on unbounded quantities. As we saw in chapter 2, a distribution that does not integrate to 1 is improper because it does not conform to the laws of probability. The use of improper priors eventually leads (in most cases) to an improper posterior probability distributions (Huelsenbeck *et al*., 2002). [What is more, the use of improper priors for some parameters in the phylogeny problem, such as branch lengths, is obviously inappropriate]. Since we cannot have a uniform distribution on an unbounded parameter, the distribution must be truncated. Figure 3.4 shows why this may be a problem. In the hypothetical example above we may arbitrarily truncate the branch length t at 5. If we plot this in terms of the

probability of change at a site p we see that it is skewed towards $p = 0.75$. Allowing the truncation to occur at a higher value only compounds the problem. These objections have not yet been addressed in the Bayesian phylogenetics literature and their implications are unclear (they may simply imply that "flat" priors are inappropriate, but other distributions may work quite nicely).

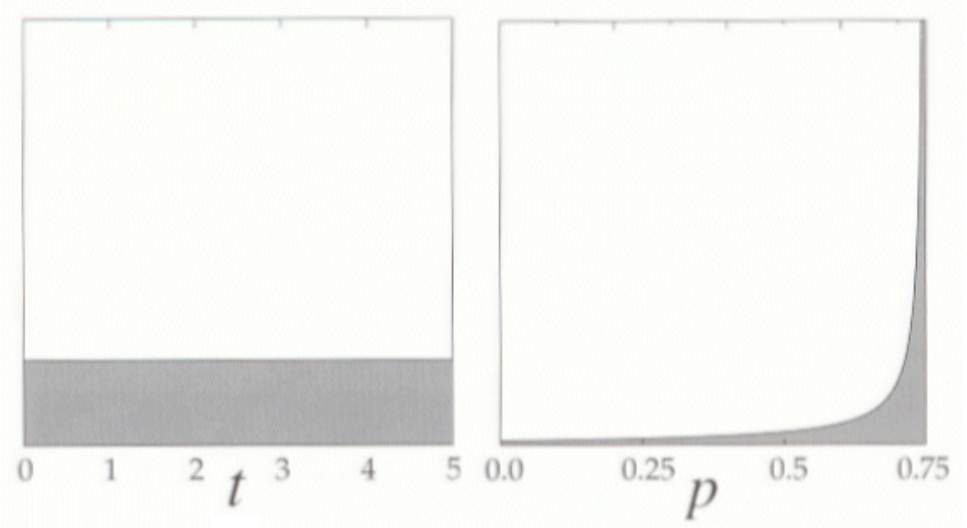

Figure 3.4: Truncation point effects (figure modified from Felsenstein, 2003). On the left is the flat prior probability specified for the branch length, arbitrarily truncated at $t = 5$, and on the right is the prior distribution for the net probability p of change at a site that the first distribution implies. Allowing the truncation point to be larger only increases the strength of the skew.

A redeeming feature of Bayesian inference is that given enough data the effect of the prior is effectively washed out, and so the results of the analysis are driven by the data (Alfaro *et al*., 2003). Huelsenbeck and Ronquist (2001b) show that priors must be extraordinarily strong to refute the information contained within a given data set. While this is somewhat relieving, we are now back to the point where the prior, rather than being a channel for prior information, is again inconsequential.

ADVANTAGES OF A BAYESIAN APPROACH TO PHYLOGENETICS

All other things being equal, the advantage of computational efficiency (speed) is more than enough reason to explain the increasing popularity in Bayesian inference of phylogeny. Douady *et al*. (2003) found that one Bayesian tree search in MrBayes ran 80 times faster than a single (heuristic) PAUP* maximum likelihood bootstrap replicate. We will see that this ratio is not constant and will change depending on the specifics of the data matrix and search strategy, but the bottom line is that there is a many-fold increase in computational efficiency in a Bayesian tree search compared to a maximum likelihood one. With the advent of PCR and automated sequencing technology we have seen an explosion in both the number and sizes of genetic data sets, and we can only expect this trend to increase as the technology becomes more widespread. Speed of analysis will therefore become an even more important (limiting) attribute than it is today, as it may mean the difference between a data set being analyzable or not. At the rate that the size of data sets are increasing we would be very optimistic indeed to believe computer speeds will keep pace.

Closely related to the computational efficiency advantage illustrated above is the ability of Bayesian analyses to handle complicated models of evolution. Complex models of nucleotide evolution are recognized as being more realistic, and this becomes increasingly true as the breadth of taxonomic scope broadens. Ideally we would like to use the stochastic model of evolution that best fits our data, and the program Modeltest (Posada and Crandall, 1998) was designed explicitly for this purpose. [Bollback (2002) designed an explicit Bayesian phylogenetic method that evaluates the adequacy of different models using posterior predictive distributions.] The problem with adding parameters to the model is that we are also adding entire dimensions to parameter space, dimensions that should be as adequately explored as the existing dimensions. The result is an enormous increase in computational burden. To boot, we saw above that maximum likelihood hill-climbing algorithms can give unreliable parameter estimates when the ratio of data points to parameters is low (Holder and Lewis, 2003). The problem of increasing model complexity is therefore two-fold: problems with practical runtimes and issues of reliability. A Bayesian approach marginalizes over all nuisance parameters and so doesn't suffer from the affliction of poor reliability.

Taking increasing complexities of data matrices and stochastic models of evolution together, it is clear that the Bayesian methodology holds much more promise for the future of molecular phylogenetics than conventional maximum likelihood approaches. However, care must be taken in the selection of the stochastic model of evolution assumed. Because current Bayesian implementations use flat priors, posterior probability distributions are largely dependent on the structure of the likelihood model. It follows that model misspecification may lead to "strong and unreliable" inferences (Buckley, 2002), and that this effect can occur whether models are either greatly underparameterized (Erixon *et al*., 2003) or overparameterized (Rannala, 2002). Clearly, then, an investigator wishing to use Bayesian inference in phylogeny must be more concerned with model specification than when working in a likelihood framework, which appears to be fairly robust the model of evolution assumed (e.g. Kuhner and Felsenstein, 1994).

Lastly is the issue of the interpretation of results from a phylogenetic analysis. A primary matter of distinction is what the two methodologies actually measure. As we have already demonstrated, maximum likelihood inference measures the probability of the data, given the model, while Bayesian inference measures the probability of the model, given the data. Clearly the latter is preferred as it gives the probabilities of different hypotheses (topologies) in the light of the data (Felsenstein, 2003), which is what the phylogeneticist is really after. A second distinction deals with how reliability estimates are constructed. Maximum likelihood analyses rely on nonparametric bootstrap proportions to derive statements of confidence. As we will see later on, a bootstrap proportion is an index of repeatability, *not* of the accuracy of the results. A Bayesian posterior probability, conversely, is a direct probability statement that the results reached are true, given the data. On the matter of interpretation of results, then, a Bayesian approach is undoubtedly ideal.

Considering both the current and future scales of phylogenetic inference we see that the Bayesian paradigm outperforms the incumbent champion maximum likelihood on most counts. These advantages, however, all rely upon a proper MCMC search of parameter space. The next chapter deals with the mechanics of MCMC methods in phylogenetic inference and the issues that the investigator should be aware of.

4. MCMC METHODS

The posterior probability distribution of trees involved in a model-based phylogenetic analysis involves both summation over all trees and (for each of these trees) integration over all possible combinations of branch lengths and model parameter values (Huelsenbeck *et al*., 2001). The result is a problem that requires evaluation of high-dimensional summations and integrals (Altekar *et al*., in press). As mentioned above, such problems cannot be solved analytically and so must rely on a stochastic simulation sampling scheme to approximate the posterior probability distribution. The most useful tools we have available for this approximation are based on Markov chain Monte Carlo (MCMC) theory; in fact, without MCMC methods it would be impossible to apply Bayesian principles to the problems of phylogenetic inference (Lewis, 2002). In addition, the advent of MCMC methods has allowed the analysis of phylogenetic problems that were previously intractable by classical statistical methods (e.g. maximum likelihood) because of the size of the data matrix or complexity of substitution model (Huelsenbeck *et al*., 2002). Such problems, impossible to tackle even with high performance computing resources over a period of months, can now be addressed within days on conventional desktop computers. The reader interested in a fuller treatment of MCMC methods is referred to Tierney (1994).

MARKOV CHAIN MONTE CARLO (MCMC)

Markov chain Monte Carlo (MCMC) is, simply, a simulated random walk through parameter space for the purpose of sampling from the posterior probability distribution of interest (Tierney, 1994). More specifically, MCMC allows phylogeneticists to sample phylogenies according to their posterior probabilities (Huelsenbeck and Ronquist, 2001a). This is accomplished by constructing a Markov chain that has as its state space the parameters of the statistical model and a stationary (equilibrium) distribution that is the probability distribution of interest (Altekar *et al*., in press).

A Markov chain is a programmed sequence (or "chain") of random samples taken at a specified interval during a walk through parameter space. Markov chains have the so-called "memoryless" property in that the probability that the chain moves from state x at time n to state y at time $n + 1$ does not depend on states visited prior to time n (Jones and Browning, 2003). Put more simply, "given the present, the past and future [samples] are independent" (Lee, 1997). A chain programmed correctly will be both aperiodic (will not get stuck in cycles) and irreducible (not having states with no paths joining them; Jones and Browning, 2003). Such a chain will produce random samples from the posterior distribution, weighted by their respective posterior probabilities.

We saw earlier that we can summarize the necessary phylogenetic statistical parameters with the variable $\theta = \{\tau, v, b, p\}$, where τ is a unique tree with a particular topology and specific combination of branch lengths, substitution parameters and gamma shape parameter. In the context of MCMC, it is easiest to think of θ as a unique point in parameter space; a perturbation in any of the statistical parameter values will necessarily alter θ to θ' and therefore define a new point in parameter space. A Markov chain works by sampling parameter value combinations (θ) as it moves randomly through parameter space. If the chain is constructed wisely (above) and is run long enough then samples from the Markov chain are valid samples from the posterior probability distribution of interest, commonly the posterior distribution of trees in phylogenetics (Altekar *et al*., in press).

Parameter values are perturbed in two ways in the MCMC implemented in MrBayes 3.0, one way for substitution model parameters and a second way for topology and branch lengths. Both methods make use of the Metropolis-Hastings algorithm, which simply determines the acceptance probabilities of new states (this will be developed verbally below). For a substitution model parameter x, a window of width δ is centered on the current value of x (Figure 4.1). The new proposed value of x is chosen through multiplying δ by a generated uniform random deviate (Lewis, 2002). If the proposed value of x lies outside of the range of x (i.e. less than zero) then it is reflected back into the range (essentially, the proposed value is the *absolute value* of the product of δ and the generated uniform random deviate).

Choice of an appropriate δ is very important, as an incorrect value may inhibit the mobility of the Markov chain. If δ is too small, then it will take the Markov chain a great (impractical) amount of time to navigate through parameter space. If, on the other hand, δ is overly high, then proposed steps will be too large and consequently will rarely be accepted. It should also be noted that each statistical model parameter will have a different window width, δ_i, as each of these parameters may have different ranges and different "malleabilities".

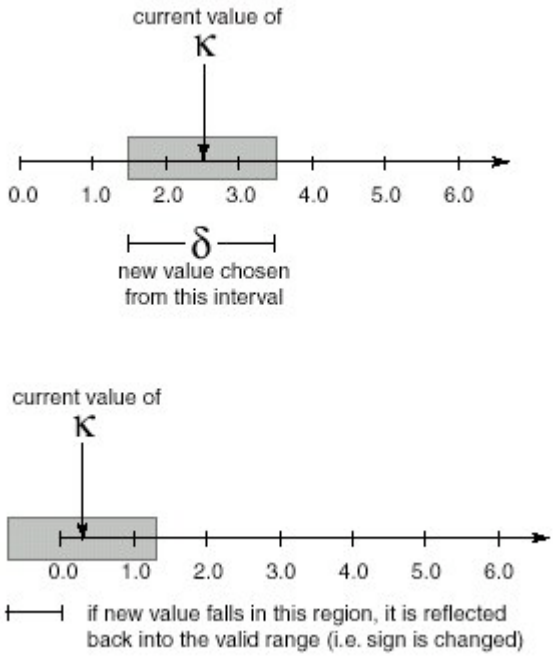

Figure 4.1: Proposal windows of width δ centered on the current value of a parameter, in this case the transition/transversion rate ratio parameter, κ (figure taken from Lewis, 2002). Only values of the parameter that lie within the shaded window are proposed. Proposed values that are less than zero are reflected back into the positive range.

Let the probability of proposing the new state (x') conditional on starting at the current state (x) be $q(x')$, and the probability of proposing the old state (x) conditional on the new state (x') be $q(x)$ (this move is never done; Altekar *et al.*, in press). Acceptance of the new state x' is determined with probability R:

$$R = \min\left[1, \frac{f(X \mid \theta')}{f(X \mid \theta)} \cdot \frac{f(\theta')}{f(\theta)} \cdot \frac{q(\theta)}{q(\theta')}\right]$$

where $\frac{f(X \mid \theta')}{f(X \mid \theta)}$ is the likelihood ratio, $\frac{f(\theta')}{f(\theta)}$ is the prior ratio, and $\frac{q(\theta)}{q(\theta')}$ is the proposal ratio.

What this equation says is that if the acceptance ratio is larger than one (meaning that the proposed state has a higher posterior probability than the current state) then the change in states is accepted, allowing $\theta = \theta'$. If the acceptance ratio is less than one, however, a uniform random variable, *U*, on the interval (0,1) is generated. If *U* is less than the acceptance ratio then the proposed state is accepted as above, otherwise θ remains unchanged and is recorded as another sample. It should be noted here that this algorithm never terminates. It is therefore up to the investigator to determine how long to run the chain, which essentially establishes the number of samples collected. Felsenstein (2003) relates this to the famous gambling casino at Monte Carlo, the namesake of MCMC methods. In a phylogenetic analysis we have the same aspirations of the house: that given enough samples the expected results will be reached. A chain that is not run sufficiently long enough (not enough samples) can therefore have drastic effects on the conclusions reached. We will return to this idea later.

Put more simply, the equation above says that proposed states of higher posterior probability are always accepted, while proposed states that lie "downhill" are accepted with a probability inversely related to the extent of the "drop" from the current state to the proposed state. This is shown visually in Figure 4.2. Proposed states that lie only slightly downhill, for example, will yield acceptance ratios near 1.0; generating a uniform random number on (0,1) that lies below this value will occur with high probability, and so acceptance of the proposed state is likely. If the proposed step takes the chain over a "cliff" in parameter space, however, the acceptance ratio will be near zero; generating a uniform random number on (0,1) that lies below this value will not occur often, and so acceptance of the proposed state is much less likely.

The result is that the Markov chain, through the acceptance probability used here, does tend to go uphill (in the high posterior probability peaks in parameter space), but the algorithm is not a strict "hill-climber" as in conventional phylogenetic heuristic algorithms. The reason for this is simple: Bayesian inference is interested in the *shape* of the posterior landscape, not merely in finding the highest peak (Lewis and Swofford, 2001). Such downhill steps are required if valleys are to be traversed and new peaks explored. The ability of a Markov chain to explore isolated, high probability peaks is called the "mixing" of the chain. A chain that gets stuck on a particular peak thus exhibits poor mixing, and is not sampling all highly probable states. Poor mixing can lead to skewed results, the reason being alternate states (potentially of equal or higher probability) are not sampled. It should be noted that steps over cliffs in parameter space, though of very low probability of acceptance, *will* occur eventually given that the Markov chain is constructed correctly and is run for a sufficient amount of time. However, such events may take a prohibitive amount of time. We will return to the problem of poor mixing in later sections.

Topology and branch lengths are perturbed together and in a very different manner from the substitution model parameters. The process is illustrated in Figure 4.3 and is described in Larget and Simon (1999) as "LOCAL WITHOUT A MOLECULAR CLOCK". First, an internal branch on the current tree is randomly selected along with two of its neighbours, for a total contiguous length of *m*. Second, the entire segment is expanded or shrunk by a small random amount according to

the equation $m^* = m \cdot e^{\lambda(U-0.5)}$ where U is a uniform random variable on (0,1) and λ is a tuning parameter. Third, one of the two branches that intersect with the selected segment is selected with equal probability and is detached from the tree. Lastly, the detached subtree is reattached to the original segment at a point chosen through the use of another uniform random variable. Such a move may result in a change of topology (and three branch lengths) if the subtree is reattached on the opposite side of the internal node than before (Lewis, 2002). If it is reattached on the same side of the internal node then the resulting tree will have the same topology as before but with slightly different branch lengths. The end result of this entire process is a new proposal state for the Markov chain; acceptance of this change in state is subject to the same acceptance probabilities as above.

In a typical Markov chain only one or a few parameters are perturbed at a time. The reason for this is that proposals requiring excessive changes to the state (θ) of a Markov chain are generally accepted with very low probability. However, sometimes more complex proposals are required. A prime example of this is when there exists a significant co-linearity between two

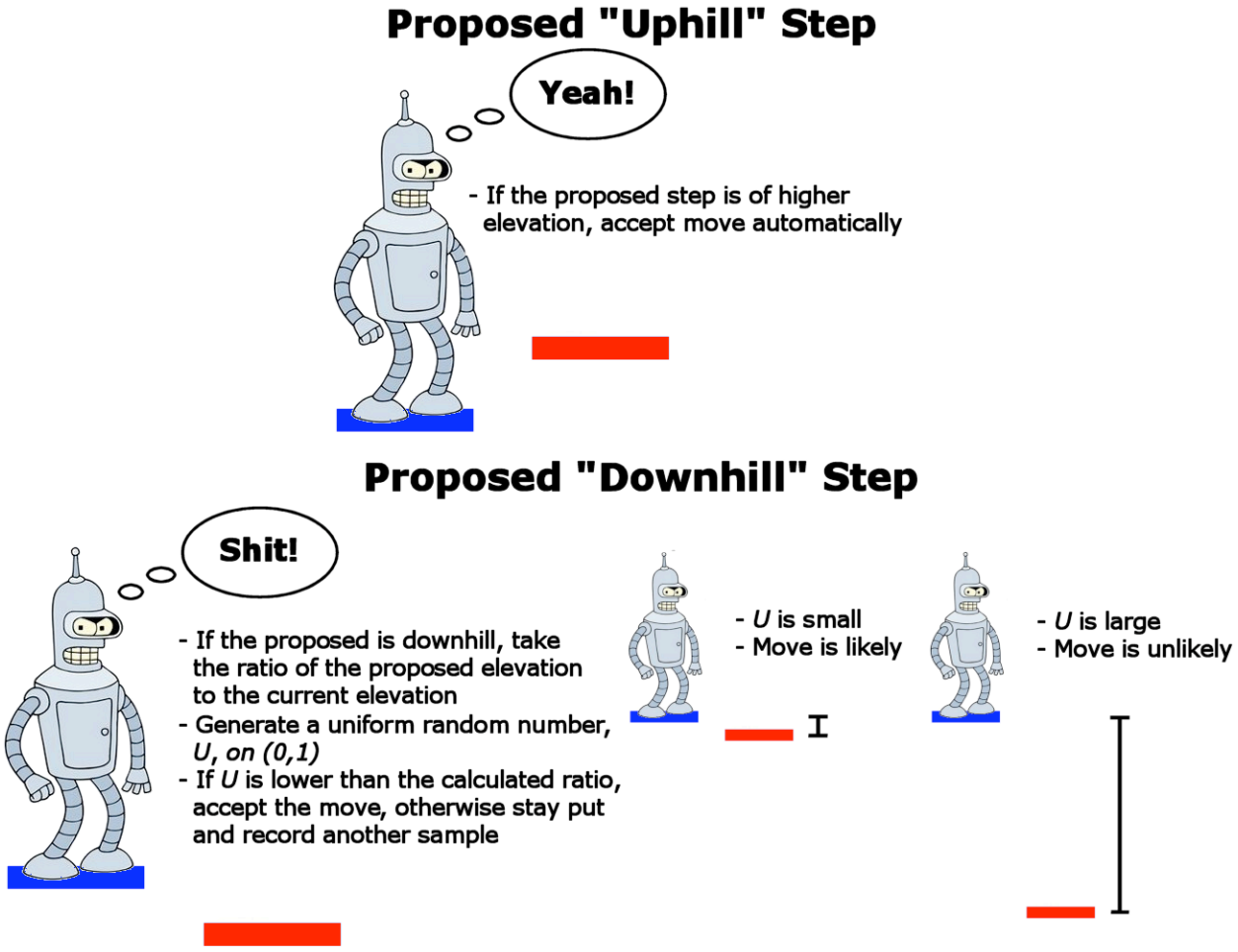

Figure 4.2: The Metropolis-Hastings algorithm, which determines the acceptance probabilities of new (proposed) states. Proposed "uphill" steps are always accepted (as we are interested in regions of high probability), but "downhill" steps are accepted with a probability inversely related to the extent of the "drop" from the current state to the proposed state. Using this algorithm, drastic drops in elevation are unlikely (but not impossible). Allowing suboptimal state changes allows the chain to traverse valleys in parameter space, and therefore permits more thorough exploration.

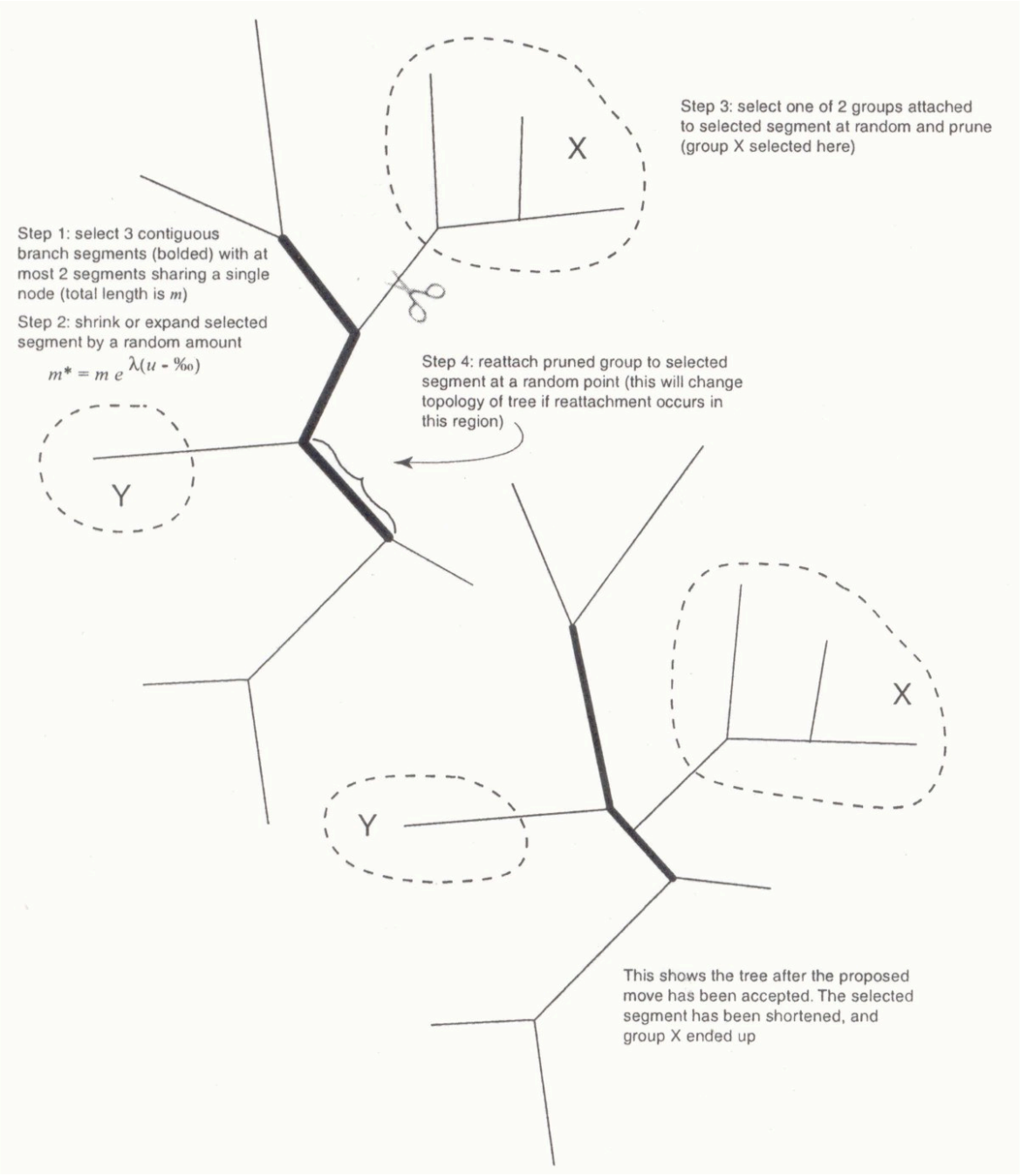

Figure 4.3: The LOCAL WITHOUT A MOLECULAR CLOCK algorithm of Larget and Simon (1999), used for navigating through tree space using (figure taken from Lewis, 2002). Acceptance of a proposed state using this algorithm will entail, minimally, a slight change in branch lengths and maximally a change in both branch lengths and topology.

parameters (Figure 4.4). In this situation, altering a single parameter value will require stepping off a cliff in parameter space. Such a step will be taken with extremely low acceptance probability. As a result the chain will remain in a single position and will not explore the rest of the high posterior probability peak. Adequate searching of parameter space in this case would necessitate a proposal scheme where both parameters are altered together.

MCMC has the desirable property that once having reached the stationary (equilibrium) distribution, the values of the parameter of interest (e.g. tree topology) are visited in proportion to their posterior probabilities. Because topology is a discrete parameter, one need only count the number of occurrences of a particular topology in the MCMC sample and divide this by the total number of trees (not topologies) sampled to obtain an estimate of the posterior probability of that particular topology. Recall from above that parameters of no direct interest (nuisance parameters) are integrated out. Therefore, the posterior probability of a particular tree topology is: 1) a directly interpretable probability of that particular topology being correct, and 2) takes into account uncertainty in *all* other parameters involved in the statistical model.

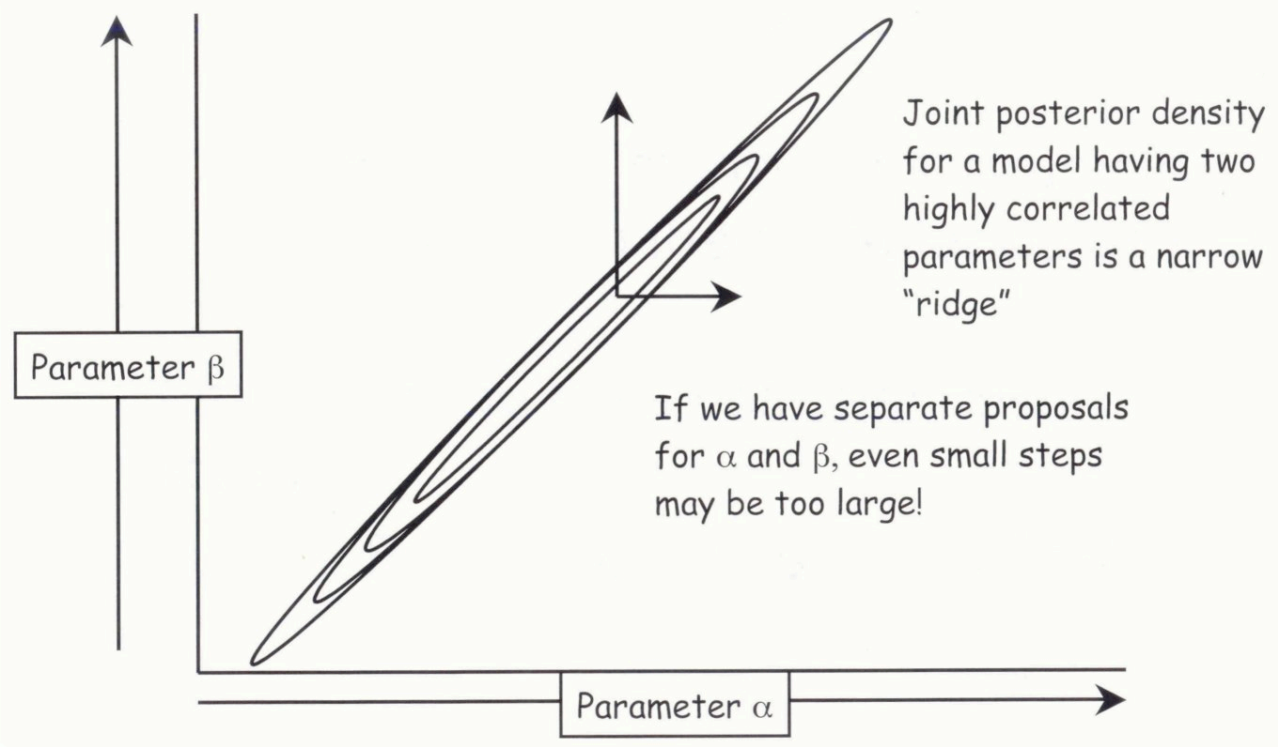

Figure 4.4: The effect of co-linearity between model parameters (figure taken from Holder, 2003). Taking a step in either one direction will entail an extreme drop in posterior probability and consequently will not be accepted with high probability. Efficient exploration of the high posterior probability ridge will require varying both parameter values simultaneously.

A properly constructed Markov chain will eventually sample from the stationary distribution, but this may take many, many generations to occur. The reason for this is that a chain is usually (ideally) started at a random point in parameter space which may be a fair distance from the posterior probability peak of interest. The samples taken by the Markov chain prior to reaching the peak are clearly not from the distribution of interest (have essentially zero probability) and so are discarded prior to sample summary (Lewis, 2001). This discarded portion is referred to as the "burnin" of the chain.

The level of burnin required is determined *post hoc* through the use of history plots (Figure 4.5). History plots typically have log likelihood (or log probability) on the y-axis and steps (iterations) on the x-axis. Likelihood is typically low at the beginning of a chain reflecting the fact that it was not initiated in the distribution of interest. As the number of steps increases the likelihood begins to climb rapidly. This climb in likelihood reflects the Markov chain converging on the

stationary distribution. Convergence of independent runs (starting from random positions in parameter space) on the same results is a way to verify findings. At a particular point the likelihood will plateau and at this point it can be assumed that the Markov chain is sampling from the stationary distribution. The likelihood will bounce around about this point as samples from the posterior distribution are taken (periods of constant likelihood indicate poor mixing; Lewis, 2002). Burnin is thus all of the samples taken before the plateau is reached. Care must be taken that the chain has been run long enough – a history plot of a chain getting stuck in a local optima will look very much like a chain sampling from the posterior probability distribution.

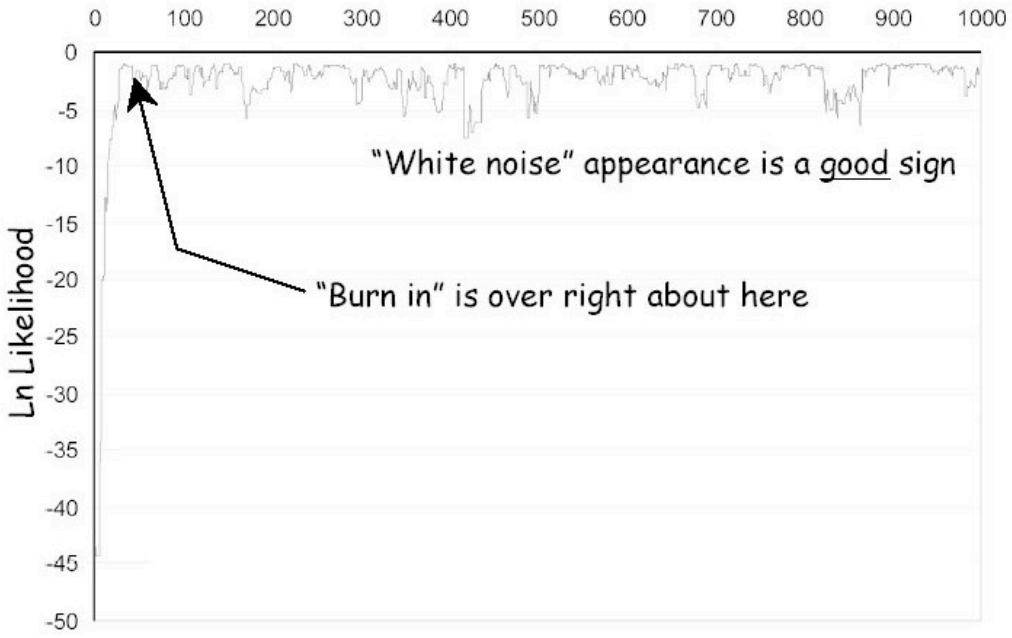

Figure 4.5: A history plot for a Markov chain, with log likelihood on the y-axis and number of steps (iterations) in the chain on the x-axis (figure from Holder, 2003). Likelihood climbs initially as the chains progresses towards peaks of high posterior probability. When the likelihood plateaus then the chain is sampling from the distribution of interest. Samples recorded before this plateau is reached is the burnin of the chain and are discarded. The bouncing around of the likelihood after reaching the plateau indicates that the chain is exploring the high posterior probability peak and not getting stuck in a particular point in parameter space.

Despite the "memoryless" property of the Markov chain, due to the sampling nature in the MCMC approximation of the posterior probability distribution samples from the MCMC chain are somewhat autocorrelated. The degree of autocorrelation depends largely on the construction of the Markov chain, in particular the choice of proposal window widths, δ_i. If, for example, δ_i is very small, then successive samples from the Markov chain will have very similar parameter values for the parameter of interest (i.e. samples will have a high level of autocorrelation). What is more, different parameters may be autocorrelated to different degrees. Samples from an MCMC chain are therefore valid, albeit *dependent*, samples from the distribution of interest (Altekar *et al*., in press). If samples are overly autocorrelated then a substantially larger sample must be taken in order to draw valid conclusions (Tierney, 1994).

Autocorrelation among samples can be combated by "thinning" the Markov chain. Thinning an MCMC chain means simply that not all samples are recorded; rather, samples are recorded

periodically at a rate that can be specified by the investigator. For example, an investigator might choose to record every 100th sample in their Markov chain (a "lag" of 100 steps). Recording non-successive samples decreases autocorrelation (increases independence) between samples. Some applications make use of an autocorrelation plot (Figure 4.6) to determine optimal thinning. An autocorrelation plot informs the investigator when correlation (for a particular parameter) between samples has decreased to an acceptable level (Jones and Browning, 2003). MrBayes does not presently have a simple way to determine optimal thinning, but investigators can "play it safe" by conservatively thinning the Markov chain (e.g. every 50 or 100 steps). This does not come without a price, however. Thinning a Markov chain necessitates that the chain be run that much longer to obtain a sample of equivalent size. This can be understood through the use of a simple equation:

$$chain\ length = \#\ recorded\ samples \times (\#\ skipped\ samples + 1)$$

The result of thinning is a near-linear increase in computation time needed to complete the analysis, which could mean an unreasonable increase in runtime for computationally expensive problems. Ideally, then, thinning should be optimized for *each* parameter through use of autocorrelation plots. Hopefully such a tool will become available soon for use with MrBayes.

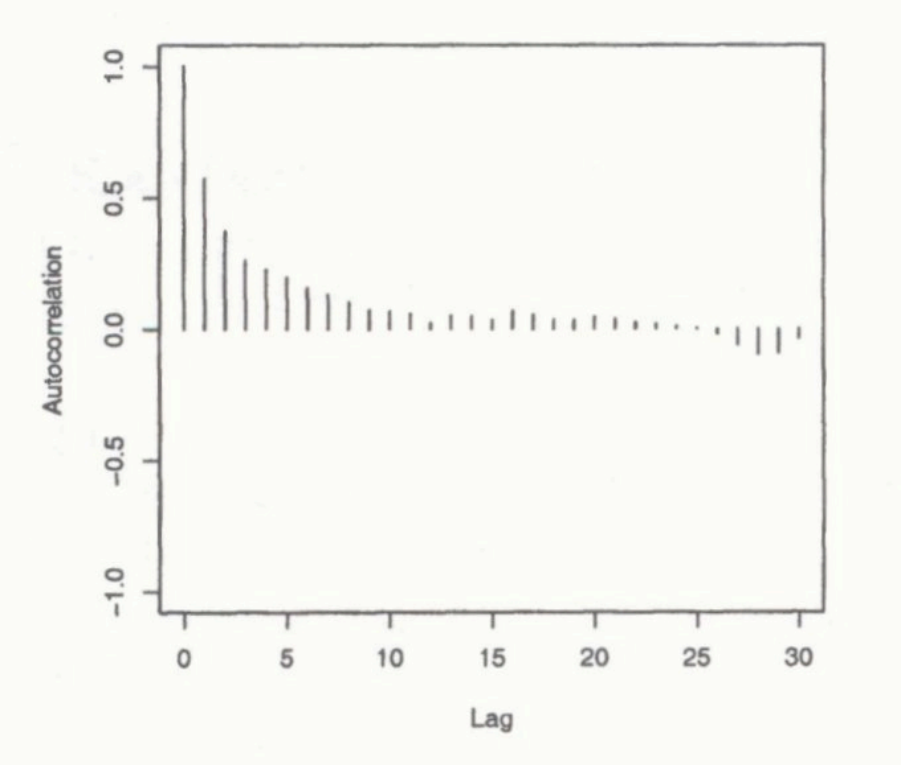

Figure 4.6: An autocorrelation plot (figure from Jones and Browning, 2003). For a particular parameter, samples from a Markov chain are autocorrelated to some degree. This tool is used to determine optimal thinning of a Markov chain. The investigator determines what level of autocorrelation is acceptable and sets the lag in their Markov chain accordingly.

However, MCMC is not necessarily the panacea it first appears. As Geyer (1999) states (quoted in Huelsenbeck *et al*., 2002) "MCMC is a complex mixture of computer programming, statistical theory, and practical experience. When it works, it does things that cannot be done any other way, but it is good to remember that it is not foolproof." The first concern is that enough

samples are taken to satisfactorily approximate the posterior distribution. A chain that is run too short will not sample all of the probable states densely enough, and may lead to skewed (or blatantly inaccurate) results.

A second concern is that MCMC can potentially get stuck in local optima like heuristic searches used to find optimal trees in other frameworks (Altekar *et al.*, in press). If a chain is sampling from one peak of high probability it may be difficult to cross deep valleys in tree space to search other peaks. If the Markov chain is constructed well then it will eventually cross the deep valleys *if* the chain is run long enough. However, because the steps required would occur with such low probability, it would take an exorbitant amount of time to explore a particularly rugged posterior probability distribution of trees; in other words, the chain would suffer from poor mixing. As a result, many peaks (trees) may not be visited. Metropolis coupled MCMC, $(MC)^3$, appears to be an effective method for improving the mixing of MCMC.

METROPOLIS COUPLED MARKOV CHAIN MONTE CARLO [$(MC)^3$]

Metropolis coupled Markov chain Monte Carlo, $(MC)^3$, is a variant of the MCMC algorithm above, and involves n Markov chains running concurrently, $n - 1$ of which are heated (Huelsenbeck and Ronquist, 2001a). Each chain computes the posterior probability for the currently sampled value for the parameter of interest (usually topology in phylogenetic studies) and then raises the posterior probability to a power β (Altekar *et al.*, in press). β is the heat value, or "temperature", of the chain and takes on the values $0 < \beta < 1$. In MrBayes chains are incrementally heated and β is calculated as follows:

$$\beta = \frac{1}{1 + T(i)}$$

where i is the labeled Markov chain ($i = 0, 1, 2, \ldots, n - 1$) and T is a temperature parameter that is set to an appropriate value (0.2 by default in MrBayes, but this value may be changed by the investigator; Huelsenbeck and Ronquist, 2001b). The unheated, or "cold", chain is labeled 0, and as such is raised to the power 1 making it unaffected by heating; heated chains are raised to powers between 0 and 1. Heated chains thus have a posterior probability distribution of the form $\Pr(\theta|X)^\beta$. The process of heating a chain effectively "melts down" the posterior probability landscape, making valleys less deep and peaks less high, even though all chains are exploring the same parameter space. As an extreme example, a chain raised to the power $\beta = 0$ would be exploring a completely flat landscape as the posterior probability value at every point in parameter space would be equal to 1 (Lewis, 2002). The effect of heating can be seen visually in Figure 4.7.

Heating of Markov chains ultimately has the effect of increasing the acceptance probability of new states (Altekar *et al.*, in press), as can be seen in the equation below:

$$R = \min\left[1, \left[\frac{f(X|\theta')}{f(X|\theta)}\frac{f(\theta')}{f(\theta)}\right]^{\beta}\left[\frac{q(\theta)}{q(\theta')}\right]\right]$$

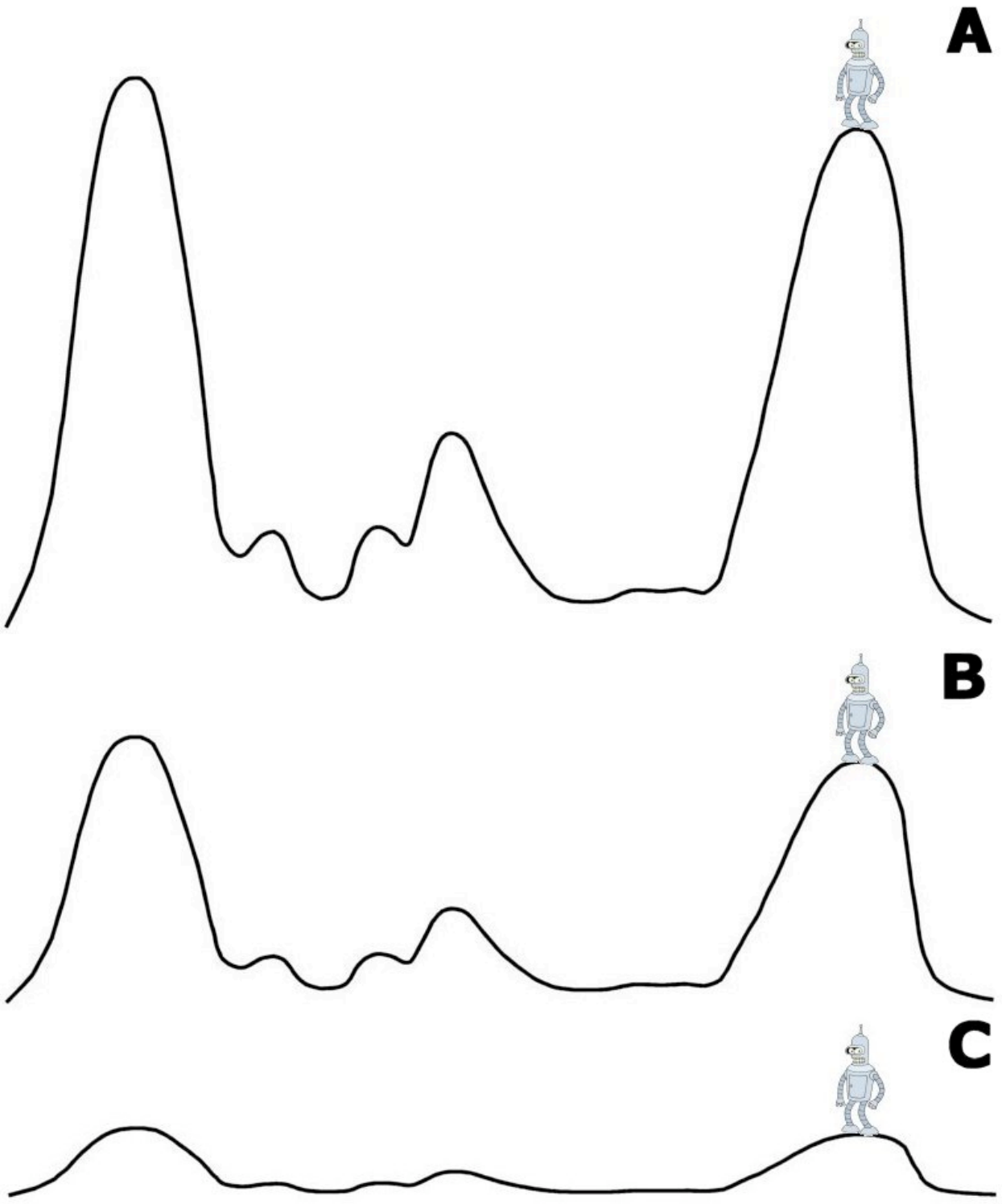

Figure 4.7: "Heating" of a hypothetical posterior probability surface. Robots represent Markov chains exploring the probability surface. Figure A represents a cold surface: the posterior probability of the chain exploring this surface is raised to the power 1 and hence in unaffected by heating. Figures B and C represent successively heated landscapes; chains exploring these surfaces would have their respective posterior probability ratios raised to a power $0 < \beta < 1$. Each chain in A, B, and C are exploring the same parameter space. Heating has the effect of "melting" the posterior probability landscape, decreasing extremes in peaks and valleys. A heated chain can more easily cross valleys because the required "downhill" steps are smaller in magnitude.

Again, a uniformly distributed random number is generated on the interval (0,1); if the random number is less than the acceptance probability above then proposed state is accepted. As a result heated chains tend to accept new states more readily than a cold chain, and so can more easily cross valleys in the posterior probability landscape because the landscape has lesser "altitudinal" extremes relative to the same landscape for the cold chain (Altekar *et al*., in press). As such the heated chains can more easily move between isolated hilltops than cold chains that may get stuck on local optima, effectively better exploring parameter space.

Despite the increased mobility of heated chains, their sole function is to provide the cold chain with intelligent proposals of new states (Huelsenbeck *et al*., 2002). The reason for this is that only the cold chain records samples from the posterior probability distribution; heated chains act merely as scouts, searching the surface of the posterior probability distribution for isolated areas (peaks) of high probability. As such the chains must be periodically in communication with one another, adding yet another level of complexity to the analysis. Some researchers like to think of the various chains in an $(MC)^3$ analysis as separate robots exploring the landscape in tree space,

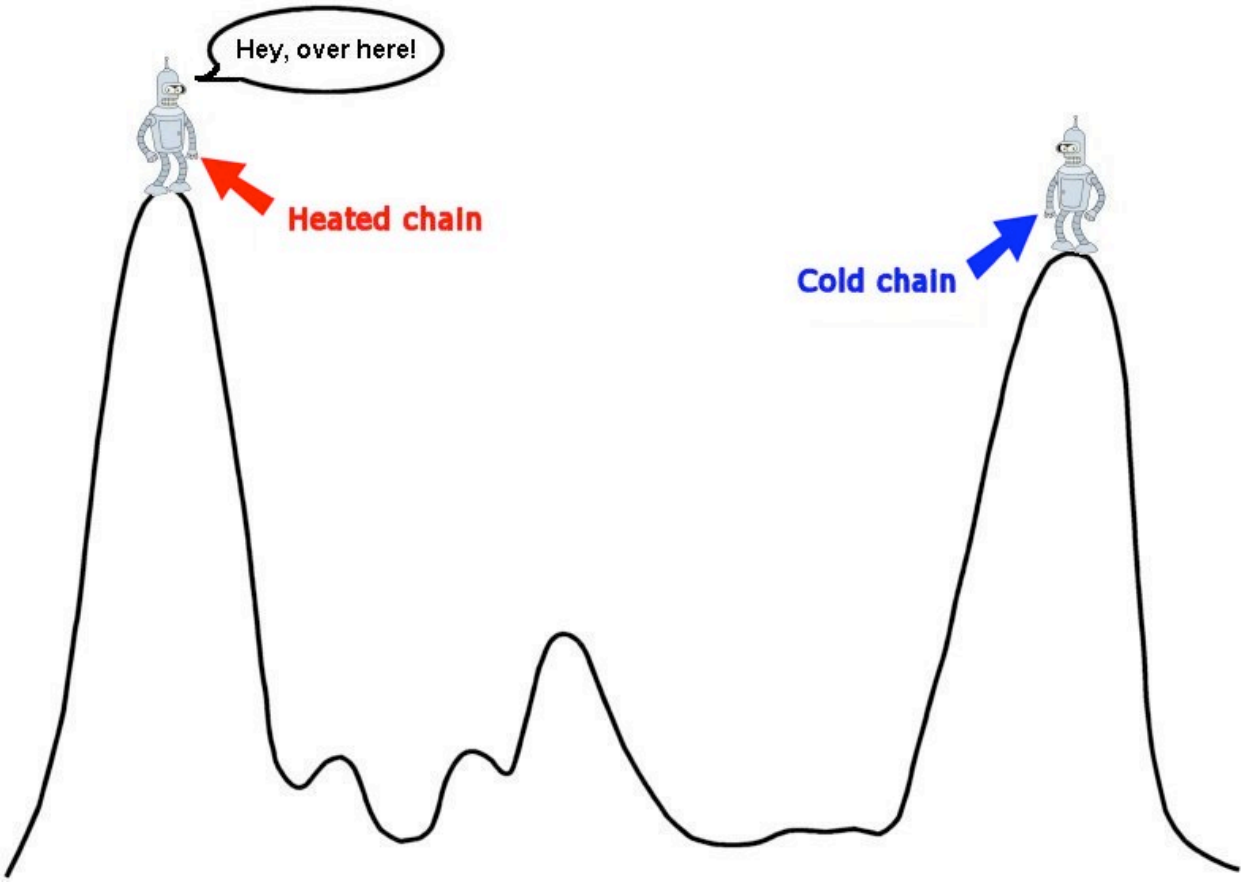

Figure 4.8: Communication between Markov chains in an $(MC)^3$ analysis. The function of heated chains is to provide the cold chain with intelligent proposals of new states. Because heated chains can more easily cross valleys in the posterior landscape, they can identify regions of substantial posterior probability that may be quite far away from the region simultaneously being explored by the cold chain. The result is that by swapping states with a heated chain, a cold chain may traverse a particularly deep valley in one step rather than the many (unlikely) steps required otherwise. Swapping thus acts to increase the mixing ability of the cold chain.

each equipped with a walkie talkie to communicate their respective "altitudes" to one another, as in Figure 4.8.

The communication referred to above occurs at a set interval (say, each iteration) and only involves two chains at a time. At the set period two chains, *j* and *k*, are chosen at random to exchange states, commonly referred to as "swapping". The chains communicate their state information and accept a swap with probability:

$$R = \min\left[1, \frac{f(\theta_k \mid X)^{\beta_j} f(\theta_j \mid X)^{\beta_k}}{f(\theta_j \mid X)^{\beta_j} f(\theta_k \mid X)^{\beta_k}}\right]$$

This equation simply compares the product of the likelihoods of the two states before and after the proposed swap. As with the acceptance probabilities above, if a uniformly distributed number on (0,1) is less than quantity calculated above then *j* and *k* swap states; otherwise the chains proceed to the next iteration with unchanged state information. Swaps can occur between heated chains or between a heated chain and a cold chain. The potential for swapping is truly the beauty of Metropolis coupling in MCMC analyses. In $(MC)^3$ a cold chain that may be stuck on a particular peak can jump to another isolated peak in one generation, a process that may take a prohibitive amount of time (depending on the breadth of the valley separating the peaks) in the absence of Metropolis coupling (Lewis, 2002). The function of Metropolis coupling is thus to facilitate the mixing of the cold chain within the posterior probability landscape. The addition of several heated chains can allow a more appropriate sampling of the posterior probability distribution by the cold chain, sampling peaks (trees) that may not have been visited in a conventional MCMC analysis (at least, not within an allowable period of time).

The advent of Metropolis coupling had an immediate impact on the field of phylogenetic inference (amongst others), and continues to do so. Huelsenbeck *et al*. (2002) remark how the limits of phylogenetic analysis have been extended because of $(MC)^3$. The largest $(MC)^3$ model-based phylogenetic analysis conducted to date (Huelsenbeck *et al*., 2001) integrated over a "tree space that is several hundred orders of magnitude larger than the tree spaces that have been successfully analyzed without Metropolis coupling." Such an analysis could have never been performed using MCMC alone. The size and complexity of problems that can be handled by MCMC are determined primarily by convergence and mixing of the chains (Huelsenbeck *et al*. 2002), so methods that improve in these two areas will extend the scope of phylogenetic inference. In the coming years we will without a doubt see the sizes of analyzable data matrices increase far beyond the largest data sets of today (let alone those analyzable by maximum likelihood), and this is due in large part to Metropolis coupling.

The benefits of Metropolis coupling in MCMC analyses are undoubtedly great, but they also come with a price. Each additional heated chain added to the analysis considerably increases the time to completion (Lewis, 2002). The reason is simple: within each chain, each iteration requires the calculation of a computationally expensive likelihood function; running *n* chains therefore requires *n* calculations of the likelihood function each iteration. What is more, each chain requires a burnin, which is wasted computing effort (Jones and Browning, 2003). This constraint forced investigators to consider the tradeoff between the necessity for running multiple heated chains (at least 4 chains are required for sufficient mixing; Altekar *et al*., in press) to better explore parameter space and the requirement of running the cold chain long enough to

obtain a sufficiently valid sample from the posterior probability distribution from which to draw meaningful conclusions. The recent advent of parallel computing, however, has greatly diminished this conflict (Ronquist and Huelsenbeck, in press).

PARALLEL METROPOLIS COUPLED MARKOV CHAIN MONTE CARLO [P(MC)^3]

Altekar *et al*. (in press) describe the implementation of a parallel version of MrBayes in order to explore the relationship between the increase in the number of chains (processors) for better mixing and the scalability of the computations. The beauty of parallel Metropolis coupled Markov chain Monte Carlo, hereafter referred to as p(MC)^3, is that the Markov chains are spread across processors (one chain per processor) and in essence can all run simultaneously rather than sequentially as on a conventional desktop computer. The actual implementation of p(MC)^3, however, is somewhat more complex than simply spreading chains across processors, as swapping between chains (the purpose of running multiple chains, and the way to achieve better mixing) must be accounted for. In order for successful swapping, chains must both synchronize and communicate state information. Since chains are being run on different processors, communication between chains necessitates communication between processors. If this communication cost is too severe then it will degrade the scalability of the parallel analysis.

To minimize costs Altekar *et al*. (in press) used a number of programming tricks to decrease the amount and frequency of information transmitted between processors. First, rather than exchanging state information between chains, "temperatures" were exchanged instead. The reason for this is two-fold. State information of a chain includes tree data structures and associated likelihoods, resulting in several megabytes worth of data compared with a few bytes of information when exchanging heat values. Additionally, exchanging of temperatures requires only one round of communication while exchanging of state information require two bouts. This is accomplished by communicating the random number used in decision making with the swap acceptance information (both chains must make identical swap acceptance decisions). The result of this process is that state information is effectively swapped between chains through exchange of a few bytes of data.

The second programming trick involves synchronization of the chains, which is a requirement for swapping. The implementation of p(MC)^3 in the parallel version of MrBayes makes use of a point-to-point exchange scheme, minimizing idle processor time. In this scheme it is recognized that only two chains are involved in a swap at one time. The identity of the chains involved in the swap in each iteration is predetermined by a pseudo-random number sequence available to each processor. A chain can therefore check to see if is involved in a swap in the present generation (and, indeed, generations to come) by checking the random number sequence. The result is that chains uninvolved in the present swap can proceed to the next generation, rather than all chains being synchronized each generation (global exchange scheme; Figure 4.9). What is more, because swapping involves only two chains, the amount of synchronization is not a function of the number of chains (processors). This ultimately makes the analysis more computationally efficient by greatly reducing idle time of processors waiting for swaps to complete.

To test the speedup of the p(MC)^3 algorithm in MrBayes Altekar *et al*. (in press) observed run times for both large (# of species) and small data sets with various numbers of chains. Speedup was determined by comparison of parallel run times with analogous sequential run times. Small

data sets have comparably simpler likelihood calculations, so communication between chains is more important and can be costly. Large data sets, on the other hand, have more complicated likelihood functions, and so communication between chains is less costly. In both data sets swaps were made a requirement every generation to obtain worst-case communication costs.

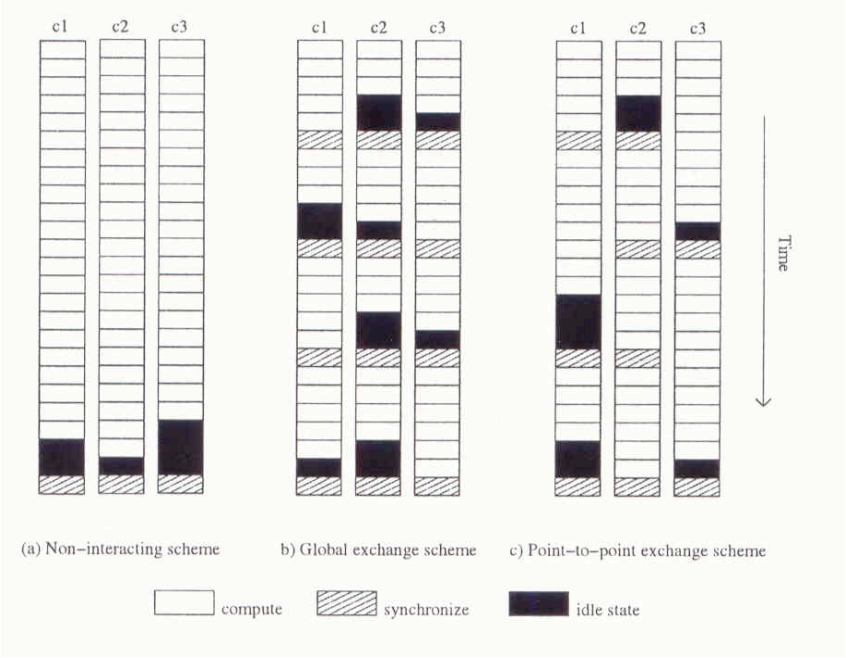

Figure 4.9: Possible communication schemes between parallel MCMC chains (figure taken from Altekar *et al*., in press). In the non-interacting scheme, idle time is minimized but chains do not communicate with one another, thereby not increasing mixing efficiency. Such a scheme could be implemented if results from one Markov chain needed verification. Practical phylogeneticists would not use the non-interacting scheme. In the global exchange scheme, all chains are synchronized at a predetermined set interval. This scheme ensures that chains are in step, a requirement for swapping state information. The point-to-point exchange scheme capitalizes on the fact that only two chains are involved in a swap at a time, and so synchronization of all chains effectively wastes valuable computation time. Chains not involved in a swap are allowed to proceed to the next iteration.

Figure 4.10 shows that comparable near linear speedups were found for both data set sizes. This high scalability permits an investigator to run multiple chains (far more than previously possible)and thus allows for better mixing and exploring of parameter space – a necessity for particularly rugged landscapes.

It is quite conceivable that the implementation of $p(MC)^3$, with its high level of concurrency and low communication costs, will have as large an impact on Bayesian phylogenetic inference as did Metropolis coupling, further extending the limits of modern model-based phylogenetic analysis. The only obstacle to this appears to be the present limited access to high performance computing resources.

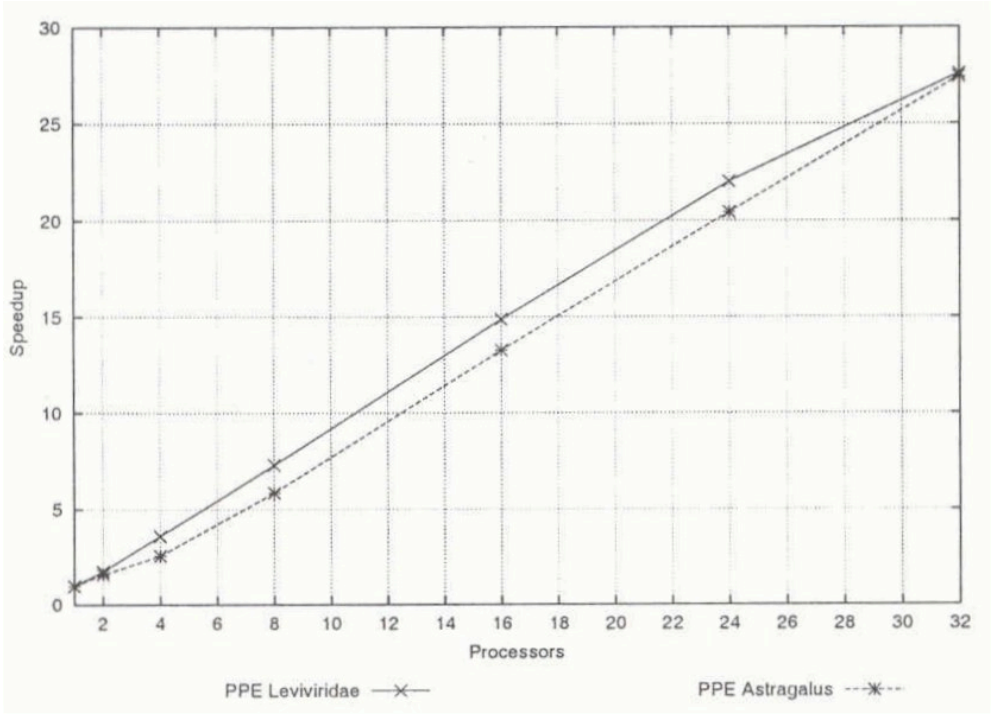

Figure 4.10: Achieved speedup of MrBayes 3.0 through use of parallel Metropolis coupled Markov chain Monte Carlo (p(MC)^3; figure taken from Altekar *et al.* in press). "Speedup" is measured by comparing runtime using x processors with the same analysis (# generations, chains, etc.) run serially. Leviviridae represents a small (# of species) data set while Astragalus represents a large data set. Both data sets scale near-linearly.

5. ASSESSING RELIABILITY IN PHYLOGENETIC TREES

The goal of many a phylogenetic investigation is ultimately a tree. However, a best estimate of a tree is rarely good enough; some level of confidence in the tree must also be expressed. Support indices are important, especially when trees serve as the conceptual framework for study of the evolution of particular a trait (Alfaro *et al*., 2003). As will be seen below, Frequentist and Bayesian frameworks construct and interpret reliability estimates in very different ways. Despite using identical stochastic substitution models and measuring similar attributes, the estimates are *not* interchangeable and cannot be directly compared (Douady *et al*., 2003). Nevertheless, a contrast can be useful to understand exactly what each statistic is measuring, how each can be interpreted, and why they might behave differently.

THE BEST TREE

In the traditional maximum likelihood approach, all parameters (those of interest plus the nuisance parameters) are jointly estimated and the MLE is obtained by finding the values of the parameters which maximize the likelihood function (Huelsenbeck and Ronquist, 2001b). If an exact search has been performed then the MLE will represent the global maximum of the likelihood surface. The best tree, then, is that tree which highest (most positive) log likelihood score. It should be noted here that no level of confidence whatsoever is attached to the MLE tree; though the tree may have the best likelihood score, there may be several (or even hundreds) of trees that have scores that are not significantly different from the MLE tree.

A Bayesian best estimate is obtained in much the same way as above, except here we are dealing with a posterior probability distribution rather than a likelihood surface. Typically the "best tree" is that with the maximum *a posteriori* probability, referred to as the MAP tree (Rannala and Yang, 1996). Three major distinctions should be made between MLE and MAP trees. First, while maximum likelihood uses jointly estimated parameter values, a Bayesian approach marginalizes over the nuisance parameters, and hence takes into consideration the uncertainty in the nuisance parameter values when determining tree posterior probabilities. Second, while one is guaranteed to find the best (MLE) tree in a traditional exact tree search (exhaustive or branch-and-bound), this is not necessarily true in the Bayesian case. The degree of approximation of the posterior probability distribution is determined by the MCMC search, most importantly by the length (number of samples) of the run. Because an MCMC analysis deals with random samples, a run that is too short will miss many trees, possibly even trees with very high posterior probability. This can be remedied, however, by running the analysis multiple times from random positions in parameter space. The final distinction deals with confidence in the best tree. We saw above that no confidence can be attached to the MLE tree, but the MAP tree is different. Because we are dealing with a (discrete) posterior probability distribution (i.e. all trees are recorded and sum to probability 1), every tree sampled in the MCMC search has a probability attributed to it. Thus, unlike the MLE tree, in the Bayesian approach it is possible to attach confidence to a point estimate.

CONFIDENCE IN A TREE

Ideally we would like to secure confidence estimates in trees and more importantly in clades of direct interest. Such indices of support (or reliability, or robustness) are important because they quantitatively illustrate how well results are jointly supported by both the data and the chosen

model of evolution (Cummings *et al*., 2003). We saw above that it is not possible to obtain these estimates in a conventional maximum likelihood approach, and the posterior probability for a MAP tree may be extremely small depending on the extent of the MCMC sampling and the number of candidate trees with probability greater than zero. So it is clear that confidence in a tree must be obtained in a different way.

In the traditional framework, statistical confidence is most often inferred through use of the nonparametric bootstrap, introduced to the estimation of phylogenetic trees by Felsenstein (1985). The procedure is summarized as follows: the original data matrix is sampled with replacement to produce *n* pseudo-replicate matrices, and a tree search is performed on each replicate. Groupings (taxon bipartitions) within the *n* trees are kept track of and the proportional frequency (multiplied by 100) of a particular grouping is equal to its bootstrap percentage, or simply bootstrap score. For example, if within 1000 bootstrap trees we find 850 of which contain the grouping of A and B, the bootstrap score for that grouping would be 85. Bootstrap scores are typically summarized on a majority-rule consensus tree, as in Figure 5.1. The bootstrap is an enormously useful tool that can be applied to virtually any type of analysis (Holder and Lewis, 2003). As such, the bootstrap procedure can be applied to any parameter in the phylogeny problem, though its use is limited almost exclusively to topology (Cummings *et al*., 2003). Despite the widespread use of the bootstrap, its interpretation is not directly intuitive and hence has been periodically subjected to upwellings of debate. A bootstrap proportion does *not* give an indication whether the result is correct; rather, it gives an idea of the repeatability of the results (i.e. if *n* more similar data sets were sampled, what percentage of them would we expect to contain the particular taxon bipartition?). Put another way, the distribution of pseudoreplicates around the observed data is a valid approximation of the distribution of observed data sets on the true, unknown process that generates the data (Alfaro *et al*., 2003). Nevertheless, most authors still tend to interpret bootstrap proportions as direct measures of endorsement for a particular bipartition. That said, the nonparametric bootstrap, though it does not deliver direct clade probabilities, is generally agreed by phylogeneticists to be a conservative measure of support.

Aside from the indirect interpretation of bootstrap proportions, a far greater concern deals with the computational burden involved (Holder and Lewis, 2003). If an analysis involves *n* bootstrap replicates it will effectively take *n* times as long as a single tree search on the original data matrix. For data matrices that currently take weeks or months to find a tree of highest likelihood, such a computation burden (e.g. 1000 bootstrap replicates) is clearly impractical and limits the application of the nonparametric bootstrap to model based phylogenetic problems (Douady *et al*., 2003). Investigators have typically taken two approaches to the problem of excessively long runs. First, rather than run exact searches (exhaustive or branch-and-bound) on each pseudoreplicate, a heuristic search is performed instead. Heuristic searches, though much quicker, suffer from failing that the tree with the highest likelihood is not guaranteed to be found. Consequentially, another source of variance, that of not finding the tree of maximum likelihood, is added to the analysis (Cummings *et al*., 2003). Secondly, researchers have turned to high performance computing laboratories for data analysis. Unfortunately such facilities are still the exception rather than the rule, and exact runs may still take several weeks to complete.

Posterior clade (or split) probabilities are obtained in a far different manner than bootstrap proportions. As reiterated throughout this paper, Bayesian statistics is interested in the shape of the posterior probability distribution rather than maximizing an objective function (Lewis and

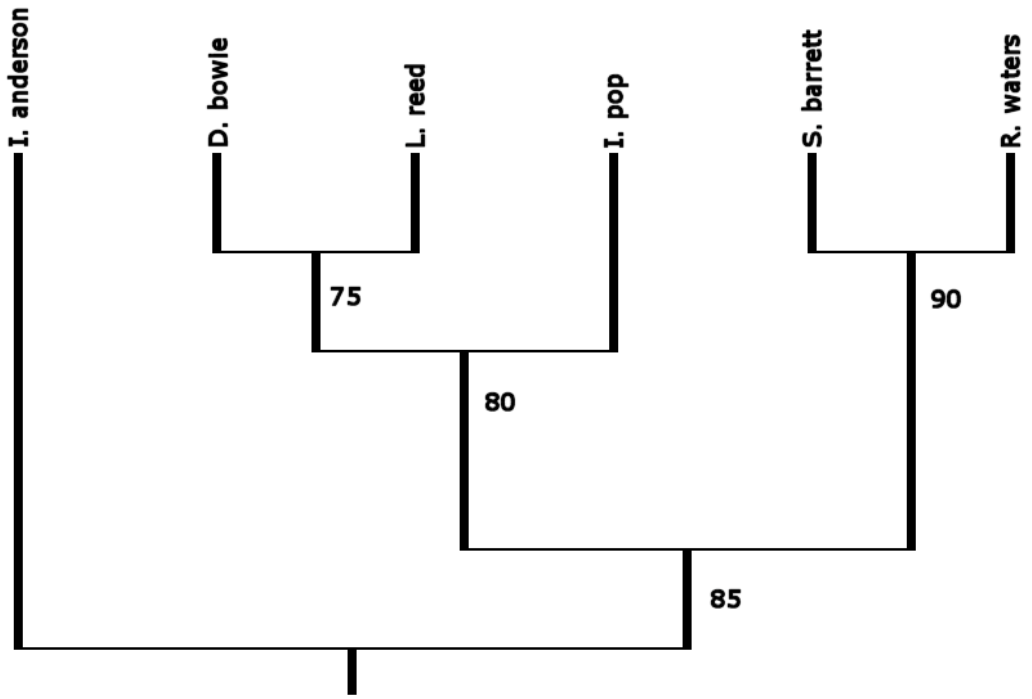

Figure 5.1: An example of a majority-rule consensus tree illustrating nonparametric bootstrap proportions for clades within the tree. The tree itself, as the names implies, is the one which occurs the most frequently throughout the bootstrap replicates. Numbers at the nodes indicate the proportional frequency that a particular grouping occurs across bootstrap replicate trees.

and Holder, 2003). The result is a simultaneous estimation of both phylogeny and support (Douady *et al*., 2003). As we have seen, Bayesians typically construct credibility intervals around parameters of interest based on the posterior probability distribution; this same procedure is used to generate reliability intervals on trees and clades. Summary of reliability estimates can take many forms. One way, first hinted at by Felsenstein (1968), creates a credible set of trees by starting with the MAP tree and adding trees in order of decreasing probability until the cumulative probability is, say, 0.95 (Huelsenbeck *et al*., 2002). The most popular, and perhaps intuitive, form of summary, however, involves applying the results of the MCMC analysis on the MAP tree (Larget and Simon, 1999). This is accomplished through computing posterior probabilities for the clades on the MAP tree from the sample of trees recorded during the MCMC analysis. Recall from above that posterior probabilities for discrete parameters (e.g. topology or taxon bipartition) are calculated simply as a proportional frequencies. In this respect the calculation of a posterior clade probability from the MCMC sample of trees is comparable to the calculation of a bootstrap score for the same clade from the bootstrap sample of trees. Summary of results in this form allows simultaneous display of both bootstrap proportions (or Bremer decay values, if that be your taste) and posterior clade probabilities.

Though both indices can be illustrated for the same clade, ostensibly measuring a similar quantity, some important distinctions between bootstrap proportions and posterior clade probabilities should be noted here. Firstly, posterior probabilities, unlike bootstrap proportions, are *not* measures of repeatability; rather, they can be directly interpreted as the probability that the underlying clade in question has been correctly recovered (Wilcox *et al*., 2002). Recall the interpretation of bootstrap proportions above (i.e. if *n* more similar data sets were sampled, what

percentage of them would we expect to contain the particular taxon bipartition?). A posterior clade probability, conversely, answers the question: given the data observed, what is the probability that the clade of interest is present in the true tree? (Lewis, 2002). The second distinction deals with uncertainty in tangent (nuisance) parameters. Likelihood methods, we have seen, jointly estimate all parameters; an MLE tree, then, deals with fixed values for the parameters of the statistical model. Bayesian methods, on the other hand, by virtue of exploring the shape of the posterior probability distribution rather than maximizing an objective function, marginalizes over nuisance parameters. Thus Bayesian methods explicitly incorporate uncertainty in these tangential parameters, which is clearly of preference because the values for these parameters are not known without error. Finally, bootstrap proportions and posterior clade probabilities differ in their respective computational burdens. While likelihood analyses require a run length of $n \prod t$ to construct bootstrap reliability (where n is the number of bootstrap replicates and t is the time required to complete one tree search), the Bayesian approach does not suffer from this syndrome. This is because of simultaneous estimation of phylogeny and support. We saw above that Douady *et al*. (2003) found that one Bayesian tree search in MrBayes ran 80 times faster than a single (heuristic) PAUP* maximum likelihood bootstrap replicate. Clearly this ratio will change depending on the data matrices and substitution models involved in the analysis (as well as the MCMC specifications – an analysis can be run arbitrarily long), but clearly the Bayesian approach has an enormous advantage in terms of computational efficiency. This property alone accounts for much of the excitement surrounding Bayesian phylogenetics, partly because it allows reliability estimates to be calculated for trees that are all but intractable in a likelihood framework.

DIFFERENCES IN RELIABILITY ESTIMATES

Given that it is clear that split probabilities and nonparametric bootstrap proportions measure fundamentally different attributes of the data, it is not surprising that discrepancies can arise when analyzing a data set using both frameworks. Instinctively, we would hope that the values are in agreement most of the time, and expect only slight differences of no real consequence. In fact, *Efron et al*. (1996) noted that while the theory for each index is largely independent, bootstrap proportions and posterior clade probabilities should be equivalent. However, many studies have shown that the indices in fact do differ to a large degree in many circumstances. In fact, Huelsenbeck *et al*. (2002) stated that "Perhaps the most vexing mystery is the observed discrepancy between Bayesian posterior probabilities and nonparametric bootstrap support values." This mystery has been the focus of several papers since, and the findings to date are explained below.

In empirical circles it is almost a ubiquitous result that posterior probabilities are consistently higher than analogous bootstrap proportions (Buckley *et al*., 2002; Leaché and Reeder, 2002; Whittingham *et al*., 2002), however it wasn't until Wilcox *et al*. (2002) that the relationship between these two indices were rigorously examined. They generated 120 data sets of 500 bp in length based on the maximum likelihood estimated tree, and bootstrap proportions for subsequent likelihood searches were compared to analogous posterior probabilities for each data set. The two methods were then contrasted for phylogenetic estimation accuracy *sensu* Hillis and Bull (1995), defined as the probability of reconstructing the correct bipartitions. They found that the average support values were consistently higher (sometimes twice as high) for Bayesian compared to likelihood analyses, and that the Bayesian posterior probabilities were much better indicators of phylogenetic accuracy than the corresponding bootstrap scores (Figure 5.2).

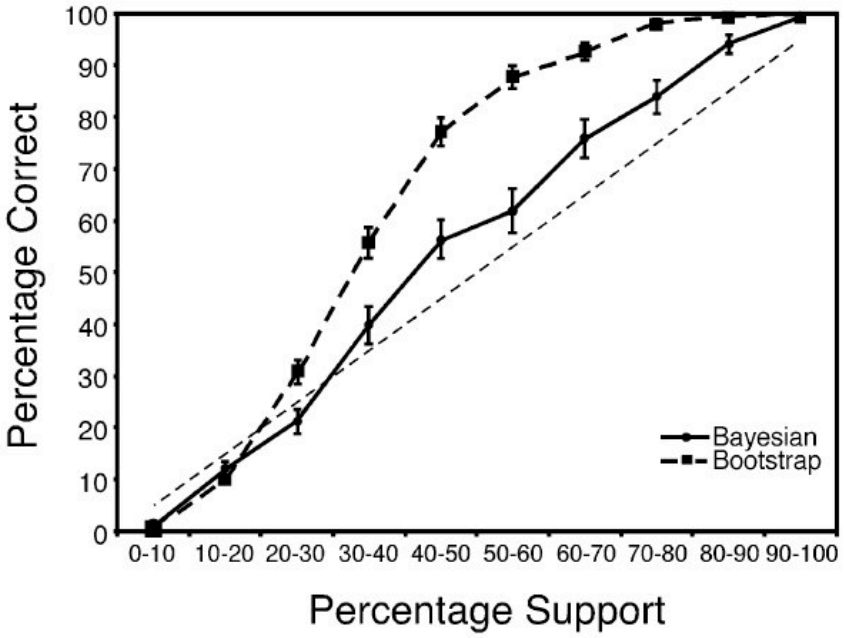

Figure 5.2: A comparison of phylogenetic estimation accuracy *sensu* Hillis and Bull (1995; figure from Wilcox *et al*., 2002). The y-axis shows the phylogenetic accuracy (percentage of correct partitions found) and the x-axis shows the percentage support values for the respective bipartitions for both Bayesian and likelihood analyses of 120 simulated data sets. The diagonal represents a perfect correspondence between phylogenetic accuracy and support. Both the posterior probabilities and the bootstrap scores underestimate accuracy at greater than 30% support, but the Bayesian values are much better estimates of phylogenetic accuracy (closer to the diagonal).

Based on these findings Wilcox *et al*. (2002) recommended that Bayesian posterior probabilities be used in preference to nonparametric bootstrap proportions to assess support for estimated clades in phylogenetic trees as both are indices are conservative, but posterior probabilities are closer to the truth.

Suzuki *et al*. (2002) object to the findings of Wilcox *et al*. (2002) above for two reasons. First of all, the model used in the phylogenetic reconstructions was the same model use to generate the data, which they see as ineffectual as the substitution model used for reconstruction will realistically never be the same as the true substitution pattern. Secondly, they object with the findings because no clearly defined null hypothesis is given. They therefore performed a simulation study of their own. In this case, sequences of 5000 bp were generated according to the three different topologies possible for the four taxon case (Figure 5.3). These sequences were then concatenated to form one 15000 bp sequence that should (barring stochastic error) generate the three topologies with equal probability.

This process was repeated 50 times each for 6 combinations of transition/transversion ratios, internal and external branch lengths. The resulting 600 data sets were each analyzed using Bayesian, neighbour-joining, and maximum likelihood approaches. An examination was then performed on the relative false-positive rate of the three methods in order to discern whether the bootstrap is too conservative or the posterior probability too liberal. Because each topology should be returned with equal probability, significant support (bootstrap ≥ 95% or posterior

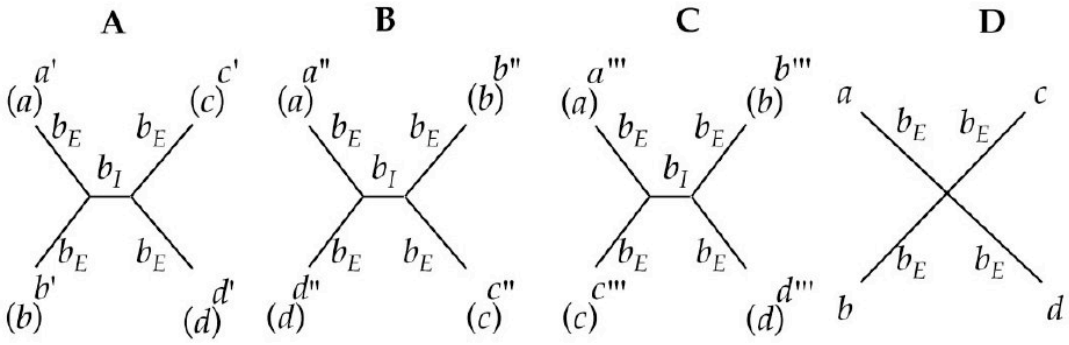

Figure 5.3: Figure from Suzuki *et al*. (2002). Sequences of length 5000 bp were generated according to the three possible topologies for four taxa (A, B, and C). b_E and b_I represent external and branch lengths, respectively. The sequences were then concatenated to form a single 15000 bp sequence that would equally support topologies A-C, and is represented above as a star tree (D).

probability ≥ 0.95) for a particular topology represents a false positive. Table 5.1 summarizes these results.

Table 5.1: Results of the simulation study of Suzuki *et al*. (2002). The sequences analyzed were generated in such a way that all three possible topologies (A-C) should be equally likely (i.e. no one topology should receive significant support). Bayesian posterior probabilities show both a higher frequency of false-positives (expected 5%) and higher averages across replications.

			Bayesian posterior probability					NJ bootstrap probability					ML bootstrap probability				
b_E	b_I	R	A	B	C	All	$\bar{P}$	A	B	C	All	$\bar{P}$	A	B	C	All	$\bar{P}$
0.05	0.005	0.5 (0.5)	8 / 16	5 / 14	8 / 20	21 / 50	85	1 / 17	0 / 13	1 / 20	2 / 50	63	0 / 16	0 / 14	0 / 20	0 / 50	64
0.1	0.01	0.5 (0.5)	6 / 18	7 / 20	7 / 12	20 / 50	85	1 / 16	0 / 19	0 / 15	1 / 50	64	0 / 18	0 / 20	0 / 12	0 / 50	65
0.05	0.005	5 (0.5)	14 / 19	13 / 19	9 / 12	36 / 50	91	0 / 17	0 / 20	0 / 13	0 / 50	62	0 / 19	0 / 19	0 / 12	0 / 50	63
0.1	0.01	5 (0.5)	12 / 18	14 / 16	11 / 16	37 / 50	95	0 / 19	1 / 18	1 / 13	2 / 50	68	0 / 18	0 / 16	1 / 16	1 / 50	69
0.05	0	5 (0.5)	11 / 18	14 / 21	6 / 11	31 / 50	89	1 / 18	0 / 22	1 / 10	2 / 50	68	1 / 18	0 / 21	1 / 11	2 / 50	66
0.1	0	5 (0.5)	11 / 18	13 / 16	15 / 16	39 / 50	95	0 / 16	0 / 18	1 / 16	1 / 50	69	0 / 18	0 / 16	1 / 16	1 / 50	68

The number of false-positive cases (replications) is given above the line, and the number of replications that supported tree *A*, *B*, or *C* (see Fig. 1) is given below the line. All, results for all replications; $\bar{P}$, average probability for all replications; *R*, transition/transversion ratio used for generating sequence data (*R* value used for phylogenetic inference is given in parentheses).

As can be seen from Table 5.1, Bayesian posterior probabilities exhibited an immensely higher false-positive rate when compared to both neighbour-joining and maximum likelihood analyses. Suzuki *et al*. (2002) speculate that the maximum likelihood tree (or set of trees) is visited again and again in the MCMC search, the result being that a stochastic result is sampled many times and so receives a high posterior probability. Neighbour-joining and maximum likelihood analyses, conversely, actually had a false-positive rate lower than the expected 5% (confidence level was 95%), corroborating previous studies that have found the nonparametric bootstrap to be a conservative measure of support. Additionally, the average posterior probability score across

replicates was much higher than corresponding bootstrap scores. Figure 5.4 illustrates the direct pairwise relationships between the three nodal scores for several combinations of transition/transversion ratios, internal and external branch lengths.

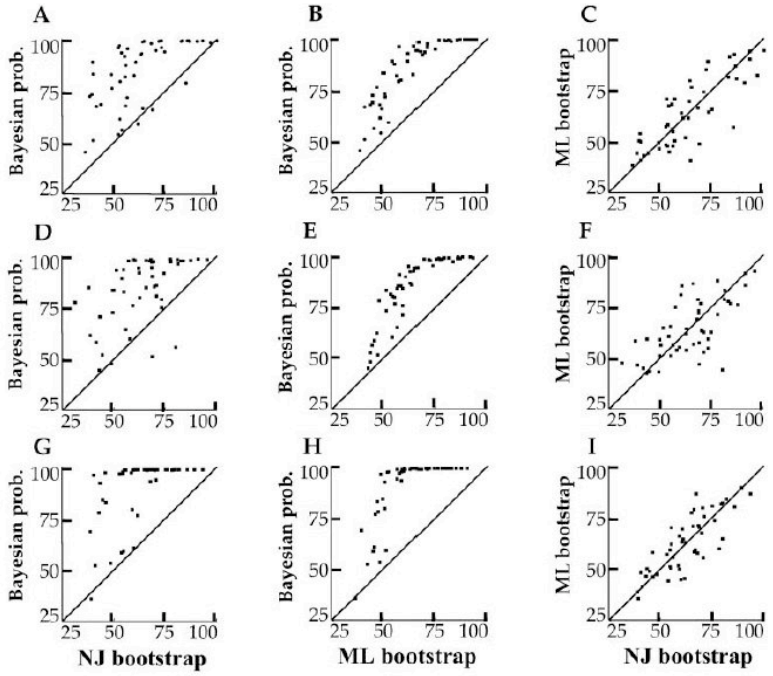

Figure 5.4: Corresponding pairwise relationships of the three nodal scores (partial figure from Suzuki *et al*., 2002). The groups (ABC), (DEF), and (GHI) represent different combinations of transition/transversion ratios, internal and external branch lengths. Bayesian posterior probabilities are significantly higher than both maximum likelihood and neighbour-joining bootstrap proportions (additional scenarios, not shown, show even stronger trends), but the two bootstrap scores show a high level of correlation.

Based on their findings Suzuki *et al*. (2002) came to the conclusion that posterior probabilities are "excessively liberal" and that conservative methods (i.e. bootstrap proportions) should be preferable to an overly lax method in phylogenetic analysis because conclusions are drawn exclusively from statistical analyses without experimentation. They go on to say that empirical phylogenetic studies using Bayesian inference may suffer from overcredibility and should thus be viewed with caution. Cummings *et al*. object to these findings because the use of the Jukes-Cantor model for analyzing data generated using a Kimura model confounded the effects attributable to the general properties of the underlying analytical methods.

Douady *et al*. (2003) investigated the relationship between nodal scores obtained from bootstrapped data matrices analyzed using maximum likelihood and Bayesian approaches for both empirical and simulated datasets. One hundred bootstrap pseudoreplicates were generated from each of eight original empirical data sets spanning different kinds of characters, types of sequences, genomic compartments, and taxonomic groups. This study differs from the previous two in that (in addition to comparing posterior clade probabilities and bootstrap proportions) here a "Bayesian bootstrap proportion" score is obtained and compared directly to the corresponding bootstrap score from maximum likelihood for the same data matrix. They found three results of note. First, posterior probabilities were found to be consistently higher than corresponding

bootstrap proportions for maximum likelihood for *both* true and false nodes (Figure 5.5). Reiterating previous studies they concluded that bootstrap proportions might be less prone to strongly supporting a false phylogenetic hypothesis.

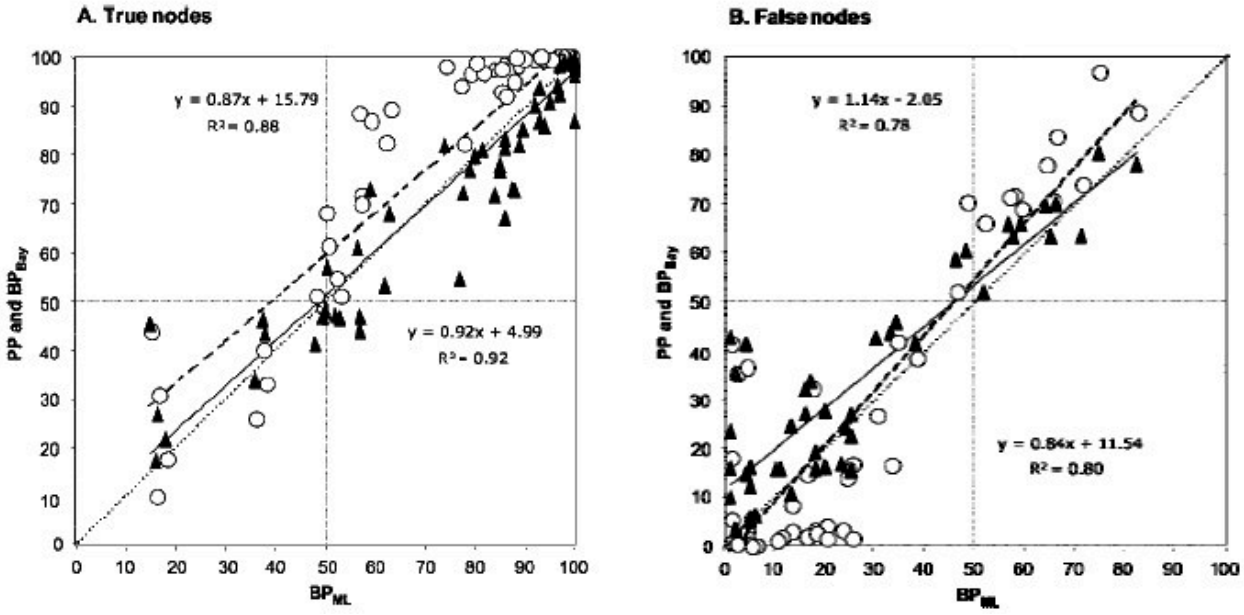

Figure 5.5: A linear correlation of nodal scores obtained from Bayesian and maximum likelihood analyses on simulated data for both true and false nodes (figure taken from Douady *et al*., 2003). PP represents the Bayesian posterior clade probability, BP_{Bay} represents bootstrapped Bayesian posterior clade probability, and BP_{ML} represents the bootstrap clade proportion using maximum likelihood inference of phylogeny. Circles illustrate the relationship between PP and BP_{ML}, while triangles illustrate the relationship between BP_{Bay} and BP_{ML}. Bayesian indices of support are consistently higher than their maximum likelihood counterparts, for both true and false nodes.

Secondly, they found that, while posterior clade probabilities and maximum likelihood scores can show significant correlation, the strength of this correlation is highly variable and sometimes very low (Figure 5.6). Bayesian inference, they conclude, may be sensitive to small model misspecifications, a sentiment shared by Huelsenbeck *et al*. (2002) and Erixon *et al*. (2003). Lastly, Douady *et al*. (2003) found that the bootstrap scores from respective Bayesian and maximum likelihood analyses were very highly correlated ($0.95 < r^2 < 0.99$). Bayesian analysis of bootstrapped data was much faster than corresponding maximum likelihood analyses, and gives comparable results. This technique, they propose, may be used to explore the range of node support estimates. My criticism of this paper is that they did not explicitly define what a Bayesian bootstrap proportion actually measures and consequently I regard these as nebulous results.

Alfaro *et al*. (2003) simulated 17-taxa topologies under 18 evolutionary scenarios to compare bootstrap proportions from maximum parsimony and maximum likelihood to posterior clade probabilities (using vague priors) for both correct and incorrect nodes. The 18 scenarios differed in tree-shape, and were meant to represent a spectrum of possible tree topologies (Figure 5.7). The results were largely congruent to those above: posterior clade probabilities were consistently higher for both correct and incorrect nodes, though overall assignment of incorrect nodes was

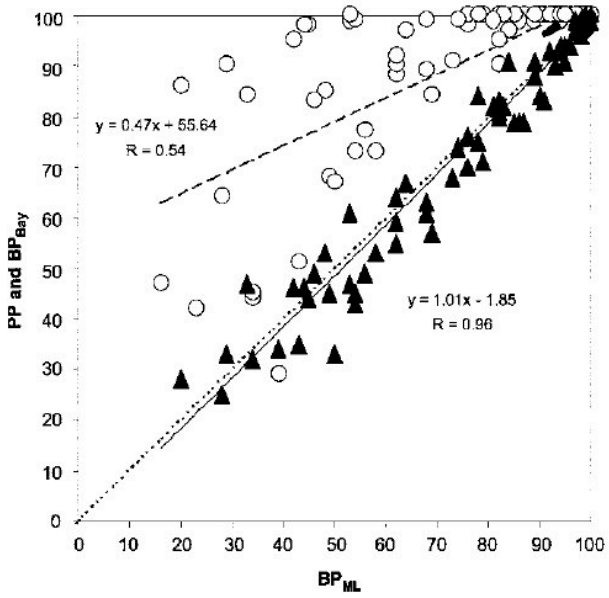

Figure 5.6: A linear correlation of nodal scores obtained from Bayesian and maximum likelihood analyses for both true and false nodes (figure taken from Douady *et al*., 2003). Symbols are explained in Figure 5.5. Posterior probabilities are significantly higher than maximum likelihood bootstrap proportions, though bootstrapped posterior probabilities show a near 1:1 relationship.

low. Posterior probabilities were found to have lower type 1 error rates (the frequency of rejecting true monophyletic groups) than bootstrapping, but when jointly considering support values for wrong monophyletic groups both methods performed similarly in recovering correct internodes. Like Wilcox *et al*. (2002) they also found that posterior probabilities are better estimators of phylogenetic accuracy than bootstrap scores and reiterate that phylogenetic accuracy is not a quantity that bootstrapping tests. They also found that posterior probabilities and bootstrap proportions, when they diverged, tended to differ most on short internodes. They attribute this finding to the putatively greater sensitivity of Bayesian methods to the signal in a data set, echoing Douady *et al*. (2003).

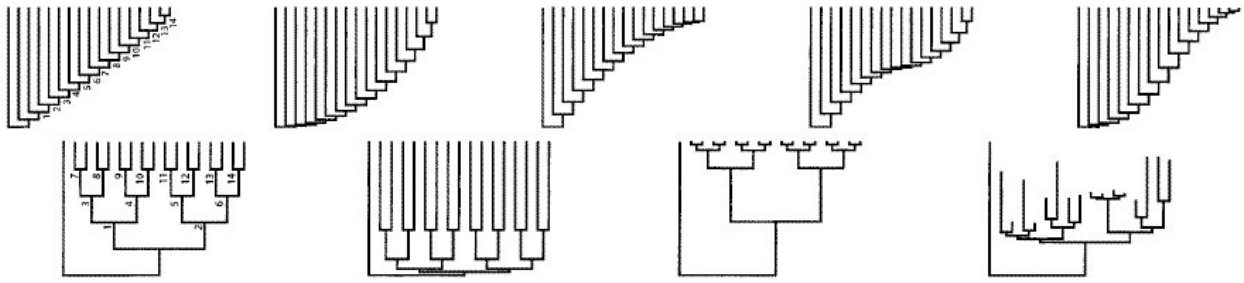

Figure 5.7: Some of the tree topologies (pectinate, above; symmetrical, below) used by Alfaro *et al*. (2003) to simulate data sets for the comparison of bootstrap proportions and posterior clade probabilities.

What really sets apart the study by Alfaro *et al*. (2003) from those above is the direct examination of the effects of increasing character number on nodal support value for both clocklike and nonclocklike symmetric topologies. They found that Bayesian posterior probabilities assigned 95% support to all internodes with a smaller number of characters than either bootstrapping method. In some cases posterior probabilities reached support values of

95% or higher with fewer characters than maximum parsimony required to reach support values of 70%. This is an appealing attribute of Bayesian inference of phylogeny because of the ubiquity of short internodes in empirical data sets that receive low bootstrap scores. However, because of the increased sensitivity of Bayesian methods mentioned above this may lead to high confidence in incorrect nodes when character sampling has not been sufficient to recover the correct topology. Bootstrap proportions, which have an inherent lower sensitivity when few characters contribute to a particular node, may then be preferable.

Alfaro *et al*. (2003) conclude by reaffirming that bootstrap proportions and posterior clade probabilities measure different, informative features of the data. They give an example of an internode with high posterior probability and moderate bootstrap support. Such a node should be interpreted as having a high probability of being correct (conditional on the data that have been collected and the model of evolution), but also being highly dependent on the particular data matrix and thus may not be observed when further data are gathered. To decide between Bayesian posterior probabilities and classical bootstrap proportions, then, the investigator should have in mind what they would like their confidence method to measure.

Finally, Cummings *et al*. (2003) have performed the most rigorous comparison of bootstrap proportions and posterior clade probabilities yet attempted and, like Suzuki *et al*. (2002), they dealt with the four taxon case. They simulated data sets of 1000 bp each using the GTR + Γ model of sequence evolution to test the null hypothesis of:

$$E(\text{maximum likelihood bootstrap proportion}) = E(\text{Bayesian posterior probability})$$

for a myriad of tree shapes. Sequences were generated according to various relationships of the 5 requisite branches which can be displayed on a two dimensional graph where the axes represent branch length combinations (Figure 5.8). The experimental model space of Cummings

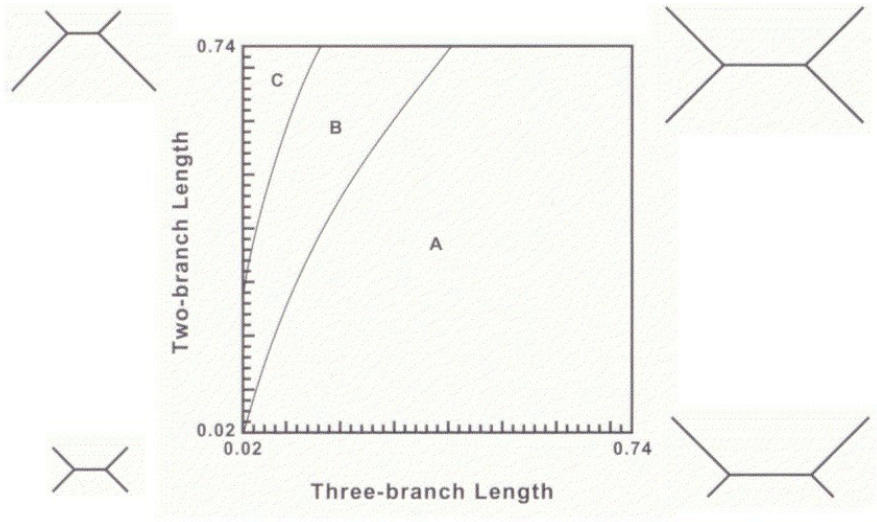

Figure 5.8: Model space of Cummings *et al*. (2003). The y-axis represents the lengths of three branches, the internal and the upper two, while the x-axis represents the other two external branches. Region A is referred to as the "neutral zone", B the "near two-branch corner", and C the "two-branch corner". The model space contains 1369 elements for each of which 1000 data sets of 1000 bp were generated.

et al. (2003) contained 1369 elements, or 1369 unique combinations of the 5 branches. For each element 1000 replicate data sets were generated. The breadth of tree shape is thought to encompass the range of realistic empirical topologies. Maximum likelihood analyses were performed with the branch-and-bound methodology to ensure that the maximally likely tree was guaranteed to be found for each replicate. Bayesian and maximum likelihood analyses were paired to eliminate a source of variance due to analysis of different sequences.

Three topologies were recognized: the model topology (τ_1), the attractive topology (a result of long branch attraction; τ_2), and the remaining third possible topology (τ_3; Figure 5.9).

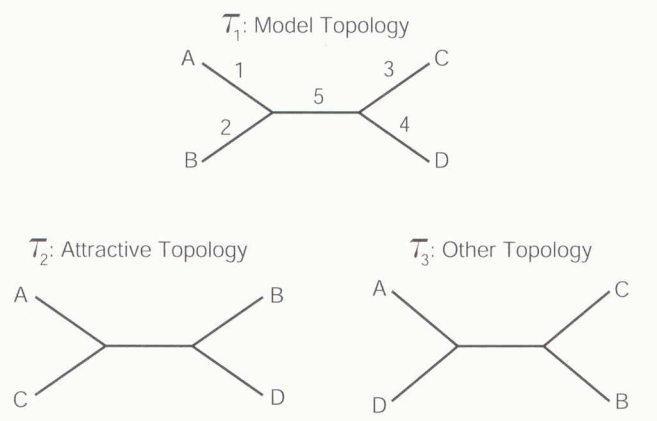

Figure 5.9: The three possible topologies for 4 taxa (figure from Cummings *et al*., 2003).

Cummings *et al*. (2003) plotted separately for each of the three topologies the difference, *d*, between the mean proportion of bootstrap replicates and the mean posterior probability values for each element in the model space (Figure 5.10). Plotting the results in this manner is important as the three topologies are interrelated; high bootstrap support for one topology necessarily means low support for the two alternative topologies. Permutation tests were performed to

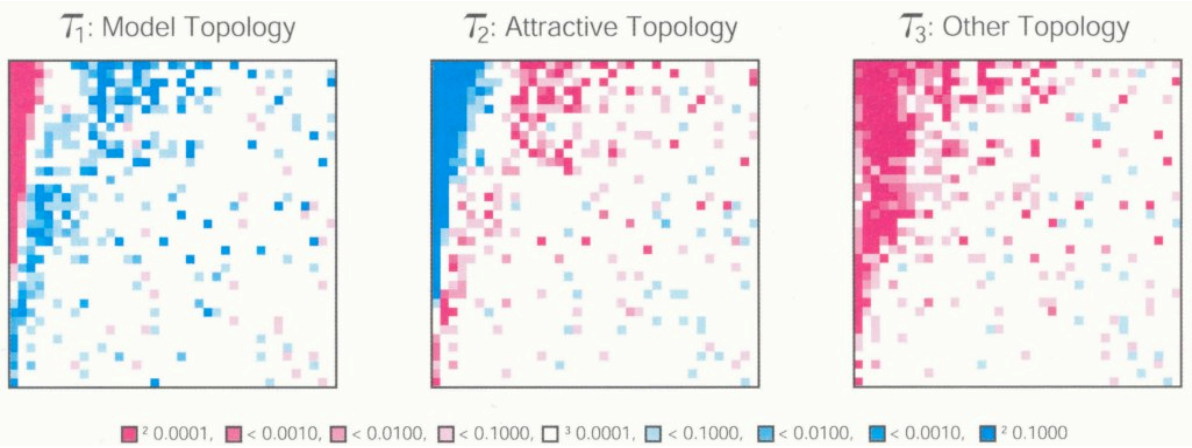

Figure 5.10: Permutation test results to determine the significance for values of *d*, the difference between the mean proportion of bootstrap replicates and the mean posterior probability values for each element in the model space (figure from Cummings *et al*., 2003). Magenta denotes elements where bootstrap values are significantly greater, cyan denotes elements where posterior probabilities are greater, and white denotes no significant difference in the two support indices.

determine the significance of *d* for each element in the model space. From these plots we can see that three apparent regions in the model space. In the "neutral zone" (from the lower left corner to the upper right), results are mixed in each plot as to which support index is higher, and results from the permutation tests show that the differences that arise are not significant.

The "two-branch corner" (upper left corner) is where the effects of long branch attraction can be seen. The third region is the "near two-branch corner" and is adjacent to both the neutral zone and the two-branch corner. It can be seen from Figure 5.10 above that mean bootstrap scores are significantly higher than pairwise mean posterior probabilities in the two-branch corner for the model topology, but that the opposite result is seen for the attractive topology. Cummings *et al*. (2003) describe these findings by the fact that the branch-and-bound search allowed the maximally likely tree to be found for every maximum likelihood replicate while the Bayesian search visited both the attractive topology and the model topology. Bayesian inference of phylogeny, it follows, appears to be susceptible to long branch attraction.

Cummings *et al*. (2003) also plotted quantile-quantile plots to compare maximum support scores with theoretical maximal values for the null model of a star topology (Figure 5.11). Both indices differ significantly from expectation over much of the range. Bootstrap scores tend to be markedly higher than expectation from $0.6 < \max(P_{boot}) < 0.7$, but are conservative for $\max(P_{boot}) > 0.95$. Bayesian posterior probabilities, on the other hand, differ from expectation from about a score of 0.6 upwards, and especially from 0.85 – 1.0. The relationship between bootstrap proportions and posterior probabilities, they conclude, is a complex one (depending on the underlying tree space, fit of the likelihood model to the data, tree search specifics, etc.), but in general posterior probabilities are excessively high, corroborating many of the results above.

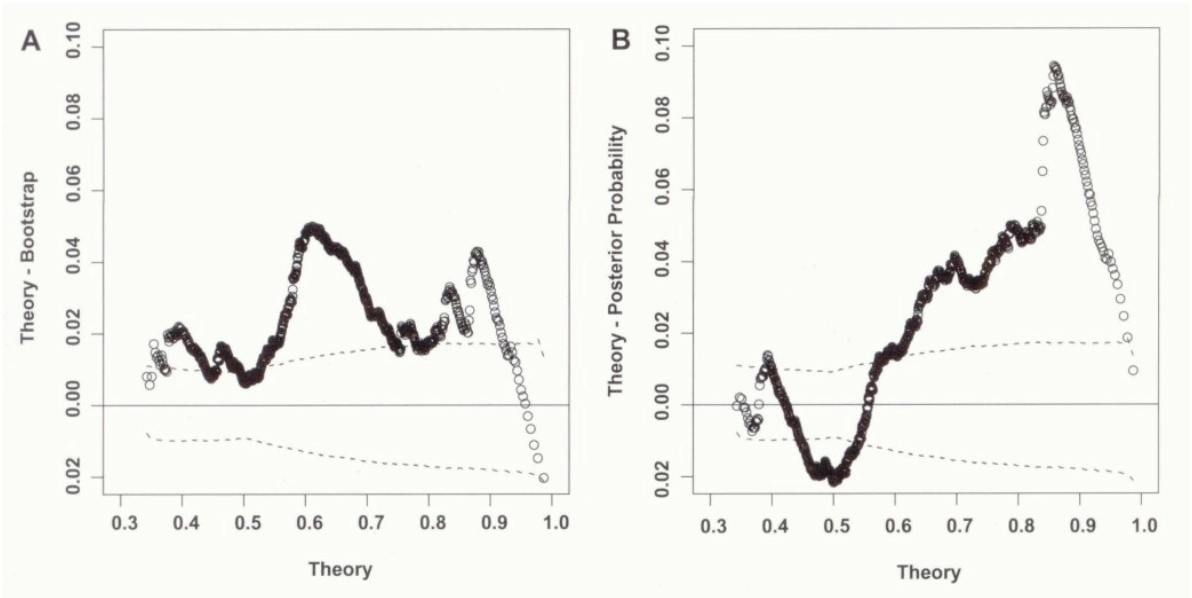

Figure 5.11: Permutation test results to determine the significance for values of *d*, the difference between the mean proportion of bootstrap replicates and the mean posterior probability values for each element in the model space

The problems of excessively high support outlined above constitute the major force opposing mainstream Bayesian phylogenetic inference. With all of the promising advantages of a Bayesian approach to the phylogenetic problem, posterior probability calculations are, I believe, the last stumbling block to this emerging revolution in phylogenetics; people can get used to the idea and design of prior probabilities, but they will never tolerate excessively high posterior probabilities. It remains to be seen, then, whether appropriate renovations can be made to the currently implemented Bayesian phylogenetic inference methodologies, or if the inflated support indices will mean the young demise of this promising field.

6. POTENTIAL DRAWBACKS

We have touched on most of the potential drawbacks to the Bayesian inference of phylogeny throughout this paper, but it is useful to reiterate them here.

MCMC ISSUES

The issues surrounding the use of Markov Chain Monte Carlo techniques for phylogenetic inference are no different than in any other field. Recall that purpose of using MCMC algorithms is to approximate the posterior probability distribution when numerical integration is impossible. Clearly, the more samples that are taken, the better approximation. But how many is enough? The truth is that we can never be absolutely sure that we have collected an appropriate number of samples (Lewis, 2002). In other words, it is not possible to determine suitable run-lengths theoretically, and so this requires some experimentation on the part of the user. The concerns here are: 1) is the MCMC sample representative of distribution that it was sampled from, and 2) are enough samples collected to estimate a particular parameter with reasonable precision (i.e. low variance; Jones and Browning, 2003). Roughly speaking, Monte Carlo error decreases as the square root of the number of iterations (e.g. to reduce error by a factor of 10, increase the number of iterations by a factor of 100). As a rule of thumb I recommend running sequential analyses of different chain lengths (i.e. different number of samples from the posterior distribution); if runs of various sample sizes give you roughly the same results, it would seem logical that you have approximated the distribution appropriately.

A strict *number* of samples is not the only concern in an MCMC search, however. As we saw in chapter 4, samples from an Markov chain are autocorrelated (Jones and Browning, 2003). In other words, the *absolute* number of samples taken is far greater than the *effective* number of samples. There are two strategies to get around this: 1) take a far greater number of samples, or 2) thin your Markov chain. Option 2 appears to be the preferred method, in part because autocorrelation can be directly measured and thus controlled for. Using an autocorrelation plot (Figure 4.6) from a pilot run, the lag time (n, the number of iterations) required for effective independence of samples can be determined. The user can then take samples every n iterations in the actual analysis and be confident that samples are independent. Regardless of the strategy taken, dealing with autocorrelation of MCMC samples requires far greater run times.

The two issues above assume that the stationary distribution has been found and is being sampled. How can be sure of this? Again, we can never be completely sure that chains are run long enough (Lewis, 2002). Two related considerations must be made when determining the adequacy of a Markov chain: burnin and convergence. Burnin refers to the samples taken by the Markov chain en route to the stationary distribution. As these samples have essentially zero probability we do not want them in our MCMC sample or they will skew our results; therefore they are discarded prior to sample summary (Holder, 2003). The problem is determining the burnin cutoff. This is accomplished through use of a history plot (Figure 4.5). Once values begin to plateau it can be inferred that the stationary distribution has been reached. However, it must be noted that getting stuck in a local optima will resemble *exactly* the stationary distribution.

This last point deals directly with the idea of convergence. Though we can never *prove* that we are sampling from the bone fide stationary distribution, there are steps we can take to increase

our confidence (Tierney, 1994). First, we can run the analysis multiple times starting from random points in parameter space; if the same results are obtained across runs we can be fairly confident that the chain has converged on the desired distribution. Secondly, we can run multiple chains simultaneously with communication between the chains (Altekar *et al*., in press). This latter method not only increases the likelihood of convergence, but also increases the mixing ability of the chain (i.e. explores isolated peaks in parameter space).

MCMC methods are undoubtedly powerful tools that allow us to evaluate problems in a Bayesian context that would otherwise be intractable. However, to reiterate Geyer (1999; quoted in Huelsenbeck *et al*., 2002), "MCMC is a complex mixture of computer programming, statistical theory, and practical experience. When it works, it does things that cannot be done any other way, but it is good to remember that it is not foolproof." Care must be taken, then, to scrutinize our results as best we can. Unfortunately, the diagnostic tools currently in use are all qualitative, but they are all we have.

BAYESIAN ISSUES

Some of the contentious issues deal not with MCMC sampling, but instead with the underlying Bayesian methodology. As with any Bayesian analysis, the priors specified are the most controversial. Presently only vague priors are used in phylogenetic inference (Lewis, 2002), partly because of the lack of empirical data with which to construct priors and partly to placate skeptics. Prior probabilities have been a litigious issue since Bayes' theorem was first published 150 years ago, and, in the words of Joseph Felsenstein (2003), "Nothing that biologists say is going to settle the matter." The issue here is whether usable prior probabilities exist. Personally I am optimistic that, with the explosion in the amount of genetic data currently being generated, within the coming decade informative priors will be available to apply to the problem of phylogenetic inference.

Of far more practical concern is the discrepancy between posterior clade probabilities and maximum likelihood nonparametric bootstrap proportions (Huelsenbeck *et al*., 2001, 2002). Even the most adamant Bayesian devotee will admit that this is cause of great concern. Reasons for the marked incongruence are somewhat unclear, possibly dealing with prior or model misspecifications, or perhaps to inappropriate MCMC elements. Regardless of the cause, the effect is intolerable. Until this issue is resolved it appears that phylogeneticists may need to resort to generating bootstrapped posterior clade probabilities (Douady *et al*., 2003), which have a much better correlation with maximum likelihood nonparametric bootstrap proportions. With this, however, comes two additional concerns: 1) bootstrapping Bayesian analyses will increase the computational effort required immensely (speed is currently a major advantage of Bayesian inference over likelihood), and 2) no explicit interpretation of a bootstrapped posterior clade probability has been given.

Lastly, model misspecification has been indicated by several studies to lead to erroneous results. Buckley (2002) notes that because current Bayesian implementations use flat priors, posterior probability distributions are largely dependent on the structure of the likelihood model. It follows that strong and unreliable posterior inferences can be made if the model is misspecified. It appears that this effect occurs whether models are either grossly underparameterized (Erixon *et al*., 2003) or overparameterized (Rannala, 2002). Clearly, then, an investigator wishing to use Bayesian inference in phylogeny must be more concerned with model specification than when

working in a likelihood framework, which appears to be fairly robust the model of evolution assumed (e.g. Kuhner and Felsenstein, 1994). The program Modeltest (Posada and Crandall, 1998, 2001) was designed explicitly for the purpose of finding the stochastic model of evolution that best fits the data, and Bollback (2002) designed a Bayesian phylogenetic method that evaluates the adequacy of different models using posterior predictive distributions.

All in all, Bayesian inference of phylogeny does require much more thoughtfulness on the part of the investigator than conventional maximum likelihood inference where, minimally, a model is selected and the data are subjected to nonparametric bootstrap analysis. I believe, however, that this is not cause for concern. MCMC diagnostics, though qualitative, do exist, and model choice should be of utmost importance to any practicing phylogeneticist. The next logical step, I think, is to combine model selection and phylogenetic inference into the same analysis. This, however, has not even been hinted to in the literature, and because of the enormous computational burden that this would require (i.e. moving between parameter spaces of various dimensions), it is unlikely that anything like this would be feasible for at least a decade.

7. CONCLUSIONS AND THE FUTURE OF BAYESIAN PHYLOGENETICS

Bayesian phylogenetics using Markov chain Monte Carlo technology has unquestionably revolutionized the field of phylogenetic systematics, and represents the first treatment of topology as a random variable (Huelsenbeck *et al*., 2001; Holder and Lewis, 2003; Lewis and Swofford, 2001). Bayesian inference has allowed successful analysis of some of the largest and most complex empirical data sets ever compiled, including examining the ordinal relationships of placental using complete mitochondrial genomes for 44 taxa (Murphy *et al*., 2001), and inferring the closest living relatives of land plants using more than 5000 bp for 40 taxa (Karol *et al*., 2001). These data sets were could not have been analyzed a scant five years ago, and, though computer processor speeds have advanced considerably since then, it is due chiefly to Bayesian methodologies that they can be analyzed today. True, such data sets *could* be analyzed in a likelihood framework, but it would require enormous computer facilities that the average phylogeneticist does not have access to. Part of the appeal of Bayesian phylogenetics, then, is that reconstructions can be performed on conventional desktop computers in a reasonable amount of time.

Bayesian inference of phylogeny has a seemingly endless number of advantages over conventional likelihood methods. Speed, mentioned above, seems to top the list, as it allows the investigator to use far more complex (realistic) models of evolution, and more thorough exploration of parameter space, than ever before possible. Not only can these complex models be applied to standard genetic data sets, but also to novel heterogeneous data sets that contain information from morphological, paleontological, genetic, and behavioural sources. If improved Bayesian implementations show to be concordant with analogous likelihood methods, it is clear that an investigator will turn to the former as it ensures that he/she will get their damn tree before their NSERC runs out.

Related to speed is the way that Bayesian phylogenetic inference is carried out. The simultaneous estimation of parameter values and support gives not only a speed benefit, but also provides the investigator with more information than in a likelihood search (i.e. the shape of the posterior probability distribution, rather than a point estimate). I agree with Buckley (2002) that "systematists should be more concerned with identifying the total set of trees that can be reasonably supported by the data, rather than focusing on point estimates of topology." Bayesian inference is the first methodology in phylogenetics that generates probabilistic calculations of such sets, and the probabilistic properties involved allow for direct interpretation that has until now been sorely missing from phylogenetic systematics.

But Bayesian inference has been utilized for purposes other than strictly inferring phylogenetic relationships. Many comparative studies in evolutionary biology require control for phylogeny. A major benefit of using Bayesian inference is that it explicitly accounts for uncertainty in phylogeny when estimating parameters of interest, rather than assuming phylogeny is known without error. This technique has been successfully applied to estimating divergence times, identifying recombination points, testing molecular clocks, and detecting selection, to name only a few examples (see Huelsenbeck *et al*., 2001; Holder and Lewis, 2003).

If you are anything like me then you probably experienced a range of emotions during your trek through this paper. First, you may have been curious about the Bayesian methodology and interested in the inner workings. Next you may have been impressed by the seemingly vast

advantages of a Bayesian approach as compared to conventional maximum likelihood methods. You may have then progressed from awe of the complexity of MCMC implementations to uneasiness with the realization that posterior clade probabilities, the values that most phylogeneticists are interested in, can be "excessively liberal" (Suzuki *et al*., 2002). I apologize for the emotional roller coaster but, as I stated in the introduction, I will not pay your shrink bills.

I am, admittedly, very optimistic about Bayesian inference, and see it as the future of phylogenetic systematics. Like with all new techniques, the initial mania has subsided as investigators have become aware of the limitations inherent in this analytical method. Regardless of whether I have convinced you, it is clear that the present Bayesian revolution has changed forever the way we think about phylogeny, and that it will have "a lasting and profound impact on the future of evolutionary biology" (Lewis and Swofford, 2001).

8. GLOSSARY

Bayes' theorem: Published posthumously in Bayes' 1763 paper *Essay Towards Solving a Problem in the Doctrine of Chances*. Bayesian statistics is a formal method for incorporating prior evidence into the inference of the probability that an event occurs together with a consideration of the current data. Let D represent the data and H represent a hypothesis. Bayes' theorem can thus be expressed as:

$$\Pr(H \mid D) = \frac{\Pr(H) \cdot \Pr(D \mid H)}{\Pr(D)}$$

In the equation above, $\Pr(H \mid D)$ is the posterior probability of the hypothesis given the data, is the $\Pr(D \mid H)$ is the likelihood of the data given the hypothesis, and $\Pr(H)$ is the prior probability, or simply "prior", of the hypothesis and represents our state of knowledge (or ignorance) about the truth of a hypothesis *before* we have observed the data (Sivia, 2002). The final quantity, $\Pr(D)$, is the marginal probability of the data given the model; this is simply the sum of the numerators over all possible hypotheses. The posterior probability is sometimes thought of as an updated version of the prior probability in the light of the data. Bayes' theorem thus fundamentally encapsulates the process of learning (Sivia, 2002)

Burnin: A Markov chain Monte Carlo (MCMC) term. A Markov chain typically starts from a random position in parameter space. This random position is likely to be some distance from areas of high posterior probability (peaks in parameter space) and consequently of lower posterior probability. Initial samples taken by the Markov chain en route to the peak have essentially zero probability since they are not from the distribution of interest. These samples are therefore discarded before summarizing the sample results. This discarded portion of the Markov chain makes up the "burnin" of the chain. The extent of burnin required is determined through use of a history plot.

Cold chain: A Metropolis coupled Markov chain Monte Carlo [$(MC)^3$] term. In an $(MC)^3$ analysis several Markov chains are run simultaneously. Each chain computes the posterior probability for the currently sampled value for the parameter of interest and then raises the posterior probability to a power, β ($0 < \beta < 1$), which is the heat of the chain. The "cold" chain is raised to the power 0 (and hence is unaffected) while the other "heated" chains have posterior probability distributions of the form $\Pr(\theta \mid X)^{\beta}$. The cold chain is the only chain that records samples from the posterior probability distribution, though it communicates with the heated chains for tips on areas of high posterior probability to sample in parameter space.

Convergence: A Markov chain Monte Carlo (MCMC) term. "Convergence" of a Markov chain is its ability to converge upon the posterior probability peak of interest when starting from a random position in parameter space. As the chain approaches the stationary distribution the likelihood climbs rapidly until a plateau is reached. At this point the high posterior probability peak of interest is being sampled. Monitoring of convergence (via a history plot) can be used as a diagnostic tool. Convergence of independent runs (starting from random positions in parameter space) on the same results can be used to verify results.

Flat (uniform) prior: All possible values of the parameter of interest are given a uniform prior probability before the data are observed. Also called an uninformative prior, though this is inaccurate because priors that are much more vague can be designed. Though at first glance this may appear to represent the more "objective" choice of prior (i.e. removing the subjectivity in choice of prior by the investigator), upon closer inspection it is clear that in many cases a flat prior is a poor choice as it gives excessive weight to extremely unlikely possibilities. When a flat prior is used the posterior probability is directly proportional to the likelihood.

Heated chains: A Metropolis coupled Markov chain Monte Carlo [$(MC)^3$] term. In an $(MC)^3$ analysis, n Markov chains are run concurrently, $n - 1$ of which are heated. Each chain computes the posterior probability for the currently sampled value for the parameter of interest and then raises the posterior probability to a power, λ ($0 < \lambda < 1$), which is the heat value of the chain. Heated chains thus have a posterior probability distribution of the form $\Pr(\theta|X)^{\lambda}$. Heating a chain effectively "melts down" the posterior probability landscape, making valleys shallower and peaks lesser in height. Chains of higher "temperature" thus explore a more flattened landscape and so are more able to cross particularly deep valleys. Despite the increased mobility, the sole function of heated chains is to provide the cold chain with intelligent proposals of new states. Heated chains do not record samples themselves, and therefore act merely as scouts, searching the surface of the posterior probability distribution for isolated areas (peaks) of high probability.

History plot: A plot of the number of steps (iterations; x-axis) by likelihood (y-axis), used in MCMC analyses. History plots allow for monitoring of convergence and mixing of the Markov chain, as well as determining the appropriate amount of burnin.

Likelihood: The conditional probability of the data given a particular model. In the classical approach all parameters are jointly estimated to maximize the likelihood function. In the Bayesian paradigm the likelihood for a parameter of interest (e.g. topology) is a marginal likelihood over all other parameters of the statistical problem.

Marginalization: Also referred to as "integrating out". Given a nuisance parameter, Z, marginalization effectively means to take account of all possible values of Z when evaluating the parameter(s) of interest. Marginalization is of utmost importance for all Bayesian probability inference: the information about a subset of the system's variables is derived by integrating out all nuisance parameters. More generally, given parameters X, Y, and Z, marginalization is the process to derive information about X and Y, given all possible values of Z, as in the following equation:

$$\Pr(X,Y) = \int_{z} \Pr(X,Y,Z)dz$$

In a phylogenetic context an investigator could focus on one parameter (e.g. topology) while marginalizing over all other parameters (transition rate parameters, gamma shape parameter, ancestral states, etc.).

MCMC (Markov chain Monte Carlo): A stochastic simulation sampling scheme used in most Bayesian and some maximum likelihood analyses; in a Bayesian framework it allows integration over high-dimensional parameter spaces. No objective function is maximized; instead, the shape of the posterior probability distribution is of interest rather than locating the highest point on the

likelihood surface. A Markov chain samples parameter value combinations and moves randomly through parameter space with no memory of where it has been; future samples are dependent only on the immediately prior sample. Put another way, "given the present, the past and future are independent" (Lee, 1997). In the context of Bayesian phylogenetic inference, MCMC has the desirable property that, once having reached the stationary distribution, the values of the parameter of interest (e.g. tree topology) are visited in proportion to their posterior probabilities. MCMC thus acts to approximate the posterior probability distribution of the parameter of interest.

MCMCMC ($[MC]^3$; Metropolis coupled Markov chain Monte Carlo): An MCMC analysis where *n* Markov chains are run concurrently, $n - 1$ of which are heated. The addition of heated chains enhances the mixing ability of the cold chain, thereby more efficiently exploring parameter space. The advent of $(MC)^3$ has drastically extended the limits of model-based phylogenetic inference (Huelsenbeck *et al*., 2001).

Mixing: A Markov chain Monte Carlo (MCMC) term. "Mixing" describes the ability of a Markov chain to explore isolated, high posterior probability peaks in parameter space. A poorly mixing Markov chain is one that gets stuck on a particular peak, and consequently is not sampling all highly probable states. Poor mixing is dangerous in that it can lead to skewed results, the reason being alternate states (potentially of equal or higher probability) are not sampled. While steps requiring large drops in posterior probability (for example, while traversing a valley between two peaks) *will* occur eventually (given that the Markov chain is constructed correctly and is run for sufficiently long enough), the time required may be impractical. Poor mixing is best combated through running multiple Markov chains concurrently, some of which are heated.

Nuisance parameter: A parameter that is required to evaluate a problem (e.g. a likelihood equation) but is not itself of direct interest. In a phylogenetic problem where topology may be the ultimate goal, parameters of the substitution model (transition rate parameters, gamma shape parameter), ancestral states, etc. would be considered nuisance parameters.

Parallelization: The running of an analysis simultaneously across several computer processors rather than serially on one processor. Parallelization of $(MC)^3$ analyses have shown near linear scaling with the number of processors available (Altekar *et al*., in press).

Posterior probability: A conditional probability, the posterior probability is the probability of the hypothesis given the data, and is proportional to the product of the likelihood and the prior probability of the hypothesis. The posterior probability thus represents our state of knowledge about the truth of a hypothesis in the light of the data. Sometimes described as an updated version of the prior probability after having observed the data. The posterior probability is a direct measure of uncertainty (unlike in the classical framework) and may or may not represent a long-term frequency.

Prior probability: The prior probability, or simply "prior", represents our state of knowledge (or ignorance) about the truth of a hypothesis before we have analyzed the current data (Sivia, 2002). The prior is modified by data through the likelihood function to yield the posterior probability. Viewed by classical statisticians as a weakness of Bayesian inference because of its inherent subjectivity. Viewed by Bayesian statisticians as a strength as prior information can

enter the analysis. Priors may be proper or improper (distribution fails to integrate to 1), informative or uninformative. Prior probabilities can be based on previous experiments or theoretical expectations and must be defensible.

Probability: "Probability" has different meanings in the different schools of statistics. In the Frequentist school (championed by Sir Roland Fisher) probability is interpreted as the long-run relative frequency with which an event occurs in many repeated similar trials. To a Frequentist, probability lies objectively in the world, not in the observer. In the Bayesian school of statistics (founded on work by Reverend Thomas Bayes) probability is interpreted as a measure of one's degree of uncertainty about an event. This may or may not represent a long term frequency. To a Bayesian, probability lies in the mind of the observer and may be different for people having different information or different past experiences.

Random variable: A variable whose values are random but whose statistical distribution is known.

Reverend Thomas Bayes (1702-1761): An English Presbyterian minister and mathematician. His posthumously published paper *An Essay Towards Solving a Problem in the Doctrine of Chances* (1763) is the basis of modern Bayesian inference.

Swapping: A Metropolis coupled Markov chain Monte Carlo [$(MC)^3$] term. Periodically in an $(MC)^3$ analysis two Markov chains are given the opportunity to trade or "swap" state information. Swaps can involve heated chains or a heated chain and the cold chain. Swapping is the beauty of Metropolis coupling in an MCMC analysis as it allows the cold chain to effectively traverse a deep valley in parameter space in one step rather than the many (unlikely) steps required otherwise. Swapping thus acts to increase the mixing ability of the cold chain.

Thinning: Due to the sampling nature in the MCMC approximation of the posterior probability distribution, samples from the MCMC chain are somewhat autocorrelated (the degree of autocorrelation depends on the construction of the Markov chain, in particular the choice of proposals). Thinning an MCMC chain means that not all samples are recorded; rather, samples are recorded periodically at a rate that can be specified by the investigator (say, every 100^{th} sample). Recording non-successive samples decreases autocorrelation (increases independence).

9. LITERATURE CITED

Alfaro, Michael E., Stefan Zoller, and François Lutzoni. 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(2): 255-266.

Altekar, Gautam, Sandhya Dwarkadas, John P. Huelsenbeck, and Fredrik Ronquist. In press. Parallel Metropolis-coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics.

Bayes, T. 1763. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society of London 53: 370-418.

Bollback, Jonathan P. 2002. Bayesian model adequacy and choice in phylogenetics. Molecular Biology and Evolution 19(7): 1171-1180.

Buckley, Thomas R. 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. Systematic Biology 51(3): 509-523.

Buckley, Thomas R., Peter Arensburger, Chris Simon, and Geoffrey K. Chambers. 2002. Combined data, Bayesian phylogenetics, and the origin of the New Zealand cicada genera. Systematic Biology 51(1): 4-18.

Bullard, Floyd. 2001. A brief introduction to Bayesian statistics. NCTM 2001 lecture notes, The North Carolina School of Science and Mathematics. Available online at: http://courses.ncssm.edu/math/TALKS/PDFS/BullardNCTM2001.pdf

Cummings, Michael P., Scott A. Handley, Daniel S. Myers, David L. Reed, Antonis Rokas, and Katarina Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Systematic Biology 52(4): 477-487.

Douady, Christophe J., Frédéric Delsuc, Yan Boucher, W. Ford Doolittle, and Emmanuel J. P. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Molecular Biology and Evolution 20(2): 248-254.

Efron, Bradley, Elizabeth Halloran, and Susan Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proceeding of the National Academy of Science 93: 7085-7090.

Erixon, Per, Bodil Svennblad, Tom Britton, and Bengt Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Systematic Biology 52(5): 655-673.

Felsenstein, Joseph. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39: 783-791.

Gaut, Brandon S., and Paul O. Lewis. 1995. Success of maximum likelihood phylogenetic inference in the four taxon case. Molecular Biology and Evolution 12(1): 152-162.

Holder, Mark. 2003. Bayesian phylogenetics. Lecture notes from the 2003 Workshop on Molecular Evolution, Woods Hole, Massachusetts.

Holder, Mark, and Paul O. Lewis. 2003. Phylogeny estimation: traditional and Bayesian approaches. Nature Reviews Genetics 4: 275-284.

Huelsenbeck, John P. 1995. Performance of phylogenetic methods in simulation. Systematic Biology 44(2): 17-48

Huelsenbeck, John P. 2000. Reverend Bayes and phylogenetic trees. Sample chapter. Available online at: http://www.cse.sc.edu/~fengx/phylogeny/Huelsenbeck-2000.pdf

Huelsenbeck, John P., Bret Larget, Richard E. Miller, and Fredrick Ronquist. 2002. Potential applications and pitfalls of Bayesian inference in phylogeny. Systematic Biology 51(5): 673-688.

Huelsenbeck, John P., and Bruce Rannala. 1997. Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science 276: 227-232.

Huelsenbeck, John P., Bruce Rannala, and John P. Masly. 2000. Accommodating phylogenetic uncertainty in evolutionary studies. Science 288: 2349-2350.

Huelsenbeck, John P., and David M. Hillis. 1993. Success of phylogenetic methods in the four taxon case. Systematic Biology 42(3): 247-264.

Huelsenbeck, John P., and Fredrick Ronquist. 2001a. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8) 754-755.

Huelsenbeck, John P., and Fredrick Ronquist. 2001b. MRBAYES: A program for the Bayesian inference of phylogeny. MrBayes version 2.01 manual.

Huelsenbeck, John P., Fredrick Ronquist, Rasmus Nielsen, and Jonathan P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294: 2310-2314.

Huelsenbeck, John P., and Keith A. Crandall. 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annual Review of Ecology and Systematics 28: 437-466.

Jones, Beatrix, and Sharon Browning. 2003. Markov chain Monte Carlo for geneticists. Lecture notes from the 2003 Summer Institute in Statistical Genetics, North Carolina State University, Raleigh, North Carolina.

Karol, Kenneth G., Richard M. McCourt, Matthew T. Cimino, and Charles Delwiche. 2001. The closest living relatives of land plants. Science 294: 2351-2353.

Kuhner, Mary, and Joseph Felsenstein. 1994. A simulation comparison of phylogenetic algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11(3): 459-468.

Larget, Bret, and Donald L. Simon. 1999. Markov Chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Molecular Biology and Evolution 16(6): 750-759.

Leaché, Adam D., and Tod W. Reeder. 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): a comparison of parsimony, likelihood, and Bayesian approaches. Systematic Biology 51(1): 44-68.

Lee, Peter M. 1997. Bayesian Statistics: An Introduction (Second Edition). Arnold Publishers, London.

Lewis, Paul O. 2001. Phylogenetic systematics turns over a new leaf. Trends in Ecology and Evolution 16(1): 30-37.

Lewis, Paul O. 2002. Class 22: Bayesian approach to phylogenetics. EEB 372 Lectures. Available online at: http://www.eeb.uconn.edu/Courses/EEB372/class22.pdf

Lewis, Paul O., and David L. Swofford. 2001. Back to the future: Bayesian inference arrives in phylogenetics. Trends in Ecology and Evolution 16(11): 600-601.

Li, Shuying. 1996. Phylogenetic tree construction using Markov chain Monte Carlo. Ph. D. Dissertation, Ohio State University, Columbus.

Li, Shuying, Dennis K. Pearl, and Hani Doss. 2000. Phylogenetic tree construction using Markov chain Monte Carlo. Journal of the American Statistical Association 95: 493-508.

Mau, Bob, Michael A. Newton. 1997. Phylogenetic inference on dendrograms using Markov chain Monte Carlo methods. Journal of Computational Graphical Statistics 6: 122-131.

Mau, Bob, Michael A. Newton, and Bret Larget. 1999. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Biometrics 55: 1-12.

Mau, Robert. 1996. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. Ph. D. Dissertation, University of Wisconsin-Madison.

McGuire, Gráinne, Michael C. Denham, and David J. Balding. 2001. MAC5: Bayesian inference of phylogenetic trees from DNA sequences incorporating gaps. Bioinformatics 17(5): 479-480.

Murphy, William J., Eduardo Eizirik, Stephen J. O'Brien, Ole Madsen, Mark Scally, Christophe J. Douady, Emma Teeling, Oliver A. Ryder, Michael J. Stanhope, Wilfried W. de Jong, and Mark S. Springer. 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294: 2348-2351.

Newton, M. A., B. Mau, and B. Larget. 1999. Markov chain Monte Carlo for the Bayesian analysis of evolutionary trees from aligned molecular sequences. Pages 143-162 *in* Statistics in Molecular Biology, Volume 33 (F. Seillier-Moseiwitch, T. P. Speed, and M. Waterman, editors). Institute of Mathematical Statistics.

Posada, David, and Keith A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14(9): 817-818.

Posada, David, and Keith A. Crandall. 2001. Selecting the best-fit model of nucleotide substitution. Systematic Biology 50(4): 580-601.

Rannala, Bruce. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. Systematic Biology 51(5): 754-760.

Rannala, Bruce, and Ziheng Yang. 1996. Probability distribution of molecular evolutionary trees : a new method of phylogenetic inference. Journal of Molecular Evolution 43: 304-311.

Ronquist, Fredrik and John P. Huelsenbeck. In Press. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. Available online at: http://morphbank.ebc.uu.se/mrbayes/manual.php

Shoemaker, Jennifer S, Ian S. Painter, and Bruce S. Weir. 1999. Bayesian statistics in genetics: a guide for the uninitiated. Trends in Genetics 15(9): 354-358.

Sivia, D. S. 2002. Data analysis: a Bayesian tutorial. Oxford University Press Inc., New York.

Suzuki, Yoshiyuki, Galina V. Glazko, and Masatoshi Nei. 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. Proceeding of the National Academy of Science 99(25): 16138-16143.

Tierney, Luke. 1994. Markov chains for exploring posterior distributions. The Annals of Statistics 22(4): 1701-1762.

Whittingham, Linda A., Beth Slikas, David W. Winkler, and Frederick H. Sheldon. 2002. Phylogeny of the tree swallow genus, *Tachycineta* (Aves: Hirundinidae), by Bayesian analysis of mitochondrial DNA sequences. Molecular Phylogenetics and Evolution 22(3): 430-441.

Wilcox, Thomas P., Derrick J. Zwickl, Tracy A. Heath, and David M. Hillis. 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. Molecular Phylogenetics and Evolution 25: 361-371.

Yang, Ziheng, and Bruce Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov Chain Monte Carlo method. Molecular Biology and Evolution 14(7): 717-724.