A Method for Detecting Positive Selection at Single Amino Acid Sites

Yoshiyuki Suzuki and Takashi Gojobori

Center for Information Biology, National Institute of Genetics, Mishima, Japan

A method was developed for detecting the selective force at single amino acid sites given a multiple alignment of protein-coding sequences. The phylogenetic tree was reconstructed using the number of synonymous substitutions. Then, the neutrality was tested for each codon site using the numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree. Computer simulation showed that this method accurately estimated the numbers of synonymous and nonsynonymous substitutions per site, as long as the substitution number on each branch was relatively small. The false-positive rate for detecting the selective force was generally low. On the other hand, the true-positive rate for detecting the selective force depended on the parameter values. Within the range of parameter values used in the simulation, the true-positive rate increased as the strength of the selective force and the total branch length (namely the total number of synonymous substitutions per site) in the phylogenetic tree increased. In particular, with the relative rate of nonsynonymous substitutions to synonymous substitutions being 5.0, most of the positively selected codon sites were correctly detected when the total branch length in the phylogenetic tree was ≥ 2.5. When this method was applied to the human leukocyte antigen (*HLA*) gene, which included antigen recognition sites (ARSs), positive selection was detected mainly on ARSs. This finding confirmed the effectiveness of the present method with actual data. Moreover, two amino acid sites were newly identified as positively selected in non-ARSs. The three-dimensional structure of the HLA molecule indicated that these sites might be involved in antigen recognition. Positively selected amino acid sites were also identified in the envelope protein of human immunodeficiency virus and the influenza virus hemagglutinin protein. This method may be helpful for predicting functions of amino acid sites in proteins, especially in the present situation, in which sequence data are accumulating at an enormous speed.

Introduction

Natural selection is one of the evolutionary mechanisms, in which relative frequencies of genotypes change according to their relative fitnesses in the population. The natural selection can be divided into positive and negative selections. Positive selection is the evolutionary mechanism whereby newly produced mutants have higher fitnesses than the average in the population, and the frequencies of the mutants increase in the following generations. On the other hand, negative selection is the evolutionary mechanism whereby newly produced mutants have lower fitnesses than the average in the population, and the frequencies of the mutants decrease in the following generations. According to the neutral theory of molecular evolution, the great majority of evolutionary changes at the molecular level are caused not by positive selection, but by random drift of selectively neutral or nearly neutral mutants (Kimura 1983).

However, positive selection operating at the amino acid sequence level has been detected for many protein-coding genes, such as mammalian major histocompatibility complex (Hughes and Nei 1988, 1989) and *sry* (Whitfield, Lovell-Badge, and Goodfellow 1993), sea urchin *bindin* (Metz and Palumbi 1996), abalone sperm *lysin* (Lee and Vacquier 1992), and envelope (*env*) of human immunodeficiency virus type 1 (HIV-1) (Seibert et al. 1995; Yamaguchi and Gojobori 1997). It has been proposed that about 0.5% of the 3,595 gene groups so far available in the international DNA data banks (DDBJ/EMBL/GenBank) may have experienced positive selection at one or more amino acid sites (Endo, Ikeo, and Gojobori 1996).

It is well known that different amino acid sites have different biological functions. For example, cell tropism and the syncytium-inducing phenotype of HIV-1 (Chesebro et al. 1992; Fouchier et al. 1992), color vision of mammals (Yokoyama and Yokoyama 1990), and foregut fermentation of colobine Old World monkeys (Stewart, Schilling, and Wilson 1987) are controlled by a few amino acid sites which are separately located in the envelope, opsin, and lysozyme proteins, respectively. Moreover, the functional motifs in PROSITE (Bairoch and Bucher 1994) and the evolutionary motifs in SODHO (Tateno et al. 1997), which are defined as the highly conserved and functionally important amino acid sites in proteins, often consist of a single amino acid site or a region in which single amino acid sites are scattered in many nonconserved and unimportant amino acid sites. Therefore, the types and strengths of selective forces operating on different amino acid sites should be different.

Selective forces operating at the amino acid sequence level have been detected mainly by comparing the number of nonsynonymous substitutions per site with that of synonymous substitutions per site (Hughes and Nei 1988, 1989; Endo, Ikeo, and Gojobori 1996; Tsunoyama and Gojobori 1998). Generally speaking, the excess number of synonymous substitutions was considered to be the result of negative selection, whereas that of nonsynonymous substitutions was attributed to positive selection (Crow and Kimura 1970). In the present paper, we will also use this criterion to detect selective forces.

Key words: positive selection, negative selection, synonymous substitution, nonsynonymous substitution, phylogenetic tree, multiple alignment.

Address for correspondence and reprints: Takashi Gojobori, Center for Information Biology, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka-ken 411-8540, Japan.
E-mail: tgojobor@genes.nig.ac.jp.

Mol. Biol. Evol. 16(10):1315–1328. 1999

Several methods have been developed for estimating numbers of synonymous and nonsynonymous substitutions (Miyata and Yasunaga 1980; Li, Wu, and Luo 1985; Nei and Gojobori 1986; Li 1993; Pamilo and Bianchi 1993; Goldman and Yang 1994; Muse and Gaut 1994; Comeron 1995; Ina 1995). However, these methods require the use of many codon sites to avoid a large variance for the estimate, which is computed as an average over a particular length of codons. Consequently, positive selection has always been assigned to an amino acid region of a particular length. Therefore, if positive selection had operated on an amino acid site, as long as the average number of nonsynonymous substitutions was smaller than that of synonymous substitutions over the analyzed region, it could not be identified. Moreover, when positive selection was detected for the analyzed region, it was impossible to identify the exact site of selection from only the conventional sequence analyses.

So far, however, some attempts have been made to detect positive selection at single amino acid sites. Fitch et al. (1997) used a multiple alignment of protein-coding sequences to reconstruct a phylogenetic tree. Then, for each codon site, they compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes to detect positively selected amino acid sites. However, their method assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. Nielsen and Yang (1998) developed a posterior probability method using the maximum-likelihood approach to detect positively selected amino acid sites. However, they assumed that the relative rate of nonsynonymous substitutions to synonymous substitutions was the same for all positively selected codon sites, which did not seem realistic.

In the present study, we developed a new method for detecting the selective force at single amino acid sites. The method does not rely on the assumptions mentioned above. The effectiveness of this method was confirmed by conducting a computer simulation and analyzing the human leukocyte antigen (*HLA*) gene. We also applied this method to the HIV-1 *env* gene and the influenza virus hemagglutinin (*HA*) gene to identify positively selected amino acid sites.

Materials and Methods

Theory

Let us assume that we have a multiple alignment of protein-coding sequences. The new method for detecting the selective force at single amino acid sites consists of the following five steps (fig. 1).

First, a phylogenetic tree was reconstructed using the number of synonymous substitutions, as the number of synonymous substitutions was thought to be roughly proportional to the evolutionary time which was used in the later computation. The neighbor-joining method (Saitou and Nei 1987) was used for reconstructing a phylogenetic tree, with the method of Nei and Gojobori (1986) being used to estimate the number of synonymous substitutions. Throughout this paper, when we refer to the method of Nei and Gojobori (1986), we mean method I of Nei and Gojobori (1986). In the phylogenetic tree, the branch length was defined as the number of synonymous substitutions per site for the branch.

Second, for each codon site, the ancestral codon was inferred at each node of the phylogenetic tree. For inference of ancestral sequences, the maximum-parsimony method (Fitch 1971; Hartigan 1973) and the maximum-likelihood method (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996; Schultz, Cocroft, and Churchill 1996; Zhang and Nei 1997) have been developed. Simulation studies have indicated that the maximum-likelihood method produced more reliable results than the maximum-parsimony method when the sequences compared were distantly related to one another (Yang, Kumar, and Nei 1995; Zhang and Nei 1997). However, these studies have also indicated that both methods produced similarly reliable results when the sequences compared were closely related to one another (Yang, Kumar, and Nei 1995; Zhang and Nei 1997). In the present study, we used the maximum-parsimony method (Hartigan 1973) for reconstructing ancestral codons because in most cases, the degree of divergence among sequences used in this study was within the range at which both methods produced similarly reliable results in the simulation studies (Yang, Kumar, and Nei 1995; Zhang and Nei 1997). When more than one codon was inferred at a node, we assumed that they had existed with the same probability. When only the termination codon was inferred at some nodes, we excluded that codon site from further analyses because it should have destructed the protein function. Furthermore, when the number of combinations for possible ancestral codons over all nodes exceeded 10,000, that site was also excluded from further analyses because of time restriction.

Third, the average numbers of synonymous and nonsynonymous sites throughout the phylogenetic tree were estimated for each codon site. The numbers of synonymous and nonsynonymous sites for a particular codon were defined as the sums of the fractions of synonymous and nonsynonymous changes at three positions of the codon, respectively (Nei and Gojobori 1986). Namely, if we denote by f_i the fraction of synonymous changes at the ith position of a particular codon ($i = 1, 2, 3$), the numbers of synonymous (s) and nonsynonymous (n) sites for that codon were given by $s = \Sigma_{i=1}^{3} f_i$ and $n = (3 - s)$, respectively. For each codon site, the average numbers of synonymous and nonsynonymous sites on each branch were computed as follows. When more than one position was different between two codons at the ends of a branch, we took into account the possible intermediate codons in the computation. For example, if TTT and TCC occupied the ends of a branch, there were two possible intermediate codons, TTC and TCT. The average numbers of synonymous ($s_{b(TTT,TCC)}$) and nonsynonymous ($n_{b(TTT,TCC)}$) sites at the branch connecting TTT and TCC were computed as

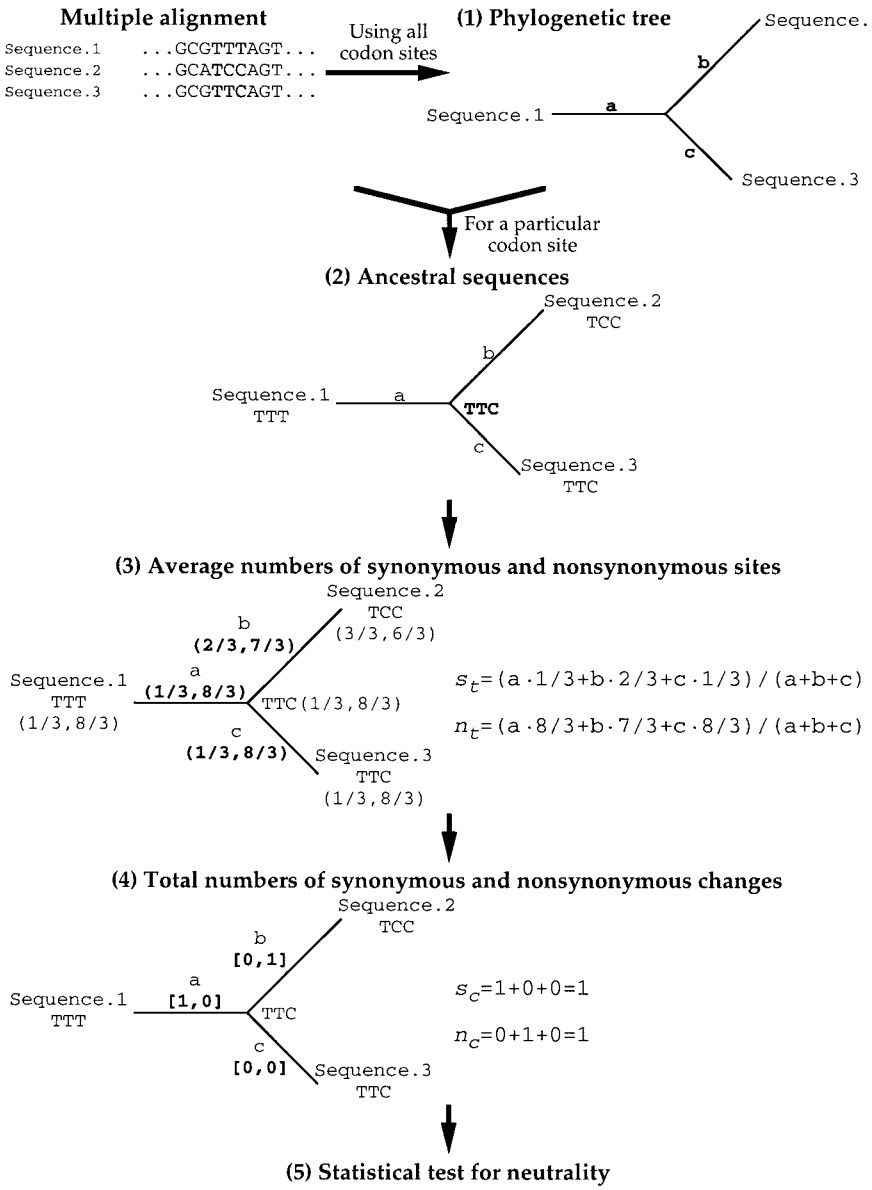

FIG. 1.—Schematic representation of the new method for detecting the selective force at single amino acid sites. The method used a multiple alignment of protein-coding sequences and consisted of five steps. In this example, the number of OTUs was assumed to be three. The numbers of synonymous and nonsynonymous sites for codons and branches in the phylogenetic tree are indicated in parentheses (synonymous, nonsynonymous). The numbers of synonymous and nonsynonymous changes on branches in the phylogenetic tree are indicated in brackets [synonymous, nonsynonymous]. s_t and n_t represent the average numbers of synonymous and nonsynonymous sites throughout the phylogenetic tree for one codon site, respectively. s_c and n_c represent the total numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree for one codon site, respectively.

$$s_{b(TTT,TCC)} = \{(s_{TTT} + s_{TTC} + s_{TCC})/3 + (s_{TTT} + s_{TCT} + s_{TCC})/3\}/2 \quad (1)$$

$$n_{b(TTT,TCC)} = \{(n_{TTT} + n_{TTC} + n_{TCC})/3 + (n_{TTT} + n_{TCT} + n_{TCC})/3\}/2, \quad (2)$$

where s_{TTT} and n_{TTT} (and so forth) are the numbers of synonymous and nonsynonymous sites for codon TTT (and so forth), respectively. When no positions or one position was different between two codons, the numbers of synonymous and nonsynonymous sites for those two codons were averaged. The average numbers of synonymous (s_t) and nonsynonymous (n_t) sites throughout the phylogenetic tree were computed by averaging each number over all branches with weights proportional to the evolutionary time. The evolutionary time was approximated by the branch length in the phylogenetic tree. That is,

$$s_t = \sum_{b=1}^{N} s_b \cdot l_b/l_t \quad (3)$$

$$n_t = \sum_{b=1}^{N} n_b \cdot l_b/l_t, \quad (4)$$

where N is the total number of branches, l_b is the length

of branch b, and l_t is the total branch length in the phylogenetic tree. When the average number of synonymous sites was zero at a codon site, that codon site was excluded from the computation of the number of synonymous substitutions per site and the test of neutrality because of inapplicability.

Fourth, the total numbers of synonymous (s_c) and nonsynonymous (n_c) changes throughout the phylogenetic tree were counted for each codon site. The numbers of synonymous and nonsynonymous changes were defined as the numbers of synonymous and nonsynonymous differences, respectively, between two codons compared (Nei and Gojobori 1986). When one position was different between two codons, we could immediately decide whether the change was synonymous or nonsynonymous. When two or three positions were different between two codons, there were two or six possible pathways by which to obtain the differences, respectively. The numbers of synonymous and nonsynonymous changes between those two codons were computed as the averages of those changes over all pathways.

Finally, the statistical test for neutrality was conducted for each codon site. If no selective force was operating on a codon site, the equations

$$\frac{s_c}{s_c + n_c} = \frac{s_t}{s_t + n_t} \tag{5}$$

$$\frac{n_c}{s_c + n_c} = \frac{n_t}{s_t + n_t} \tag{6}$$

should hold. The numbers of synonymous and nonsynonymous changes throughout the phylogenetic tree were rounded off to the integer, in order to calculate the exact binomial probability (p) of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes for each codon site. $s_t/(s_t + n_t)$ and $n_t/(s_t + n_t)$ were used as the probabilities for the occurrences of synonymous and nonsynonymous changes for each codon site, respectively. The significance level was set at 5%. When the number of synonymous changes was significantly larger than that of nonsynonymous changes, negative selection was considered to have operated on that site. In the opposite situation, on the other hand, positive selection was assigned.

Computer Simulation

The computer simulation was conducted to investigate the accuracies of the estimates of the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree and of the inference for the selective force at single amino acid sites. For the former purpose, however, we investigated the accuracy of the estimates for the numbers of synonymous (s_s) and nonsynonymous (n_s) substitutions, which were defined as the numbers of synonymous and nonsynonymous changes per site, respectively. That is,

$$s_s = \frac{s_c}{s_t} \tag{7}$$

$$n_s = \frac{n_c}{n_t}. \tag{8}$$

The numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree should be different among codon sites, depending on the codon in the ancestral sequence. In contrast, the numbers of synonymous and nonsynonymous substitutions per site should be constant for all codon sites if the operating selective forces were the same for all codon sites.

The simulation method used in this study was originally established by Gojobori (1983) and Ina (1995). First, we constructed the mutation matrix for four nucleotides. In this study, a one-parameter model (Jukes and Cantor 1969) was adopted. That is, the probability of mutation (λ_{ij}) from nucleotide i (T, C, A, or G) to the different nucleotide j was assumed to be the same for all combinations of i and j. λ_{ii} was defined as

$$\lambda_{ii} = 1 - \sum_{j \neq i} \lambda_{ij}. \tag{9}$$

Then, the 61 × 61 codon substitution matrix (excluding termination codons) was constructed from the mutation matrix and the coefficient f, the relative rate of nonsynonymous substitutions to synonymous substitutions. For example, the probability of substitution ($p_{TTT,TCC}$) from TTT to TCC was computed as

$$p_{TTT,TCC} = \lambda_{TT} \cdot \lambda_{TC} \cdot \lambda_{TC} \cdot f. \tag{10}$$

If the amino acids encoded by two codons were the same, f was set to 1.0, whereas if they were different, f was set to 1.0 for no selection scheme, 0.2 and 0.5 for a negative selection scheme, and 2.0 and 5.0 for a positive selection scheme. p_{ii} was defined as

$$p_{ii} = 1 - \sum_{j \neq i} p_{ij}, \tag{11}$$

where i and j are different codons.

The equilibrium frequencies of 61 codons were all set to 1/61. The expected numbers of synonymous ($E(S_s)$) and nonsynonymous ($E(N_s)$) sites at one codon site were computed as

$$E(S_s) = \sum_{i=\mathrm{TTT}}^{\mathrm{GGG}} s_i/61 \tag{12}$$

$$E(N_s) = \sum_{i=\mathrm{TTT}}^{\mathrm{GGG}} n_i/61, \tag{13}$$

where s_i and n_i are the numbers of synonymous and nonsynonymous sites at codon i, respectively, as defined by Nei and Gojobori (1986).

The expected numbers of synonymous ($E(S_c)$) and nonsynonymous ($E(N_c)$) changes between two codons for one unit of time were computed as

$$E(S_c) = \sum_{i=\mathrm{TTT}}^{\mathrm{GGG}} \sum_{j=\mathrm{TTT}}^{\mathrm{GGG}} s_{ij} \cdot p_{ij}/61 \quad (14)$$

$$E(N_c) = \sum_{i=\mathrm{TTT}}^{\mathrm{GGG}} \sum_{j=\mathrm{TTT}}^{\mathrm{GGG}} n_{ij} \cdot p_{ij}/61, \quad (15)$$

where s_{ij} and n_{ij} are the numbers of synonymous and nonsynonymous changes between codons i and j, respectively, as defined by Nei and Gojobori (1986).

Finally, the expected numbers of synonymous ($E(S_d)$) and nonsynonymous ($E(n_d)$) substitutions per site for one unit of time were computed as

$$E(S_d) = \frac{E(S_c)}{E(S_s)} \quad (16)$$

$$E(N_d) = \frac{E(N_c)}{E(N_s)}. \quad (17)$$

In the present simulation, one unit of time was set so that the expected number of synonymous substitutions was 0.01. This was achieved by iteratively calculating $E(S_d)$ until

$$(0.01 - E(S_d))^2 \leq 10^{-30} \quad (18)$$

held, refining λ_{ij} by multiplying it by the ratio of 0.01 to $E(S_d)$. Actually, λ_{ij} was set as 0.0035 ($f = 0.2$) to 0.0032 ($f = 5.0$); the value decreased as f increased, which was expected from equation (10).

The ancestral sequence having a codon length of 300 was constructed from the equilibrium codon frequencies using pseudorandom numbers. The number 300 was used because the average number of amino acids in a protein has been reported to be about 300 (Ina 1995). The ancestral sequence was evolved along the artificial phylogenetic tree according to the codon substitution matrix using pseudorandom numbers. f was set to be the same for all codon sites in one sequence. For a phylogenetic tree, we assumed a symmetrical topology with 64 and 128 extant operational taxonomy units (OTUs) and the same branch lengths (b) of 0.01, 0.02, and 0.03. The extant sequences obtained were subjected to our method. In the simulation, we assumed two situations: one in which the phylogenetic relationship was known, and one in which it was unknown. The simulation scheme was iterated 200 times for each parameter set, yielding a total of 60,000 codon sites.

Application to the *HLA* Gene

HLA is one of the proteins expressed on the surface of antigen-presenting cells in humans. The protein binds to an antigenic peptide and presents it to T lymphocytes. Amino acid sites important for peptide binding have been identified and are called antigen recognition sites (ARSs). Hughes and Nei (1988) demonstrated that positive selection had operated on ARSs by comparing the average number of nonsynonymous substitutions with that of synonymous substitutions over all ARSs using pairs of *HLA* sequences. They also demonstrated that negative selection had operated on non-ARSs (Hughes and Nei 1988). To investigate whether our method could produce the same result, we analyzed the *HLA* gene. However, the number of sequences used by Hughes and Nei (1988) was 21, which was considered too small to obtain conclusive results with our method. Then, we collected more sequence data from a World Wide Web site (The Japanese Society for Histocompatibility and Immunogenetics; http://square.umin.ac.jp/JSHI/hla_data/data.html). On that site, nucleotide sequence data for *HLA* (*HLA-A*, *HLA-B*, and *HLA-C*) genes are deposited. We used 228 sequences (which did not include any gaps in 819 nucleotides) that corresponded to exons 2–4 of the *HLA* gene. The average (total) branch length in the phylogenetic tree was 0.002 (1.06).

Application to the V3 Region of the HIV-1 *env* Gene

HIV-1 is the causative agent of acquired immunodeficiency syndrome in humans. The envelope protein of HIV-1 is the major target of the immune response from the host. The V3 region of this protein determines the cell tropism and syncytium-inducing capacity of HIV-1. In addition, the V3 region is entirely covered with monoclonal antibody and cytotoxic T lymphocyte epitopes. Yamaguchi and Gojobori (1997) analyzed sequence data of the V3 region obtained from single patients at different time points (Wolfs et al. 1991; Holmes et al. 1992; McNearney et al. 1992). They found that the number of amino acid substitutions was significantly larger at five amino acid sites. They speculated that those sites might have been positively selected. Theoretically, however, the larger number of amino acid substitutions does not necessarily suggest the operation of positive selection. The larger number of nonsynonymous sites and the higher mutation rate for those codon sites can also explain the above phenomenon. We analyzed the same data set as Yamaguchi and Gojobori (1997) to detect the positively selected amino acid sites in the V3 region. Partial *env* sequences with no gaps in the 162 nucleotides, 105 of which encoded the V3 region and 57 its upstream, were obtained from six patients (patients A–F in Yamaguchi and Gojobori 1997). The numbers of sequences from patients A–F were 78, 39, 47, 16, 14, and 17 (total 211), respectively. The average (total) branch lengths in the phylogenetic trees reconstructed for patients A–F were 0.003 (0.51), 0.003 (0.19), 0.004 (0.33), 0.001 (0.03), 0.011 (0.27), and 0.004 (0.12), respectively. The numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree were estimated for each codon site in each of the six phylogenetic trees. Then, those numbers from the six phylogenetic trees were combined to yield the numbers of synonymous and nonsynonymous sites and the numbers of those changes for each codon site over six phylogenetic trees. The average (total) branch length over six phylogenetic trees was 0.004 (1.45).

Application to the Influenza A Virus *HA* Gene

Influenza A virus is the causative agent of the acute respiratory illness known as influenza. HA is an envelope protein which is responsible for the adsorption and penetration of viral particle and is the major target of the immune response from the host. This protein is

Table 1
Expected Numbers of Synonymous (Syn.) and Nonsynonymous (Non.) Substitutions per Site Throughout the Phylogenetic Tree for One Codon Site

		f[a]				
b[b]		0.2	0.5	1.0	2.0	5.0
64 OTUs						
0.01 . . .	Syn.	1.26 ± 1.28[c]	1.26 ± 1.28	1.26 ± 1.28	1.26 ± 1.28	1.26 ± 1.28
	Non.	0.26 ± 0.34	0.64 ± 0.53	1.26 ± 0.75	2.48 ± 1.05	5.88 ± 1.62
0.02 . . .	Syn.	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81	2.52 ± 1.81
	Non.	0.51 ± 0.48	1.27 ± 0.76	2.52 ± 1.06	4.95 ± 1.49	11.75 ± 2.29
0.03 . . .	Syn.	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22	3.78 ± 2.22
	Non.	0.77 ± 0.59	1.91 ± 0.93	3.79 ± 1.30	7.43 ± 1.82	17.63 ± 2.81
128 OTUs						
0.01 . . .	Syn.	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82	2.54 ± 1.82
	Non.	0.52 ± 0.48	1.28 ± 0.76	2.54 ± 1.07	4.99 ± 1.49	1.85 ± 2.30
0.02 . . .	Syn.	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58	5.08 ± 2.58
	Non.	1.03 ± 0.68	2.57 ± 1.07	5.09 ± 1.51	9.98 ± 2.11	23.69 ± 3.26
0.03 . . .	Syn.	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16	7.62 ± 3.16
	Non.	1.55 ± 0.83	3.85 ± 1.31	7.63 ± 1.85	14.97 ± 2.59	35.54 ± 3.99

[a] The relative rate of nonsynonymous substitutions to synonymous substitutions.
[b] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.
[c] The standard error was calculated with the assumption of a Poisson distribution for the substitution number.

cleaved into HA_1 and HA_2 upon infection. Fitch et al. (1997) analyzed 254 sequences for the *HA_1* gene and proposed that 25 codon sites might have been positively selected. They compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes for each codon site. The exact binomial probability of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes were calculated for each codon site. However, Fitch et al. (1997) assumed that the probabilities for the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. We analyzed the same data set as Fitch et al. (1997) to detect positively selected codon sites in the *HA_1* gene. We used 248 of the 254 sequences, because 6 contained gaps or ambiguous characters. Each sequence consisted of 987 nucleotides. The average (total) branch length in the phylogenetic tree was 0.004 (1.88).

Results

Computer Simulation

In the computer simulation, we assumed two situations: one in which the phylogenetic relationship was known, and one in which it was unknown. This assumption allowed us to investigate the effect of information about the phylogenetic relationship on the overall results. In fact, the results were almost identical in both situations (data not shown). Therefore, we present only the results obtained with the assumption that the phylogenetic relationship was known.

The expected and estimated numbers of synonymous and nonsynonymous substitutions per site throughout the phylogenetic tree at one codon site are summarized in Tables 1 and 2, respectively. The expected numbers of synonymous and nonsynonymous substitutions (table 1) may be regarded as the true values to be estimated. It was clear that the numbers of synonymous and nonsynonymous substitutions were estimated accurately in many situations. However, under the strong positive selection scheme ($f = 5.0$), the number of synonymous substitutions tended to be overestimated, whereas that of nonsynonymous substitutions tended to be underestimated. These tendencies became obvious when the expected number of nonsynonymous substitutions on a branch exceeded 0.1. These findings probably resulted from the branches in the phylogenetic tree containing multiple nonsynonymous substitutions which were not corrected for by the maximum-parsimony method. Moreover, some of the multiple pathways between two codons, which were different at more than one position due to multiple nonsynonymous substitutions, probably contained artificial synonymous substitutions in the computation. The variance was generally larger for the estimation than for the expectation, probably due to errors accompanying the inference of ancestral codons and the estimation of substitution numbers. However, when f was 5.0, the variance for nonsynonymous substitution was smaller. This may be explained by the saturation effect on the number of nonsynonymous substitutions due to the use of the maximum-parsimony method.

In the test of neutrality, we excluded codon sites at which only the termination codon was inferred at some nodes in the phylogenetic tree. We also excluded sites at which the number of combinations for possible ancestral codons over all nodes exceeded 10,000 and those at which the average number of synonymous substitutions throughout the phylogenetic tree was zero. For all parameter sets except one, the number of testable sites was close to 60,000 (53,340~60,000), which was the total number of codon sites in a simulation. However, a dramatic decline (20,611) was observed for 128 OTUs with strong positive selection ($f = 5.0$) and long branch

Table 2
Expected Numbers of Synonymous (Syn.) and Nonsynonymous (Non.) Substitutions per Site Throughout the Phylogenetic Tree for One Codon Site[a]

		f[b]				
b[c]		0.2	0.5	1.0	2.0	5.0
64 OTUs						
0.01 . . .	Syn.	1.26 ± 1.97	1.27 ± 2.39	1.28 ± 1.91	1.30 ± 1.52	1.45 ± 1.45
	Non.	0.25 ± 0.34	0.64 ± 0.54	1.26 ± 0.76	2.47 ± 1.06	5.81 ± 1.58
0.02 . . .	Syn.	2.48 ± 2.72	2.46 ± 2.51	2.49 ± 2.07	2.56 ± 1.96	3.05 ± 1.99
	Non.	0.51 ± 0.49	1.27 ± 0.76	2.49 ± 1.06	4.81 ± 1.45	10.75 ± 1.98
0.03 . . .	Syn.	3.66 ± 2.98	3.66 ± 2.81	3.70 ± 2.42	3.82 ± 2.30	4.72 ± 2.37
	Non.	0.77 ± 0.60	1.90 ± 0.93	3.70 ± 1.28	7.01 ± 1.69	14.69 ± 2.07
128 OTUs						
0.01 . . .	Syn.	2.53 ± 3.20	2.53 ± 3.04	2.55 ± 2.30	2.60 ± 2.09	2.91 ± 2.03
	Non.	0.52 ± 0.49	1.29 ± 0.77	2.54 ± 1.09	4.99 ± 1.52	11.72 ± 2.27
0.02 . . .	Syn.	5.00 ± 4.54	4.99 ± 3.21	5.01 ± 2.89	5.18 ± 2.75	6.06 ± 2.76
	Non.	1.03 ± 0.70	2.56 ± 1.09	5.04 ± 1.52	9.70 ± 2.08	21.35 ± 2.74
0.03 . . .	Syn.	7.37 ± 4.43	7.36 ± 3.67	7.45 ± 3.41	7.73 ± 3.25	8.97 ± 3.19
	Non.	1.56 ± 0.86	3.83 ± 1.33	7.45 ± 1.84	14.13 ± 2.44	28.19 ± 2.78

[a] The phylogenetic relationship was assumed to be known.
[b] The relative rate of nonsynonymous substitutions to synonymous substitutions.
[c] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.

length ($b = 0.03$). The majority of excluded sites had more than 10,000 combinations of possible ancestral codons.

The results for the detection of the selective force at single amino acid sites are summarized in table 3. In general, the false-positive rate for detection of the selective force was low. Namely, the rate was at most 2% under no selection scheme, which was expected with significance level of 5%. The rate declined to almost zero when positive and negative selections operated. This tendency was not related to the strength of the selective force, the number of OTUs, or the branch length in the phylogenetic tree. On the other hand, the true-positive rate for detection of the selective force depended on the parameter values. The rate improved as the selective force, the number of OTUs, and the branch length in the phylogenetic tree increased. The increase in the latter two factors corresponded to the increase in the total branch length in the phylogenetic tree. In particular, most of the sites with strong positive selection ($f = 5.0$) were correctly detected when the total branch length was ≥2.5. The negatively selected sites were less well detected than the positively selected sites. That is, a total branch length of ≥5.0 was needed to correctly detect most of the sites with strong negative selection ($f = 0.2$), probably because the total number of nucleotide changes throughout the phylogenetic tree for one codon site was smaller in the negative selection scheme. However, our method should still be useful for detecting selective forces at single amino acid sites, because the false-positive rate was generally low.

Table 3
Frequencies of Codon Sites on Which Negative (Neg.) and Positive (Pos.) Selections Were Detected[a]

		f[b]				
b[c]		0.2	0.5	1.0	2.0	5.0
64 OTUs						
0.01 . . .	Neg.	0.08	0.05	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.01	0.21
0.02 . . .	Neg.	0.22	0.09	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.05	0.47
0.03 . . .	Neg.	0.33	0.12	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.08	0.59
128 OTUs						
0.01 . . .	Neg.	0.23	0.10	0.02	0.00	0.00
	Pos.	0.00	0.00	0.00	0.06	0.57
0.02 . . .	Neg.	0.43	0.15	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.15	0.84
0.03 . . .	Neg.	0.56	0.20	0.02	0.00	0.00
	Pos.	0.00	0.00	0.01	0.21	0.92

[a] The phylogenetic relationship was assumed to be known.
[b] The relative rate of nonsynonymous substitutions to synonymous substitutions.
[c] The branch length, represented as the number of synonymous substitutions per site, for each branch in the phylogenetic tree.

Application to the *HLA* Gene

The results for detecting the selective force at single amino acid sites in the HLA protein are described in figure 2. Of the 57 ARSs, 17 were inferred as positively selected but none were inferred as negatively selected. Of the remaining 216 non-ARSs, 2 were inferred as positively selected and 16 were inferred as negatively selected. The χ^2 test and Fisher's exact test clarified that a significantly larger fraction of sites were positively selected in ARSs than in non-ARSs (table 4). Furthermore, we investigated the selective force operating on ARSs and non-ARSs. For each region, the total number of nonsynonymous changes throughout the phylogenetic tree over all codon sites was compared with that of synonymous changes. As a result, the number of nonsynonymous changes was significantly ($P = 4.8 \times 10^{-59}$) larger than that of synonymous changes in ARSs, whereas the inverse ($P = 3.9 \times 10^{-10}$) was true for non-

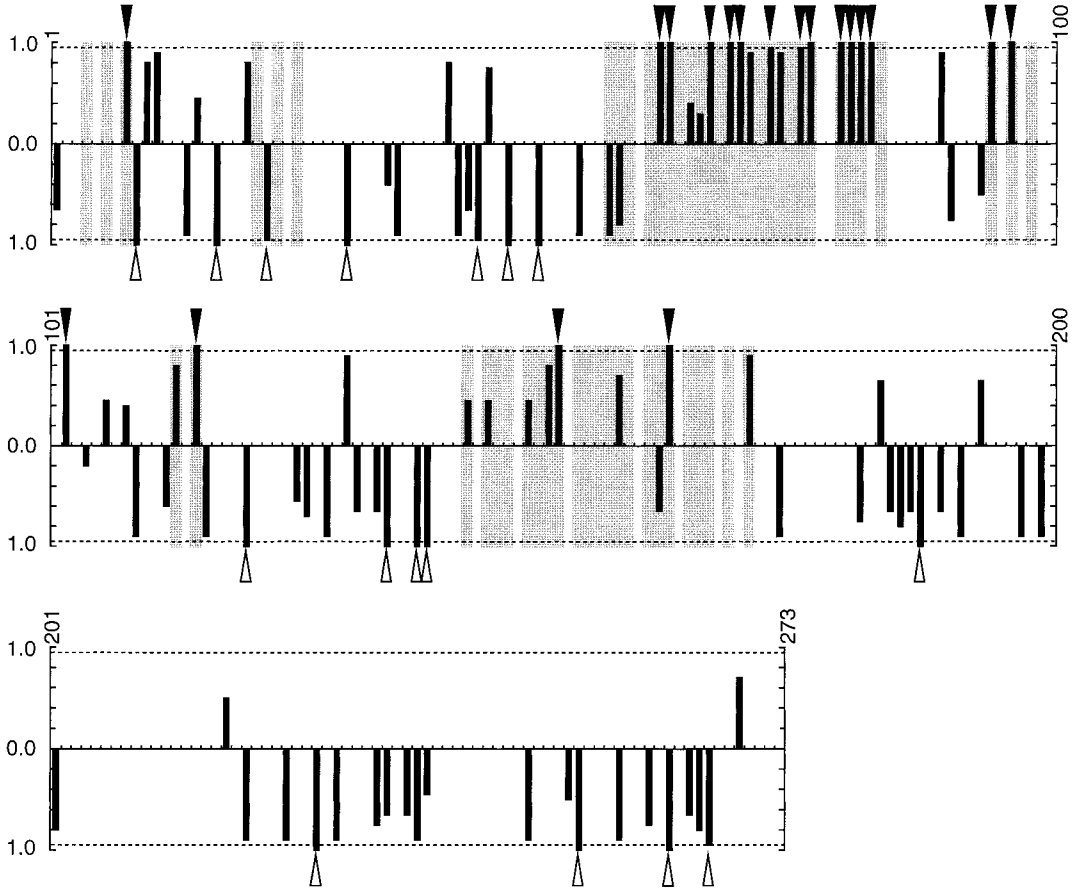

FIG. 2.—Amino acid sites at which positive and negative selections were detected in the HLA protein. The abscissa indicates the amino acid site counted from the N-terminus of exon 2. The ordinate indicates the value of $1 - P$ for each amino acid site (see the text). When the number of nonsynonymous substitutions per site is larger than that of synonymous substitutions per site, the value is indicated above the abscissa. In the opposite situation, the value is indicated below the abscissa. Dotted lines indicate the 5% significance level. Positive (filled arrowhead) or negative (open arrowhead) selection was assigned to the amino acid site when the corresponding value exceeds that of the dotted line. ARSs are indicated with shading. See also table 4 for statistical analyses.

ARSs. Similar results were obtained when Hughes and Nei's (1988) data set was analyzed (data not shown). These results were consistent with those of Hughes and Nei (1988), confirming the effectiveness of our method with actual data.

Table 4
Numbers of Codon Sites on Which Positive and Negative Selections Were Detected for ARSs and Non-ARSs in the *HLA* Gene[a]

	Positive Selection	Negative Selection	No Selection[b]	Excluded[c]
ARS.....	17	0	37	3
non-ARS.....	2	16	188	10

[a] The χ^2 test and Fisher's exact test were conducted for the 2×2 contingency table of positive selection and negative plus no selections, versus ARS and non-ARS. The χ^2 value was 58.8 with one degree of freedom ($P < 0.005$), and the Fisher's exact probability was 3.2×10^{-11}.

[b] This category included codon sites for which a statistically significant difference was not detected between the numbers of synonymous and nonsynonymous changes.

[c] This category included codon sites for which the statistical test could not be conducted.

It should be noted that positive selection was detected at two amino acid sites in non-ARSs (fig. 2). When we examined the three-dimensional structure of the HLA molecule (PDBid: 1HLA; Bjorkman et al. 1987), all positively selected amino acid sites, including those in non-ARSs, faced the cleft for antigen recognition (fig. 3). Thus, two positively selected amino acid sites in non-ARSs might also be involved in antigen recognition. In contrast, most of the negatively selected amino acid sites did not face the cleft for antigen recognition (fig. 3).

Application to the V3 Region of the HIV-1 *env* Gene

The results for detection of the selective force at single amino acid sites in the V3 region of the HIV-1 envelope protein are described in figure 4. Among the five sites (positions 11, 13, 18, 20, and 25) at which the number of amino acid substitutions was larger (Yamaguchi and Gojobori 1997), positive selection was detected at two sites (positions 13 and 18), but not at the other three (positions 11, 20, and 25). However, for all of the latter three sites, the numbers of nonsynonymous substitutions per site were larger than those of synony-

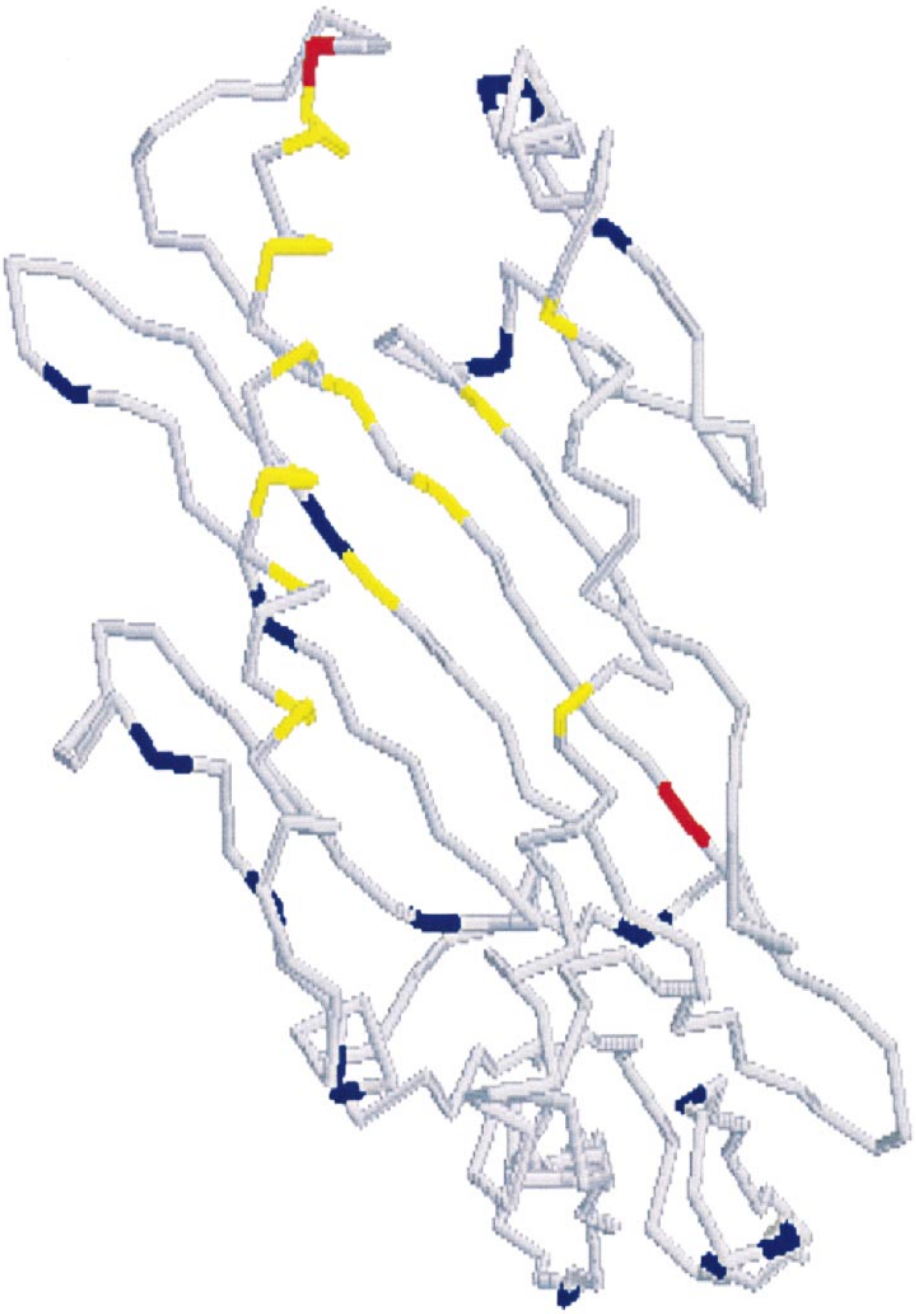

FIG. 3.—Three-dimensional structure of the HLA molecule (PDBid: 1HLA; Bjorkman et al. 1987). Positively selected amino acid sites in ARSs, those in non-ARSs, and negatively selected amino acid sites are colored yellow, red, and blue, respectively.

mous substitutions per site (fig. 4). Therefore, we did not rule out the possibility that those sites may also be positively selected. It should be noted, however, that positive selection was detected at two sites (positions 22 and 24) at which the number of amino acid substitutions was not larger in Yamaguchi and Gojobori (1997).

Positions 13 and 24 have been related to antigenic variation (Wolfs et al. 1991; Shioda et al. 1994) and cell tropism and syncytium-inducing capacity (Chesebro et al. 1992; Fouchier et al. 1992) of HIV-1, respectively. However, for positions 18 and 22, no particular functions have been assigned. The entire V3 region is covered with monoclonal antibody and cytotoxic T lymphocyte epitopes (HIV Molecular Immunology Database; http://hiv-web.lanl.gov/immunology/index.html). Therefore, those sites might be important for recognition by the immune system of the host.

Application to the Influenza A Virus *HA* Gene

In the HA_1 gene of influenza A virus, positive selection was detected at three codon sites (positions 138, 196, and 226). All of these sites were included in the 25 sites proposed to be positively selected by Fitch et al. (1997). The discrepancy in the number of sites de-

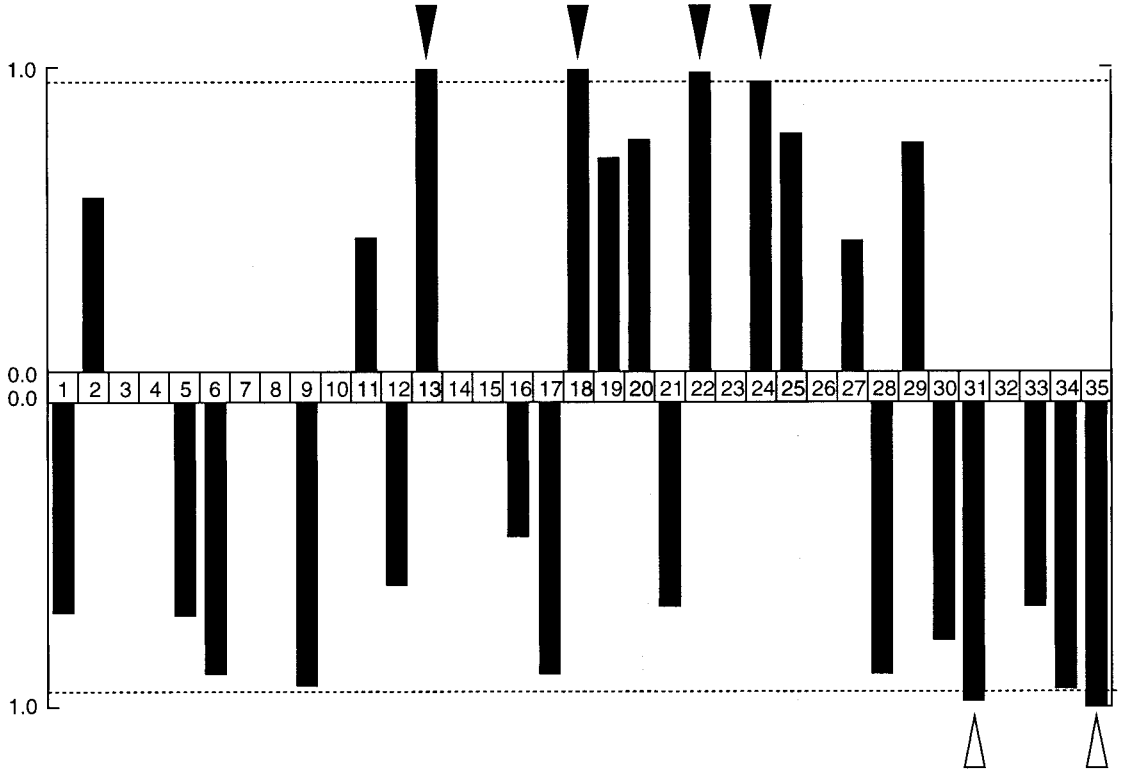

FIG. 4.—Amino acid sites at which positive and negative selections were detected in the V3 region of the HIV-1 envelope protein. The abscissa indicates the amino acid site counted from the N-terminal cysteine residue in the V3 region. The ordinate indicates the value of $1 - P$ for each amino acid site (see the text). When the number of nonsynonymous substitutions per site is larger than that of synonymous substitutions per site, the value is indicated above the abscissa. In the opposite situation, the value is indicated below the abscissa. Dotted lines indicate the 5% significance level. Positive (filled arrowhead) or negative (open arrowhead) selection was assigned to the amino acid site when the corresponding value exceeds that of the dotted line. Amino acid sites at which the substitution number has been reported to be larger (Yamaguchi and Gojobori 1997) are indicated with shading.

tected in the previous (Fitch et al. 1997) and present studies may have resulted from the difference in the methodologies. That is, Fitch et al. (1997) assumed that the probabilities of the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, whereas we did not make that assumption. Indeed, when that assumption was made for our method, 17 amino acid sites (positions 88, 121, 133, 135, 137, 138, 145, 156, 159, 186, 188, 190, 193, 194, 226, 275, and 276) were detected as positively selected. An additional 18 sites (positions 53, 75, 78, 94, 124, 140, 157, 158, 163, 172, 174, 189, 196, 201, 214, 219, 310, and 312) were detected when the probabilities for the occurrences of synonymous and nonsynonymous changes for all codon sites were assumed to be 0.565 and 0.435, respectively, which were the values used by Fitch et al. (1997). These sites include most of the 25 sites detected by Fitch et al. (1997). Moreover, when the probabilities used by Fitch et al. (1997) were adopted in our computer simulation, up to 75%, 31%, and 5% of sites with f values of 1.0, 0.5, and 0.2, respectively, were falsely detected as positively selected. Therefore, the results of Fitch et al. (1997) might contain some false-positives.

It has been reported that positions 138 and 226 are involved in selection during growth in eggs (Meyer et al. 1993; Rocha et al. 1993; Gubareva et al. 1994; Hardy et al. 1995) and positions 196 and 226 are involved in antibody recognition and neutralization (Wiley, Wilson, and Skehel 1981; Bizebard et al. 1995) of influenza A virus. Therefore, these functions may be the causes of positive selection for those sites. Interestingly, all positively selected amino acid sites were located close together in the three-dimensional structure of the HA_1 molecule (PDBid: 1HGF; Sauter et al. 1992), whereas negatively selected sites were evenly scattered in the molecule (fig. 5).

Discussion

The method described in the present study detects the selective force by comparing the number of nonsynonymous changes with that of synonymous changes, assuming that the synonymous change was almost neutral. However, some reports have indicated that the selective force may operate on the synonymous change. As for the cause of the selective force, the codon usage bias (Akashi 1995) and the secondary structure of the messenger RNA (mRNA) (Smith and Simmonds 1997) have been hypothesized. However, our method may still be useful in those situations, because the selective force operating on the codon usage bias may be weak (Akashi 1995). Moreover, a similar degree of selective constraint to that on the synonymous change may also operate on

FIG. 5.—Three-dimensional structure of the influenza A virus HA_1 molecule (PDBid: 1HGF; Sauter et al. 1992). This protein constitutes a trimer in the virion. Positively and negatively selected amino acid sites are colored yellow and blue, respectively.

the nonsynonymous change to maintain the secondary structure of mRNA.

In the computer simulation, values ranging from 0.2 to 5.0 were used for f, the relative rate of nonsynonymous substitutions to synonymous substitutions. f was, in general, considered approximately equal to $4N_e s$, with positive genic selection in the diploid organism, where N_e is the effective population size and s is the selective coefficient (Crow and Kimura 1970). To investigate whether the values used for f were realistic, we computed the ratio (r) of the number of nonsynonymous substitutions per site to that of synonymous substitutions per site in the results of Hughes and Nei (1988). r was, by definition, almost equivalent to f. As a result, r ranged from 0.06 to 5.08, encompassing the values of f used in the simulation. These findings would suggest that our simulation schemes were not unrealistic.

The computer simulation and the application to the *HLA* gene confirmed the effectiveness of our method for detecting the selective force at single amino acid sites. The efficiency of the method increased as the selective force, the number of OTUs, and the branch length in the phylogenetic tree increased. The increase in the latter two factors corresponded to the increase in the total branch length in the phylogenetic tree. However, it should be recalled that in the case of 128 OTUs with strong positive selection ($f = 5.0$) and long branch length ($b = 0.03$), the number of codon sites on which the statistical test could be conducted was small, mainly due to the sites with more than 10,000 combinations of possible ancestral codons over all nodes in the phylogenetic tree. This result suggests that overall accuracy in detecting the selective force was low in that situation, in spite of the long total branch length in the phylogenetic tree. Moreover, the results for 128 OTUs with a b value of 0.01 were generally better than those for 64 OTUs with a b value of 0.02, although the total branch length in the phylogenetic tree was almost the same, probably because the longer branch included more multiple substitutions which were not corrected for by the maximum-parsimony method. These results indicate that merely increasing the total branch length in the phylogenetic tree is not sufficient to improve the efficiency of the method. In conclusion, to improve the efficiency of the method, the total branch length should be increased, with the individual branches kept relatively short in the phylogenetic tree (for 128 OTUs, b should be less than 0.03).

Furthermore, another problem may arise if we use many sequences which are closely related to each other. That is, the topology of the phylogenetic tree may not be reliable, which may lead to the incorrect estimation of the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree for each codon site. That may eventually cause the incorrect inference of the selective force operating on each codon site. However, the computer simulation indicated that the results were almost identical in both the situation in which the phylogenetic relationship was assumed to be known and the situation in which it was assumed to be unknown (data not

shown). Therefore, the lack of information about the phylogenetic relationship may not affect the results seriously. However, there were two topologies examined in this study. Further extensive simulation studies with many topologies may be conducted to obtain final conclusions about the influence of the topology on the accuracy of detection of selective forces at single amino acid sites.

Some previous attempts have been made to detect positive selection at single amino acid sites. Fitch et al. (1997) used a multiple alignment of protein-coding sequences to reconstruct a phylogenetic tree. Then, for each codon site, they compared the total number of nonsynonymous changes throughout the phylogenetic tree with that of synonymous changes. They computed the exact binomial probability of obtaining the observed or more biased numbers of synonymous and nonsynonymous changes for each codon site. In their analyses, however, it was assumed that the probabilities of the occurrences of synonymous and nonsynonymous changes were constant for all codon sites, which may not hold in general. Specifically, to obtain their results, Fitch et al. (1997) computed the total numbers of synonymous (s_T) and nonsynonymous (n_T) changes throughout the phylogenetic tree over all codon sites. Then, $s_T/(s_T + n_T)$ and $n_T/(s_T + n_T)$ were used as the probabilities of the occurrences of synonymous and nonsynonymous changes for all codon sites, respectively. Nielsen and Yang (1998) developed a method using the maximum-likelihood approach. They divided codon sites into three categories: negatively selected, neutral, and positively selected. Then, they calculated the posterior probability that a particular codon site belonged to the positive-selection category. Their method could be used for distantly related sequence data. However, they assumed that the relative rate of nonsynonymous substitutions to synonymous substitutions was the same for all positively selected codon sites, which did not seem realistic. The method developed in the present study does not rely on the assumptions mentioned above. Moreover, the new method may require less computation and can handle larger data sets.

However, our method may be improved, to some extent, in the following manners. First, the transition/transversion rate bias and the base/codon frequency bias may be considered in the estimation of the numbers of synonymous and nonsynonymous sites and the numbers of those changes throughout the phylogenetic tree for each codon site, because those factors may affect the accuracy of the estimates (Ina 1995; Yang and Nielsen 1998). Indeed, our preliminary simulation studies taking these factors into account indicated that the number of synonymous substitutions tended to be overestimated, and that of nonsynonymous substitutions tended to be underestimated (data not shown), as was indicated previously (Ina 1995; Yang and Nielsen 1998). As a result, the true- and false-positive rates for detection of negative selection increased, and those for detection of positive selection decreased. However, our method should still be useful in these situations, particularly for detecting positively selected amino acid sites, because the false-positive rate for detection of positive selection would be very small. Second, the likelihood approach, instead of the maximum-parsimony method, may be used for inference of the most plausible ancestral codon at each node of the phylogenetic tree to reduce the number of combinations for possible ancestral codons over all nodes for one codon site (Yang, Kumar, and Nei 1995; Koshi and Goldstein 1996; Schultz, Cocroft, and Churchill 1996; Zhang and Nei 1997). Third, multiple substitutions on the long branches may be corrected for, to apply this method to distantly related sequence data.

Moreover, there are some restrictions associated with our method. In this method, the total number of nonsynonymous changes throughout the phylogenetic tree is compared with the number of synonymous changes for each codon site. Therefore, if positive selection has operated in an episodic manner on some branches in the phylogenetic tree (Messier and Stewart 1997), our method may fail to detect it. Our method is considered most effective in cases where positive selection has operated continuously or very strongly at some evolutionary period.

The simulation study indicated that a total branch length in the phylogenetic tree of ≥ 2.5 was sufficient to detect most of the positively selected amino acid sites. However, it was also shown that the false-positive rate for detection of the selective force was low, regardless of the number of OTUs and the branch length in the phylogenetic tree. These observations suggest that the method may be safely applied to gene sequences which do not have such a long total branch length in the phylogenetic tree. Indeed, the effectiveness of the method was supported by its application to the *HLA* gene, which had a total branch length of 1.06. On the other hand, sequence data in the international DNA data banks (DDBJ/EMBL/GenBank) are accumulating exponentially (Tateno and Gojobori 1997), and data concerning the diversity within a single species are systematically collected (Cavalli-Sforza et al. 1991). Therefore, we would have more gene sequences which have a long total branch length in the phylogenetic tree. We believe our method will become increasingly more useful in the future, particularly for predicting functions of amino acid sites in proteins.

Acknowledgments

We thank Dr. Ziheng Yang at the University College London, England, Dr. Andrew J. Leigh Brown at the University of Edinburgh, Scotland, Dr. Naoko Takezaki, Dr. Yumi Yamaguchi, and Ms. Rose Chapman at the National Institute of Genetics, Japan, and two anonymous reviewers for their helpful comments on this work. This work was supported, in part, by grants from the Ministry of Education, Science, Sports, and Culture of Japan. Y.S. was supported by the JSPS Research Fellowships for Young Scientists.

LITERATURE CITED

AKASHI, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in Drosophila DNA. Genetics **139**:1067–1076.

BAIROCH, A., and P. BUCHER. 1994. PROSITE: recent developments. Nucleic Acids Res. **22**:3583–3589.

BIZEBARD, T., B. GIGANT, P. RIGOLET, B. RASMUSSEN, O. DIAT, P. BOSECKE, S. A. WHARTON, J. J. SKEHEL, and M. KNOSSOW. 1995. Structure of influenza virus haemagglutinin complexed with a neutralizing antibody. Nature **376**:92–94.

BJORKMAN, P. J., M. A. SAPER, B. SAMRAOUI, W. S. BENNETT, J. L. STROMINGER, and D. C. WILEY. 1987. Structure of the human class I histocompatibility antigen, HLA-A2. Nature **329**:506–512.

CAVALLI-SFORZA, L. L., A. C. WILSON, C. R. CANTOR, R. M. COOK-DEEGAN, and M.-C. KING. 1991. Call for a worldwide survey of human genetic diversity: a vanishing opportunity for the human genome project. Genomics **11**:490–491.

CHESEBRO, B., K. WEHRLY, J. NISHIO, and S. PERRYMAN. 1992. Macrophage-tropic human immunodeficiency virus isolates from different patients exhibit unusual V3 envelope sequence homogeneity in comparison with T-cell-tropic isolates: definition of critical amino acids involved in cell tropism. J. Virol. **66**:6547–6554.

COMERON, J. M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41**:1152–1159.

CROW, J. F., and M. KIMURA. 1970. An introduction to population genetics theory. Harper and Row, New York, Evanston, and London.

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

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.

FOUCHIER, R. A., M. GROENINK, N. A. KOOTSTRA, M. TERSMETTE, H. G. HUISMAN, F. MIEDEMA, and H. SCHUITEMAKER. 1992. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. J. Virol. **66**:3183–3187.

GOJOBORI, T. 1983. Codon substitution in evolution and the ''saturation'' of synonymous changes. Genetics **105**:1011–1027.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

GUBAREVA, L. V., J. M. WOOD, W. J. MEYER, J. M. KATZ, J. S. ROBERTSON, D. MAJOR, and R. G. WEBSTER. 1994. Codominant mixtures of viruses in reference strains of influenza virus due to host cell variation. Virology **199**:89–97.

HARDY, C. T., S. A. YOUNG, R. G. WEBSTER, C. W. NAEVE, and R. J. OWENS. 1995. Egg fluids and cells of the chorioallantoic membrane of embryonated chicken eggs can select different variants of influenza A (H3N2) viruses. Virology **211**:302–306.

HARTIGAN, J. A. 1973. Minimum mutation fits to a given tree. Biometrics **29**:53–65.

HOLMES, E. C., L. Q. ZHANG, P. SIMMONDS, C. A. LUDLAM, and A. J. LEIGH BROWN. 1992. Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient. Proc. Natl. Acad. Sci. USA **89**:4835–4839.

HUGHES, A. L., and M. NEI. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**:167–170.

———. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. Proc. Natl. Acad. Sci. USA **86**:958–962.

INA, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**:190–226.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KIMURA, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

KOSHI, J. M., and R. A. GOLDSTEIN. 1996. Probabilistic reconstruction of ancestral protein sequences. J. Mol. Evol. **42**: 313–320.

LEE, Y.-H., and V. D. VACQUIER. 1992. The divergence of species-specific abalone sperm lysins is promoted by positive Darwinian selection. Biol. Bull. **182**:97–104.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36**: 96–99.

LI, W.-H., C.-I. WU, and C.-C. LUO. 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.

MCNEARNEY, T., Z. HORNICKOVA, R. MARKHAM, A. BIRDWELL, M. ARENS, A. SAAH, and L. RATNER. 1992. Relationship of human immunodeficiency virus type 1 sequence heterogeneity to stage of disease. Proc. Natl. Acad. Sci. USA **89**:10247–10251.

MESSIER, W., and C.-B. STEWART. 1997. Episodic adaptive evolution of primate lysozymes. Nature **385**:151–154.

METZ, E. C., and S. R. PALUMBI. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein Bindin. Mol. Biol. Evol. **13**:397–406.

MEYER, W. J., J. M. WOOD, D. MAJOR, J. S. ROBERTSON, R. G. WEBSTER, and J. M. KATZ. 1993. Influence of host cell-mediated variation on the international surveillance of influenza A (H3N2) viruses. Virology **196**:130–137.

MIYATA, T., and T. YASUNAGA. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J. Mol. Evol. **16**: 23–36.

MUSE, S. V., and B. S. GAUT. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11**:715–724.

NEI, M., and T. GOJOBORI. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol. Biol. Evol. **3**:418–426.

NIELSEN, R., and Z. YANG. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.

PAMILO, P., and O. N. BIANCHI. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. Mol. Biol. Evol. **10**:271–281.

ROCHA, E. P., X. XU, H. E. HALL, J. R. ALLEN, H. L. REGNERY, and N. J. COX. 1993. Comparison of 10 influenza A (H1N1 and H3N2) haemagglutinin sequences obtained directly from clinical specimens to those of MDCK cell- and egg-grown viruses. J. Gen. Virol. **74**:2513–2518.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SAUTER, N. K., J. E. HANSON, G. D. GLICK, J. H. BROWN, R. L. CROWTHER, S. J. PARK, J. J. SKEHEL, and D. C. WILEY.

1992. Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography. Biochemistry **31**:9609–9621.

SCHULTZ, T. R., R. B. COCROFT, and G. A. CHURCHILL. 1996. The reconstruction of ancestral character states. Evolution **50**:504–511.

SEIBERT, S. A., C. Y. HOWELL, M. K. HUGHES, and A. L. HUGHES. 1995. Natural selection on the *gag, pol,* and *env* genes of human immunodeficiency virus 1 (HIV-1). Mol. Biol. Evol. **12**:803–813.

SHIODA, T., S. OKA, S. IDA, K. NOKIHARA, H. TORIYOSHI, S. MORI, Y. TAKEBE, S. KIMURA, K. SHIMADA, and Y. NAGAI. 1994. A naturally occurring single basic amino acid substitution in the V3 region of the human immunodeficiency virus type 1 env protein alters the cellular host range and antigenic structure of the virus. J. Virol. **68**:7689–7696.

SMITH, D. B., and P. SIMMONDS. 1997. Characteristics of nucleotide substitution in the hepatitis C virus genome: constraints on sequence change in coding regions at both ends of the genome. J. Mol. Evol. **45**:238–246.

STEWART, C.-B., J. W. SCHILLING, and A. C. WILSON. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature **330**:401–404.

TATENO, Y., and T. GOJOBORI. 1997. DNA Data Bank of Japan in the age of information biology. Nucleic Acids Res. **25**: 14–17.

TATENO, Y., K. IKEO, T. IMANISHI et al. (13 co-authors). 1997. Evolutionary motif and its biological and structural significance. J. Mol. Evol. **44**:S38–S43.

TSUNOYAMA, K., and T. GOJOBORI. 1998. Evolution of nicotinic acetylcholine receptor subunits. Mol. Biol. Evol. **15**:518–527.

WHITFIELD, L. S., R. LOVELL-BADGE, and P. N. GOODFELLOW. 1993. Rapid sequence evolution of the mammalian sex-determining gene SRY. Nature **364**:713–715.

WILEY, D. C., I. A. WILSON, and J. J. SKEHEL. 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature **289**:373–378.

WOLFS, T. F. W., G. ZWART, M. BAKKER, M. VALK, C. L. KUIKEN, and J. GOUDSMIT. 1991. Naturally occurring mutations within HIV-1 V3 genomic RNA lead to antigenic variation dependent on a single amino acid substitution. Virology **185**:195–205.

YAMAGUCHI, Y., and T. GOJOBORI. 1997. Evolutionary mechanisms and population dynamics of the third variable envelope region of HIV within single hosts. Proc. Natl. Acad. Sci. USA **94**:1264–1269.

YANG, Z., S. KUMAR, and M. NEI. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141**:1641–1650.

YANG, Z., and R. NIELSEN. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46**:409–418.

YOKOYAMA, R., and S. YOKOYAMA. 1990. Convergent evolution of the red- and green-like visual pigment genes in fish, *Astyanax fasciatus,* and human. Proc. Natl. Acad. Sci. USA **87**:9315–9318.

ZHANG, J., and M. NEI. 1997. Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. J. Mol. Evol. **44**:S139–S146.

MASAMI HASEGAWA, reviewing editor

Accepted May 26, 1999