*Syst. Biol.* 51(5):729–739, 2002
DOI: 10.1080/10635150290102393

# Mapping Mutations on Phylogenies

RASMUS NIELSEN

*Department of Biological Statistics, Cornell University, 439 Warren Hall, Ithaca, New York 14853-7801, USA; E-mail: rn28@cornell.edu*

*Abstract.*—Mapping of mutations on a phylogeny has been a commonly used analytical tool in phylogenetics and molecular evolution. However, the common approaches for mapping mutations based on parsimony have lacked a solid statistical foundation. Here, I present a Bayesian method for mapping mutations on a phylogeny. I illustrate some of the common problems associated with using parsimony and suggest instead that inferences in molecular evolution can be made on the basis of the posterior distribution of the mappings of mutations. A method for simulating a mapping from the posterior distribution of mappings is also presented, and the utility of the method is illustrated on two previously published data sets. Applications include a method for testing for variation in the substitution rate along the sequence and a method for testing whether the $d_N/d_S$ ratio varies among lineages in the phylogeny. [Data augmentation; DNA sequence evolution; Markov chain Monte Carlo; phylogenetics.]



The mapping of mutations on trees has been one of the most popular methods for data analysis in phylogenetics and molecular evolution. The basic idea is that if we know the location of mutations on a phylogenetic tree we can easily make inferences about the causes of molecular evolution. For example, we may want to test whether some types of mutations occur at a higher rate than other types of mutations, e.g., transitions versus transversions or nonsynonymous versus synonymous mutations. The classical approach to this problem has been to infer the location of mutations on the phylogeny using maximum parsimony. The inferred mutations are then treated as pseudodata to which classical statistical methods are applied. For example, such methods have been used to infer the ratio of nonsynonymous to synonymous substitution rates (e.g., Templeton, 1996; Bush et al., 1999) and the frequency of conservative versus radical amino acid substitutions (Wyckoff et al., 2000). Unfortunately, it is difficult to analyze inferred mappings of mutations in a rigorous statistical framework. Doing so would require taking into account the uncertainty in the estimation of the mapping, and no statistical method for doing this has been proposed. By only focusing on possible parsimony mappings, we ignore many possible mappings that are expected to occur with high probability. The mappings selected by focusing only on maximum parsimony mappings are the ones requiring the fewest possible mutations. Estimates of the numbers and rates of mutations using maximum parsimony are therefore biased. The cost of using maximum parsimony is twofold: We ignore the variance caused by the estimation of the mutational mapping, often making statistical tests using the pseudodata anticonservative, and we bias our results toward low rates of substitution. As an illustration of this problem, consider the phylogeny and the data in Figure 1. The branch lengths of the phylogeny are scaled, as is tradition, in terms of the expected number of substitutions per site. For simplicity, I assume a Jukes and Cantor (1969) model of evolution and calculate the probability of each possible mapping. The probability of a mapping ($M$) given the observed nucleotide data ($D$) is

$$\Pr(M \mid D) = \frac{\Pr(M, D)}{\Pr(D)}, \tag{1}$$

and the numerator and the denominator can be calculated using well-known methods. Assuming a time-reversible Markov model of evolution and the tree topology in Figure 1, the denominator is

$$\Pr(D) = \sum_{i\in\{A,C,T,G\}} \sum_{j\in\{A,C,T,G\}} \pi_i P_{iT}(0.1) P_{iG}(3) \times P_{ij}(3) P_{jG}(3) P_{jG}(0.1), \tag{2}$$

where $\pi_i$ is the stationary frequency of nucleotide $i$. Under the Jukes and Cantor (1969) model $\pi_i = 1/4$ for all $i$, and

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

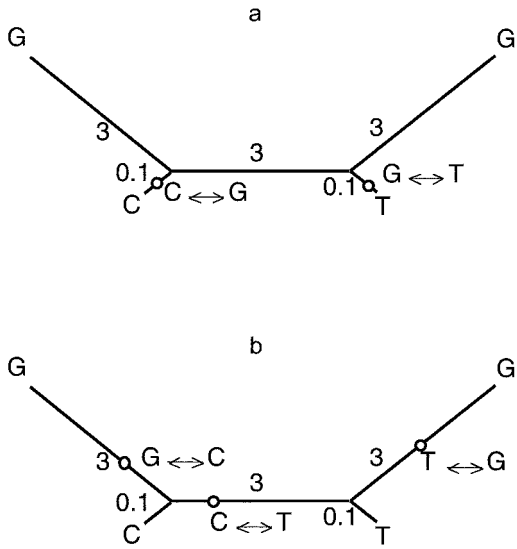


FIGURE 1. Two alternative mutational mappings. The numbers along the lineages indicate the length of the lineages measured in expected number of substitutions. The circles indicate mutations. (a) Parsimony mapping, which requires only two mutations. (b) The alternative mapping is almost 1,000 times as likely to be the right mapping.

for all $i$ and $j$. The parameter $t$ is the length of the lineage measured in expected number of mutations. We then get $\Pr(D) = 3.84 \times 10^{-3}$ for the data and phylogeny of Figure 1. For each particular possible mapping of mutations, we can also calculate $\Pr(M, D)$. For example, for the mapping in Figure 1a,

$$\Pr(M_a, D) = e^{-3}(0.1e^{-0.1}/3)e^{-3}e^{-3} \times (0.1e^{-0.1}/3)/4 = 2.81 \times 10^{-8}. \quad (4)$$

This follows from the definition of the substitution process as a continuous-time Markov chain. Each of the five factors in Equation (4) correspond to the probability of the path of the Markov chain along a lineage of the phylogeny. For example, because the number of substitutions is Poisson distributed on each lineage under the Jukes and Cantor (1969) model, the conditional probability of observing exactly one mutation between G and T on a lineage of length 0.1, given that the ancestral state is G, is the probability of observing exactly one mutation multiplied by the probability that this mutation is from G to T, i.e., $0.1e^{-0.1} \times (1/3)$.

We can now calculate the probability of each possible mapping of mutations on the phylogeny. For example, $\Pr(M_a \mid D) = 2.81 \times 10^{-8}/3.84 \times 10^{-3} = 7.31 \times 10^{-6}$, and for the mappings in Figure 1b, $\Pr(M_b \mid D) = 2.53 \times 10^{-5}/3.84 \times 10^{-3} = 8.89 \times 10^{-3}$. Mapping $M_a$ is the parsimony mapping, because it is the only mapping compatible with the data requiring only two mutations. Mapping $M_b$ is not a parsimony mapping because it assumes three mutations. However, mapping $M_b$ is almost 1,000 times as likely to be the right mapping as the parsimony mapping. Not only is the parsimony mapping very unlikely, there are nonparsimonious mappings that are much more likely than the most parsimonious mapping. Clearly, it seems unreasonable in this case to focus just on the parsimony mappings. Instead it might be more reasonable to consider a larger collection of possible mappings and to weight them by their probabilities. Here, I describe such a method, how it can be established using simulations for a fixed phylogeny, and how it can be combined with a method for sampling genealogies from the posterior distribution of genealogies to provide a framework for statistical inference in molecular evolution. I illustrate the method by analyzing a previously published data set of DNA sequences from the influenza virus (Fitch et al., 1997).

### *Sampling a Mapping of Mutations*

We are interested in making inferences based on the distribution of mutations on a phylogeny. In very general terms, we might be interested in evaluating the function $h_\Theta(D)$, but we can only calculate $h_\Theta(M, D)$ directly. The subscript $\Theta$ is a vector of parameters, including the tree topology and branch lengths. We might, for example, be interested in estimating the number of nonsynonymous mutations, the proportion of the amino acid substitutions that are radical, or the ratio of the number of mutations to arginine in one part of the phylogeny to the number occurring in another part of the phylogeny. These are not observable quantities; we cannot simply deduce which mutations occurred simply from observing the

sequence data $D$. However, we can evaluate the conditional expectation of $h_\Theta(D, M)$:

$$E[h_\Theta(D, M) \mid D] = \sum_{M \in \Psi} h_\Theta(M, D) \Pr_\Theta(M \mid D), \quad (5)$$

where $\Psi$ is the set of possible mappings. For example, if we are interested in finding the number of mutations leading to an arginine in a certain part of the phylogeny, a natural estimator of this quantity is the posterior expectation of the number of mutations to arginine:

$$E[n_{\text{ARG}} \mid D] = \sum_{M \in \Psi} n_{\text{ARG}}(M, D) \Pr_\Theta(M \mid D), \quad (6)$$

where $n_{\text{ARG}}$ is the number of mutations to arginine and $n_{\text{ARG}}(M, D)$ is the observed number of mutations to arginine occurring in a particular mapping of mutations ($M$). In general, we cannot evaluate the sums in Equations 5 and 6 directly because the set $\Psi$ is not of finite size. However, we can approximate it by simulating $N$ mappings of mutations, $M_1, \ldots, M_N$, from the distribution $\Pr_\Theta(M \mid D)$ and evaluating

$$E[h_\Theta(D, M) \mid D] \approx \frac{1}{N} \sum_{i=1}^{N} h_\Theta(M_i, D), \quad (7)$$

where the right side of the equation converges to the left side of the equation by the law of large numbers as $N$ becomes large. If we can simulate mappings of mutations from the distribution $\Pr_\Theta(M \mid D)$, we can evaluate $E[h_\Theta(D, M) \mid D]$.

### *Model and Simulations*

In the following, I assume that the substitution process can be modeled using a continuous-time Markov chain model with infinitesimal generator

$$\begin{bmatrix} -q_{\text{A}} & q_{\text{AC}} & q_{\text{AT}} & q_{\text{AG}} \\ q_{\text{CA}} & -q_{\text{C}} & q_{\text{CT}} & q_{\text{CG}} \\ q_{\text{TA}} & q_{\text{TC}} & -q_{\text{T}} & q_{\text{TG}} \\ q_{\text{GA}} & q_{\text{GC}} & q_{\text{GT}} & -q_{\text{G}} \end{bmatrix}, \quad (8)$$

where $q_i = \Sigma_{j:j \neq i} q_{ij}$. The rest of this section describes an algorithm for sampling mappings of mutations ($M$) from the distribution $\Pr_\Theta(M \mid D)$ for a single site. Simulations for an entire sequence can be completed by applying this algorithm to all sites in the sequence. The algorithm has three basic steps: (1) calculate the fractional likelihood ($f_{ki}$) of all four nucleotides ($i$ = A, C, T, G) at each node ($k$) in the rooted phylogeny; (2) simulate a set of ancestral states ($Y_k$) at each node of the phylogeny, using the fractional likelihoods calculated in step 1; and (3) simulate the mutational history of each lineage of the phylogeny conditional on the ancestral states.

Step 1 is the most time-consuming step, but it is easily completed using the familiar Felsenstein (1981) algorithm. The fractional likelihood of nucleotide $i$ in node $k$ is $f_{ki} = \Pr(D_k \mid Y_k = i)$, where $D_k$ is the data in the leaf nodes (tips in the tree) that are eventual descendents of node $k$ and $Y_k$ is the nucleotide in node $k$.

Step 2 is completed by first simulating the ancestral state at the root of the tree. If the index for the root node is $r$, then

$$\Pr_\Theta(Y_r = i \mid D) = \frac{f_{ri}\pi_i}{\sum_{j \in \{\text{A,C,T,G}\}} f_{rj}\pi_j}. \quad (9)$$

The ancestral state at the root node can, therefore, easily be sampled from the posterior distribution by simulation.

The ancestral states at a node right above the root, say node $r - 1$, can similarly be simulated by sampling from the distribution

$$\Pr_\Theta(Y_{r-1} = j \mid Y_r = i, D) = \frac{f_{r-1,i} P_{ij}(t_{r-1})}{\sum_{k \in \{\text{A,C,T,G}\}} f_{r-1,k} P_{ik}(t_{r-1})}, \quad (10)$$

where $t_{r-1}$ is the length of the lineage leading to node $r - 1$ in the rooted tree. This step can be repeated recursively for nodes higher in the tree, resulting in a sample from the distribution $\Pr_\Theta(Y_1, Y_2, \ldots, Y_r \mid D)$.

In step 3, the mutational history of a lineage in the tree is simulated. A lineage is attached to two nodes, an ancestral node that has been assigned nucleotide $Y_A$ and a descendent node assigned nucleotide $Y_E$.

The mutational path of the lineage ($M_j$) must then be simulated from the distribution $\Pr_\Theta(M_j \mid Y_A, Y_E)$. Simulations from this distribution can be obtained in various ways. A very simple method that is quite efficient is simply to simulate a realization of the continuous-time Markov chain defined by the mutational model from the initial state $Y_A$. If the last state visited is not $Y_E$, a new independent simulation of the Markov chain is completed. Simulations are repeated until a path of the chain visiting $Y_E$ last has been found. The simulations are completed by simulating the waiting times between mutations according to the process given by Equation 8. For example, if the current state is nucleotide C, the waiting time to the next mutation is simulated from an exponential distribution with rate $q_C$. If the total simulation time is larger than the length of the lineage ($t$), the simulation scheme is terminated. Otherwise, a new mutation is mapped on the tree according to the distribution $\Pr(i) = \frac{q_{Ci}}{q_C}$, $i \in \{A, T, G\}$.

This simulation scheme is efficient assuming that the rates of change between all nucleotides are reasonably large, except when $Y_A \neq Y_E$, and $t$ is very short. To improve computational efficiency in this case when $Y_A \neq Y_E$, the time to the first mutation, is simulated from the density

$$f(t_1 \mid t_1 < t) = \frac{q_{Y_A} e^{-q_{Y_A} t_1}}{1 - e^{-q_{Y_A} t}}, \quad 0 \leq t_1 < t, \quad (11)$$

where $t_1$ is the time of the first mutation. Simulations from this density can easily be performed using the inverse transformation method (e.g., Nielsen, 2001). Simulating the time of the first mutation from this density is equivalent to simulating the process conditional on at least one mutation occurring during time $t$.

### *Computational Speed*

To illustrate the computational efficiency of this method, 1,000 data sets, each containing 20 sequences, were simulated under the Jukes and Cantor (1969) model. The model tree was simulated assuming a uniform prior over all topologies and equal length of all lineages. The total tree length (expected number of substitutions per site) was set to 5, and there were 500 nucleotide sites in each sequence, corresponding to a highly saturated data set. For each of the simulated data sets, a random mapping was simulated from $\Pr_\Theta(M \mid D)$. The total computational time required to simulate all 1,000 data sets was 57 seconds on a 800-MHz Pentium III desktop computer. The mean number of mutations per site in the simulated mutational mappings was 4.9967.

### *Integrating over Genealogies*

The algorithm described above provides a reasonably efficient method for simulating mappings of mutations from the distribution $\Pr_\Theta(M \mid D) = \Pr(M \mid D, \Theta)$. However, in real data analysis the exact values of the parameters, $\Theta$, are not known. We would instead like to be able to sample mutational mappings from the distribution

$$\Pr(M \mid D) = \int_{\Theta \in \Omega} \Pr(M \mid D, \Theta) p(\Theta \mid D) d\Theta, \quad (12)$$

where $\Omega$ is the set of possible values of $\Theta$. We can think of this integral as a sum over all possible tree topologies and a multiple integral over all possible branch lengths and values of the parameters of the mutational process. By evaluating this integral, we can effectively take uncertainty regarding the nuisance parameters in $\Theta$ into account. One method for simulating from this distribution was described by Nielsen (2001). However, for highly divergent data sets, that method is not computationally efficient. Instead, we can use the following simulation scheme: (1) simulate a value of $\Theta$ ($\Theta^*$) from the density $p(\Theta \mid D)$, and (2) simulate a mapping of mutations ($M^*$) from the distribution $\Pr(M \mid D, \Theta^*)$. Then $M^*$ is a sample from the marginal posterior distribution of $M$, $\Pr(M \mid D)$. In the previous section, a method for performing step 2 was described. Several authors, including Yang and Rannala (1997), Larget and Simon (1999), and Huelsenbeck et al. (2000), have presented methods for simulating from $p(\Theta \mid D)$ using Markov chain Monte Carlo (MCMC). In these methods, a Markov chain is defined with state space given by the possible values of $\Theta$ and stationary distribution $p(\Theta \mid D)$. The Markov chain is simulated using the

Metropolis–Hastings method (Metropolis et al., 1953; Hastings, 1970), and correlated samples from $p(\Theta \mid D)$ are obtained by sampling the Markov chain at stationarity. For example, using the computer program MrBayes (Huelsenbeck and Ronquist, 2001) it is possible to obtain samples of $p(\Theta \mid D)$ under a general time-reversible model and assuming uniform priors for branch lengths, topology, and parameters of the mutational process. We are therefore now capable of realizing samples from $\Pr(M \mid D)$ in a computationally efficient manner by repeating the simulating scheme described in steps 1 and 2. Such samples can be used for making inferences and testing hypotheses regarding the distribution and abundance of mutations on the phylogeny.

## APPLICATIONS

To illustrate the utility of the method, I here analyze two previously published data sets. The first data set contains 28 sequences of the hemagglutinin (HA) gene of human influenza virus A. This data set was also analyzed in Yang et al. (2000) and contains a subset of the sequences originally analyzed by Fitch et al. (1997). There has been some interest in this data set because it represents a case of positive selection. As argued by Fitch et al. (1997) and Yang et al. (2000), there seems to be an excess of nonsynonymous over synonymous substitutions in some codons in this data set. Furthermore, the larger data set originally analyzed by Fitch et al. (1997) has also been used to make predictions regarding which strains will cause epidemics. The basic idea is that the distribution of nonsynonymous mutations on the genealogy can be used to make inferences regarding where positive Darwinian selection is acting. Sites targeted by positive selection might be the sites for which the virus depends for immune avoidance. Changes in these sites might then provide important clues as to which strain is most likely to cause the next epidemic. The 28-sequence data set was used to illustrate how inferences about the pattern of mutation can be made using the methods for mapping mutations.

The second data set includes vertebrate $\beta$-globin sequences described by Yang et al. (2000) and contains 17 nucleotide sequences, 432 bp in length. This data set was originally compiled using a GenBank database search. Yang et al. (2000) showed that positive Darwinian selection has been influencing the evolution of these sequences.

### *Number of Mutations in the Phylogeny*

To evaluate the difference in results obtained using parsimony and using the Bayesian method for mapping mutations, the parsimony number of mutations was determined, optimizing for topology using PAUP* (Swofford, 2001). The parsimony scores were 757 for the $\beta$-globin data set and 272 for the influenza data set. Next, the posterior distribution of the number of mutations in the phylogeny was obtained using the new Bayesian method for mapping mutations. For each data set, the following protocol was used to generate the distributions. One thousand samples from the posterior distribution of $\Theta$ (mutational parameters and tree) were generated using the computer program MrBayes (Huelsenbeck and Ronquist, 2001). The model of nucleotide substitution was a general time reversible model (GTR; e.g., Tavaré, 1986), with a Dirichlet (1, 1, 1, 1) prior for the base frequencies and a uniform (0, 100) prior for the substitution rates between nucleotides (fixing the rate of substitution between nucleotides G and T to 1 to avoid overparameterization). A uniform prior on all possible topologies was assumed. Branch lengths were assumed to be uniformly distributed between 0 and 100 (expected number of substitutions). To generate the 1,000 samples, 1,100,000 cycles were simulated in the Markov chain, and values of $\Theta$ were sampled every 1,000 cycles from cycle 100,000 to cycle 1,100,000. For each of the 1,000 samples from the posterior distribution of $\Theta$, a mapping of mutations was simulated from the distribution $\Pr(M \mid D, \Theta)$ under the same model of nucleotide change.

The posterior distributions of the number of mutations for the two data sets are shown in Figures 2 and 3. In the case of the influenza data set, using parsimony would result in underestimating the number of mutations by approximately 60 mutations, or $>20\%$. In the case of the $\beta$-globin data set, the number of mutations would be underestimated by almost 200, or about 25%. The mutations that would be missed are not just random mutations. The tendancy would be to preferentially miss mutations in quickly evolving sites, mutations in internal

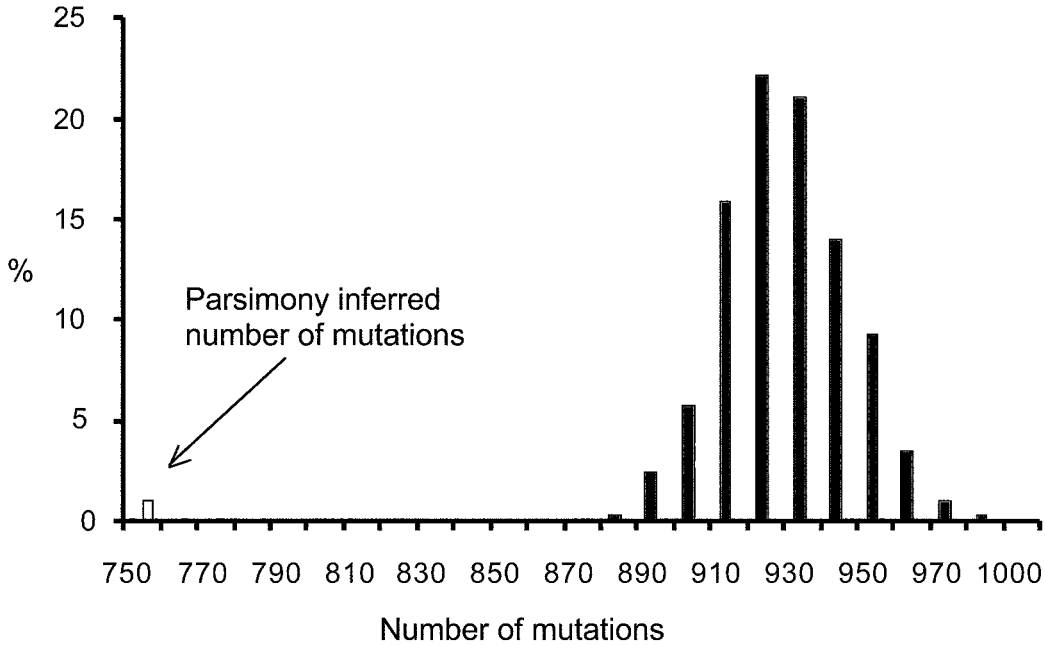


FIGURE 2. The posterior distribution of the number of mutations (solid) and the parsimony-based estimate of the number of substitutions (open) for the $\beta$-globin data set.

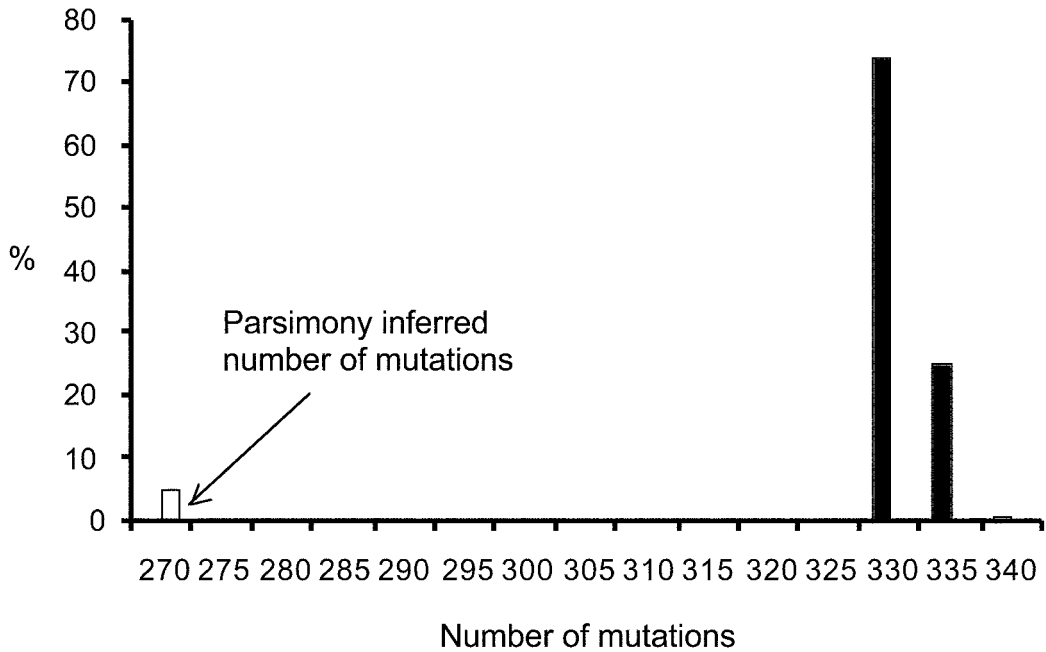


FIGURE 3. The posterior distribution of the number of mutations (solid) and the parsimony-based estimate of the number of substitutions (open) for the influenza HA data.

lineages of the phylogeny, transitions if there is a transition/transversion bias, and synonymous mutations if they occur at a higher rate than nonsynonymous mutations. This effect should not be ignored when making inferences in molecular evolution.

To illustrate the utility of the new method for mapping mutations on the phylogeny, I consider two biological questions of interest: (1) Does the rate of substitution vary among sites? and (2) Is the ratio of the rate of synonymous to nonsynonymous mutations the same in all lineages of the tree?

### *Posterior Predictive Distributions*

To examine each of these problems, some functions of both the observed and the missing data that will take on the role of test statistics will be defined. The methods described above can be used to approximate the posterior distribution of these functions given the data. To examine whether the posterior distribution differs from what would be expected under a particular null model, it can be compared with the posterior predictive distribution. The posterior predictive distribution of a statistic, $T(\cdot)$, is defined as

$$p[T(D^{rep}) \mid D] = \int_{\Theta \in \Omega} p[T(D^{rep}) \mid \Theta] \times p(\Theta \mid D)\, d\Theta, \quad (13)$$

where $D^{rep}$ denotes a replication of the data $D$, i.e., a future observation. The posterior predictive distribution provides the distribution that would be expected given what has been learned about the parameters from the observed data. The use of posterior predictive distributions for inference purposes is well known in Bayesian statistics (e.g., Rubin, 1984). In the present case, the statistics of interest are also functions of the missing data, $M$. The posterior predictive distribution of interest is therefore

$$p[T(D^{rep}, M^{rep}) \mid \mathrm{D}] = \int_{\Theta \in \Omega} p[T(D^{rep}, M^{rep}) \mid \Theta] \times p(\Theta \mid D)\, d\Theta. \quad (14)$$

The posterior predictive distribution can be evaluated using simulations. Data sets can be simulated under a particular value of $\Theta$, using a method identical to the method used for simulating missing data from the distribution $\Pr(M \mid D, \Theta)$. The only difference being that we now do not condition on the observed data when simulating mutational mappings on the branches of the phylogeny.

### *Rate Variation*

The first question to examine is whether there is rate variation among sites in the two data sets. This is a relatively easy question to answer. There already exists a variety of tests of rate heterogeneity. For example, in a likelihood framework a model that assumes a gamma distribution of rates among sites can be compared with a model that allows no rate variation among sites. If the difference in the maximum likelihood value between the two models is sufficiently large, the hypothesis of no rate variation can be rejected. In the present approach, we can simply observe whether the simulated distributions of mutations are compatible with the hypothesis of no rate variation. For present purposes, it will suffice simply to consider the variance in the number of substitutions per site ($V_K$).

To estimate the posterior and the predictive distribution of the variance, 1,000 mutational mappings were simulated from the posterior and the predictive distribution of mappings. The simulation method and the model assumptions were as described in the *Number of Mutations in the Phylogeny* section. No rate variation among sites is assumed.

The posterior distribution and the predictive distribution of $V_K$ is plotted in Figure 4 for the influenza data set and in Figure 5 for the $\beta$-globin data set. The distribution of the observed value of $V_K$ in the influenza data set lies well outside the predictive distribution. In the $\beta$-globin data set, there is very little overlap between the two distributions. Thus, we can reject the null hypothesis of no rate variation. Clearly, the variance in the number of substitutions per site is much larger than expected under a model of no rate variation.

### *Distribution of Nonsynonymous and Synonymous Substitutions*

Testing hypotheses regarding the distribution of nonsynonymous and synonymous substitutions has been a commonly used tool for examining hypotheses regarding

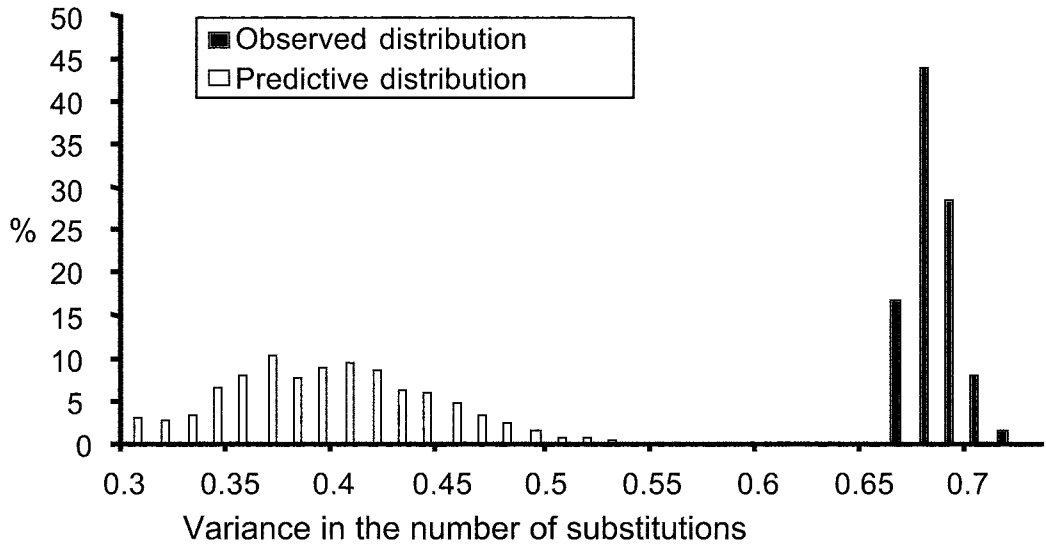


FIGURE 4. The posterior and predictive distributions of the variance in the number of substitutions per site in the influenza HA data.

selection at the molecular level. Such methods are not well justified if substitutions are inferred using parsimony or related methods. The most common statistical solution to this problem is to model the substitution process at the codon level using a continuous-time Markov chain model with state space on the set of all possible codons (e.g., 61 codons in the universal genetic code). The rate ratio of nonsynonymous to synonymous substitutions, which is a proxy for the strength of selection at the amino acid level, is an explicit parameter in this model and can be estimated using maximum likelihood. These methods have been very successful, especially for identifying genes undergoing positive

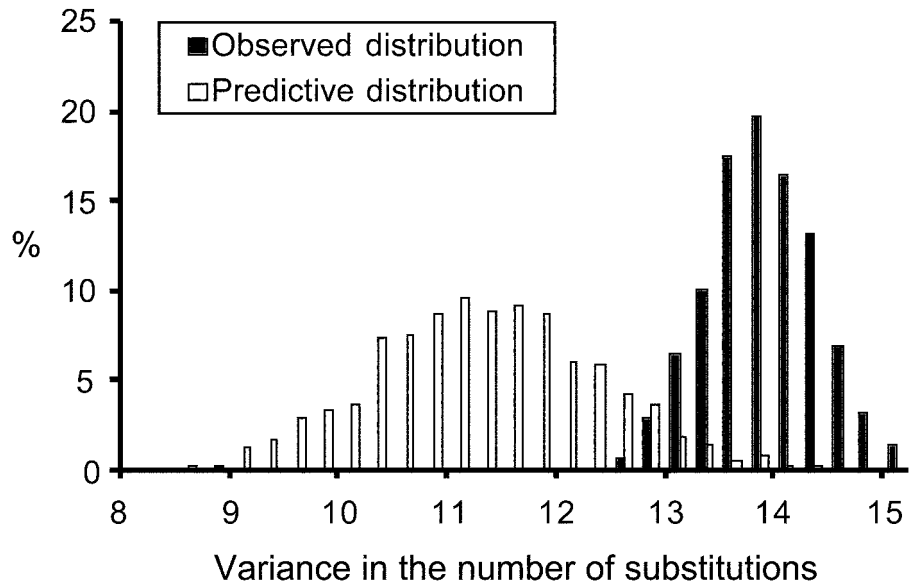


FIGURE 5. The posterior and predictive distributions of the variance in the number of substitutions per site in the $\beta$-globin data set.

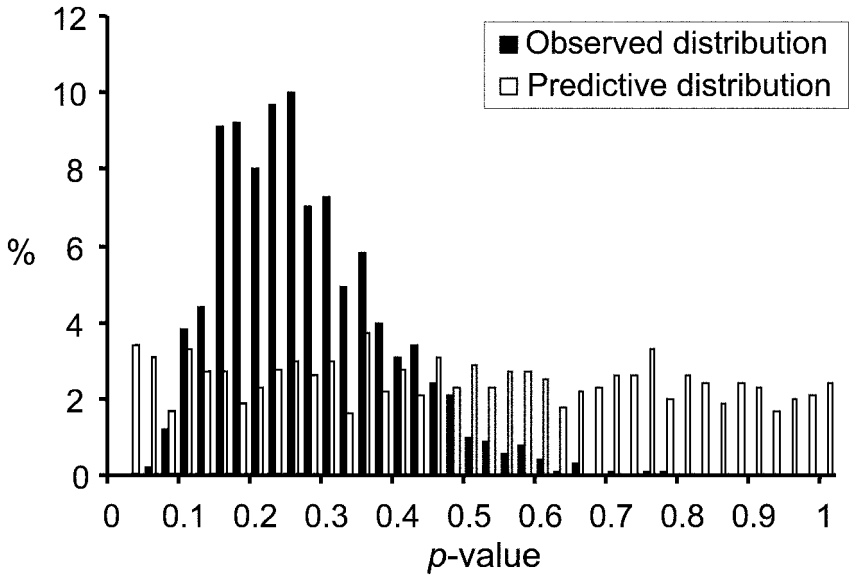


FIGURE 6. The posterior and predictive distributions of the $P$ value from a test of homogeneity of the number of nonsynonymous and synonymous mutations on internal and external lineages for the influenza HA data.

selection. However, they are usually very computationally intensive and they also usually rely on the assumption that the topology of the genealogy is known (or can be estimated with great accuracy). An alternative is to use the new method for mapping mutations on the genealogy to make statistical inferences regarding the distribution of nonsynonymous and synonymous substitutions.

In the new method, nonsynonymous and synonymous can be directly inferred for any particular mapping. It is not necessary to use explicit codon-based models of substitutions. However, in some cases it may

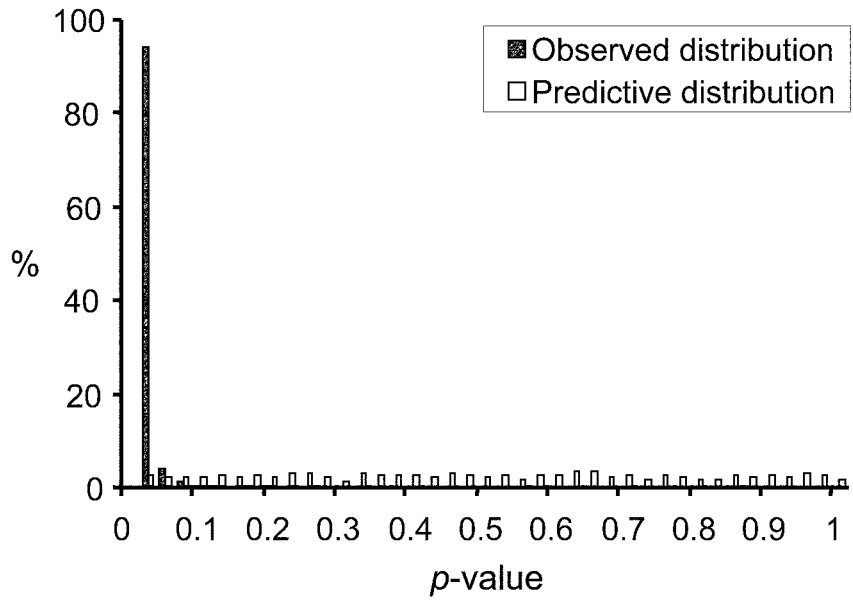


FIGURE 7. The posterior and predictive distributions of the $P$ value from a test of homogeneity of the number of nonsynonymous and synonymous mutations on internal and external lineages for the $\beta$-globin data set.

still be more appropriate to use codon-based models because they are more realistic in incorporating differences in the rates of nonsynonymous and synonymous substitutions. For now, we settle for the GTR model, i.e., it is assumed that $d_N/d_S = 1$ and that nonsynonymous and synonymous mutations are identically distributed among lineages. As always, the models may not incorporate all complexities of molecular evolution, and the results should be interpreted accordingly.

The first question is to what degree the $d_N/d_S$ ratio varies among lineages in the genealogy. The statistic used to examine this problem is the $P$ value from a test of homogeneity in a $k \times 2$ contingency table, with $k$ rows for the $k$ lineages of the phylogeny and two columns for nonsynonymous and synonymous substitutions. The $P$ value in the current representation is a function of the mutational mapping and is therefore a random variable, even after the data have been observed. The results for the two data sets are shown in Figures 6 and 7. The predictive distributions are approximately uniform as expected. For influenza data set, the $P$ values are of moderate size; there seems to be very little variation in the $d_N/d_S$ ratio among lineages. However, for the $\beta$-globin data set, the $P$ values are very low; almost all (982/1,000) were $<0.05$. This finding strongly suggests that there is significant variation in the $d_N/d_S$ ratio among lineages in the $\beta$-globin data set.

## Conclusions

Inferences based on mapping mutations on a phylogeny using parsimony to construct pseudodata sets of mutational events is a common practice in molecular evolution. However, from a statistical standpoint, this approach is not well justified. Instead, it is possible to use the Bayesian method for mapping mutations on the phylogeny as presented. Inferences are then based on the posterior distribution of mutational mappings. This approach can readily be adapted to address a variety of questions. Here, methods for analyzing the degree of rate variation and the distribution of nonsynonymous and synonymous substitutions were explored. Other obvious applications include the analysis of correlated evolution among sites and inferences in amino acid–based models.

An obvious extension of the method is the evaluation of posterior predictive $P$ values (e.g., Meng, 1994) for the purpose of hypothesis testing. Although the evaluation of such $P$ values would be very computationally intensive in the present approach because it would involve integrating over the set of possible mappings for each predictive data set, it would not be computationally unfeasible. The use of posterior predictive $P$ values might be an attractive alternative to the use of the parametric bootstrap in molecular evolution in that nuisance parameters are dealt with using integration instead of optimization. More research on the statistical properties of such tests is needed.

## Acknowledgment

This research was support by NSF grant DEB-0089487.

## References

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. 16:1457–1465.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA 94:7712–7718.

Hastings, W. K. 1970. Monte Carlo sampling methods using Markov chains and their applications. Biometrika 57:97–109.

Huelsenbeck, J. P., B. Rannala, and B. Larget. 2000. A Bayesian framework for the analysis of cospeciation. Evolution 54:353–364.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes 2.0. Available from http://brahms.biology.rochester.edu/software.html

Jukes, T. H., and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Larget, B., and D. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Meng, X.-L. 1994. Posterior predictive $p$-values. Ann. Stat. 22:1142–1160.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equations of state calculations by fast computing machines. J. Chem. Phys. 21:1087–1091.

Nielsen, R. 2001. Mutations as missing data: Inferences on the ages and distributions of nonsynonymous and synonymous mutations. Genetics (in press).

Rubin, D. B. 1984. Bayesianly justifiable and relevant frequency calculations for the applied statisticians. Ann. Stat. 12:1151–1172.

Swofford, D. L. 2001. PAUP* 4.0b2. Sinauer, Sunderland, Massachusetts.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lect. Math. Life. Sci. 17:57–86.

TEMPLETON, A. R. 1996. Contingency tests of neutrality using intra/interspecific gene trees: The rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. Genetics 144:1263–1270.

WYCKOFF, G. J., W. WANG, AND C. I. WU. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309.

YANG, Z., R. NIELSEN, N. GOLDMAN, AND A.-M. K. PEDERSEN. 2000. Codon-substitution models for variable selection pressure at amino acid sites. Genetics 155:431–449.

YANG, Z., AND B. RANNALA. 1997. Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. Mol. Biol. Evol. 14:717–724.

*First submitted 27 June 2001; reviews returned 5 February 2002; final acceptance 9 April 2002*
*Associate Editor: John Huelsenbeck*