Spatial Distribution of Selection Pressure on a Protein Based on the Hierarchical Bayesian Model

Teruaki Watabe*1 and Hirohisa Kishino2
1Center of Medical Information Science, Kochi University, Kochi, Japan
2Laboratory of Biometrics, Graduate School of Agriculture and Life Science, University of Tokyo, Tokyo, Japan
*Corresponding author: E-mail: twatabe-mi@umin.ac.jp.
Associate Editor: Jeffrey Thorne

Abstract

Proteins adapt to novel environments and/or gain function by substitution in amino acid sequences. Therefore, mutations in protein-coding genes are subject to selection pressure. The strength and character of selection pressure may vary among the regions of the protein. Thus, the spatial distribution of selection pressure provides information on the adaptive evolution of the protein. We developed a hierarchical Bayesian model that detects the spatial distribution of selection pressure on a protein. We expressed selection pressure by the substitution rate ratio of nonsynonymous to synonymous substitutions in the DNA sequence. The Potts model describes the prior distribution of spatial aggregation of selection pressure. The hyperparameters that define the strength and range of spatial clustering are estimated by maximizing the marginal likelihood. Because our prior distribution is un-normalized, we calculated the log marginal likelihood by “thermodynamic integration.” We applied the method to historical data on the influenza hemagglutinin protein, comparing the estimated spatial distribution of the substitution rate ratio with that of antigenic sites A–E. The amino acid residues with higher substitution rate ratios, representing diversifying selection pressure, overlapped the antigenic sites.

Key words: molecular evolution, selection pressure, spatial distribution, hierarchical Bayesian model, Potts model.

Introduction

Understanding the variation in evolutionary rates among sites is a fundamental problem in the study of molecular evolution. In protein evolution, the selection pressure on the (micro-)structural change of a protein can be identified by investigating the distribution of the substitution rate ratio \( \frac{dN}{dS} \) of nonsynonymous to synonymous substitutions (Kimura 1983; Yang et al. 2000). If the induced change in the amino acid sequence undermines the structural stability of the protein, that mutation might be subject to purifying selection. In this context, the rate of nonsynonymous substitution is lower than that of synonymous substitution; however, if a nonsynonymous mutation increases the fitness of the protein, it will have a higher probability of fixation than synonymous mutations (\( \frac{dN}{dS} > 1 \)).

For example, influenza virus uses the hemagglutinin (HA) protein to bind the host-cell receptor to initiate infection. The epitopes recognized by immunoglobulin (Ig) are often located on the receptor-binding (RB) region on the HA surface. Those epitopes overlap antigenic sites A, B, and D. Ig binding thus prevents the virus from invading a host cell. Amino acid mutations in the RB region that reduce the binding affinity of the Ig might become fixed because they may allow the virus particle to evade immune attack; however, such mutations might also reduce the binding affinity to the host-cell receptor. When the threat of the immune attack posed by a specific Ig has gone, nonsynonymous mutations in the HA gene that recover binding affinity to the host-cell receptor may become fixed (Watabe et al. 2007). Furthermore, the attachment of an oligosaccharide to an N-glycosylation site prevents antibody binding and thus changes the selection pressure on the surrounding sites (Kobayashi and Suzuki 2012). The evolution of influenza virus is the result of interactions between evolutionary and ecological processes driven by the host’s acquired immunity (Grenfell et al. 2004; Rambaut et al. 2008; Russell et al. 2008; Pybus and Rambaut 2009). Estimation of the variation in the \( \frac{dN}{dS} \) ratio by location in the protein could be used to monitor these dynamics.

In their pioneering works, Nielsen and Yang (1998) and Yang et al. (2000) considered the heterogeneity of the \( \frac{dN}{dS} \) ratio across sites. Based on an analysis of 10 proteins, they demonstrated that strong diversifying selection is often overlooked if heterogeneity is taken into account. In the absence of structural information, their prior distribution for the \( \frac{dN}{dS} \) ratio assumed independence among sites. We call this model the “independent-\( \omega \) prior model.” Due to protein folding, the spatial aggregation of the sites under diversifying selection does not indicate the aggregation of the corresponding sites in the primary sequence. Therefore, a prior assumption of correlation of \( \omega \) among neighboring sites in a primary sequences may yield false positives and negatives.

To more powerfully identify the spatial aggregation of amino acid sites under positive selection, Suzuki (2004) developed an approach based on a local likelihood method (Tibshirani and Hastie 1987). It maximizes, for each site, the probability of an oligosaccharide to an N-glycosylation site prevents antibody binding and thus changes the selection pressure on the surrounding sites (Kobayashi and Suzuki 2012). The evolution of influenza virus is the result of interactions between evolutionary and ecological processes driven by the host’s acquired immunity (Grenfell et al. 2004; Rambaut et al. 2008; Russell et al. 2008; Pybus and Rambaut 2009). Estimation of the variation in the \( \frac{dN}{dS} \) ratio by location in the protein could be used to monitor these dynamics.

In their pioneering works, Nielsen and Yang (1998) and Yang et al. (2000) considered the heterogeneity of the \( \frac{dN}{dS} \) ratio across sites. Based on an analysis of 10 proteins, they demonstrated that strong diversifying selection is often overlooked if heterogeneity is taken into account. In the absence of structural information, their prior distribution for the \( \frac{dN}{dS} \) ratio assumed independence among sites. We call this model the “independent-\( \omega \) prior model.” Due to protein folding, the spatial aggregation of the sites under diversifying selection does not indicate the aggregation of the corresponding sites in the primary sequence. Therefore, a prior assumption of correlation of \( \omega \) among neighboring sites in a primary sequences may yield false positives and negatives.

To more powerfully identify the spatial aggregation of amino acid sites under positive selection, Suzuki (2004) developed an approach based on a local likelihood method (Tibshirani and Hastie 1987). It maximizes, for each site, the probability of an oligosaccharide to an N-glycosylation site prevents antibody binding and thus changes the selection pressure on the surrounding sites (Kobayashi and Suzuki 2012). The evolution of influenza virus is the result of interactions between evolutionary and ecological processes driven by the host’s acquired immunity (Grenfell et al. 2004; Rambaut et al. 2008; Russell et al. 2008; Pybus and Rambaut 2009). Estimation of the variation in the \( \frac{dN}{dS} \) ratio by location in the protein could be used to monitor these dynamics.

In their pioneering works, Nielsen and Yang (1998) and Yang et al. (2000) considered the heterogeneity of the \( \frac{dN}{dS} \) ratio across sites. Based on an analysis of 10 proteins, they demonstrated that strong diversifying selection is often overlooked if heterogeneity is taken into account. In the absence of structural information, their prior distribution for the \( \frac{dN}{dS} \) ratio assumed independence among sites. We call this model the “independent-\( \omega \) prior model.” Due to protein folding, the spatial aggregation of the sites under diversifying selection does not indicate the aggregation of the corresponding sites in the primary sequence. Therefore, a prior assumption of correlation of \( \omega \) among neighboring sites in a primary sequences may yield false positives and negatives.

To more powerfully identify the spatial aggregation of amino acid sites under positive selection, Suzuki (2004) developed an approach based on a local likelihood method (Tibshirani and Hastie 1987). It maximizes, for each site, the probability of an oligosaccharide to an N-glycosylation site prevents antibody binding and thus changes the selection pressure on the surrounding sites (Kobayashi and Suzuki 2012). The evolution of influenza virus is the result of interactions between evolutionary and ecological processes driven by the host’s acquired immunity (Grenfell et al. 2004; Rambaut et al. 2008; Russell et al. 2008; Pybus and Rambaut 2009). Estimation of the variation in the \( \frac{dN}{dS} \) ratio by location in the protein could be used to monitor these dynamics.
site. The size of the three-dimensional window is a key to the success of the procedure, but as yet, no objective criterion for window size is available.

To achieve statistical rigor, we developed a hierarchical Bayesian model that detects the spatial distribution of $\omega$ in a protein. The prior spatial aggregation of selection pressure is expressed by the Potts model. The Potts model is a generalization of the Ising model that has a theoretical framework based on magnetic material physics. The Ising model has been used in image restoration (Inoue and Tanaka 2001) and forest gap dynamic simulations (Schlicht and Iwasa 2004). The hyperparameters that define the strength and range of the spatial clustering of $\omega$ are estimated by maximizing the marginal likelihood. We applied the method to the influenza HA protein, and using the estimated hyperparameters, we obtained the spatial distribution of the selection pressure on the influenza HA protein.

**New Approaches**

**Marginal Likelihood**

We consider the conditional probability of the sequences given the three-dimensional protein structure as follows:

$$Z = P(A^{(1)}, \ldots, A^{(N)} | X).$$

Here, $A^{(i)}$ denotes the $i$th codon sequence of length $L$, $A^{(i)} = (a_1^{(i)}, \ldots, a_L^{(i)})$, where $a_k^{(i)}$ corresponds to the $k$th codon of the $i$th sequence, $N$ is the number of sequences, $X$ denotes the positions of the $C^\alpha$ atoms of the amino acid residues, and $X = (x_1, \ldots, x_k)$, where $x_k$ is the spatial coordinates of the $k$th residue's $C^\alpha$ atom and $x_k = (x_k^1, x_k^2, x_k^3)$. In this study, we intend to reveal the spatial distribution of the selection pressure on the protein. We expressed the selection pressure using the $d_{\text{H}}/d_5$ ratio ($\omega = dN/dS$) as follows:

$$Z = \int P(A^{(1)}, \ldots, A^{(N)} | \omega, X) P(\omega | X) d\omega.$$

The $d_{\text{H}}/d_5$ ratio of each codon site is embedded in the vector $\omega = (\omega_1, \ldots, \omega_k)$.

In this study, we assumed that the evolution of a codon sequence is influenced by the three-dimensional protein structure through the $d_{\text{H}}/d_5$ ratios only. Hence, the likelihood of $\omega$, given the structure $X$, is expressed by the product of contributions from individual residues as follows:

$$P(A^{(1)}, \ldots, A^{(N)} | \omega, X) = \prod_{k=1}^L P(a_k | \omega_k),$$

where $a_k$ contains the $k$th elements of the $N$ sequences and $a_k = (a_1^{(k)}, \ldots, a_L^{(k)})$.

To consider the prior distribution of the $d_{\text{H}}/d_5$ ratio, $P(\omega | X)$, we discretized the $\omega$-value into $K$ states:

$$P(\omega | X) = \sum_s \int P(\omega | s, \hat{\omega}, X) P(s, \hat{\omega} | X) d\hat{\omega}.$$  

Here, $s$ is a vector consisting of states at the codon sites, $s = (s_1, \ldots, s_L)$, and $s_k$ varies from 1 to $K$. $\hat{\omega}$ is a vector consisting of discretized $\omega$-values linked to the state, $\hat{\omega} = (\hat{\omega}_1, \ldots, \hat{\omega}_k)$. The summation of the index $s$ includes all configurations of the states at codon sites. The value of $\omega$ at the $k$th codon site is restricted to $\hat{\omega}_k$:

$$P(\omega | s, \hat{\omega}, X) = \prod_{k=1}^L \delta(\omega_k - \hat{\omega}_k).$$

Here, the delta notation expresses the Dirac delta function that satisfies $\int_{-\infty}^{\infty} \delta(x) dx = 1$ and $\int_{-\infty}^{\infty} f(x) \delta(x - y) dx = f(y)$. The marginal likelihood (eq. 1) is now rewritten as follows:

$$Z = \sum_s \int \left( \prod_{k=1}^L P(a_k | \hat{\omega}_k) \right) P(s | X) d\hat{\omega}.$$

**Expectation Value of $\omega_k$**

The marginal likelihood (eq. 2) can be interpreted as a partition function of the thermodynamic system that may occupy a specific configuration $s$ with specific values of $\hat{\omega}$. By using the partition function, we can compute the expectation (expected) value of $\omega_k$:

$$\langle \omega_k \rangle = \frac{1}{Z} \sum_s \int \hat{\omega}_k \left( \prod_{k=1}^L P(a_k | \hat{\omega}_k) \right) P(s | X) d\hat{\omega}.$$

**Thermodynamic Integration**

When the prior distribution of the states, $P(s | X)$, is unnormalized, the marginal likelihood should be divided by the distribution summed over all configurations of $s$ with all possible values of $\hat{\omega}$:

$$Z = \sum_s \int \left( \prod_{k=1}^L P(a_k | \hat{\omega}_k) \right) P(s | X) d\hat{\omega}.$$  

Here the factor $V$ represents the volume of $\hat{\omega}$ space. In general, however, summing all configurations of $s$ is unrealistically expensive. Ogata (1989) solved this problem in a general framework by using the Monte Carlo method. This method has broad applications and has recently been called "thermodynamic integration" (Gelman and Meng 1998; Larillot and Philippe 2006).

We calculated the logarithm of the marginal likelihood (eq. 4) and introduced the parameter $\beta$ to express the difference between the two logarithms on the right-hand side of the equation by integration over $\beta$:

$$\ln Z = \int_0^1 \frac{d}{d\beta} \left\{ \ln \sum_s \int \left( \prod_{k=1}^L P(a_k | \hat{\omega}_k) \right) P(s | X) d\hat{\omega} \right\} d\beta.$$

By differentiating the integrand with respect to $\beta$, we obtained a formula that can be calculated without summing all configurations of $s$ as follows (see supplementary material, Supplementary Material online):

$$\ln Z = \int_0^1 E_\beta \left( \sum_{k=1}^L \ln P(a_k | \hat{\omega}_k) \right) d\beta.$$
The thermodynamic integration requires us to evaluate the expectation value of the log-likelihood of \( \omega \) in the system characterized by the introduced parameter \( \beta \), rather than summing all configurations of \( s \). Here, the formula \( E_\beta[\cdot|\cdot] \) stands for the posterior mean with the quasi-likelihood:

\[
d_\beta(s, \hat{\omega}) \propto \left\{ \prod_{k=1}^{l} P(a_k | \hat{\omega}_k) \right\}^\beta P(s | X).
\]

We employed \( d_\beta(s, \hat{\omega}) \) as a Gibbs sampler. To evaluate the expectation, \( E_\beta[\cdot|\cdot] \), we employed the Gibbs sampling method. To simplify the calculations, we substituted the specific values of \( \hat{\omega}_k \) at the \( K \) states \((s = 1, \ldots, K)\) that maximized the likelihood, \( \prod_{k=1}^{l} P(a_k | \hat{\omega}_k) \), at each sampling step. Using the formula \( E_\beta[\cdot|\cdot] \), the expectation value of \( \omega_k \) (eq. 3) is described by the formula \( \langle \omega_k \rangle = E_{\beta=1}[\hat{\omega}_k] \).

### The Potts Model

The prior distribution of the states is modeled by the Potts model with \( K \) categorizations, given by

\[
P(s | X) = \exp\left\{ \lambda \sum_{i>j} Q_{ij}\exp(-\alpha r_{ij}) \right\},
\]

where hyperparameters \( \lambda \) and \( \alpha \) are introduced. The three-dimensional protein structure, \( X \), is represented by the distance between the \( C^\alpha \) atoms of an amino acid residue pair, \( r_{ij} \). \( Q \) is a \( K \)-by-\( K \) matrix, of which the diagonal elements are positive, \(-1/K\), and the off-diagonal elements are negative, \(-1/K\). A pair of amino acid residues in the same state contributes the positive component of the logarithm of the prior distribution, whereas a pair of residues in different states contributes the negative component. The distance dependence in equation (6) guarantees that an amino acid residue pair with a short distance between them is more highly correlated than amino acid residues with a longer distance between them.

### Update of \( \hat{\omega}_s \) Values

The values of \( \hat{\omega}_s \) at the \( K \) states \((s = 1, \ldots, K)\) that maximized the likelihood, \( \prod_{k=1}^{l} P(a_k | \hat{\omega}_k) \), were stored at each sampling step. A sampling step consisted of updates of \( s_k \) at all codon sites. The values of \( \hat{\omega}_s \) were updated at each sampling step with the average of the past 1,000 steps.

### Results

#### Estimation of the Hyperparameters

By using thermodynamic integration, we maximized the marginal likelihood and determined the hyperparameters \((\lambda, \alpha)\). We carried out the thermodynamic integration by the numerical approach. We divided the integration range \((0 to 1)\) of \( \beta \) into 40 intervals and performed Gibbs sampling at each point of division, \( \beta_i \) \((i = 0, \ldots, 40)\). In figure 1A, we show the \( \beta \) dependency of \( E_\beta[\ln P] \) (the integrand of eq. 5) in the case of \( K = 3 \) and \( \lambda = 3.0 \) for various \( \alpha \). The \( \beta \) dependency of \( E_\beta[\ln P] \) was very smooth, and these results gave the \( \alpha \) dependency of \( \ln Z \). In the case of \( K = 3 \) and \( \lambda = 3.0 \), the maximum was obtained around \( \alpha = 0.475 \) (ln \( Z = -5721.44 \)). In figure 2, the \( \lambda \) dependency of the maximum value in the range of \( \lambda = 10^{-1} - 10^4 \) is shown. By setting the interval of \( \Delta \lambda = 1.0 \) in the range of \( \lambda = 1.0 - 10.0 \), we found the maximum marginal likelihood at \( \lambda = 3.0 \). We named the parameter sets \( K = 3 \) with \((\lambda, \alpha) = (3.0, 0.475)\) model \( M^{(K)} \). We also estimated the hyperparameters in the case of \( K = 10 \) and obtained \((\lambda, \alpha) = (4.0, 0.450)\) giving the maximum marginal likelihood of \( Z = -5713.51 \). The confidence intervals of the estimated hyperparameters are given in the supplementary material, Supplementary Material online. The likelihood ratio test (LRT) suggests that model \( M^{(K)} \) is slightly better than model \( M^{(K)} \): \( 2\Delta \ln Z = 15.87 \). The \( P \) value for this comparison is 0.026, compared with the \( \chi^2 \) distribution with seven degrees of freedom (df) (table 1).

#### Distribution of the \( d_{ni}/d_S \) Ratio

In figure 3, we show the scatter plot of the expectation values \( \langle \omega_k \rangle \) estimated in \( M^{(K)} \) and in \( M^{(K)} \). The features of the obtained distributions in the two models were very similar. In figure 4, we show the spatial distribution of the \( d_{ni}/d_S \) ratio. By comparing the distribution of the \( d_{ni}/d_S \) ratio with the antigenic sites A–E (see supplementary fig. S1, Supplementary Material online), we found that the residues with higher \( d_{ni}/d_S \) ratio \((\omega > 1.0)\) overlapped the antigenic sites. Among the antigenic sites, the proportion of the residues with \( \omega > 1.0 \) in model \( M^{(K)} \) was 14.5% (19/131), whereas it was 7.6% (21/275) in the region overall (sites 41–315 of the HA1 domain). In figure 5, the RB sites were enhanced. The antigenic sites A, B, and D overlapped the RB region (fig. 5B). The RB sites were identified in the three-dimensional protein structures of HA in a complex with cell-receptor analogs (Weis et al. 1988). In the RB region, the proportion of residues with \( \omega > 1.0 \) was 40.0% (6/15), higher than in the region overall.

#### Testing Homogeneity and Independence of \( \omega \)

As the limit of hyperparameter \( \alpha \) approaches zero, the contribution of a pair of amino acid residues to the prior...
distribution is independent of the distance between the residues. Thus, a residue in a different state than other residues with respect to $\omega$ significantly decreases the value of the prior distribution. Thus, most of the residues tend to be in the same state at the limit, and the obtained distribution of the $d_N/d_S$ ratio is the same as that resulting from an analysis in which all sites are under the same selection pressure (model $M_{\omega \rightarrow 0}$).

The marginal likelihood at the limit was found to be $\ln Z = -5868.62$ when $K = 3$ and $\lambda = 3.0$ (fig. 1B).

As the limit of hyperparameter $\alpha$ approaches infinity, the prior distribution does not affect the residues’ states. Thus, the state of each residue is decided independently. At the limit, the independence of $\omega$ among codon sites is realized (model $M_{\omega \rightarrow \infty}$). The marginal likelihood in this limit was determined to be $\ln Z = -5742.44$ in the case of $K = 3$ and $\lambda = 3.0$ (fig. 1B).

With $K = 3$ and $\lambda = 3.0$, we compared these two models with $M^{(\omega)}$. The Bayes factors were evaluated as follows:

$$K = \frac{Z_{M^{(\omega)}}}{Z_{M_{\omega \rightarrow 0}}} = 8.3 \times 10^{63},$$

and

$$K = \frac{Z_{M^{(\omega)}}}{Z_{M_{\omega \rightarrow \infty}}} = 1.3 \times 10^9.$$ 

This finding suggests strong support for the model in which the $d_N/d_S$ ratio varies among the residues and that the $d_N/d_S$ ratio at each residue correlates with the ratios at other residues, depending on the distance. In table 1 we summarized model comparison. Furthermore, comparing the maximum value at $\lambda = 3.0$ with the value at the other $\lambda$ in the range of $\lambda = 10^{-1} - 10^3$, the Bayes factor ($K = 1.9 \times 10^4$ at $\lambda = 10^{-1}$) strongly supported the model at $\lambda = 3.0$.

Comparison to Independent-$\omega$ Prior Models

We also analyzed the data (206 sequences in the HA1 domain, each consisting of 328 codons) using independent-$\omega$ prior models implemented in PAML (Yang 2007). We employed five models: 1) model M3-3, which uses an unconstrained discrete distribution for $\omega$ among sites in three categories; 2) model M3-10, which uses 10 categories; 3) model M5, which assumes a gamma distribution; 4) model M7, which assumes a beta distribution; and 5) model M8, which adds one class of sites to model M7. Model $M_{\omega \rightarrow \infty}$ in our framework corresponds to model M3, with the prior probabilities set to $1/K$ for all $K$ states. In supplementary table S1, Supplementary Material online, we show the maximum likelihood and estimated parameters for each model. Model M7 did not fit the data. The LRT statistics for comparing the models are shown in supplementary table S2, Supplementary Material online. Models M3-3 and M8 were slightly better than M5. Model M3-10 did not show any significant advantage. The maximum likelihood in supplementary table S1,
Table 1. The Maximum Marginal Likelihood and Model Comparison Are Summarized.

<table>
<thead>
<tr>
<th>Model</th>
<th>$K$</th>
<th>$\ln Z$</th>
<th>$\lambda$</th>
<th>$\alpha$</th>
<th>Comparison among Models</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^{(3)}$</td>
<td>3</td>
<td>$-5721.44$</td>
<td>3.0</td>
<td>0.475</td>
<td></td>
</tr>
<tr>
<td>$M^{(10)}$</td>
<td>10</td>
<td>$-5713.51$</td>
<td>4.0</td>
<td>0.450</td>
<td>$2(\ln Z_{M^{(10)}} - \ln Z_{M^{(3)}}) = 15.87$</td>
</tr>
<tr>
<td>$M_{s=0}$</td>
<td>3</td>
<td>$-5868.62$</td>
<td>3.0</td>
<td>0</td>
<td>$Z_{M^{(10)}/M_{s=0}} = 8.3 \times 10^{63}$.</td>
</tr>
<tr>
<td>$M_{s=\infty}$</td>
<td>3</td>
<td>$-5742.44$</td>
<td>3.0</td>
<td>$\infty$</td>
<td>$Z_{M^{(10)}/M_{s=\infty}} = 1.3 \times 10^{5}$.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. The Average Values of $\omega_k$ Estimated During the Sampling of 80,000 Steps After a Burn-in of 20,000 Initial Steps.

<table>
<thead>
<tr>
<th>Model</th>
<th>$s = 1$</th>
<th>$s = 2$</th>
<th>$s = 3$</th>
<th>$s = 4$</th>
<th>$s = 5$</th>
<th>$s = 6$</th>
<th>$s = 7$</th>
<th>$s = 8$</th>
<th>$s = 9$</th>
<th>$s = 10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M^{(3)}$</td>
<td>$\hat{\omega_k}$</td>
<td>0.120</td>
<td>0.766</td>
<td>1.648</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>$f_s$</td>
<td>0.714</td>
<td>0.183</td>
<td>0.103</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>$M^{(10)}$</td>
<td>$\hat{\omega_k}$</td>
<td>0.080</td>
<td>0.245</td>
<td>0.370</td>
<td>0.477</td>
<td>0.577</td>
<td>0.678</td>
<td>0.784</td>
<td>0.907</td>
<td>1.091</td>
</tr>
<tr>
<td></td>
<td>$f_s$</td>
<td>0.588</td>
<td>0.045</td>
<td>0.044</td>
<td>0.045</td>
<td>0.046</td>
<td>0.044</td>
<td>0.048</td>
<td>0.050</td>
<td>0.054</td>
</tr>
</tbody>
</table>

Note.—The posterior probabilities $f_s$ of state occupation by codon sites are also presented.

Supplementary Material online, cannot be compared with $\ln Z$ in table 1 because we analyzed the full-length sequences in PAML, while we trimmed the edge of the HA1 domain to avoid effects of possible structural ambiguity. Furthermore, simply by numerical reason, our program produces the log likelihood of a codon In 61 larger than that of PAML.

In figure 6, we compare the distributions of $\langle \omega_k \rangle$ in our models with that in M5. In supplementary figure S2C, Supplementary Material online, we show the spatial distribution of $\langle \omega_k \rangle$ obtained in M5. Although model M5 employed the prior distribution with independence of $\omega$ among sites, it gave a distribution of $\langle \omega_k \rangle$ similar to that seen in $M^{(3)}$. The LRT suggested that M3-3 is slightly better than M5. In figure 7A, we show the scatterplot of the expectation values $\langle \omega_k \rangle$ in $M^{(3)}$ and M3-3. Both the models discretized the $\omega$-value into three states; however, the obtained distributions in $M^{(3)}$ and M3-3 were very different in the region of higher $\langle \omega_k \rangle$. In supplementary figure S2A, Supplementary Material online, we show the spatial distribution of the $d_{\alpha}/d_{\gamma}$ ratio obtained in M3-3. The distribution in M3-3 looks very different from that in $M^{(3)}$. Most likely, the prior distribution led to the difference between the results of the two models. In M3-3, the prior probabilities of the three states were parameterized and applied to all codon sites (supplementary fig. S3, Supplementary Material online). The codon sites in the independent-$\omega$ prior models independently decide their states in accordance with the prior probability. In $M^{(3)}$, the prior probability depended on the codon site and on the configuration of the states of the other sites. In figure 8, we show the prior probabilities in $M^{(3)}$. We averaged the prior probabilities sampled during 80,000 steps after a 20,000-step burn-in period. The prior probabilities in figure 8 were normalized at each codon site. The prior probabilities show considerable site dependency.

In figure 7B, we compare the distribution of $\langle \omega_k \rangle$ in $M^{(10)}$ with that in M3-10. Supplementary figure S2B, Supplementary Material online, shows the spatial distribution of the $d_{\alpha}/d_{\gamma}$ ratio in M3-10, with the same differences observed in the comparison of $M^{(3)}$ and M3-3. In supplementary figure S4, Supplementary Material online, we compare the distributions of $\langle \omega_k \rangle$ in our models with that in M8. In supplementary figure S2D, Supplementary Material online, we show the spatial distribution of $\langle \omega_k \rangle$ obtained in M8. The distribution in M8 was different from that found in our models, whereas it was similar to that found in M3-3.

Discussion

We inferred the spatial distribution of selection pressure on the influenza HA protein. We compared our models ($M^{(3)}$ and $M^{(10)}$) with the independent-$\omega$ prior models (M3-3, M3-10, M5, and M8). Model M5 reproduced results similar to those in $M^{(3)}$. In our framework, the LRT suggests that model $M^{(10)}$ is slightly better than $M^{(3)}$. In the independent-$\omega$ prior framework, models M3-3 and M8 appeared slightly better than M5. In both frameworks, better models
reproduced the $\omega$ distributions with higher $\langle \omega_k \rangle$ in the region $\omega > 1.5$ (fig. 3 and supplementary fig. S5, Supplementary Material online); however, in the independent-$\omega$ prior models, the codon sites with the highest $\omega$ values arose suddenly, likely because of the independence of $\omega$ among sites.

The higher $d_{HN}/d_S$ ratio, representing diversifying selection pressure, was preferentially distributed in the antigenic sites. Furthermore, 40% of the RB sites consisted of sites with higher $d_{HN}/d_S$ ratios. The RB region may lose its ability to bind to the host-cell receptor if most of the constituent sites are subject to diversifying selection pressure. In fact, a reduction in the avidity of H3N2 viruses for the human receptor analog was observed between 2001 and 2004 (Lin et al. 2012). To understand how the virus lost its ability to bind to the host-cell receptor, we must consider the phylogenetic tree to infer when changes in binding ability arose. If the shift in binding ability occurs at a deeper node, it may be more difficult for the virus to recover the lost binding ability. The temporal distribution of the selection pressure on the HA protein may also be inferred from evolutionary data. Early in the virus’s

---

**Fig. 4.** The spatial distributions of selection pressure (A) in model $M^{(K1)}$ and (B) in model $M^{(K10)}$. The colors indicate the value of the $d_{HN}/d_S$ ratio, as shown in the legend.

**Fig. 5.** (A) The spatial distribution of selection pressure in model $M^{(K10)}$. (B) The distributions of the antigenic sites A–E. In both figures, the RB sites were enhanced.

---

2719

Spatial Distribution of Selection Pressure · doc.10.1093/molbev/mst151

doi:10.1093/molbev/mst151

MBE
adaptation to human hosts, selection pressure on the viral protein can lead to rapid evolution (Nelson and Holmes 2007). Subsequently, although the genotype of the HA protein evolves gradually, phenotypic (antigenic) changes are punctuated (Smith et al. 2004). The phylogenetic tree of the HA protein shows restricted genetic diversity and consists of concatenated antigenic clusters. Within an antigenic cluster, the sequences of the HA differ by a substantial number of amino acids and form a neutral network (Koelle et al. 2006). This neutral network brings the antigenic similarity of HA...
proteins within a cluster. Studies of how selection pressure on the HA proteins in the antigenic clusters changes over time may give valuable insight into the evolution of influenza. In this study, we inferred the spatial distribution of selection pressure by using an established procedure that can be improved to infer the spatiotemporal distribution of selection pressure. We leave these challenges to future works.

Materials and Methods

Codon Model for the Sequence Evolution

We allowed residues to have their own $d_{ij}$/d$_{i}$ ratios as described by $\omega_k$, for the $k$th residue. Thus, the instantaneous rate of substitution at the $k$th residue is given by the following formula (Nielsen and Yang 1998):

$$ q_k^{(j)} = \begin{cases} 0 & \text{for more than one nucleotide} \\ u\pi^{(i)} & \text{for a synonymous transversion} \\ u\omega_k\pi^{(i)} & \text{for a synonymous transition} \\ u\omega_k\pi^{(i)} & \text{for a nonsynonymous transversion} \\ u\omega_k\pi^{(i)} & \text{for a nonsynonymous transition} \end{cases}$$

Here, the instantaneous rate of substitution from codon $i$ to codon $j$ is described. Codon $j$ accounts for the portion $\pi^{(i)}$ of the N sequences. $\kappa$ denotes the substitution rate ratio of transition to transversion in the base sequence. The rate of substitution between identical codons is given by $q_k^{(i)} = \sum_{j\neq i} q_k^{(j)}$. $u$ affects the number of synonymous substitutions during a unit of time and is determined by $\sum_i \pi^{(i)} \sum_{j\neq i} q_k^{(j)} = 1$ with $\omega_k = 0$.

The probability of substitution from codon $i$ to codon $j$ at the $k$th residue is given by $p_k^{(j)}(t) = \{\exp(q_k(t))\}_{j\neq i}$ where $t$ is the number of synonymous substitutions and $q_k$ is the matrix representation of the instantaneous rate of substitution (eq. 7). The likelihood of the $d_{ij}$/d$_{i}$ ratio, given the codon polymorphism at the $k$th residue, $P(a_k | \omega_k)$, consists of the probabilities on all branches of the phylogenetic tree and is computed by summing over the states at all internal nodes (Felsenstein 1981).

Genome Sequences and Protein Structures

We analyzed the HA1 domain of 206 HA genes of human influenza A (H3N2) virus selected by Smith et al. (2004). The original set of HA genes contains 253 sequences. We omitted the HA genes with sequences identical to other genes. We used the three-dimensional protein structure of HA (Sauter et al. 1992: the Protein Data Bank code is 1HGF) to detect the spatial distribution of selection pressure. We trimmed the edges of the HA1 domain and analyzed sites 41–315.

Phylogenetic Tree of the 206 HA Genes

We analyzed the base sequences of the 206 HA1 genes using PHYLIP (Felsenstein 2005) to obtain the phylogenetic tree. We used this topology to evaluate the branch lengths that maximized the likelihood $\prod_{k=1}^{L} P(a_k | \omega_k)$ under the constraint that the values of $\omega_k$ at all residues were the same. The parameter $\kappa$ was obtained at 4.34.

Supplementary Material

Supplementary material, figures S1—S5, and tables S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank the two anonymous reviewers for their constructive comments. This study was supported by the JSPS KAKENHI Grant-in-Aid for Scientific Research (B) 22300095 and Scientific Research (C) 24570249.

References


