Reliabilities of Parsimony-based and Likelihood-based Methods for Detecting Positive Selection at Single Amino Acid Sites

Yoshiyuki Suzuki and Masatoshi Nei

Institute of Molecular Evolutionary Genetics, Department of Biology, The Pennsylvania State University

The reliabilities of parsimony-based and likelihood-based methods for inferring positive selection at single amino acid sites were studied using the nucleotide sequences of human leukocyte antigen (*HLA*) genes, in which positive selection is known to be operating at the antigen recognition site. The results indicate that the inference by parsimony-based methods is robust to the use of different evolutionary models and generally more reliable than that by likelihood-based methods. In contrast, the results obtained by likelihood-based methods depend on the models and on the initial parameter values used. It is sometimes difficult to obtain the maximum likelihood estimates of parameters for a given model, and the results obtained may be false negatives or false positives depending on the initial parameter values. It is therefore preferable to use parsimony-based methods as long as the number of sequences is relatively large and the branch lengths of the phylogenetic tree are relatively small.

Introduction

Positive selection is an evolutionary event in which a wild-type allele at a locus is replaced by a mutant allele with a higher fitness. At the DNA level, positive selection may be detected by comparing the rate of nonsynonymous (amino acid–altering) nucleotide substitution per nonsynonymous site (r_N) with that of synonymous substitution per synonymous site (r_S) (Hughes and Nei 1988, 1989). In a protein molecule, different amino acid sites have different biochemical functions, indicating that the type and strength of natural selection may vary among different amino acid sites. It is therefore important to detect positive selection at single amino acid sites.

For this purpose, Suzuki and Gojobori (1999) developed parsimony-based methods, and Yang et al. (2000) developed likelihood-based methods (see also Nielsen and Yang 1998) for comparing r_N and r_S at single codon sites by using a phylogenetic tree for protein-coding gene sequences. By using these methods, positively selected amino acid sites have been inferred in the human immunodeficiency virus envelope (Nielsen and Yang 1998; Yamaguchi-Kabata and Gojobori 2000), human leukocyte antigen (HLA) (Suzuki and Gojobori 1999), influenza virus hemagglutinin (Suzuki and Gojobori 1999; Yang 2000*a*; Yang et al. 2000), and others (e.g., Yang et al. 2000; Yang, Swanson, and Vacquier 2000; Swanson et al. 2001).

However, the reliabilities of the inference of these methods are not well understood, and the biochemical interpretation of the results obtained is sometimes difficult, particularly when positive selection is inferred at the amino acid sites without known functions. If the reliabilities of these methods are high enough, the inferred sites are likely to be important for adaptation, and experimental studies should be conducted to confirm the prediction. If the reliabilities of the methods are low, however, the inferred sites may simply represent false positives (incorrectly identified, positively selected sites) and further investigation may not be rewarding. We have therefore examined the reliabilities of the methods of both Suzuki and Gojobori (SG) and Yang et al. (Yang).

One approach for studying the reliabilities of these methods is to conduct computer simulation. Actually, Suzuki and Gojobori (1999) did such a simulation and showed that the probability of occurrence of false positives was generally low in their method and that of identifying truly selected sites increased as the strength of selection and the total branch length of the phylogenetic tree increased. Unfortunately, however, this approach is not applicable to the Yang method because it requires an enormous amount of computer time when the number of sequences is relatively large (Yang et al. 2000). Another approach is to examine the nucleotide sequences of a real protein for which positively selected amino acid sites are known from other information. Although only a small number of such proteins exist, HLA seems to be particularly suited for this purpose.

The HLA-A, -B, and -C proteins are expressed on the surface of most adult somatic cells in humans. The mature protein consists of three extracellular domains ($\alpha1$, $\alpha2$, and $\alpha3$), a transmembrane region, and a cytoplasmic region. This protein binds to an intracellularly processed antigenic peptide and presents it to $CD8^+$ T lymphocytes for eliciting immune responses (Klein and Horejsi 1997, pp. 87–159). Fifty-seven amino acid sites that are responsible for peptide binding have been identified in the $\alpha1$ and $\alpha2$ domains from the three-dimensional structure and called the antigen recognition sites (ARS) (Bjorkman et al. 1987*a*, 1987*b*). Biological and statistical evidence strongly suggests that positive selection is operating at the ARS (for review, see Hughes 1999, pp. 54–89). That is, as the ARS recognizes a wide variety of foreign antigenic peptides, amino acid mutations in the ARS may increase the fitness of an individual through overdominant selection (Doherty and Zinkernagel 1975). Moreover, Hughes and Nei (1988) showed that r_N is significantly higher than r_S at the codon sites encoding the ARS, whereas r_N is significantly lower than r_S at the codon sites encoding the non-ARS

Key words: positive selection, human leukocyte antigen, parsimony, likelihood.

Address for correspondence and reprints: Yoshiyuki Suzuki, Institute of Molecular Evolutionary Genetics, Department of Biology, The Pennsylvania State University, 328 Mueller Laboratory, University Park, Pennsylvania 16802. E-mail: yis1@psu.edu.

Mol. Biol. Evol. 18(12):2179–2185. 2001

region. Parham et al. (1988) also reported a high degree of amino acid polymorphism at the ARS. The *HLA* gene thus provides a unique opportunity for studying the reliabilities of the statistical methods of detecting positive selection.

In the present paper, we examined the reliabilities of detecting positively selected amino acid sites in the ARS by the SG and the Yang methods.

Materials and Methods

Methods

The SG and the Yang methods are intended to detect positive selection at single amino acid sites using a phylogenetic tree for protein-coding gene sequences. In the former method, the following algorithm is applied to each codon site (Suzuki 1999; Suzuki and Gojobori 1999). First, we infer ancestral codons at all the interior nodes of the phylogenetic tree by the maximum parsimony method (Fitch 1971; Hartigan 1973). We then compute the total numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions per codon site as well as the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites per codon site for the entire phylogenetic tree. The probabilities of occurrence of synonymous and nonsynonymous substitutions are approximated by $s_S/(s_S + s_N)$ and $s_N/(s_S + s_N)$, respectively, and the null hypothesis of selective neutrality is tested under the assumption that c_S and c_N are binomially distributed. If this null hypothesis is rejected and the relationship $c_N/s_N > c_S/s_S$ is observed, positive selection is inferred.

In the Yang method, the ratio of r_N to r_S for a given codon site is denoted by ω, and ω is assumed to follow a certain probability distribution among different codon sites. Fourteen different probability distributions of ω (M0–M13) have been proposed to be used, but M2, M3, and M8 were reported to give more reliable results than the others (Nielsen and Yang 1998; Yang et al. 2000). Model M2 is used in combination with M0 or M1. In M0, all codon sites are assumed to have the same ω value. In M1, codon sites are classified into categories 0 and 1, and ω is assumed to have 0 and 1 with probabilities P_0 and P_1 ($=1 - P_0$), respectively. M2 assumes an additional category (category 2) for which ω takes the value ω_2 with probability P_2 ($=1 - P_0 - P_1$). Under each model, a likelihood function is formulated using the codon substitution model, and free parameters are estimated by maximizing the likelihood (Yang and Nielsen 1998). If ω in M0 is estimated to be smaller than 1 and ω_2 in M2 is greater than 1, we test whether M2 fits the data better than M0 by the likelihood ratio test (LRT). We can also conduct the LRT for M2 using M1 as a null model. If the test is significant, we conclude that positively selected amino acid sites exist in the sequence. We then compute the posterior probability that a given codon site belongs to category 2 in M2. If the probability is higher than a given confidence probability level ($=1 -$ significance level), positive selection is inferred.

In model M3, a more general probability distribution of ω is used than in M2, and codon sites are classified into three categories with ω_0, ω_1, and ω_2, which are assumed to exist with probabilities P_0, P_1, and P_2 ($=1 - P_0 - P_1$), respectively. If any of the ω values estimated is greater than 1, we conduct the LRT against M0 or M1, and if the test is significant, positively selected amino acid sites are inferred in a way similar to that for M2.

In model M7, ω is assumed to follow a beta distribution with $0 \leq \omega \leq 1$. M8 is assumed to have an additional category, in which ω takes the value ω_1 with probability P_1 ($=1 - P_0$), where P_0 is the proportion following a beta distribution. If ω_1 is estimated to be greater than 1, we conduct the LRT against M0 or M7, and if the test is significant, positively selected amino acid sites are inferred.

Data Analysis

We used the same set of *HLA* sequences as that used by Suzuki and Gojobori (1999). The original data set was composed of 228 nucleotide sequences from the *HLA-A*, *-B*, and *-C* loci. Each sequence consisted of 273 codon sites that encoded the α1, α2, and α3 domains of HLA proteins. After excluding the identical sequences, we made a multiple alignment for a total of 218 sequences by the computer program CLUSTAL W (Thompson, Higgins, and Gibson 1994). The alignment did not contain any gaps. In the present paper, the amino acid positions in HLA are numbered according to Bjorkman et al. (1987*a*, 1987*b*).

In order to analyze *HLA* sequences by the SG method, a neighbor-joining (NJ) tree (Saitou and Nei 1987) was constructed with the number of synonymous substitutions (Nei and Gojobori 1986). It should be noted that the phylogenetic tree obtained may have topological and branch-length errors, because of the stochastic process of nucleotide substitution and the intralocus (Belich et al. 1992; Watkins et al. 1992) and interlocus (Pease et al. 1983; Weiss et al. 1983) gene conversions at the *HLA-A*, *-B*, and *-C* loci. However, it has been shown that the effect of interlocus gene conversion on the diversification of *HLA-A*, *-B*, and *-C* genes is generally small (Parham et al. 1988; Gu and Nei 1999), and minor errors in the phylogenetic tree do not affect the reliability of the SG method seriously (Suzuki and Gojobori 1999). We used the models of Jukes and Cantor (1969) and Kimura (1980) to compute s_S and s_N (Suzuki 1999). In Kimura's model, the transition-transversion ratio (R) was estimated as the ratio of the average number of transitional substitutions ($\bar{s}$) to that of transversional substitutions ($\bar{v}$) over all pairs of sequences which were estimated by Kimura's two-parameter method. The values of $\bar{s}$ and $\bar{v}$ were 0.033 and 0.035 per nucleotide site, respectively, and R was 0.95. It should be noted that the model of Jukes and Cantor is a special case of Kimura's model where $R = 0.5$. The significance level for rejecting selective neutrality was 5%.

In the Yang method, a maximum likelihood (ML) tree is supposed to be constructed. However, we failed to produce the ML tree because it required an enormous

Table 1
Numbers of Positively Selected Amino Acid Sites in HLA Identified by the SG and the Yang Methods[a]

Method	Initial ω Value	ln *L*	Number of Positively Selected Amino Acid Sites		ARS Index[d]
			ARS (57)[b]	Non-ARS (216)[c]	
SG (0.5, 0.95)[e]	N.A.[f]	N.A.	17	3 [0][g]	30.2
Yang (M0)[h]	0.2, 0.4, 0.6, 0.8, 1, 2, 3, 3.14, 4, 5	−9114.23	N.A.	N.A.	N.A.
Yang (M1)[i]	N.A.	−7759.34	N.A	N.A.	N.A.
Yang (M2)[j]	**0.2, 0.4, 0.6, 0.8, 1**	**−7637.75**	**0**	**0**	N.A.
	2, 3, 3.14, 4, 5	**−7593.90**	**14**	**4 [0]**	**17.3**
Yang (M3)[k]	0.2	−8332.38	36	42 [14]	7.1
	0.4	−8668.06	0	0	N.A.
	0.6	−8549.94	0	0	N.A.
	0.8	−8180.63	27	19 [4]	9.3
	1	−8327.32	29	19 [4]	10.7
	2	−8464.39	0	0	N.A.
	3	−8479.37	0	0	N.A.
	3.14	−8611.37	0	0	N.A.
	4	−8333.05	35	37 [9]	7.7
	5	−8378.32	31	26 [6]	8.7
Yang (M7)[l]	N.A.	−7803.40	N.A.	N.A.	N.A.
Yang (M8)[m].	0.2	−8260.79	0	0	N.A.
	0.4	−8114.22	0	0	N.A.
	0.6	**−7694.07**	**19**	**6 [0]**	**17.5**
	0.8	**−7732.55**	**17**	**6 [0]**	**14.9**
	1	**−7759.36**	**0**	**0**	N.A.
	2	−7831.09	25	6 [0]	27.3
	3	**−7770.29**	**22**	**4 [0]**	**33.3**
	3.14	−7857.20	25	19 [4]	8.1
	4	**−7772.56**	**24**	**5 [0]**	**30.7**
	5	**−7775.08**	**22**	**5 [0]**	**26.5**

[a] Results from Yang (M2) and Yang (M8), which show higher ln *L* values than Yang (M1) and Yang (M7), respectively, are boldfaced; results from Yang (M2), Yang (M3), and Yang (M8), which show the highest ln *L* value in each model, are underlined.

[b] The total number of codon sites in the ARS.

[c] The total number of codon sites outside the ARS.

[d] The ratio of the odds of detecting positively selected amino acid sites inside the ARS to those of detecting them outside the ARS. For the formula, see the text.

[e] SG method with the models of Jukes and Cantor ($R = 0.5$) and Kimura ($R = 0.95$).

[f] Not applicable.

[g] The number of positively selected amino acid sites identified in the α3 domain is indicated in brackets.

[h] Yang method with model M0.

[i] Yang method with model M1.

[j] Yang method with model M2.

[k] Yang method with model M3.

[l] Yang method with model M7.

[m] Yang method with model M8.

amount of computer time. For the same reason, we also failed to estimate even the branch lengths for a given topology using the ML method. We therefore constructed an NJ tree following Yang (2000*a*). The evolutionary distance was estimated by Kimura's method, and the branch lengths of the phylogenetic tree were multiplied by three because in the Yang method, the branch lengths are measured in terms of the number of nucleotide substitutions per codon site. It should be noted that minor errors in the phylogenetic tree do not affect the reliability of the Yang method seriously (Yang 2000*a*), as in the case of the SG method. The data analysis for the Yang method was conducted using the computer program PAML 3.0 (Yang 2000*b*). The observed codon frequencies were used as the equilibrium codon frequencies, and the ratio of the transition-transversion rates was estimated by maximizing the likelihood. Yang (2000*b*) noted that even with the same mathematical model, different results may be obtained depending on the initial ω values used, apparently because there are multiple local optima on the likelihood surface. He then recommended that two different initial ω values, one greater and the other smaller than unity, be used and the result with a higher likelihood value be regarded as the final result. In the present analysis, we used 0.2, 0.4, 0.6, 0.8, 1, 2, 3, 3.14, 4, and 5 as the initial ω values in each model. The significance level for the LRT was 5%, and the confidence probability level for inferring positive selection at single amino acid sites was 95%.

Results

The numbers of positively selected amino acid sites in HLA identified by the SG and the Yang methods are presented in table 1. The SG method with the model of Jukes and Cantor inferred 20 positively selected sites. These results appeared biologically reasonable because most (17) of the positively selected sites were located in the ARS, and all of the remaining 3 sites were in the α1 and α2 domains near the ARS (tables 1 and 2). When

Table 2
Positions of Positively Selected Amino Acid Sites Identified by the SG and the Yang Methods[a]

Domain	α1																								
Position	**9**	11	12	17	21	**24**	45	46	52	**62**	**63**	**66**	**67**	**69**	**70**	**71**	**73**	**74**	**76**	**77**	**80**	**81**	**82**	83	90
SG (0.5, 0.95)[b]	×									×	×		×	×	×		×		×	×	×	×	×	×	×
Yang (M2)[c]	×					×	×					×	×	×	×	×	×					×	×	×	
Yang (M3)[d]	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
Yang (M8)[e]	×	×				×	×			×		×	×	×	×	×	×	×	×		×	×	×	×	

[a] For the Yang method, only the results with the highest ln *L* value in each model are presented. Positively selected amino acid sites are indicated by "×". The amino acid sites involved in the ARS are boldfaced. The amino acid positions are numbered according to Bjorkman et al. (1987a, 1987b).

[b] SG method with the models of Jukes and Cantor ($R = 0.5$) and Kimura ($R = 0.95$) models.

[c] Yang method with model M2 using the initial ω values of 2, 3, 3.14, 4, and 5.

[d] Yang method with model M3 using the initial ω value of 0.8.

[e] Yang method with model M8 using the initial ω value of 0.6.

Table 2
Extended

Domain	α2																			α3			
Position	94	**95**	**97**	**99**	103	105	109	113	**114**	**116**	131	**143**	**147**	**151**	**152**	**156**	**163**	**167**	**171**	194	245	253	270
SG (0.5, 0.95)[b] ...		×	×		×					×					×		×						
Yang (M2)[c]	×							×	×					×	×			×					
Yang (M3)[d]	×		×	×	×	×	×	×	×		×	×	×	×	×	×	×	×	×	×	×	×	×
Yang (M8)[e]	×			×	×			×	×					×	×			×					

Kimura's model was used, positive selection was indicated at the same sites, although in this method, the observed value of R (=0.95) rather than $R = 0.5$ was used.

In the Yang method, different models produced different results. Moreover, even in a given model, different initial ω values produced different results, suggesting the existence of multiple local optima on the likelihood surface. When model M2 was used, all the initial ω values gave log-likelihood (ln L) values significantly greater than those in M1 ($P < 0.0001$), suggesting that M2 fitted the data better than M1 in all cases. The initial ω values of 2, 3, 3.14, 4, and 5 produced the same results and indicated that 18 amino acid sites were positively selected. These results appeared reasonable because most (14) of the positively selected sites were located in the ARS and none in the α3 domain. In contrast, the initial values of 0.2, 0.4, 0.6, 0.8, and 1 suggested that no positively selected sites existed in the HLA molecule. These results are obviously erroneous because we know that such sites exist in HLA. To sum up, half the initial ω values produced erroneous results.

When model M3 was used, all the initial ω values gave results with the ln L values significantly smaller than those in M1, suggesting that positively selected amino acid sites were absent in the sequence. These results suggest that the search for the ML value was trapped by local optima because M3 is a more general model than M1 and should give an ML value higher than that in M1. When we ignored the results from the LRT and attempted to detect positively selected sites, no such sites were inferred with half the initial ω values (0.4, 0.6, 2, 3, and 3.14). Positively selected sites were inferred with the remaining initial values (0.2, 0.8, 1, 4, and 5), but some of these sites were located in the α3 domain which was unrelated to the ARS (tables 1 and 2). These observations further suggest that all the initial ω values used here produced erroneous results.

In model M8, the initial ω values of 0.6, 0.8, 1, 3, 4, and 5 gave ln L values significantly greater than those in M7 ($P < 0.0001$). Positive selection was inferred at 22–29 amino acid sites for the initial values of 0.6, 0.8, 3, 4, and 5, and the results appeared reasonable. The initial value of $\omega = 1$, in contrast, indicated that positively selected sites were absent. The initial values of 0.2, 0.4, 2, and 3.14 gave ln L values significantly smaller than those in M7, suggesting the absence of positively selected sites. When we ignored this result and attempted to detect the positively selected sites, the initial value of $\omega = 2$ produced a reasonable result; but 0.2 and 0.4 indicated the absence of such sites, and 3.14 indicated the presence of positively selected sites in the α3 domain. Overall, half the initial ω values generated erroneous results.

In the above analyses by the Yang method, we used M1, M1, and M7 as null models for M2, M3, and M8, respectively, and the erroneous results in the inference of positively selected amino acid sites were all false negatives. However, when we used M0 as a null model, all the initial ω values in M2, M3, and M8 produced ln L values significantly greater than those in M0 ($P < 0.0001$). As the positively selected sites inferred in M3 with the initial ω values of 0.2, 0.8, 1, 4, and 5 and those in M8 with the initial $\omega = 3.14$ included false positives (in the α3 domain), any one of reasonable, false negative, or false positive results may be obtained as the final result in this case.

The reliabilities of the SG and the Yang methods were measured by the ARS index, which was defined as $n_{ARS}(216 - n_{non\text{-}ARS})/n_{non\text{-}ARS}(57 - n_{ARS})$, where n_{ARS} and $n_{non\text{-}ARS}$ denote the numbers of positively selected amino acid sites inside and outside the ARS, respectively (table 1). This index measures the ratio of the odds of detecting positively selected sites inside the ARS to those of detecting them outside the ARS (Sokal and Rohlf 1995, pp. 685–793). In the SG method, it was 30.2. In the Yang method, positively selected sites were inferred for five, five, and seven out of the 10 initial ω values used in models M2, M3, and M8, respectively. In M2, the ARS index is 17.3, which is lower than that in the SG method. M3 also produced the indices 7.1–10.7, which were all lower than that in the SG method. In M8, various ARS indices (8.1–33.3) were obtained, but the highest ln L value was associated with an index of 17.5, which was again lower than that in the SG method. These results indicate that the reliability of the SG method is similar to or higher than that of the Yang method.

The positively selected amino acid sites inferred by the SG and the Yang methods are presented in table 2. For the Yang method, only the results for the highest ln L value in each model are presented. The positively selected sites for model M2 were a subset of those for M8, which in turn were a subset of those for M3. Many of the positively selected sites inferred by the SG method were also inferred by the Yang method, and vice versa. These sites are likely to be truly selected because they are inferred by different methods based on different principles. It is interesting that all three positively selected sites inferred outside the ARS by the SG method were also inferred by the Yang method. In contrast, some of the positively selected sites inferred outside the ARS by the Yang method were not inferred by the SG method. These results further suggest that the SG method has a higher reliability than the Yang method.

Discussion

In the Yang method, we have to use multiple initial ω values to obtain the ML estimates of parameters for a given model, because of the presence of multiple local optima on the likelihood surface. The number of local optima appeared to increase as the number of free parameters increased. It has been recommended that two different initial ω values, one greater and the other smaller than unity, be used and the result with a higher likelihood value be regarded as the final result (Yang 2000*b*). However, this procedure does not necessarily work well. In the present analysis, all the 10 different initial ω values, of which five were greater and five were equal to or smaller than 1, gave different results in models M3 and M8. Thus, we may have to use as many

different initial ω values as possible to obtain the ML estimates, although there is still no guarantee that we shall always obtain the ML estimates, as shown with M3 in the present study. Moreover, the Yang method gives any one of reasonable, false negative, or false positive results that are statistically significant. We therefore have to be cautious about the results obtained by this method.

Even if we have obtained a putative ML estimate by using many different initial ω values, the reliability of detecting positively selected amino acid sites in the Yang method appears to be similar to or lower than that of the SG method, as mentioned earlier. It should be noted that many of the models used in the Yang method appear unrealistic. For example, ω is assumed to follow a certain probability distribution among different codon sites. Fourteen different probability distributions (M0–M13) have been proposed, of which M2, M3, and M8 were reported to give more reliable results than others (Nielsen and Yang 1998; Yang et al. 2000). However, the real distribution of ω is not known for any real protein and is likely to be more complicated than any of the proposed probability distributions. In particular, M2, M3, and M8 assume that all positively selected sites have the same ω value. This assumption is unrealistic because different amino acid sites have different biochemical functions in a protein, and thus the extent of positive selection should also vary among different amino acid sites. It should be noted that even if a more realistic model is developed, the reliability of the Yang method may not necessarily improve because such a model should include a large number of free parameters, as well as a large number of local optima on the likelihood surface. Moreover, in the Yang method, the pattern of codon substitution is assumed to be the same at all codon sites in a given category for the entire evolutionary time involved in the phylogenetic tree. Similarly, the equilibrium codon frequencies are assumed to be the same for all codon sites regardless of the category and evolutionary time. These assumptions are also unrealistic and unlikely to be correct (e.g., Zhang, Rosenberg, and Nei 1998). In addition, it has been shown that the LRT for molecular evolutionary hypotheses is often too liberal or too conservative when incorrect models are used (Zhang 1999), indicating the possibility that the LRT in the Yang method is biased.

In contrast, the SG method does not use any specific models but allows for any probability distribution for ω. This method also allows for different patterns of codon substitution and different equilibrium codon frequencies at different codon sites at any evolutionary time. This is because Suzuki and Gojobori use essentially model-free parsimony methods. A potential problem in this approach is that we do not take into account multiple substitutions at a nucleotide site for each branch. However, as long as the nucleotide sequences are relatively closely related, the number of multiple substitutions for a branch may be sufficiently small so that the results obtained appear to be reliable (Saitou 1989). Indeed, in the present analysis, the branch lengths of the phylogenetic tree were very small (on an average, 0.0025 per synonymous site), and the SG method produced reasonable results. In addition, underestimation of the number of nucleotide substitutions is likely to make the test of positive selection conservative by reducing the sample size. The underestimation is also likely to decrease the difference between the numbers of synonymous and nonsynonymous substitutions. Because a conservative test is generally more favorable than a liberal test in the study of molecular evolution (Nei and Kumar 2000, pp. 51–71), underestimation of the number of nucleotide substitutions may not be a serious problem.

In the present study, we have analyzed a large number of sequences. In order to see whether our observations are also applicable when the number of sequences is small, we conducted small-scale computer simulation. A nucleotide sequence with 300 codon sites was allowed to evolve following a symmetrical phylogenetic tree with 16 sequences, of which all branch lengths were set to be 0.1 per synonymous site. We assumed $\omega = 0$ at half the sites and $\omega = 1$ at the other half. Sixteen sequences were analyzed by the SG and the Yang methods. This procedure was repeated 200 times. In the SG method, positive selection was not inferred at a total of 30,000 amino acid sites with the true value of $\omega = 0$, but it was falsely inferred at 35 (0.1%) sites with the true value of $\omega = 1$. When we used M1, M1, and M7 as the null models for M2, M3, and M8, respectively, in the Yang method, the existence of positively selected sites was inferred in 2 (1%), 1 (0.5%), and 42 (21%) out of 200 replications, respectively. Positive selection was inferred at 0, 0, and 2 sites with the true value of $\omega = 0$ and at 0, 109 (0.4%), and 4,116 (13.7%) sites with the true value of $\omega = 1$ by M2, M3, and M8, respectively. When M0 was used as the null model, the existence of positively selected sites was inferred in 129 (64.5%), 143 (71.5%), and 147 (73.5%) replications by M2, M3, and M8, respectively. Positive selection was inferred at 0, 5, and 2 sites with the true value of $\omega = 0$, and at 5,394 (18.0%), 15,655 (52.2%), and 11,780 (39.3%) sites with the true value of $\omega = 1$ by M2, M3, and M8, respectively. These results indicate that the SG method is conservative, whereas the Yang method tends to give false positives with high probabilities, even when the number of sequences is small. Details of this study will be published elsewhere.

In conclusion, the reliability of parsimony-based methods is generally higher than that of likelihood-based methods as long as the number of sequences is relatively large and the branch lengths of the phylogenetic tree are relatively small. The likelihood-based methods also tend to give a substantial number of false positives for selective sites. It is also important to note that the results obtained from these methods are statistical inferences based on given models. The validity of the results obtained should be examined from the biological point of view and tested by experimental studies before drawing final conclusions.

Acknowledgments

We thank Ziheng Yang, Chen Su, and two anonymous reviewers for their valuable comments. This work

was supported by a grant from the National Institutes of Health (GM20293) to M.N. Y.S. is supported by the JSPS Research Fellowships for Young Scientists.

LITERATURE CITED

BELICH, M. P., J. A. MADRIGAL, W. H. HILDEBRAND, J. ZEMMOUR, R. C. WILLIAMS, R. LUZ, M. L. PETZL-ERLER, and P. PARHAM. 1992. Unusual *HLA-B* alleles in two tribes of Brazilian Indians. Nature **357**:326–329.

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

———. 1987*b*. The foreign antigen binding site and T cell recognition regions. Nature **329**:512–518.

DOHERTY, P. C., and R. M. ZINKERNAGEL. 1975. Enhanced immunological surveillance in mice heterozygous at the H-2 gene complex. Nature **256**:50–52.

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

GU, X., and M. NEI. 1999. Locus specificity of polymorphic alleles and evolution by a birth-and-death process in mammalian MHC genes. Mol. Biol. Evol. **16**:147–156.

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

HUGHES, A. L. 1999. Adaptive evolution of genes and genomes. Oxford University Press, Oxford.

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.

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. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KLEIN, J., and V. HOREJSI. 1997. Immunology. 2nd edition. Blackwell Science, Oxford.

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

NEI, M., and S. KUMAR. 2000. Molecular evolution and phylogenetics. Oxford University Press, Oxford.

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.

PARHAM, P., C. E. LOMEN, D. A. LAWLOR, J. P. WAYS, N. HOLMES, H. L. COPPIN, R. D. SALTER, A. M. WAN, and P. D. ENNIS. 1988. Nature of polymorphism in HLA-A, -B, and -C molecules. Proc. Natl. Acad. Sci. USA **85**:4005–4009.

PEASE, L. R., D. H. SCHULZE, G. M. PFAFFENBACH, and S. G. NATHENSON. 1983. Spontaneous H-2 mutants provide evidence that a copy mechanism analogous to gene conversion generates polymorphism in the major histocompatibility complex. Proc. Natl. Acad. Sci. USA **80**:242–246.

SAITOU, N. 1989. A theoretical study of the underestimation of branch lengths by the maximum parsimony principle. Syst. Zool. **38**:1–6.

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

SOKAL, R. R., and F. J. ROHLF. 1995. Biometry. 3rd edition. Freeman, New York.

SUZUKI, Y. 1999. Molecular evolution of pathogenic viruses. Doctoral dissertation, The Graduate University for Advanced Studies, Hayama, Japan.

SUZUKI, Y., and T. GOJOBORI. 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16**:1315–1328.

SWANSON, W. J., Z. YANG, M. F. WOLFNER, and C. F. AQUADRO. 2001. Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA **98**:2509–2514.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

WATKINS, D. I., S. N. MCADAMS, X. LIU et al. (13 co-authors). 1992. New recombinant *HLA-B* alleles in a tribe of South American Amerindians indicate rapid evolution of MHC class I loci. Nature **357**:329–333.

WEISS, E. H., A. L. MELLOR, L. GOLDEN, K. FAHRNER, E. SIMPSON, J. HURST, and R. A. FLAVELL. 1983. The structure of a mutant *H-2* gene suggests that the generation of polymorphism in *H-2* genes may occur by gene conversion-like events. Nature **301**:671–674.

YAMAGUCHI-KABATA, Y., and T. GOJOBORI. 2000. Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. J. Virol. **74**:4335–4350.

YANG, Z. 2000*a*. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J. Mol. Evol. **51**:423–432.

———. 2000*b*. Phylogenetic analysis by maximum likelihood (PAML). Version 3.0. University College London, London.

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

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

YANG, Z., W. J. SWANSON, and V. D. VACQUIER. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17**:1446–1455.

ZHANG, J. 1999. Performance of likelihood ratio tests of evolutionary hypotheses under inadequate substitution models. Mol. Biol. Evol. **16**:868–875.

ZHANG, J., H. F. ROSENBERG, and M. NEI. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc. Natl. Acad. Sci. USA **95**:3708–3713.

NARUYA SAITOU, reviewing editor

Accepted August 13, 2001