Pacific Symposium on Biocomputing 7:576-588 (2002)

DETECTING POSITIVELY SELECTED AMINO ACID SITES USING POSTERIOR PREDICTIVE P-VALUES

R. NIELSEN
Department of Biometrics, Cornell University, 439 Warren Hall, Ithaca, NY 14853-7801, (rn28@cornell.edu)

J. P HUELSENBECK
Department of Biology, University of Rochester, Rochester, NY 14627-0211, (johnh@brahms.biology.rochester.edu)

Identifying positively selected amino acid sites is an important approach for making inference about the function of proteins; an amino acid site that is undergoing positive selection is likely to play a key role in the function of the protein. We present a new Bayesian method for identifying positively selected amino acid sites and apply the method to a data set of hemagglutinin sequences from the Influenza virus. We show that the results of the new methods are in accordance with results obtained using previous methods. More importantly, we also demonstrate how the method can be used for making further inferences about the evolutionary history of the sequences. For example, we demonstrate that sites that are positively selected tend to have a preponderance of conservative amino acid substitutions.

1 Introduction

1.1 The d_N/d_S ratio

The degree to which an amino acid site is free to vary is strongly dependent on its structural and functional importance. An amino acid that plays a critical role—perhaps as a member in a functionally important structure—is unlikely to change over evolutionary time. In fact, most methods aimed at detecting regions or sites of functional importance in amino acid or DNA sequences are based on detecting regions of low variability. However, very high levels of variability also signify functional importance. For example, many viruses experience positive diversifying selection in their coat proteins to avoid immune recognition[1]. The regions that have been targeted by selection are hypervariable, having an excess of amino acid altering substitutions (nonsynonymous substitutions) compared to what would be expected if all substitutions at the DNA level occur at the same rate. The evidence for selection is in the d_N/d_S ratio. The d_N/d_S ratio is the ratio of the rate of nonsynonymous substitutions per nonsynoymous sites (d_N) to the rate of synonymous substitutions (non-amino acid altering) per synonymous site (d_S). If no Darwinian selection is acting on the DNA sequences, we would expect $d_N = d_S$. If there is negative selection (selection against new amino acid variants) $d_N < d_S$, and if there is positive selection (selection for new variants) $d_N > d_S$. The d_N/d_S

ratio is a proxy for the strength of selection and can, therefore, be used to search for regions of functional importance. For example, in some viruses the amino acid sites that are important for interactions between the virus and the host can be identified by finding the sites undergoing positive selection.

1.2 The maximum likelihood method

Recently, several new methods have been developed for detecting positively selected amino acid sites. The methods of Nielsen and Yang[2] and Yang *et al.*[3] are based on modeling the evolution of a nucleotide sequence as a continuous time Markov chain with state space on the set of possible codons. In these models the d_N/d_S ratio is a parameter and it can be estimated using maximum likelihood (i.e., by finding the value for the d_N/d_S ratio that maximizes the probability of observing the data). The instantaneous rate of change from codon i to codon j in site k is given by

$$q_{ij}^{(h)} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \omega_h\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \omega_h\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases} \quad (1)$$

where ω_h is the value of $\omega = d_N/d_S$ in the hth site, κ is the transition/transversion rate ratio, and π_j is the stationary frequency of codon j. The value of ω_h at a site is considered to be unknown, but drawn from some parametric distribution. Parameters of this distribution can be estimated using maximum likelihood. The details of the calculations can be found in Felsenstein[4] and Nielsen and Yang[2]. In brief, the likelihood function is calculated by superimposing the substitution process along the branches of the phylogenetic tree. The pruning algorithm of Felsenstein[4] can then be used to calculate the likelihood function for parameters in the ω distribution and other parameters such as the branch lengths of the tree. Because of the use of a phylogenetic tree, the method is applicable to multiple aligned sequences. However, at this time it can only be applied to one, or a few, trees because of computational limitations.

Formulating the problem of detecting positive selection in an explicitly statistical framework has a number of advantages. An important strength of the likelihood-based methods is the ease with which hypotheses can be tested. Yang *et al.*[3] showed how the method can be used to test if there is evidence for positive

selection in a particular data set. For example, in one of the models (M7) in Yang *et al.*[3] it was assumed that ω follows a beta distribution. A beta distributed random variable is defined in the interval between 0 and 1; no positive selection is allowed under this model. A slightly more complicated model can be made by assuming that ω for a site is either drawn from a beta distribution (as before) or is under positive selection ($\omega > 1$). The maximum likelihood values obtained under the more general model and the model assuming a beta distribution can be compared, and the hypothesis of no sites undergoing positive selection can be tested using a likelihood ratio test. If the beta distribution is rejected in favor of a model allowing positively selected sites, it is also possible to predict which sites have values of $\omega > 1$ using an empirical Bayes method. The positively selected sites can be identified by calculating the posterior probability that a particular site has a value of $\omega > 1$ under the parameter estimates obtained for the model allowing for positive selection.

1.3 Mapping mutations on phylogenies

In this manuscript we will explore an alternative approach. This approach is based on explicitly mapping mutations on a phylogeny. In Nielsen[5] an approach was described in which inferences regarding molecular evolution can be made using the posterior distribution of mappings of mutations. Let D be the data, in our case a set of aligned nucleotide sequences, and let M be a possible mapping of mutations on the phylogeny. Figure 1 shows an example of some observations (our data, D) at the tips of a tree with one possible realization of mutations (a mapping, M) that could have led to the observations at the tips of the tree.

We are typically interested in evaluating some function of the mapping of mutations. For example, we could be interested in the number of nonsynonymous mutations in a particular amino acid (codon) site or the distribution of such mutations along the sequence or along the phylogeny. If we knew the correct mapping of mutations we could easily evaluate this function. The problem is that we do not know which mutational mapping is correct. Performing statistical analyses based on a single mapping, for example a parsimony mapping, might lead to serious biases[5]. Instead a more appropriate approach is to sum a statistic over all possible mappings, weighting by the posterior probability of each.

Let $h(M)$ be the value of the function we are interested in (e.g. the number of nonsynonymous mutations in a site). We assume that this function cannot be calculated directly from the data using some simple method, but that $h(M)$ easily can be evaluated if M is known. We are then interested in evaluating

$$h(D) := E\{h(M) \mid D\} = \sum_{M \in \Psi} h(M) \Pr(M \mid D). \qquad (2)$$

Here Ψ is the set of possible mappings and Pr(M | D) is the posterior probability of a mapping. In a Bayesian framework we can evaluate Pr(M | D) as

$$\Pr(M \mid D) = \int_{\Theta \in \Omega} \Pr(M \mid D, \Theta)\, p(\Theta \mid D) d\Theta \,. \tag{3}$$

Here Θ is a vector of parameters and Ω is the set of all possible values of this vector. In our case, it will include the topology of the phylogeny, the branch lengths, and parameters of the mutational process which will be detailed later. In other words, the posterior distribution of mappings is evaluated by integrating over the posterior density of Θ, $p(\Theta \mid D)$. This distribution can be specified under a particular model of sequence evolution and using appropriate prior distributions for the parameters. In practice, it is necessary to use Markov chain Monte Carlo (MCMC), as in Larget and Simon[6] or Huelsenbeck *et al*[7]., to evaluate $p(\Theta \mid D)$. In these methods, a Markov chain with state space on the possible values of Θ and stationary distribution $p(\Theta \mid D)$ is simulated using the Metropolis-Hastings algorithm[8,9]. By sampling from this chain at stationarity, (correlated) samples of Θ from $p(\Theta \mid D)$ can be obtained.

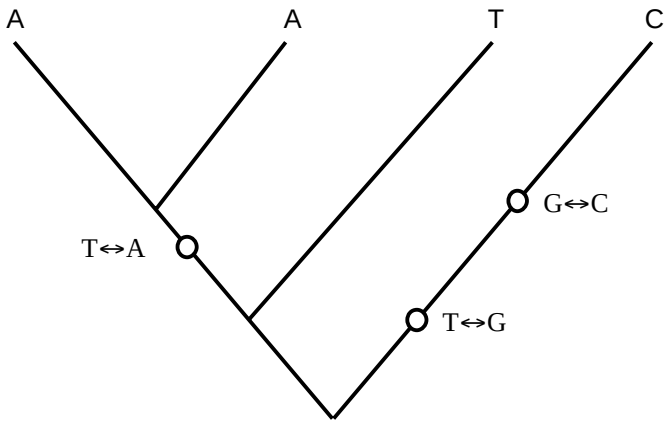

Figure 1. A Mutational mapping on a phylogeny for a single nucleotide site. Circles indicate mutations. For each mutation, the type of mutation (e.g. T↔A), the edge of the tree on which it occurred and, possibly, the time at which it occurred, is recorded in the computer memory. A mutational mapping (M) consists of these recordings for all sites in a data set, and all edges in the tree.

In Nielsen[5] an algorithm for sampling a random mapping of mutations from the distribution $\Pr(M \mid \Theta, D)$ was described. Using this algorithm it is possible to obtain samples of M from $\Pr(M \mid D)$ and thereby stochastically evaluate $h(D)$. First, k values of $\Theta, \Theta_1, \Theta_2,...,\Theta_k$, are sampled from $p(\Theta \mid D)$ using the MCMC method. In our particular applications this will be done using the program MrBayes[10]. Then, for each of these values of Θ, a mapping of mutations, M_1, $M_2,\ldots M_k$, is simulated from $\Pr(M \mid \Theta_i, D)$. By a law of large numbers for Markov chains,

$$\frac{1}{k}\sum_{i=1}^{k} h(M_i \mid D) \rightarrow h(D) \tag{4}$$

as $k \rightarrow \infty$. In Nielsen[5] this method was found to be quite computationally efficient. It provides a practical, and statistically well-justified, approach for examining patterns of molecular evolution.

1.4 Posterior predictive distributions

The idea of using posterior predictive distributions for making statistical inferences is well established in Bayesian statistics[11]. The posterior predictive distribution of a statistic is the distribution of a future (predicted) value of the statistic given the observed data. More formally, if D^{rep} denotes a replication of D, the posterior predictive distribution of a statistic, $T(\cdot)$, is given by

$$p(T(D^{rep}) \mid D) = \int_{\Theta \in \Omega} p(T(D^{rep}) \mid \Theta)\, p(\Theta \mid D) d\Theta \,. \tag{5}$$

Likewise, we can define a posterior predictive p-value[11,12] as

$$p_T = \Pr\left\{T(D^{rep}) \geq T(D) \mid D\right\} \tag{6}$$

This probability is evaluated with respect to the probability distribution given by Equation 5. A posterior predictive p-value is a hybrid between Bayesian and frequentist concepts. It involves a p-value, which is a frequentist construction that traditionally is justified by its properties in repeated sampling; however, integration over a posterior distribution of parameters is used to deal with the nuisance parameter problem. Its use can be justified both in a frequentist and a Bayesian setting. Meng[12] showed that the probability of a type I frequentist error of an α-level posterior predictive test is often close to but less than α and will never exceed

2α. Rubin[11] has argued that using posterior predictive *p*-values is Bayesian justifiable and also Bayesian relevant because of its use in model diagnosis.

A final point worth noticing in the current context, is that to perform hypothesis tests based on posterior predictive *p*-values, it is only necessary to specify the model under the null-hypothesis. We do not need to explicitly specify an alternative model.

2. Simulations

2.1 Models and data

In the following, we will apply the ideas described above to the identification of positively selected sites in a data set containing 28 sequences of the hemagglutinin protein from the Influenza virus. This data set was previously analyzed in Yang *et al.*[3] where it was shown, using the likelihood methods, that this protein is in fact subject to positive selection. We will use this data set for illustrating the new method to facilitate easy comparison with the likelihood method.

We are interested in testing the hypothesis of $\omega = 1$ against a one sided alternative of $\omega > 1$. Our null model is therefore specified by a codon substitution model in which $\omega = 1$. We notice that such a model is identical to a nucleotide substitution model with state space on the four nucleotides, with the exception of the presence of stop codons and codon usage bias. We will, therefore, for computational reasons, use a nucleotide model to closely approximate the codon substitution model. This will also allow us to use a more complex mutational model than the model assumed in Equation 1. In particular we will use the Generalized Time Reversible model[13] (GTR) to model the mutation process in the DNA sequence. We will use an uninformative prior for the base frequencies: a flat Dirichlet distribution. We will also assume uniform priors for the rest of the parameters: the other remaining parameters of the mutational model, the tree topology, and the branch lengths of the tree. To assure that the resulting posterior distribution is proper, a maximum branchlength of 100 (expected substitutions per nucleotide site) is assumed. We notice that the use of uniform priors ensures that our results are minimally influenced by the choice of priors.

2.2 Estimating the number of nonsynonymous mutations in a site

Using the computer program MrBayes[10] we ran a Markov chain for 1,100,000 cycles. After the first 100,000 cycles, we sampled a value of Θ at every 1000th cycle, eventually obtaining a total of 1000 samples, Θ_1, Θ_2, Θ,…,Θ_{1000}. These samples are valid, albeit correlated, draws from the posterior probability

distribution of Θ. For each of these samples of Θ, we sample a mutational mapping from Pr(M | Θ, D). The resulting set of mutational mappings, M_1, M_2,…,M_{1000}, are then distributed as Pr(M | D). The samples will all be correlated because they are sampled from the same Markov chain, but only weakly so, because of the sampling interval of 1000 cycles.

For each of the mappings, we calculate the posterior expectation of the number of nonsynonymous (amino acid altering) mutations in all sites. For site j

$$T(D_j) := E(n_j \mid D) \approx \frac{1}{1000}\sum_{i=1}^{1000} n_j(M_{ij}, D_j), \tag{7}$$

where n_j is the number of nonsynonymous substitutions in site j, and $n_j(M_{ij}, D_j)$ is the number of nonsynonymous mutations that occurred in site j in mutational mapping i. $T(D_j)$ is the statistic we will use to evaluate the hypothesis $\omega = 1$ in site j. Since we are only interested in detecting positive selection, we will make a one sided test, that rejects when

$$p_{T_j} = \Pr\left\{T(D_j^{rep}) \geq T(D_j) \mid D\right\} \leq \alpha. \tag{8}$$

To evaluate this probability, we need to know the distribution of $T(D_j^{rep})$. Notice that $T(D_j^{rep})$ is identically distributed for all j, because of the independence assumption of the substitution process under the null hypothesis. We simply evaluate the predictive distribution for one site to obtain the distribution for all sites.

2.3 Estimating the posterior predictive distribution

To evaluate the posterior predictive distribution we simulate 10 new sites for each of the previously simulated values of Θ from the distribution Pr(D_j | Θ), for a total of 10,000 new simulated sites, $D_1^{rep}, D_2^{rep}, ..., D_{10,000}^{rep}$. For each of these 10,000 posterior predictive site patterns, we evaluate the posterior predictive expectation of the number of nonsynonymous substitutions. For site pattern D_j^{rep}, we sample a mutational mapping for each of the 1000 values of Θ, to construct a new simulated set of mappings, sampled from the distribution

$$\Pr_{\Theta|D}(M_j^{rep} \mid D_j^{rep}) = \int_{\Theta \in \Omega} \Pr(M_j^{rep} \mid D_j^{rep}, \Theta)\, p(\Theta \mid D) d\Theta, \tag{9}$$

the subscript $\Theta \mid D$ indicates that this probability is evaluated over the posterior distribution of Θ given D. By applying Equation 7 on the set of simulated mappings for a particular predictive site pattern, and repeating this for all of the predictive site patterns, we can construct a sample of 10,000 posterior predictive values of $T(D_j)$. These values approximate the posterior predictive distribution of the test statistic (Equation 7) and the posterior predictive p-value (Equation 8) can be evaluated for all sites.

3. Results

3.1 The number of nonsynonymous mutations along the sequence

The observed distribution of the test statistic (the expected number of nonsynonymous mutations in a site given the site pattern), and its posterior predictive distribution is shown in Figure 2 for the Influenza data set.

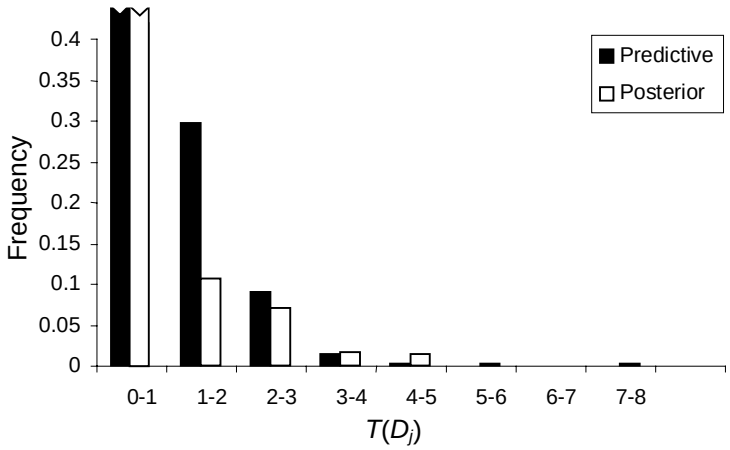

Figure 2. The predictive and the posterior distribution of the number of nonsynonymous substitutions among sites for the Influenza hemagglutinin data set.

The posterior predictive distribution has relatively fewer sites with less than one expected nonsynonymous mutation (60% predicted versus 78% observed). This is not surprising since constraints at the amino acid level will tend to lower the rate of amino acid substitution in many sites. However, we notice from figure 2 that there are also relatively more observed sites with more than 3 amino acid

substitutions than expected from the posterior predictive distribution. The posterior predictive *p*-value in a one-sided test of the hypothesis $\omega = 1$, is shown in Figure 3.

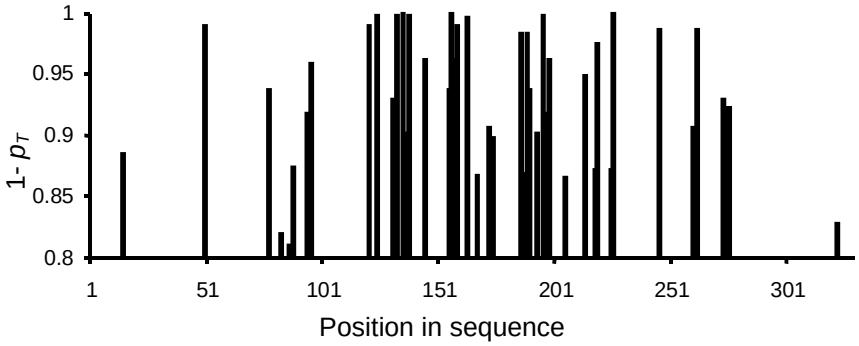

Figure 3. One minus the posterior predictive *p*-value for the hypothesis $\omega = 1$ in a one-sided test using the expected number of nonsynonymous substitutions given the data, as a test statistic.

We see that there are fairly many sites with *p*-values close to zero. There are 11 sites with posterior predictive *p*-values < 0.01. Given the number of sites, we would expect approximately 3 such sites if the null model were correct. In Yang *et al*[3]. the proportion of sites undergoing positive selection was estimated to 0.013, or approximately 4 sites (Model M8[3]).

For comparison, the posterior probability that a site is undergoing positive selection according to model M3 of Yang *et al.*[3] is shown in Figure 4. Notice the very strong correlation between this probability and $1 - p_T$.

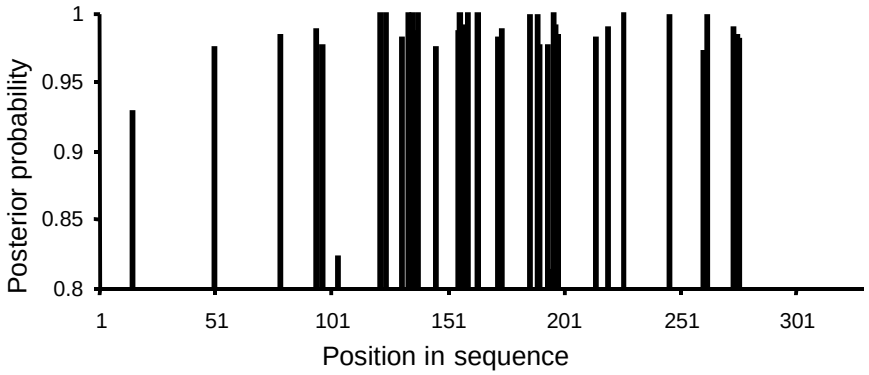

Figure 4. The posterior probability that a site is positively selected according to model M3 of Yang et al (2000).

For example, among the 11 sites with posterior predictive *p*-values less than 0.01, there are no sites with a posterior probability of being positively selected less than 0.975. The two methods essentially identify the same sites as being positively selected. This is quite remarkable given the differences in model assumptions.

3.2 The number of radical amino acid substitutions along the sequence

One of the advantages of the present approach is that extensions to other problems follow quite easily. For example, it might be of some interest whether positively selected mutations are radical mutations or tend to be conservative amino acid mutations. If positively selected substitutions tend to be radical, we can use this information when trying to identify sites undergoing positive selection. We can estimate the expected number of conservative substitutions given the data for each site in the sequences, using the very same samples of $\Pr(M \mid D)$ obtained for the purpose of identifying positively selected sites. Here, we simply define a substitution to be radical if it has a PAM100 score of less than –2 and conservative if it has a PAM100 score of more than –2. Obviously, many other measures could have been used to divide substitutions into radical and conservative substitutions. We also divide sites into sites with positive selection and sites evolving neutrally or subject to negative selection. The results are shown in Figure 5

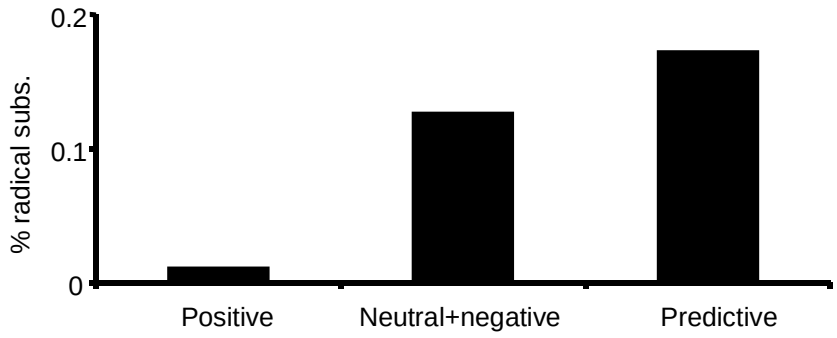

Figure 5. The proportion of radical amino acid changes in positively selected sites (Positive), sites that are evolving neutrally or subject to negative selection (Neutral) and in sites sampled from the posterior predictive distribution (Predictive).

Notice that the proportion of radical substitutions is much lower in the positively selected sites than in other sites. Moreover, the proportion of radical substitutions in positively selected sites is much less than expected from the posterior predictive distribution assuming no selection. It appears that positively selected substitutions tend to be conservative.

Discussion and conclusion

The approach based on posterior predictive *p*-values for identifying positively selected sites differs from the previous approaches[2,3] in several different ways. Most importantly, the model assumptions are quite different. In the present approach we assume prior distributions for all the nuisance parameters, including parameters related to the tree. We then base our inferences directly on the posterior distribution of mutations. In the previous approach parameters are first estimated by maximizing the likelihood function. Estimates of the posterior probabilities are then obtained based on these estimates. The cost in the new approach is an additional set of assumptions, but it does allow a proper treatment of the problem of the unknown tree topology. In the likelihood approach, maximization over tree topologies is not presently computationally feasible.

In the present application, a nucleotide model was used under the null-model to approximate a codon based model with $\omega = 1$. For some data sets it might be worrisome that this model does not take into account codon bias and the existence of stop codons. Also, the null hypothesis of $\omega = 1$ is arguably very simplistic. A more realistic null model that also allows sites in which $\omega < 1$ might be considered in future applications.

Despite the differences between the two approaches, the biological conclusions are remarkably similar in the two studies. There appears to be a small

fraction of sites in the data set undergoing positive selection, and the sites identified to be undergoing positive selection are more or less same in the two studies.

The strength of the current approach was best illustrated in the analysis of the proportion of conservative and radical substitutions. It was easily demonstrated that the positively selected amino acid substitutions tend to be conservative substitutions. This is not a trivial result. In fact, it could be hypothesized that in the sites of a coat protein that interacts with a host immune system, any substitution is favorable. In particular, very radical substitutions that would change the binding affinities of antibodies and other components of the host immune system, should be favored. However, as shown here, this does not appear to be the case, at least not in the hemagglutinin protein of the Influenza virus.

Although this question could also have been addressed using explicit modeling in the likelihood framework, this would have required the computer implementation of such a model. In the present case the analysis could be done by simply reading of the results from the simulated mutational mappings. The strength of the present approach is that it allows for this type of exploratory data analysis in a rigorous statistical framework. Earlier studies have also used a mutation-mapping approach, where the mapping is performed using the parsimony method. A parsimony-based approach, however, suffers from a number of problems; the method only considers the mapping with the minimum number of changes (thereby underestimating the total number of changes) and treats the mapping as an observation in further analysis. The Bayesian approach discussed here avoids the many statistical problems associated with using parsimony and focussing on just a single mutational mapping.

Acknowledgments

This research was supported by NSF grants DEB-0089487 awarded to RN and DEB-0075406 awarded to JPH.

References

1. E.C. Holmes, L.Q. Zhang, P. Simmonds, C. A. Ludlam, A.J. Leigh Brown, “Convergent and divergent sequence evolution in the surface envelope glycoprotien of human immunodeficiency virus type 1 within a single infected patient” *Proc. Natl. Acad. Sci.* **89**, 4835 (1992)
2. R. Nielsen. and Z. Yang, “Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene” *Genetics*, **148**, 929 (1998)

3. Z. Yang, R. Nielsen, N. Goldman, and A.-M. K. Pedersen, "Codon-Substitution Models for Variable Selection Pressure at Amino Acid Sites" *Genetics* **155**, 431 (2000)
4. J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach" *J. Mol. Evol.* **17**, 368 (1981)
5. R. Nielsen, "Mapping mutations on phylogenies", *Syst. Biol.* (In review)
6. B. Larget. and D. Simon, 1999 "Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees" *Mol. Biol. Evol.* **16**, 750 (1999)
7. J. P. Huelsenbeck,., B. Rannala, and B. Larget, "A Bayesian framework for the analysis of cospeciation" *Evolution* **54**, 353 (2000)
8. N. Metropolis, A. W. Rosenbluth., M. N. Rosenbluth, A. H. Teller, and E. Teller. "Equations of state calculations by fast computing machines" *J. Chem. Phys.* **21**, 1087 (1953)
9. W. K Hastings, "Monte Carlo sampling methods using Markov chains and their applications" *Biometrika* **57**, 97 (1970)
10. J. P. Huelsenbeck F. Ronquist, *MrBayes 2.0*. Available from http://brahms.biology.rochester.edu/software.html (2001)
11. D. B. Rubin, "Bayesianly justifiable and relevant frequency calculations for the applied statisticians" *Ann. Statist.* **12**, 1151 (1984)
12. X.-L. Meng, "Posterior predictive *p*-values" *Ann. Statist*. **22**, 1142 (1994)