16

Posterior Mapping and Posterior Predictive Distributions

Jonathan P. Bollback

Section of Ecology, Behavior, and Evolution, Division of Biological Sciences,
University of California, San Diego, La Jolla, CA 92093-0116, USA,
bollback@biomail.ucsd.edu

> If we view statistics as a discipline in the service of science, and science as being an attempt to understand (i.e., model) the world around us, then the ability to reveal sensitivity of conclusions from fixed data to various model specifications, all of which are scientifically acceptable, is equivalent to the ability to reveal boundaries of scientific uncertainty. When sharp conclusions are not possible without obtaining more information, whether it be more data, new theory, or deeper understanding of existing data and theory, then it must be scientifically valuable and appropriate to expose this sensitivity and thereby direct efforts to seek the particular information needed to sharpen conclusions. (Rubin [38])

16.1 Introduction

Bayesian statistical approaches are becoming increasingly common in the field of molecular evolution and phylogenetics. Rubin [38] makes an eloquent argument for the value of Bayesian approaches through the identification of sensitivity to our assumptions and the potential uncertainty in our conclusions given our data at hand. While many may see Bayesian approaches as flawed by their dependence on prior distributions and sensitivity to model specifications, others, as with Rubin, will view this as a beneficial property of the method—not accounting for uncertainty can lead to overconfidence in the conclusions. This chapter will review two Bayesian approaches that in the last few years have seen important developments: posterior mapping of characters and posterior predictive distributions. These methods clearly identify and accommodate uncertainty while providing valuable solutions to our questions. It is this author's opinion that these methods will provide invaluable contributions to our understanding of molecular evolution and phylogenetics in the future.

16.1.1 Character Mapping

The mapping of characters on genealogies has been invaluable in answering questions in evolutionary biology since the 1970s; studies such as testing for a molecular clock [21], detecting the signature of positive selection [28], and looking for associations between characters (see [10] for a review) have all employed character mapping. Traditionally, parsimony has been the mainstay—although approaches that combine the methods of maximum likelihood and parsimony and Bayesian inference and parsimony have been developed [21, 16]. Parsimony as a method for mapping characters, while straightforward in its application, has a number of serious drawbacks. First, it underestimates the number of character transformations, often severely. This underestimation arises because parsimony does not account for evolutionary time along branches of a phylogeny: as evolutionary time increases, the number of inferred changes at a site is either zero or one. Second, parsimony underestimates the variance in ancestral states, placing all of the support on one reconstruction when they are not known with certainty. Lastly, parsimony provides no framework for accommodating uncertainty in genealogical relationships.

The drawbacks inherent in parsimony have long been recognized both by molecular evolutionists [8] and phylogeneticists [10]. For example, Langley and Fitch [21], in a study testing the molecular clock hypothesis, employed a mixed method of parsimony to assign ancestral states and maximum likelihood to estimate the rates along the branches. While this early approach acknowledged the underestimation of character changes by parsimony and accommodated it using maximum likelihood, it still left the problem of uncertainty in the phylogeny and ancestral states unresolved.

Recently, methods for accommodating uncertainty in the ancestral states and topology have been devised. For example, one approach to accommodating uncertainty in ancestral states is to use maximum likelihood to estimate the probabilities of each possible state and parsimony to reconstruct the character changes weighted by their probabilities [39, 40, 29, 34]. Uncertainty in topology has also been addressed in a number of ways. Some authors have used a set of reasonable trees and evaluated mappings on each of them (e.g., [42]). Others have evaluated mappings on trees generated under a stochastic process, such as birth-death [24, 26], or evaluated mappings on trees weighted by the probability of the tree being true [25, 33, 16].

While these approaches have made significant advances in accommodating different sources of uncertainty none of them accommodate all sources of uncertainty. In addition, due to their reliance on parsimony, none of these approaches is able to provide detailed information on the timing, order, and types of multiple changes—if any—occurring along a branch. Nielsen [30, 31] has developed a stochastic method for mapping characters using a Bayesian statistical framework. This approach of sampling from the posterior distribution of character histories (also referred to in this chapter as mappings or maps) successfully addresses the drawbacks inherent in parsimony and provides a statistically valid framework for accommodating uncertainty in the phylogeny and model parameters. This approach is the topic of the next section and will be discussed in detail.

16.1.2 Posterior Predictive Distributions

Posterior predictive distributions evolved from concerns regarding the dependence on the prior distribution in prior predictive distributions. Instead of integrating out nuisance parameters using the specified prior distribution of the parameters, the posterior approach integrates with respect to the posterior distribution of the parameters. The justification for, the particular implementation of, and other issues surrounding the use of posterior predictive distributions are rather contentious among statisticians, resulting in an active and healthy research program. Because of this there exists a diversity of different approaches—prior, posterior, and their use in approximating Bayes' factors, to name a few—and opinions regarding these predictive distributions. Much of the discussion revolves around the appropriate formulation of a p-value. The discussion here will deal mostly with posterior predictive distributions and their related p-values. Differences between the approaches and the shortcomings of the posterior method will be highlighted in the relevant places, and a brief account of the controversy will be discussed at the end of the chapter.

Within evolutionary biology, posterior predictive distributions appeared simultaneously with those of posterior mapping. While they can be used to test a variety of hypotheses, their first application was to character histories [32]. A similarity between posterior mapping and posterior predictive distributions is their ability to naturally accommodate uncertainty in the phylogeny and model parameters by treating them as nuisance parameters. (This aspect of both methods is not unique to them and has been a motivating factor behind many of the Bayesian developments in biology; see [18] for a review.) Posterior predictive distributions have in the last few years seen application to hypotheses such as detecting positive selection [32], evaluating substitution model adequacy [3], testing for nucleotide frequency heterogeneity [18], correlated character change [14], concordance between genes [46], and patterns in protein evolution.

The use of predictive distributions in Bayesian hypothesis testing in general and evolutionary biology in particular is appealing for a number of reasons. First, the generality of the approach makes it applicable to a wide variety of questions in molecular evolution and phylogenetics. Second, the method provides a rigorous statistical framework for accommodating uncertainty in model parameters and genealogical (or phylogenetic) relationships. This alone may be the strongest argument for the use of predictive distributions over methods such as the parametric bootstrap. Third, predictive probabilities (called posterior predictive p-values) are constructed using tail areas of the predictive distribution and are straightforward in their implementation and interpretation. Unlike classical frequency probabilities, posterior predictive probabilities do not evaluate observed values relative to a fixed set of values under the null model but averages over probable sets. Lastly, predictive p-values produce a Type I frequentist error at a given α similar to the expected α (often lower but never greater than 2α) [27]. While these reasons make predictive distributions appealing, a number of concerns and potential drawbacks exist and will be discussed at the end of Section 16.3. Briefly, this approach requires the description of a probabilistic model (null hypothesis), specification of a prior distribution for the model, an estimation of the model's posterior distribution, and a little ingenuity on the part of the researcher in determining appropriate test statistics (see [38] for a general review). Each of these will be dealt with in detail in Section 16.3. Of these requirements, the last is clearly the most difficult to accomplish: a good test statistic needs to be a relevant summary of the hypothesis being tested, and each question will require a different sort of test statistic. The logic behind the posterior predictive approach is similar to that underlying the parametric bootstrap. In fact, the parametric bootstrap sampling distribution may be indistinguishable from the posterior predictive sampling distribution when maximum likelihood estimates are used and the posterior is concentrated. The fit of a hypothesis is tested by comparison of the observed test statistic—often referred to as the realized value—with the distribution of that statistic under the null model. If our realized value falls within the 95% confidence region of the null distribution, we are unable to reject the null hypothesis—otherwise, we reject it.

The remainder of this chapter will explore the underlying methodology of these two approaches, review a number of their recent applications, demonstrate how posterior predictive distributions can be used to test hypotheses about character histories, and discuss how predictive distributions can be used to address a wealth of different questions in molecular evolution and phylogenetics.

16.2 Posterior Mapping

In this section, I will try to answer four questions: (1) What are character histories?; (2) How do we go about sampling character histories?; (3) How do we accommodate uncertainty in model parameters and topologies?; and (4) What types of questions can we address with posterior mapping? The second and third questions will be answered by introducing the method of posterior mapping first proposed by Nielsen [30, 31] and then later extended by Huelsenbeck et al. [14] to morphological characters. The last question will be answered by briefly reviewing examples from the literature.

First, let us tackle the question of what a character history is by providing a definition. A character history is a description of the historical pattern of state occurrences and transformations along a phylogeny. The history is more than just a simple description of the ancestral reconstructions at the internal nodes of the tree. It includes information about the placement (timing), order of states, types of character state transformations (e.g., A ⇔ G), and direction (or bias; e.g., A → G versus G → A) of transformations when the root of the phylogeny is known (see Figure 16.1d for an example of a character history). What we would like is to sample possible character histories (individual character histories will also be referred to as a map) in which they are sampled in proportion to their posterior probabilities. More often we will be interested in a function of these sampled histories and not individual histories. For example, we may wish to determine the number of radical amino acid changes relative to conservative changes [32]. In addition, we may be interested not only in the relative types of changes but also the order and timing of changes. For example, contingency tests of neutrality rely on being able to determine types of changes (silent/replacement) and their placement on the tree [30].

But, before we get into the details of the method (questions 2 and 3), we might wonder why we should not rely on parsimony and what the differences are. To illustrate these differences, we will explore four different mappings of a single site for four species shown in Figure 16.1. We will ask: (1) How does the placement of character transformations along a branch differ?; (2) How does the number of character transformations along a branch differ?; and (3) How probable are nonparsimonious mappings? Two of the trees in Figure 16.1 are parsimony mappings (trees a and b) and two are posterior mappings, one of which is consistent with parsimony (trees c and d).

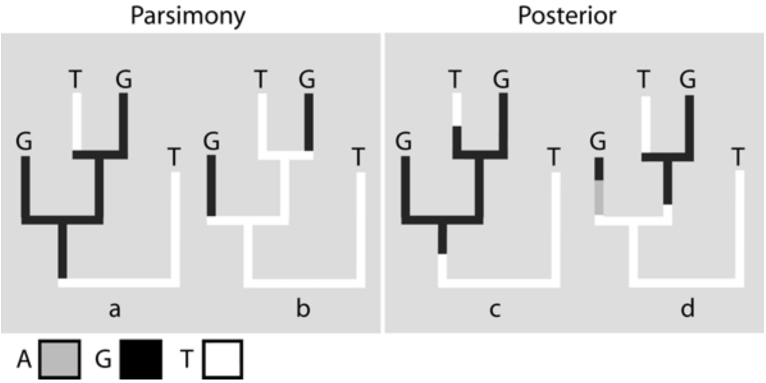

Fig. 16.1. A comparison between parsimony and two representative realizations from the posterior distribution of mappings. Trees a and b are parsimony reconstructions, while c and d are from the posterior distribution of mappings. The inferred number of changes in tree *c* is consistent with parsimony. The posterior mappings were generated with SIMMAP, a program that implements the posterior mapping method and can be downloaded at `http://www.simmap.com`.

The position along a branch at which an inferred change occurs under parsimony is shown directly following a bifurcation. This was done for convenience—we could have placed the changes equidistant along the branches. This illustrates the first difference between parsimony and posterior mappings—their placement of transformations along branches. Parsimony provides no information about the timing of changes along a branch; parsimony simply concludes that a single change has occurred. Posterior mapping, however, does provide information about placement and order of multiple changes along a branch. (In addition, the timing of changes between different sites can be compared. See the discussion on correlated character evolution, in Subsection 16.3.5 , for an example.) For example, in trees c and d, in Figure 16.1, we can clearly see when the events occurred and the order in the case of tree d. In many cases, the order of changes is of interest. For example, we might wish to know whether a burst of amino acid replacements immediately follows speciation or whether it is evenly distributed after the split.

To illustrate the difference in the number of transformations considered by each method, let us compare the posterior mapping on tree d in Figure 16.1 with the parsimony mappings (trees a and b). First, we should note that the map on tree d is not consistent with parsimony; four changes have been inferred, compared with two changes required by parsimony. Sampling from the posterior distribution of mappings has produced a map in which two additional changes have occurred. While, admittedly, I have not shown you that mappings with two additional changes have a large or small probability, it does have a probability greater than zero. Under parsimony this is not even considered plausible, let alone probable, while the posterior method is not constrained to minimizing the number of changes.

Let's consider the final difference between parsimony and posterior mappings—how probable are nonparsimonious mappings? In this example, we will compare the probability of parsimonious and nonparsimonious mappings. In effect, we will be evaluating two assumptions of parsimony: the minimization of changes and the reduction in variance associated with ancestral state reconstruction at the root. This example should also provide an introduction to the underlying logic of posterior mapping. To address this difference, we will first calculate the overall probability of the data and then conditional probabilities given the branch lengths and the number of character changes along the trees shown in Figure 16.2.

In this particular example of two species, there is only a single phylogeny relating the two sites. This is the equivalent of assuming that the tree is known in cases of more than three species. (Later, it will be shown that the method allows us to accommodate uncertainty in the phylogeny and model parameters.) To compare the mappings, we are interested in calculating

$$\Pr(M_i|D) = \frac{\Pr(M_i, D)}{\Pr(D)}, \tag{16.1}$$

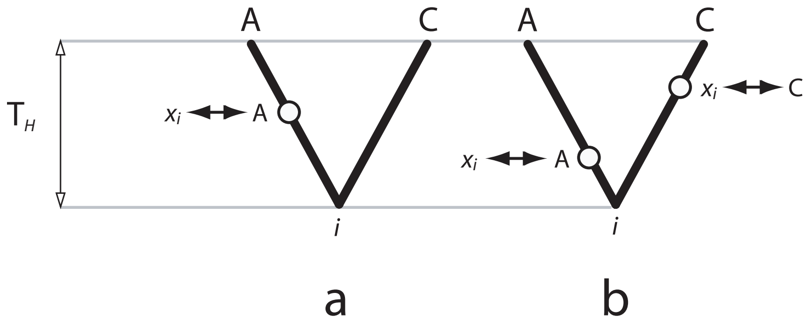

Fig. 16.2. Comparison of the probabilities associated with parsimonious and non-parsimonious character histories. T_H is the tree height, from the root, in the expected number of substitutions per site and will be used to evaluate an increase in branch lengths (0.5, 1.0, and 2.0, respectively; see the text). x_i is the state we are changing from at the root and is dependent on whether we are observing one or two changes; under one change $x_i \in \{C\}$, while under two changes $x_i \in \{G, T\}$.

where M_i is a character map and D is the observed data. This is the probability of the map given the data. Calculation of the probability of the data, $\Pr(D)$, requires a model that describes substitution probabilities from one state to the next. We will assume the Jukes and Cantor [19] model, which is a time-reversible Markov model. Under the JC69 model, the stationary nucleotide frequencies are $\pi_i = 1/4$ for all i, and the probability of a change from nucleotide i to j along a branch of length t is

$$P_{ij}(t) = \begin{cases} 1/4 + (3/4)e^{-(4/3)t} & \text{if } i = j, \\ 1/4 - (1/4)e^{-(4/3)t} & \text{if } i \neq j. \end{cases} \tag{16.2}$$

We can now calculate $\Pr(D)$ by considering all possible state assignments at the root i as

$$\Pr(D) = \sum_{i \in \{A,C,G,T\}} \pi_i P_{iA}(t) P_{iC}(t). \tag{16.3}$$

When $T_H = 0.5$, then $\Pr(D) = 0.04602$ for the data and phylogeny shown.

Next we want to calculate the probability of histories a and b conditional on the data at the tips of the trees. For the mapping shown on tree a (M_a), we want to calculate $\Pr(M_a, D)$. This can easily be done using the fact that for the JC69 model and other continuous-time Markov chain models, the number of changes along a branch is Poisson-distributed. For example, along the left lineage of tree a, the conditional probability of observing a single change is $0.5e^{-0.5} \times (1/3)$. The last term represents the probability of a change between nucleotides, which is 1/3 under the JC69 model. Therefore, we calculate $\Pr(M_a, D)$ as

$$\Pr(M_a, D) = \frac{e^{-0.5} \times (0.5e^{-0.5}/3)}{4} = 0.0153, \tag{16.4}$$

where the probability of not observing a change along a branch of length $t = 0.5$ is $e^{-0.5}$ and again the probability of observing a single change along a branch of this length is $0.5e^{-0.5} \times 1/3$ under the JC69 model.

The root state for tree a must be a C, given the states at the tips and a single change occurring along the branch leading to the state A. However, in tree b the state of the root is uncertain. An observation of a T or a G at the root of tree b would be consistent with the mapping shown and the states at the tips of the tree. Given these possible root states, we can calculate the probability as

$$\Pr(M_b, D) = \frac{(0.5e^{-0.5}/3)^2}{4} \times \frac{(0.5e^{-0.5}/3)^2}{4} = 0.0051. \tag{16.5}$$

Using these probabilities and $\Pr(D)$, we can calculate the conditional probabilities for the character histories on trees a and b as 0.333 and 0.111, respectively. The parsimony-consistent history is three times as probable. However, what happens as the time from the root to the tips increases? Table 16.1 shows the probabilities for the trees and mappings in Figure 16.2 given three different sets of branch lengths.

Table 16.1. A comparison of the probabilities associated with the parsimony consistent mapping in tree a with that of the nonparsimonious mapping of tree b (see Figure 16.2) and the cumulative probability of mappings greater than two substitutions ($\Pr(M_{i>b}|D)$).

			Tree a		Tree b	Changes > 2
T_H	$\Pr(D)$	$\Pr(M_a, D)$	$\Pr(M_a\|D)$	$\Pr(M_b, D)$	$\Pr(M_b\|D)$	$\Pr(M_{i>b}\|D)$
0.5	0.046025	0.015328	0.333	0.005109	0.111	0.556
1.0	0.058157	0.011277	0.194	0.007519	0.129	0.677
2.0	0.066239	0.003052	0.034	0.004070	0.045	0.921

A couple of things should be noticed in Table 16.1. First, as the branch lengths increase, the probability of the mapping consistent with parsimony (tree a) decreases. Second, the parsimony mapping decreases from a threefold higher probability to a probability lower than the mapping with two changes (tree b) as branch lengths increase. As expected, as time increases, the probability of multiple changes increases, making mappings with one, and even two, changes much less probable (although they probably have the largest probabilities). The cumulative probability of more than two changes increases with increasing time, reaching 0.921 at divergences of 2.0 expected substitutions per site. Hopefully, I have been able o show that even for the simplest phylogeny, nonparsimonious mappings should be considered.

16.2.1 Sampling Character Histories

How do we go about sampling character histories using the method of posterior mapping? The following is a description of simulating a map for a site. Complete gene sequences can be sampled by repeating this approach for each site. Four steps are involved in sampling a character map: (1) define a substitution model in which probabilities of state changes can be calculated; (2) calculate the conditional likelihood for each state at each node of the tree, including the root; (3) simulate ancestral states; and (4) simulate a substitution (mutational) history, conditional on the ancestral states and states at the tips of the tree. (Often the states at the tips of the tree are unknown or uncertain (e.g., N, R, etc.). This type of uncertainty can easily be accommodated by revisiting these nodes after simulating ancestral states for the internal nodes and repeating step 3 for the tips.)

First, we need to define a model of nucleotide (or morphological) change (step 1). Any number of continuous-time Markov models are available, that accommodate a variety of different plausible aspects of sequence evolution. Available models and their uses have been extensively described elsewhere, and a detailed treatment is beyond the scope of this chapter [48, 11]. Briefly, many commonly used models are special cases of the general time-reversible (GTR) model of sequence evolution [20, 37]. With this model, we can describe the instantaneous rates of changing from state i to state j using the rate matrix

$$\mathbf{Q} = \{q_{ij}\} = \begin{pmatrix} - & a\pi_C & b\pi_G & c\pi_T \\ a\pi_A & - & d\pi_G & e\pi_T \\ b\pi_A & d\pi_C & - & f\pi_T \\ c\pi_A & e\pi_C & f\pi_G & - \end{pmatrix}, \tag{16.6}$$

where a–f represent the rates of changing from one nucleotide to the next, and π_i represent the stationary nucleotide frequencies. Using this matrix, we can easily calculate substitution probabilities for a change from nucleotide i to j over a branch of length t as $\mathbf{P} = \{p_{ij}(t)\} = e^{\mathbf{Q}t}$. In many cases, such as the JC69 model described above, analytical solutions are available. In those cases in which solutions are not available, standard linear algebra approaches are available for exponentiating the matrix $\mathbf{Q}$.

With these probabilities, step 2 can be easily accomplished using the *pruning algorithm* of Felsenstein [4]. Given a tree with branch lengths τ, a set of observations D at the tips of the tree, and a vector θ containing a set of model parameters describing sequence evolution, we can calculate the conditional likelihood for each internal node and the root using a post-order traversal of the tree.

Next, we simulate a state at the root of the tree (step 3). Let us denote the root as σ and the simulated observation as d. The new state at the root will then be denoted d_σ. (All s descendant nodes and branches are indexed as $\sigma - 1, \ldots, \sigma - (2s - 3)$.) A site can be simulated by sampling from the posterior distribution

$$\Pr(d_\sigma = i|D, \tau, \theta) = \frac{l_{\sigma,i}\pi_i}{\sum_{j\in\{A,C,G,T\}} l_{\sigma,j}\pi_j}, \tag{16.7}$$

where $l_{\sigma,i}$ is the conditional likelihood of being in state i—we are conditioning on the observations at the tips of the tree, model parameters, and topology. Now, in a preorder traversal of the tree from the root, we visit a node directly above, $\sigma - 1$, and simulate an ancestral state by sampling from

$$\Pr(d_{\sigma-1} = j|d_\sigma = i, D, \tau, \theta) = \frac{l_{\sigma-1,i}P_{ij}(t_{\sigma-1})}{\sum_{k\in\{A,C,G,T\}} l_{\sigma-1,k}P_{ik}(t_{\sigma-1})}, \tag{16.8}$$

where j represents the recently simulated state at the ancestral node (in this case the root) and $P_{ij}(t_{\sigma-1})$ is the transition probability from state i to state j over a length of $t_{\sigma-1}$. We proceed with the traversal and simulate ancestral states for the remaining nodes. As noted above, often we find that a site may be unknown or uncertain for some sequences. Using this approach, we can also simulate a tip state. In this way, we treat the uncertainty at the tips in a fashion identical to that for internal nodes. Now we have sampled and assigned ancestral states from the posterior distribution for each internal node of the phylogeny.

The final step is to generate a character history for each branch of the tree given the previously simulated ancestral states and observed states at the tips of the tree (step 4). This, perhaps, is the most challenging step, and Nielsen [31] provides an elegant and computationally efficient solution. We simulate a realization of a continuous-time Markov chain conditional on the starting state and ending states along a branch. The waiting times between substitution events along a branch are drawn from an exponential distribution

$$\lambda e^{-\lambda t} \tag{16.9}$$

with the rate $\lambda = -q_{ii}$. This rate is taken from the diagonal elements of our $\mathbf{Q}$ matrix, which are interpreted as the rate of moving away from a state i. Waiting times can be obtained from this distribution using the inverse transformation method. If the exponential waiting time is longer than the branch length t and the states at each end of the branch are the same, then the process is terminated; no changes have occurred along this branch. If the waiting time is smaller than the branch length t, then a character transformation is determined by $\Pr_{ij} = \frac{q_{ij}}{-q_{ii}}$, and the process is continued with the new length, $t - t_1$, by drawing another exponential waiting time. If the next waiting time is longer than the remaining time along the branch and the states are the same, the process ends for that branch. On the other hand, if the states are different, the process is repeated from the ancestral node, not the previous simulated transformation. If we were to proceed from the previous transformation, the waiting times would no longer be exponentially distributed.

Nielsen [30] has pointed out that this approach is not computationally efficient when the reconstructed ancestral states are not the same and the length t is small. Nielsen [30] proposed conditioning the first waiting time on being less than the length of the branch as

$$f(t_1|t_1 < t) = \frac{\lambda e^{-\lambda t_1}}{1 - e^{-\lambda t}}, \quad 0 \leq t_1 < t, \tag{16.10}$$

where $\lambda = -q_{ii}$. Waiting times can also be drawn from this distribution using the inverse transformation method. This approach enhances the computational efficiency of the algorithm by reducing the number of realizations that are rejected. Using this approach, the first draw always produces a waiting time less than t and thus is consistent with at least one change occurring along the branch. The next draw uses the unconditional distribution as above. Once all internal nodes of the tree have been visited, we have successfully simulated a single realization of a map from $\Pr(M|D, \theta, \tau)$.

16.2.2 Integrating over Topologies and Model Parameters

In general, parameter values of the substitution model θ and the topology τ are not known with certainty. We would like to evaluate $\Pr(M|D)$ and not $\Pr(M|D, \theta, \tau)$. The Bayesian approach permits a natural way of accommodating uncertainty in these values. We wish to sample from

$$\Pr(M|D) = \sum_{k=1}^{\psi} \int_{v_k} \int_{\theta} \Pr(M|D, \theta, \tau) p(\tau_k, v_k, \theta|D) dv_k d\theta, \tag{16.11}$$

where ψ is the set of possible trees and v_k are the branch lengths associated with tree k. While this cannot be solved analytically due to its complexity, numerical approximations can be obtained using MCMC methods [35, 22, 17] (see Chapters 3 and 7).

In practice, how do we go about sampling character histories not dependent on fixed values for these parameters? The answer is quite simple. As described above, we have a method for sampling a map along a phylogeny. Using a program such as MrBayes or BAMBE, we can easily obtain an approximation of $p(\tau_k, v_k, \theta|D)$. With this distribution in hand, we can simulate a map for each posterior sample producing a valid approximation of $\Pr(M|D)$.

As mentioned previously, what we are most often interested in is some function of the histories, $h(M, D)$. These functions might evaluate the number of nonsynonymous substitutions, radical amino acid changes, relative timing of changes, correlation in the timing of transformations between two sites, or covariation of states between sites. We now have all the pieces necessary to evaluate any desired function and its expectation. For example, if we wish to evaluate the expected number of nonsynonymous changes, $n_{NSYN}(M, D)$, we can evaluate the expectation numerically from the distribution of character histories as

$$E[n_{NSYN}(M, D)|D] \approx \frac{1}{N} \sum_{i=1}^{N} n_{NSYN}(M_i, D), \qquad (16.12)$$

where N is the number of simulated character histories and $n_{NSYN}(M_i, D)$ is the observed number of nonsynonymous changes along map i.

16.2.3 Examples from the Literature

This section is intended to direct the reader to the most recent applications of posterior mapping in the literature. A brief overview of the specific questions addressed in the literature should provide a better understanding of the power of this approach.

The first application of this method in the literature [30] used it to address a number of questions pertinent to molecular evolution and population genetics. First, the author made inferences regarding the population parameter θ, which is the product of the population size and mutation rate, to a data set of 63 human mtDNA sequences from the Nuu-chah-Nulth tribe (see [50] in [30]) demonstrating the method's utility in population genetics. In addition, the method was applied to estimating the ages of mutations and then specifically the ages of synonymous and nonsynonymous mutations in a test of neutrality proposed by Templeton [49].

The method was further used to address how the parsimony method compared with the posterior method in estimating the number of mutations across two genes: β-globin and influenza hemagglutinin-A [31]. An analysis of the complete gene sequences found that the parsimony method greatly underestimated the total number of substitutions compared with the posterior method. Nielsen argued that the large discrepancy was likely due to differences in lineages; for example, rate heterogeneity among lineages, mutational biases among lineages, such as a transition/transversion bias, or biases among lineages in synonymous and nonsynonymous evolutionary rates. To address these questions, he tested for rate homogeneity among lineages, finding that there appeared to be considerable variance among lineages, particularly in the β-globin data set.

Finally, this method was extended to mapping morphological characters [14] using the Mk series of stochastic models [23]. While possibly of little interest to molecular evolutionists, this represents a major advancement in the phylogeneticist's ability to address questions about morphological character evolution using a statistical approach not relying on parsimony. Not only does this paper extend the method of stochastic mapping to morphological characters, using the Nielsen [31] method, but it provides a novel approach to looking for correlated character evolution using predictive distributions (see Section 16.3).

16.3 Predictive Distributions

Often we are confronted with situations in which the data, or some aspect of an analysis, do not meet the assumptions of a standard statistical test (e.g., the use of improper prior distributions in calculating Bayes factors). In cases like these in molecular evolution and phylogenetics, we rely on alternative methods, such as permutation tests (e.g., randomization tests), resampling approaches (e.g., the nonparametric bootstrap), the parametric bootstrap, and, in the Bayesian framework, predictive distributions. The latter approach is operationally analogous to the parametric bootstrap but has a number of differences and potential advantages over the traditional parametric bootstrap. This potential will hopefully become clear in the remainder of the chapter.

16.3.1 Posterior Predictive Simulations

Bayesian approaches to hypothesis testing come in two general forms: Bayes factors and predictive distributions. While hypothesis tests using Bayes factors have received a fair amount of attention in the phylogenetics literature [44, 13, 43, 45], the alternative, predictive distributions, only recently have been applied to methods in molecular evolution and phylogenetics [32, 31, 3, 46]. In this section, I will provide background on what predictive distributions are and how to use them, explore some recent applications from the literature, and discuss the pros and cons of their use. Predictive distributions provide a very general and flexible framework for Bayesian hypothesis testing, making them likely to be applied to a broad array of questions. In addition, they provide a natural way of accommodating uncertainty in the substitution model parameters and topology. This being said, the method isn't free of problems. The specifics of these issues will be reviewed at the end of this section. In evaluating a hypothesis, we would like to know how well it fits the underlying process that generated the data at hand. If a hypothesis is adequate, then it should perform well in predicting the distribution of data observations or some summary value relevant to the hypothesis being scrutinized. These distributions of future observations are called *predictive distributions* (also called reference distributions or densities). Most often we are not directly interested in the predictive distribution of the data but a summary statistic, referred to as a test statistic in this chapter, that captures relevant features of predictive data and our observed data given the hypothesis. Test statistics are dealt with in Section 16.3.2 but, for the moment let us assume we have some function, $T(\cdot)$, that summarizes an aspect of our data.

An analogy: parametric bootstrap

Before we get into the details of how to sample from posterior predictive distributions, I want to develop an operational analogy with the parametric bootstrap. Since many readers are already familiar with the use of the parametric bootstrap, it will hopefully serve as a useful heuristic to understanding predictive methods. The thought experiment will be a test of the molecular clock. While I don't advocate the test described below, as it is untested, it does provide a useful heuristic for understanding the differences between the two methods. (Note: There are numerous other well-established ways of testing the molecular clock.)

Let θ_c be a vector containing our model parameters (which include the substitution model parameters, topology, and associated node depths) under the clock hypothesis and θ_{nc} be the similar vector of parameters under the unconstrained hypothesis. Under the parametric bootstrap, these values are chosen to be the maximum likelihood estimates (MLE) for these quantities. Since we wish to test the molecular clock, we can generate our reference distribution using these $\hat{\theta}_c$ values and simulate n data sets (see Figure 16.3). These are the predictive outcomes we might expect to observe in future data collection expeditions, given that the values of $\hat{\theta}_c$ are true. Next, we need to summarize the data (observed and predictive) in some relevant way. We can use the difference in maximum likelihood estimates between the constrained (clock) and unconstrained branch length hypotheses [4], but for this example we will take an alternate approach. Let's assume that we have an outgroup that establishes the placement of the root and use the standard deviation of distance of the tips to the root under each hypothesis. The reference distribution, simulated under the clock, allows us to check the degree to which the clock would appear violated (magnitude of the standard deviation), given that the underlying process is truly clock-like. If the observed, or realized, value falls outside of this distribution, we might be inclined to reject the clock hypothesis or, more precisely, we reject that the observed deviation could have arisen under our null hypothesis—a molecular clock and the particulars of the substitution model.

In comparison, how might this be accomplished using posterior predictive simulations, and what are the possible differences in outcome with the parametric bootstrap? The first difference is immediately apparent: values of θ are not point estimates but averaged over samples from the posterior distribution of θ (see Figure 16.3) under the clock and unconstrained hypotheses. Samples from the posterior distribution under the clock model (θ_c) and unconstrained model (θ_{nc}) can be obtained using a program such as MrBayes [17]. Using these models, we can evaluate the expectation of our standard deviation test statistic, under the unconstrained hypothesis. This reveals a–second difference with the parametric bootstrap: we have accommodated uncertainty in the θ_{nc}, and therefore uncertainty in the value of the realized test statistic, by averaging over values sampled from the posterior distribution. To obtain the null distribution of the test statistic under the clock hypothesis, we will simulate data by sampling the posterior distribution of θ_c under the clock hypothesis (see Figure 16.3). For each of the predictive data sets sampled, we will need to perform another round of MCMC to sample from the posterior distributions

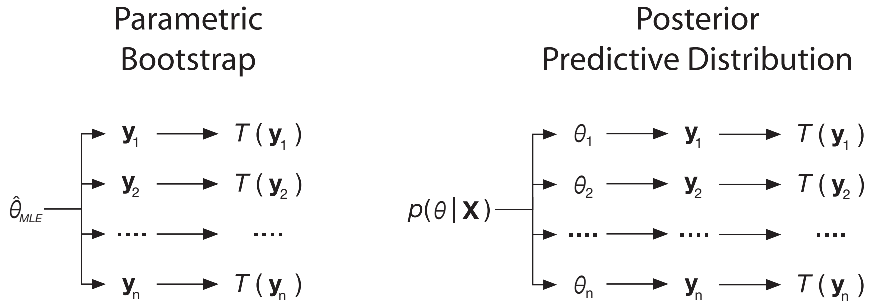

Fig. 16.3. Comparison of the parametric bootstrap and posterior predictive simulation. Values of θ are used to simulate n new data sets ($\{y_1, y_2, \ldots, y_n\}$). These are then evaluated using our chosen test statistic, $T(\cdot)$, giving us the reference distribution under the hypothesis, which is compared with the realized test statistic, $T(\mathbf{X})$.

under the unconstrained hypothesis. The null distribution is summarized from these samples. In this case, the standard deviation for each of these replicates is the predictive distribution of standard deviations expected under the clock hypothesis (conditional on the data and chosen model). As with the parametric bootstrap, we can compare the expectation of the realized deviation to the predictive values under the molecular clock. If the realized value falls outside of the predictive distribution under the clock, then we are tempted to consider the observed deviations as unexplained by a strict molecular clock.

Now, hopefully, you have a feel for the mechanics of predictive tests and some of the differences with the parametric bootstrap, and we are ready to move on and look more closely at the method of posterior predictive simulations.

Sampling from posterior predictive distributions

First, we need a method for generating the predictive distribution of the data before evaluating some function of it. Let $\mathbf{Y} = \{y_1, y_2, \cdots, y_n\}$ be a vector containing n future observations and $\mathbf{X} = \{x_1, x_2, \cdots, x_n\}$ be a vector containing our current observations. What we would like to sample is the predictive distribution of $\mathbf{Y}$ conditional on the hypothesis H,

$$p(\mathbf{Y}|H, \mathbf{X}) = \int_\theta p(\mathbf{Y}|\theta)p(\theta|\mathbf{X})d\theta, \tag{16.13}$$

where θ is a vector containing model parameters under the hypothesis under scrutiny, and $p(\theta|\mathbf{X})$ is the posterior distribution of these parameters. Unfortunately, we can't analytically determine $p(\mathbf{Y}|H, \mathbf{X})$ because the posterior distribution, $p(\theta|\mathbf{X})$, the source of a reasonable set of values for θ under the hypothesis being scrutinized, is impossible to determine analytically for all but the simplest cases in molecular evolution. Furthermore, we can use sampling methods, such as Markov chain Monte Carlo (MCMC), to sample from this distribution, providing an approximation of $p(\theta|\mathbf{X})$ [35, 22, 17] (see Chapters 3 and 7). With values of θ from the posterior distribution, we can approximate the predictive distribution by sampling using the following algorithm:

1) Draw a set of parameter values, θ_i, from the joint posterior distribution of parameters under the null model being tested. (In practice, this can be accomplished by sampling the posterior output of a program that approximates posterior distributions using MCMC, such as MrBayes [17].)
2) Using the values of θ_i (which may include values for the parameters of the substitution process, topology, branch lengths, etc.), simulate data, $\mathbf{y}_i$.
3) Repeat steps one and two N times to create a collection of data sets, $\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n$, corresponding to samples from the posterior distribution of $\theta_1, \theta_2, \cdots, \theta_n$.
4) These simulated data sets are samples from the posterior predictive distribution shown in (16.13) and can be used to evaluate our hypothesis of interest.

The precision of the sampling approximation is a function of the number of draws from the posterior distribution, the precision of our posterior estimate, and the appropriateness of the underlying prior distributions. Fortunately, we are guaranteed by the law of large numbers that we will converge on the target distribution. What exactly is "large" is not clear and is likely to be dependent on the particular parameters of the distribution.

16.3.2 Test Statistics

By sampling we now have an approximation of the posterior predictive distribution of the data simulated under the null model being scrutinized. But we are still left with the following problem: How can we use the posterior predictive distribution to assess our hypothesis H? As already mentioned, we are generally not interested in the predictive distribution of the data directly but some function of it (in our case, a function of the sampling distribution), or more concisely the predictive distribution of the function of interest. Our functions will most often be a descriptive test statistic (often referred to as a summary or discrepancy variable [5]) that quantifies some aspect of the data. The test statistic is referred to as a realized value when summarizing the observed data. In principle, an appropriate test statistic can be defined to measure any aspect of the predictive distribution of the data, but in practice the issue of defining an appropriate statistic for a given hypothesis may not be straightforward [6] and is considered contentious [2].

I follow the general notation $T(\cdot)$, where this is some function of the data. To emphasize that our interests are in sampling from the predictive distribution of $T(\cdot)$, equation (16.13) could be rewritten as

$$p[T(\mathbf{Y})|\mathbf{X}] = \int_{\theta} p[T(\mathbf{Y})|\theta]p(\theta|\mathbf{X})d\theta, \tag{16.14}$$

where $\mathbf{Y}$ is a set of future or predictive observations of the data, $\mathbf{X}$. Using the algorithm outlined above, we can sample this distribution with one additional step; for each simulated data set, we evaluate the function $T(\mathbf{Y})$. (Examples of different test statistics will be described later.) In this way, we now have a sampling approximation of the predictive distribution of the test quantity in which we are directly interested. Importantly, it should be noted that this distribution is averaged over samples from the posterior distribution, allowing us to accommodate uncertainty in our parameter estimates. This frees the test from dependence on any particular set of parameter values by evaluating them in accordance with their probabilities. Whether this is a benefit of the method is yet unclear. (The effects of accommodating uncertainty in parameters in Bayesian molecular evolution studies has not been looked at closely.) This distribution can then be compared with the realized test statistic, $T(\mathbf{X})$, which is calculated from the original data, and the predictive probability of the null hypothesis can then be evaluated.

16.3.3 Predictive *p*-Values

Recently, much research has been directed at the use, properties, and interpretation of p-values as measures for predictive distributions and we direct the reader to [2, 36]. Predictive p-values are often denoted p_T to indicate their dependence on the test statistic and have an operational interpretation similar to classical p-values, as they are both derived from tail area probabilities; values that lie in the extremes of the null distribution of the test quantity are considered significant to reject the null hypothesis. Under classical statistics, the distributions are conditioned on point estimates for model parameters. Predictive densities, on the other hand, are not because parameter values are sampled from the posterior distribution in proportion to their probabilities. This sampling scheme allows them to be treated as nuisance parameters—values not of direct interest—and to be integrated out. Samples from the predictive distribution of the test statistic allow us to evaluate the posterior predictive probability as

$$p_T = \Pr[T(\mathbf{y_{rep}}) \geq T(\mathbf{X})|\mathbf{X}, \theta]. \tag{16.15}$$

The posterior predictive p-value for the test statistic is calculated as

$$p_T = \frac{1}{N}\sum_{i=1}^{N} I(T(\mathbf{y}_i){\geq}T(\mathbf{X})), \tag{16.16}$$

where I is an indicator function that takes on the value of 1 when the equality is satisfied and 0 otherwise, $T(\mathbf{y}_i)$ is the test statistic for the ith simulated data set, and $T(\mathbf{X})$ is the realized test statistic. Probabilities less than the critical threshold, say $\alpha = 0.05$, suggest that the hypothesis under examination is inadequate. Predictive p-values are interpreted as the probability that the hypothesis would produce as extreme a test value as that observed for the data [6]. This approach evaluates the practical fit of the hypothesis to our observations and is dependent on the test statistic employed. These p-values should not be interpreted as frequentist error probabilities or as the probability of our hypothesis. Sellke, Bayarri, and Berger [41] have suggested that p-values can be calibrated to allow for a Bayes factor interpretation (i.e., the odds of H_0 to an unspecified alternative H_1),

$$B(p) = -ep\log(p), p < e^{-1}, \tag{16.17}$$

or a frequentist error probability,

$$\alpha(p) = \left(1 + [-ep\log(p)]^{-1}\right)^{-1}. \tag{16.18}$$

While an extremely powerful and appealing aspect of predictive distributions is the ease and flexibility in test statistics that can be employed, not all test statistics are appropriate. Careful consideration of the hypothesis and its underlying assumptions, and the test statistic, should be made prior to decisions about the hypothesis under scrutiny.

16.3.4 Issues Concerning the Use of Predictive Distributions

Practitioners should be aware of a number of issues surrounding the application of posterior predictive distributions. First, there is an apparent double use of the data. The data are used in the estimation of the posterior distribution during simulation of the predictive distribution and are used again during calculation of the tail area probabilities. A number of general solutions have been suggested by various authors (see [27, 6, 7]). Second, the results are dependent on the choice of test statistic. While the ability of the method to accommodate many different statistics is a benefit, poorly chosen statistics may lead to incorrect conclusions and unpredictable behavior. Third, there are concerns over the properties and interpretation of the different predictive p-values that are available (see [2, 36]), particularly in situations for which composite null models are being entertained. Finally, posterior predictive methods may be highly conservative, resulting in a failure to detect problems with, or deviations from, the null model.

16.3.5 Examples from the Literature

Predictive distributions are a new introduction to studies in molecular evolution and phylogenetics although they have been extensively discussed in the statistical literature (see [38]). Yet they have seen a rapid application to a diverse array of questions in the last few years. In this section, I will briefly review a few different applications from the literature. This should give us some insight into what types of questions have been addressed and can be addressed in the future.

Substitution model adequacy

While substitution model testing in phylogenetics and molecular evolution has been an area of extensive research, until recently little had been done within the Bayesian framework, and many researchers relied on classical approaches, such as the likelihood-ratio test (for a review, see [15]), parametric bootstrap [9], or Akaike information criterion [1], to select models for Bayesian analysis. One drawback to these approaches is that they do not easily accommodate uncertainty in parameter estimates and the topology used in the test. As we have seen, predictive distributions provide a natural approach to accommodating uncertainty. (This is not the only Bayesian approach to model testing that accommodates uncertainty; see the use of Bayes factors in model selection [44].) This approach has been applied to determining model adequacy and choice [3], testing for homogeneity of base frequencies among lineages [18], and testing for lineage rate heterogeneity [31].

Bollback [3] proposed that we could evaluate a substitution model's adequacy using predictive distributions and that this would naturally lead to selection through refinement or enhancement of the model to be used in further analysis. This approach differed most importantly from likelihood-based approaches by taking into account uncertainty in topology, branch lengths, and model parameters. Therefore, model choice has been freed from conditioning on these parameters and has resulted in a more accurate estimate of model variance. The multinomial test statistic was used to evaluate how well a model was able to generate data similar to existing data. Further, the study found that a number of factors affected an increase in the power of the test statistic: (1) increasing the number of sites; (2) increasing sequence divergence (expected number of substitutions per site); and (3) the degree of violation of a model's assumptions.

In a review of Bayesian inference, Huelsenbeck et al. [18] tested for homogeneity of nucleotide frequencies among lineages of the *Drosophila* alcohol dehydrogenase (*Adh*) locus. They used the following test statistic to evaluate the deviation from homogeneity among 58 lineages over time:

$$\chi^2 = \sum_{i=1}^{58} \sum_{j \in \{A,C,G.T\}} \frac{(f_{ij} - \bar{f}_j)^2}{f_j}. \tag{16.19}$$

The authors were able to strongly reject the null hypothesis of nucleotide frequency homogeneity among lineages.

In the final example of evaluating substitution models, Nielsen [31] evaluated lineage rate variation for two data sets: β-globin and influenza hemagglutinin-A. He used the variance in expected number of substitutions, (V_k), as the test statistic and tested the null hypothesis of homogeneity of variances among lineages. By examining the posterior and predictive distributions, he concluded that, because of their small overlap, the null hypothesis of homogeneity could be rejected. This study is important because it used the method of posterior mapping to obtain estimates of V_k for each lineage and used predictive distributions to evaluate significance.

Positive selection

A diverse array of methods for detecting positive selection at sites within a gene is available to molecular evolutionists and phylogeneticists alike, ranging from parsimony-based methods [47] to likelihood-based methods (see Chapter 5) and Bayesian methods (e.g., [32, 12]). The use of posterior mapping and predictive distributions to detect positive selection was introduced by Nielsen [32]. I will focus on this paper because it demonstrates both posterior mapping and predictive distributions to test the null hypothesis of no selection. The authors evaluated the number of nonsynonymous substitutions as their test statistic for an influenza hemagglutinin-A data set. They observed that 11 sites had significant p-values ($p_T \leq 0.01$), suggesting these sites had an excess of nonsynonymous substitutions. They concluded that these sites were under positive selection. To further strengthen their argument, they compared their results with the results of Yang et al. [51], showing a strong concordance between the posterior predictive p-values and posterior probabilities according to the M3 model. None of the 11 sites determined to be under positive selection showed posterior probabilities lower than 0.975.

Correlated character evolution

In this last section, I will review a recent study in which the authors used posterior mapping and predictive distributions to determine correlation among evolving characters [14]. Because the paper deals with morphological characters, it may seem on the surface to have little importance to studies in molecular evolution. But, quite the contrary, it demonstrates how these methods can be extended to studies of correlated molecular evolution. For example, the methods could be applied to looking for correlated change among nucleotides, such as RNA stem partners, or interactions among amino acid sites. Huelsenbeck et al. [14] analyzed the coincidence of states for two morphological characters: self-incompatibility and flower reproductive structure morphology in the family Pontederiaceae.

The phylogeny was estimated using molecular data, and then characters were mapped using the Mk class of models of Lewis [23]. In addition, because the branch lengths of the topology do not reflect the evolutionary rates of the morphological traits and the bias parameter of the morphology model is unknown, a variety of prior distributions were explored for these parameters to reduce dependence on a particular set of values. They used two different test statistics to evaluate coincidence or correlation among the states of the two traits. The first evaluated each character individually, while the second looked for coincidence summed over all state comparisons between the two characters. The basic form of the statistics is

$$d_{ij} = a_{ij}^{(o)} - a_{ij}^{(e)}, \tag{16.20}$$

where $a_{ij}^{(o)}$ is the observed coincidence and $a_{ij}^{(e)}$ is the expected coincidence. The authors found that when evaluating overall coincidence among states they were unable to detect a significant coincidence between the states of the traits. However, by looking at states individually, there was support for a strong coincidence between tristylous flowers and self-incompatibility. This demonstrates an important point about test statistics: a test statistic is only as good as it is a relevant summary of the data with respect to the hypothesis being tested. In the case of the overall coincidence measure, it masked the effect.

16.4 Conclusions

Two recent developments, posterior mapping and predictive distributions, have been developed and applied to questions on molecular evolution and phylogenetics. These methods provide a natural way to address and accommodate uncertainty in various model parameters by sampling with respect to the model's posterior distribution. Posterior mapping provides a powerful method for addressing questions in which detailed data (e.g., type, timing, and order) about the history of a character(s) is required. The dependence on the method of parsimony and its assumptions is no longer necessary. Predictive distributions offer a new approach to hypothesis testing that is general and flexible. Application of these new methods has just begun and will undoubtedly play an ever-increasing role in future studies in molecular evolution and phylogenetics.

References

[1] H. Akaike. A new look at statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

[2] M. J. Bayarri and J. O. Berger. *P* values for composite null models. *Journal of the American Statistical Association*, 95:1127–1142, 2000.

[3] J. P. Bollback. Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, 19:1171–1180, 2002.

[4] J. Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[5] D. Gelfand and X. L. Meng. Model checking and model improvement. In *Markov Chain Monte Carlo in Practice*, pages 189–198. Chapman and Hall, London, 1996.

[6] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall, London, 1995.

[7] A. Gelman, X. L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

[8] J. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, Oxford, 1991.

[9] N. Goldman. Statistical tests of models of DNA substitution. *J Mol Evol*, 36:182–198, 1993.

[10] P. H. Harvey and M. D. Pagel. *The Comparative Method in Evolutionary Biology*. Oxford University Press, 1991.

[11] J. P. Huelsenbeck and J. P. Bollback. Application of the likelihood function in phylogenetic analysis. In *Handbook of Statistical Genetics*, pages 415–439. John Wiley and Sons, Inc., New York, 2001.

[12] J. P. Huelsenbeck and K. A. Dyer. Detecting adaptive molecular evolution when selection changes over time. *Genetics*, In Press.

[13] J. P. Huelsenbeck and N. S. Imennov. Geographic origin of human mitochondrial DNA: Accommodating phylogenetic uncertainty and model comparison. *Systematic Biology*, 51:155–165, 2002.

[14] J. P. Huelsenbeck, R. Nielsen, and J. P. Bollback. Stochastic mapping of morphological characters. *Systematic Biology*, 52:131–158, 2003.

[15] J. P. Huelsenbeck and B. Rannala. Phylogenetic methods come of age: Testing hypotheses in a phylogenetic context. *Science*, 276:174–180, 1997.

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

[17] J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics Applications Note*, 17:754–755, 2001.

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

[19] T. Jukes and C. Cantor. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pages 21–132. Academic Press, New York, 1969.

[20] C. Lanavé, G. Preparata, C. Saccone, and G. Serio. A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution*, 20:86–93, 1984.

[21] C. H. Langley and W. M. Fitch. An estimation of the constancy of the rate of molecular evolution. *Journal of Molecular Evolution*, 3:161–177, 1974.

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

[23] P. O. Lewis. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, 50:913–925, 2002.

[24] J. B. Losos. An approach to the analysis of comparative data when a phylogeny is unavailable or incomplete. *Systematic Biology*, 43:117–123, 1994.

[25] J. B. Losos and D. B. Miles. Ecological Morphology: Integrative Organismal Biology. In P. C. Wainwright and S. M. Reilly, editors, *Adaptation, constraint, and the comparative method: Phylogenetic issues and methods*, pages 60–98. University of Chicago Press, Chicago, 1994.

[26] E. P. Martins. Conducting phylogenetic comparative studies when the phylogeny is not known. *Evolution*, 50:12–22, 1996.

[27] X-L. Meng. Posterior predictive p-values. *Annals of Statistics*, 22:1142–1160, 1994.

[28] W. Messier and C-B. Stewart. Episodic adaptive evolution of primate lysomzymes. *Nature*, 385:151–154, 1997.

[29] A. Ø. Mooers and D. Schluter. Support for one and two rate models of discrete trait evolution. *Systematic Biology*, 48:623–633, 1999.

[30] R. Nielsen. Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. *Genetics*, 159:401–411, 2001.

[31] R. Nielsen. Mapping mutations on phylogenies. *Systematic Biology*, 51:729–732, 2002.

[32] R. Nielsen and J. P. Huelsenbeck. Detecting positively selected amino acid sites using posterior predictive p-values. In *Pacific Symposium on Biocomputing, Proceedings*, pages 576–588. World Scientific, Singapore, 2001.

[33] M. D. Pagel. Detecting correlated evoluton on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B*, 255:37–45, 1994.

[34] M. D. Pagel. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Systematic Biology*, 48:612–622, 1999.

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

[36] J. R. Robins, A. van der Vaart, and V. Ventura. Asymptotic distribution of p-values in composite null models. *Journal of the American Statistical Association*, 95:1143–1156, 2000.

[37] F. Rodríguez, J. Oliver, A. Marín, and J. Medina. The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, 142:485–501, 1990.

[38] D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12:1151–1172, 1984.

[39] D. Schluter. Uncertainty in ancient phylogenies. *Nature*, 377:108–109, 1995.

[40] D. Schluter, T. Price, A. Ø. Mooers, and D. Ludwig. Likelihood of ancestor states in adaptive radiation. *Evolution*, 51:1699–1711, 1997.

[41] T. Sellke, M. J. Bayarri, and J. O. Berger. Calibration of p-values for precise null hypotheses. In *ISDS Discussion Paper 99-13*, Durham, NC, 1999. Duke University.

[42] B. Sillen-Tullberg. Evolution of gregariousness in aposematic butterfly larvae: A phylogenetic analysis. *Evolution*, 42:293–305, 1988.

[43] M. A. Suchard, R. E. Weiss, K. S. Dorman, and J. S. Sinsheimer. Oh brother, where art thou? A Bayes factor test for recombination with uncertain heritage. *Systematic Biology*, 51:715–728, 2002.

[44] M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer. Bayesian selection of continuous-time Markov chain evolutionary models. *Molecular Biology and Evolution*, 18:1001–1013, 2001.

[45] M. A. Suchard, R. E. Weiss, and J. S. Sinsheimer. Testing a molecular clock without an outgroup: Derivations of induced priors on branch length restrictions in a Bayesian framework. *Systematic Biology*, 52:48–54, 2003.

[46] M. A. Suchard, R. E. Weiss, J. S. Sinsheimer, K. S. Dorman, P. Patel, and E. R. B. McCabe. Evolutionary similarity among genes. *Journal of the American Statistical Association*, 98:653–662, 2003.

[47] Y. Suzuki and T. Gojobori. A method for detecting positive selection at single amino acid sites. *Molecular Biology and Evolution*, 16:1315–1328, 1999.

[48] D. L. Swofford, G. J. Olsen, P. J. Waddell, and D. M. Hillis. Phylogenetic inference. In *Molecular Systematics*, pages 407–514. Sinauer Associates, Sunderland, MA, 2nd edition, 1996.

[49] A. R. Templeton. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the cytochrome oxidase ii gene in the hominoid primates. *Genetics*, 144:1263–1270, 1996.

[50] R. H. Ward, B. L. Frazier, K. Dew-Jager, and S. Pääbo. Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences USA*, 88:8720–8724, 1991.

[51] Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedereon. Codon-substitution models for variable selection pressure at amino acid sites. *Genetics*, 155:431, 2000.