

## Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses

(synonymous and nonsynonymous substitutions/neutral theory of molecular evolution/positive selection)

YASUO INA AND TAKASHI GOJOBORI

National Institute of Genetics, Mishima 411, Japan

Communicated by Motoo Kimura, April 11, 1994

**ABSTRACT** To examine whether positive selection operates on the hemagglutinin 1 (*HA1*) gene of human influenza A viruses (H1 subtype), 21 nucleotide sequences of the *HA1* gene were statistically analyzed. The nucleotide sequences were divided into antigenic and nonantigenic sites. The nucleotide diversities for antigenic and nonantigenic sites of the *HA1* gene were computed at synonymous and nonsynonymous sites separately. For nonantigenic sites, the nucleotide diversities were larger at synonymous sites than at nonsynonymous sites. This is consistent with the neutral theory of molecular evolution. For antigenic sites, however, the nucleotide diversities at nonsynonymous sites were larger than those at synonymous sites. These results suggest that positive selection operates on antigenic sites of the *HA1* gene of human influenza A viruses (H1 subtype).

The evolution of influenza A viruses has been controversial (refs. 1 and 2 and references therein); some researchers argue that positive selection operates on the hemagglutinin 1 (*HA1*) gene of these viruses, particularly antigenic sites of the *HA1* gene product, but others claim that the *HA1* gene undergoes neutral evolution. Here we summarize the arguments of Fitch *et al.* (2) as the former and of Sugita *et al.* (3) as the latter.

Fitch *et al.* (2) concluded from the following three results that positive selection is responsible for the evolution of human influenza A viruses (H3 subtype). (i) Branch lengths in phylogenetic trees for genes of these viruses were short. (ii) The evolutionary rate of the *HA1* gene was higher than that of the nonstructural gene. It is thought that the *HA1* gene product is a target for host immune systems, whereas the nonstructural gene product is not so. (iii) Patterns of amino acid substitutions were different between surviving viruses and extinct ones. (According to the terminology of Fitch *et al.*, surviving viruses correspond to the trunk in a tree of the *HA1* gene, and extinct viruses correspond to branches in the tree.) For surviving viruses, amino acid substitutions in the *HA1* gene product occurred more frequently at antigenic sites than at nonantigenic sites. For extinct viruses, on the other hand, amino acid substitutions in the *HA1* gene product did not occur more frequently at antigenic sites than at nonantigenic sites. The difference in substitution patterns was statistically significant.

However, these three results can be also explained by other reasons. Although coalescence times or branch lengths in phylogenetic trees are shorter for advantageous mutations than for neutral ones (4), coalescence times can be small even under the assumption of neutrality when the effective population size is small. The effective population size of human influenza A viruses is possibly not so large because the number of patients infected with influenza A viruses is small out of season. Thus, result *i* can be also explained by the

neutral theory of molecular evolution (5, 6). Result *ii* can be also explained by the neutral theory. The neutral theory claims that evolutionary rates are negatively correlated with the degree of functional constraints (7). Thus, it is possible from the standpoint of the neutral theory that functional constraints are weaker for the *HA1* gene than for the nonstructural gene. To clarify whether positive selection or weaker functional constraints on the *HA1* gene are responsible for result *ii*, it is necessary to compare the numbers of synonymous ( $d_S$ ) and nonsynonymous ( $d_N$ ) substitutions per site for the *HA1* gene. It is expected that  $d_S \geq d_N$  under the neutral mutation hypothesis (6, 8). Result *iii* may not be reasonably explained by the neutral theory. By the maximum parsimony method, however, Fitch *et al.* assigned amino acid changes in the *HA1* gene product into branches or the trunk in a tree of the *HA1* gene. It is quite possible that errors were involved in the assignment. When divergent sequences are analyzed and branch lengths in a tree or evolutionary rates vary from lineage to lineage, the maximum parsimony method tends to be erroneous (9, 10). This will be the case for antigenic sites of the *HA1* gene because the amino acid sequences for these sites are quite divergent.

When Sugita *et al.* (3) estimated  $d_S$  and  $d_N$  at antigenic sites of the *HA1* gene of human influenza A viruses (H1 subtype) by the method of Nei and Gojobori (NG method; ref. 11), they did not find that the estimates of  $d_N$  were larger than those of  $d_S$  in most cases. Their results appear to be consistent with the neutral theory. However, recent computer simulation by one of us (39) has shown that the method of Miyata and Yasunaga (12), the method of Li *et al.* (13), and the NG method tend to give overestimates of  $d_S$  and underestimates of  $d_N$ . This indicates that a test by these methods is favorable for the neutral mutation hypothesis. Thus, the results of Sugita *et al.* should be reexamined by better methods for estimating  $d_S$  and  $d_N$ .

Recently, Pamilo and Bianchi (14) and Li (15) proposed a different method for estimating  $d_S$  and  $d_N$ . In this paper, we call their method the "PBL method." Alternative methods for estimating  $d_S$  and  $d_N$  also were developed by one of us (39); his methods ("Ina1 and Ina2") are extensions of the NG method, based on the assumption that nucleotide mutations and substitutions follow the two-parameter model of Kimura (16). The ratio of transitional mutation rate  $\alpha$  to transversional mutation rate  $\beta$  is estimated from the ratio of the transitional substitution rate to transversional substitution rate at the third nucleotide position of codons (method 1) or at synonymous sites (method 2). Computer simulation by one of us (39) has shown that the PBL method and the Ina1 and Ina2 methods give good estimates of  $d_S$  and  $d_N$ . Thus, by the PBL, Ina1, and Ina2 methods, we analyzed 21 nucleotide

Abbreviations:  $d_S$ , number of synonymous substitutions per site;  $d_N$ , number of nonsynonymous substitutions per site;  $L$ , length of sequences in codons;  $t$ , two times the divergence time; NG and PBL methods, methods of Nei and Gojobori and of Pamilo, Bianchi, and Li.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

sequences of the *HA1* gene of human influenza A viruses (H1 subtype) to examine whether positive selection operates on the *HA1* gene. In addition to the PBL, Ina1, and Ina2 methods, we also used the NG method to compare our results with those of Sugita *et al.*

A computer program that estimates  $d_S$  and  $d_N$  by the PBL method was provided by Wen-Hsiung Li (Center for Demographic and Population Genetics, University of Texas, Houston).

### SEQUENCE ANALYSIS

As described by Caton *et al.* (17) and Sugita *et al.* (3), we divided nucleotide sites of the *HA1* gene into antigenic and nonantigenic sites. We computed the nucleotide diversities (18) at antigenic and nonantigenic sites separately. The nucleotide diversities were computed at synonymous ( $\bar{d}_S$ ) and nonsynonymous ( $\bar{d}_N$ ) sites separately. Here we define nucleotide diversity as the average number of nucleotide substitutions per site.

Before we computed  $\bar{d}_S$  and  $\bar{d}_N$  by Ina1 and Ina2, we had estimated the ratio of transitional mutation rate ( $\alpha$ ) to transversional mutation rate ( $\beta$ ) in the following way. First, from the nucleotide sequences at nonantigenic sites, we computed the geometric mean of the  $\hat{\alpha}/\hat{\beta}$  ratio as 8.125 and 5.399 by Ina1 and Ina2, respectively. From these values, we computed  $\bar{d}_S$  and  $\bar{d}_N$  at antigenic and nonantigenic sites. Since antigenic sites are small (length  $L = 30$  codons),  $\bar{d}_S$  is probably biased unless the  $\hat{\alpha}/\hat{\beta}$  ratio is given.

Table 1 shows the nucleotide diversities at antigenic and nonantigenic sites of the *HA1* gene of human influenza A viruses (H1 subtype). The NG method gave the largest values of  $\bar{d}_S$  and the smallest values of  $\bar{d}_N$  at both antigenic and nonantigenic sites. The computer simulation results showed that probably the values of  $\bar{d}_S$  were 20–40% overestimated and the values of  $\bar{d}_N$  were about 10% underestimated by the NG method. On the other hand, the PBL, Ina1, and Ina2 methods gave similar values of  $\bar{d}_S$  and  $\bar{d}_N$ , particularly at

Table 1. Nucleotide diversities at antigenic and nonantigenic sites of the human influenza A virus (H1 subtype) *HA1* gene

	Methods			
	NG	PBL	Ina1 ( $\hat{\alpha}/\hat{\beta} = 8.125$ )	Ina2 ( $\hat{\alpha}/\hat{\beta} = 5.399$ )
Antigenic sites ( $L = 30$ codons)				
$n$	210	210	210	210
$\bar{d}_S$	0.090	0.082	0.068	0.071
$\bar{d}_N$	0.137	0.159	0.160	0.157
$\bar{d}_N/\bar{d}_S$	1.52	1.94*	2.35*	2.21*
Nonantigenic sites ( $L = 296$ codons)				
$n$	210	210	210	210
$\bar{d}_S$	0.118	0.090	0.091	0.095
$\bar{d}_N$	0.026	0.027	0.029	0.028
$\bar{d}_N/\bar{d}_S$	0.22	0.30	0.32	0.30

The number of replications was 2000 in the bootstrap test. Unreliable cases where  $\bar{d}_S \geq 10$  or  $\bar{d}_N \geq 10$  between some pairs of the nucleotide sequences were observed for the PBL, Ina1, and Ina2 methods, respectively, in 393 replications, 9 replications, and 4 replications of the bootstrap resampling. By excluding such replications, the bootstrap probabilities were computed. The bootstrap probabilities were also computed by excluding both  $\bar{d}_S$  and  $\bar{d}_N$  in a replication where  $\bar{d}_S \geq 10$  or  $\bar{d}_N \geq 10$  between a pair of the nucleotide sequences in the replication. However, both bootstrap probabilities were essentially the same. Sources of data were Hiti *et al.* (19), Winter *et al.* (20), Caton *et al.* (17), Nakajima *et al.* (21), Raymond *et al.* (22), Beklemishev *et al.* (23, 24), Concannon *et al.* (25), Robertson (26), Cox *et al.* (27), Rajakumar *et al.* (28), and Rocha *et al.* (29). GenBank accession numbers for unpublished data: D13573, D13574, and X59778.

\*,  $P > 0.95$ ;  $n$ , number of comparisons.

nonantigenic sites. It is worthwhile to point out that Ina1 and Ina2 gave essentially the same values of  $\bar{d}_S$  and  $\bar{d}_N$ , although the  $\hat{\alpha}/\hat{\beta}$  ratios used for these methods were quite different. This may be due to the fact that the  $\alpha/\beta$  ratio for the *HA1* gene is probably large.

It is clear that at nonantigenic sites,  $\bar{d}_S$  is larger than  $\bar{d}_N$ . This result indicates that even for rapidly evolving viruses, negative or purifying selection against amino acid changes operates on these sites. This is consistent with the neutral theory of molecular evolution. Hayashida *et al.* (30) and Saitou and Nei (31) also pointed out that negative selection predominates for the evolution of influenza A virus genes, although they did not separate nucleotide sites of the *HA1* gene into antigenic and nonantigenic sites. At antigenic sites, however, a different pattern was observed;  $\bar{d}_N$  is larger than  $\bar{d}_S$ . This picture does not depend on the methods used. Even by the NG method,  $\bar{d}_N$  is larger than  $\bar{d}_S$ . Note that the NG method gives overestimates of  $\bar{d}_S$  and underestimates of  $\bar{d}_N$ . Therefore, it is possible that positive selection operates on antigenic sites of the *HA1* gene of human influenza A viruses (H1 subtype).

To clarify whether the differences between  $\bar{d}_S$  and  $\bar{d}_N$  at antigenic sites can be explained by the neutral theory, it is necessary to conduct a statistical test. Since the method of Nei and Jin (32) assumes the Jukes and Cantor model (33), their method cannot be applied for computing the variances of  $\bar{d}_S$  and  $\bar{d}_N$  obtained by the PBL, Ina1, and Ina2 methods. This is because these methods do not assume the Jukes and Cantor model. Although the method of Takahata and Tajima (34) for computing the variance of nucleotide diversity does not depend on a specific substitution model, the statistical power of their method is low when the number of sites compared is small. In addition, the methods of both Nei and Jin and Takahata and Tajima assume that the topology of a reconstructed phylogenetic tree for compared nucleotide sequences and the branch lengths in the tree are correct. However, it is unlikely that this assumption holds for the present case. This is because only 30 codons are available for the present analysis of antigenic sites of the *HA1* gene. Furthermore, computer simulation has shown that when the number of sites compared is small, observed variances of  $\bar{d}_S$  and  $\bar{d}_N$  do not always agree with variances given by the  $\delta$  method (Y.I., unpublished data). This is the case even for simple methods such as the NG method. Thus, we conducted a statistical test by the bootstrap method (35). We resampled with replacement 30 codons from the alignment of the nucleotide sequences for antigenic sites of the *HA1* gene and then computed  $\bar{d}_S$  and  $\bar{d}_N$  for the resampled data by the NG, PBL, Ina1, and Ina2 methods. This procedure was repeated 2000 times, and then the bootstrap probability ( $P$ ) that  $\bar{d}_S < \bar{d}_N$  at antigenic sites of the *HA1* gene was computed.

The differences between  $\bar{d}_S$  and  $\bar{d}_N$  obtained by the PBL, Ina1, and Ina2 methods are statistically significant ( $P > 0.95$ ). The difference by the NG method is not statistically significant at the 95% bootstrap probability. However, as mentioned earlier, it is highly likely that the NG method overestimated  $\bar{d}_S$  and underestimated  $\bar{d}_N$ . Thus, the bootstrap test suggests that positive selection may operate on antigenic sites of the *HA1* gene.

Following Sugita *et al.* (3), we divided the 21 nucleotide sequences into two groups—namely, group A0 and group A1. Group A0 consists of the nucleotide sequences of three viruses that were isolated in the 1930s, whereas group A1 consists of the nucleotide sequences of the other 18 viruses. We further divided the 18 nucleotide sequences in group A1 into two subgroups; the nucleotide sequences of 11 viruses that were isolated before the first half of the 1980s and the nucleotide sequences of the other 7 viruses, which were isolated after the second half of the 1980s. For convenience, we call the former subgroup A1a and the latter subgroup A1b

in this paper. The validity of this classification was confirmed by a phylogenetic tree for the *HAI* gene, which was reconstructed by the neighbor-joining method (36) from estimates of  $d_S$  at nonantigenic sites. We computed the nucleotide diversities at antigenic sites of the *HAI* gene within and between the groups or the subgroups.

It is clear from Table 2 that the values of the  $\bar{d}_N/\bar{d}_S$  ratio within group A0 are smaller than those for antigenic sites in Table 1, whereas the values of the  $\bar{d}_N/\bar{d}_S$  ratio within group A1 are larger than those for antigenic sites in Table 1. In addition, the values of the  $\bar{d}_N/\bar{d}_S$  ratio between groups A0 and A1 are smaller than those for antigenic sites in Table 1. Moreover, Table 2 shows that the values of the  $\bar{d}_N/\bar{d}_S$  ratio

Table 2. Nucleotide diversities at antigenic sites of the human influenza A virus (H1 subtype) *HAI* gene within and between groups or subgroups

	NG	PBL	Ina1 $\hat{\alpha}/\hat{\beta} = 8.125$	Ina2 $\hat{\alpha}/\hat{\beta} = 5.399$
Within group A0				
<i>n</i>	3	3	3	3
$\bar{d}_S$	0.134	0.108	0.103	0.107
$\bar{d}_N$	0.123	0.144	0.142	0.139
$\bar{d}_N/\bar{d}_S$	0.92	1.33	1.38	1.30
Within group A1				
<i>n</i>	153	153	153	153
$\bar{d}_S$	0.022	0.032	0.016	0.017
$\bar{d}_N$	0.072	0.086	0.084	0.082
$\bar{d}_N/\bar{d}_S$	3.27*	2.69	5.25**	4.82**
Between groups A0 and A1				
<i>n</i>	54	54	54	54
$\bar{d}_S$	0.280 (0.202)	0.221 (0.151)	0.212 (0.153)	0.222 (0.160)
$\bar{d}_N$	0.320 (0.223)	0.365 (0.250)	0.377 (0.264)	0.368 (0.258)
$\bar{d}_N/\bar{d}_S$	1.14 (1.10)	1.65 (1.66)	1.78 (1.73)	1.66 (1.61)
Within subgroup A1a				
<i>n</i>	55	55	55	55
$\bar{d}_S$	0.027	0.039	0.020	0.021
$\bar{d}_N$	0.032	0.038	0.037	0.036
$\bar{d}_N/\bar{d}_S$	1.19	0.97	1.85	1.71
Within subgroup A1b				
<i>n</i>	21	21	21	21
$\bar{d}_S$	0.015	0.009	0.011	0.011
$\bar{d}_N$	0.039	0.050	0.046	0.045
$\bar{d}_N/\bar{d}_S$	2.60	5.56*	4.18	4.10
Between subgroups A1a and A1b				
<i>n</i>	77	77	77	77
$\bar{d}_S$	0.021	0.033	0.015	0.016
$\bar{d}_N$	0.110	0.131	0.128	0.125
$\bar{d}_N/\bar{d}_S$	5.24**	3.97*	8.53**	7.81**

Values in parentheses are the net differences given by equation 25 of Nei and Li (18) or the  $\bar{d}_N/\bar{d}_S$  ratios based on the net differences. The number of replications was 2000 in the bootstrap test. Unreliable cases, where  $\bar{d}_S \geq 10$  or  $\bar{d}_N \geq 10$  between some pairs of the nucleotide sequences, were observed for the PBL method for 4 replications of the bootstrap resampling within group A0, for 3 replications of the bootstrap resampling within group A1, and for 3 replications of the bootstrap resampling within subgroup A1b; unreliable cases were observed for the PBL, Ina1, and Ina2 methods, respectively, in 387 replications, 9 replications, and 4 replications of the bootstrap resampling between groups A0 and A1 and for the PBL method for 2 replications of the bootstrap resampling between subgroups A1a and A1b. By excluding such replications, the bootstrap probabilities were computed. The bootstrap probabilities were also computed by excluding both  $\bar{d}_S$  and  $\bar{d}_N$  in a replication where  $\bar{d}_S \geq 10$  or  $\bar{d}_N \geq 10$  between a pair of the nucleotide sequences in the replication. However, both bootstrap probabilities were essentially the same. \*,  $P > 0.95$ ; \*\*,  $P > 0.99$ ; *n*, number of comparisons.

within subgroup A1b are larger than those within subgroup A1a. Both of these values are smaller than the values within group A1 for all methods except the PBL method. This result implies that nonsynonymous substitutions occurred more frequently in the internodal branches leading to subgroups A1a and A1b than in the branches within these subgroups. Actually, we can see from Table 2 that for all methods but the PBL method, the values of the  $\bar{d}_N/\bar{d}_S$  ratio between subgroups A1a and A1b are larger than those within these subgroups. However, this result should be taken with caution because the values of  $\bar{d}_S$  and  $\bar{d}_N$  between subgroups A1a and A1b in Table 2 are not the net differences given by equation 25 of Nei and Li (18). Thus, these values contain not only substitutions in the internodal branches leading to subgroups A1a and A1b but also those in the branches within these subgroups. Furthermore, since the number of synonymous sites at antigenic sites is small, stochastic fluctuations for synonymous substitutions are so large that the values of  $\bar{d}_S$  within and between subgroups A1a and A1b are not so reliable. Actually, the 18 viruses in subgroups A1a and A1b were not clearly separated into two clusters by a neighbor-joining tree when estimates of  $d_S$  at antigenic sites were used. As a result, for all of the methods used, the values of  $\bar{d}_S$  within subgroup A1a are larger than those between subgroups A1a and A1b. Thus, it does not seem to be biologically meaningful to compute the net differences between these subgroups.

The differences between  $\bar{d}_S$  and  $\bar{d}_N$  within group A0, within subgroup A1a, and between groups A0 and A1 are not statistically significant at the 95% bootstrap probability by any method used. The difference within group A1 is statistically significant by the NG method ( $P > 0.95$ ) and by Ina1 and Ina2 ( $P > 0.99$ ). By the PBL method, the bootstrap probability of  $\bar{d}_S < \bar{d}_N$  is high ( $P > 0.94$ ). It is noteworthy to point out that even by the NG method, the difference is statistically significant. Thus, this result suggests that positive selection may operate on antigenic sites of the *HAI* gene of the viruses in group A1. Although the bootstrap probabilities of  $\bar{d}_S < \bar{d}_N$  within subgroup A1b are high and those between subgroups A1a and A1b are also very high, it seems to be difficult to interpret these results. This is because the problems as described earlier are involved in the values of  $\bar{d}_S$  and  $\bar{d}_N$  within and between subgroups A1a and A1b.

## DISCUSSION

Tajima (37) has shown that when the number of sites compared is small, ordinary algorithms such as the method of Jukes and Cantor (33) and the two-parameter method of Kimura (16) introduce systematic biases into estimation of the number of substitutions. Computer simulation by one of us (39) has shown that such biases were also observed for the NG, PBL, Ina1, and Ina2 methods. To confirm that the results shown in Tables 1 and 2 are not due to the biases, we conducted computer simulation by the method of Ina (39). The number of replications was 1000 in the computer simulation.

Table 3 shows that for all cases, the  $d_N/d_S$  ratio estimated by the NG method was  $< 1$  and that for all cases but  $t = 30$  (where  $t =$  two times the divergence time), the  $d_N/d_S$  ratios estimated by the PBL method and Ina2 were  $< 1$ . Table 3 also shows that for all cases but  $t = 30$  or 40, the  $d_N/d_S$  ratio estimated by Ina1 was  $< 1$ . Even in the case where the  $d_N/d_S$  ratio was  $> 1$ , the ratio was close to 1. For antigenic sites of the *HAI* gene, 180, 190, and 186 of 210  $\bar{d}_S$  values obtained by the PBL, Ina1, and Ina2 methods, respectively, were  $\leq 0.2$ . These results indicate that the expectations of the  $\bar{d}_N/\bar{d}_S$  ratio by the NG, PBL, Ina1, and Ina2 methods are  $< 1$  under the assumption of neutrality, even though estimates of  $d_S$  and  $d_N$  by these methods are biased. Note that the values of  $\bar{d}_S$  at antigenic sites in Table 1 are  $< 0.1$  and the values of  $\bar{d}_S$  in

Table 3. Means and standard deviations of  $\hat{d}_S$  and  $\hat{d}_N$  obtained by the NG, PBL, Ina1, and Ina2 methods under the simulation scheme of influenza virus gene mutation, no selection, and L = 30 codons

	Expectation	Methods			
		NG	PBL	Ina1	Ina2
		$t = 10$			
$n$	—	1000	1000	1000	1000
$\hat{d}_S$	0.100	0.145 ± 0.093	0.117 ± 0.081	0.109 ± 0.070	0.112 ± 0.071
$\hat{d}_N$	0.100	0.095 ± 0.038	0.110 ± 0.045	0.109 ± 0.044	0.108 ± 0.044
$\hat{d}_N/\hat{d}_S$	1.000	0.659	0.941	0.995	0.966
		$t = 20$			
$n$	—	1000	994	1000	1000
$\hat{d}_S$	0.201	0.293 ± 0.161	0.226 ± 0.124	0.226 ± 0.123	0.231 ± 0.126
$\hat{d}_N$	0.201	0.186 ± 0.054	0.218 ± 0.068	0.216 ± 0.067	0.215 ± 0.066
$\hat{d}_N/\hat{d}_S$	1.000	0.635	0.962	0.956	0.931
		$t = 30$			
$n$	—	999	989	1000	999
$\hat{d}_S$	0.301	0.411 ± 0.201	0.325 ± 0.175	0.319 ± 0.163	0.326 ± 0.162
$\hat{d}_N$	0.301	0.284 ± 0.071	0.336 ± 0.095	0.339 ± 0.094	0.335 ± 0.092
$\hat{d}_N/\hat{d}_S$	1.000	0.690	1.034	1.062	1.026
		$t = 40$			
$n$	—	1000	959	997	995
$\hat{d}_S$	0.402	0.572 ± 0.292	0.455 ± 0.209	0.445 ± 0.250	0.454 ± 0.247
$\hat{d}_N$	0.402	0.365 ± 0.085	0.444 ± 0.128	0.451 ± 0.124	0.447 ± 0.122
$\hat{d}_N/\hat{d}_S$	1.000	0.638	0.976	1.015	0.984
		$t = 50$			
$n$	—	991	917	985	982
$\hat{d}_S$	0.502	0.742 ± 0.394	0.603 ± 0.266	0.567 ± 0.297	0.580 ± 0.311
$\hat{d}_N$	0.502	0.444 ± 0.095	0.560 ± 0.163	0.566 ± 0.155	0.560 ± 0.152
$\hat{d}_N/\hat{d}_S$	1.000	0.599	0.929	0.998	0.965

Expectation refers to the expected values of  $\hat{d}_S$ , the expected value of  $\hat{d}_N$ , or the ratio of the expected value of  $\hat{d}_N$  to the expected value of  $\hat{d}_S$ . Means and standard deviations of  $\hat{d}_S$  and  $\hat{d}_N$  were calculated by excluding inapplicable cases. Values for the Ina1 and Ina2 methods were obtained when the  $\hat{\alpha}/\hat{\beta}$  ratio was given. The  $\hat{\alpha}/\hat{\beta}$  ratios were estimated by Ina1 and Ina2 under the corresponding simulation scheme for L = 290 codons. The geometric means of  $\hat{\alpha}/\hat{\beta}$  for 100 pairs of nucleotide sequences were used to estimate  $\hat{d}_S$  and  $\hat{d}_N$  by Ina1 and Ina2. n, number of applicable cases; t, two times the divergence time.

Table 2 are <0.3. Thus, it seems to be unlikely that the results shown in Tables 1 and 2 are artifacts due to biased estimates of  $\hat{d}_S$  and  $\hat{d}_N$  by the NG, PBL, Ina1, and Ina2 methods.

Our results show that for antigenic sites of the HA1 gene, the number of nonsynonymous substitutions is larger than that of synonymous substitutions, whereas the results of Sugita *et al.* (3) showed that even for antigenic sites of the HA1 gene, the number of nonsynonymous substitutions was not larger than that of synonymous substitutions. The difference is mainly explained by the following fact. Although Sugita *et al.* used 21 nucleotide sequences of the HA1 gene making the number of possible comparisons 210 (= 21 × 20/2), they showed only 26 points representing the values of  $\hat{d}_S$  and  $\hat{d}_N$  in their figure 5 of ref. 3. All of the 26 points show that  $\hat{d}_S > 0$  and  $\hat{d}_N > 0$ . However, the value of  $\hat{d}_S$  is 0 for some comparisons, particularly comparisons within groups. This is because the number of synonymous sites is small, as described earlier. Actually, when we compared the 18 nucleotide sequences within group A1, the value of  $\hat{d}_S$  was 0 for 70 comparisons. On the other hand, the value of  $\hat{d}_N$  was 0 for 7 comparisons of the nucleotide sequences within group A1. Sugita *et al.* eliminated points of  $\hat{d}_S = 0$  and  $\hat{d}_N > 0$ . As a result, the data of Sugita *et al.* based on the 26 points as shown in figure 5 of ref. 3 are substantially biased. Furthermore, when they compared nucleotide sequences between groups A0 (two sequences) and A1 (19 sequences), only 14 comparisons were made, although the number of possible comparisons is 38 (= 2 × 19). On the other hand, we used all of the 21 nucleotide sequences. Thus, our results are considered to be more reliable than those of Sugita *et al.* In addition to the above fact, there is another reason for the difference between our results and those of Sugita *et al.* Some of the nucleotide

sequences used in the present study are different from those used in the Sugita *et al.* study. We used the nucleotide sequences of seven viruses that had been isolated after the second half of the 1980s, whereas Sugita *et al.* did not use these sequences.

We showed the following three results in Tables 1 and 2. (i) For antigenic sites of the HA1 gene of 21 human influenza A viruses (H1 subtype), the values of  $\hat{d}_N$  were larger than those of  $\hat{d}_S$ , and the difference between the values of  $\hat{d}_S$  and  $\hat{d}_N$  was statistically significant (by the PBL, Ina1, and Ina2 methods). (ii) The values of the  $\hat{d}_N/\hat{d}_S$  ratio within group A1 were larger than those within group A0 and between groups A0 and A1. The difference between the values of  $\hat{d}_S$  and  $\hat{d}_N$  was statistically significant only within group A1 (by the NG, Ina1, and Ina2 methods). (iii) The values of the  $\hat{d}_N/\hat{d}_S$  ratio between subgroups A1a and A1b were larger than those within these subgroups. Furthermore, the values of the  $\hat{d}_N/\hat{d}_S$  ratio within subgroup A1b were much larger than those within subgroup A1a. The difference between the values of  $\hat{d}_S$  and  $\hat{d}_N$  was statistically significant within subgroup A1b (by the PBL method) and between subgroups A1a and A1b (by all methods). These three results suggest the possibility that the evolutionary pattern of the HA1 gene varies from virus to virus. Positive selection may have been responsible for the evolution of the HA1 gene of only viruses belonging to group A1, particularly to viruses belonging to subgroup A1b since the divergence of subgroups A1a and A1b occurred. Nonsynonymous substitutions may have occurred much more frequently in the internodal branches between subgroups A1a and A1b than in the branches within these subgroups, whereas nonsynonymous substitutions may not have occurred much more frequently in the internodal

branches between groups A0 and A1 than in the branches within these groups. This is inconsistent with the arguments of Fitch *et al.* (2). The inconsistency between our results and those of Fitch *et al.* may represent the difference of the evolution between subtypes H1 and H3. However, the above interpretation of results *i-iii* should be taken with caution because the problems as described earlier are involved in result *iii*.

Results *i-iii* can also be interpreted as follows. As pointed out by Fitch *et al.*, the result that the number of nonsynonymous substitutions is not statistically larger than that of synonymous substitutions does not preclude the possibility of positive selection. This is because positive selection may not operate on all antigenic sites of the *HA1* gene. It is possible that functional constraints exist in some of the antigenic sites of the *HA1* gene. Actually, the affinity of hemagglutinin for viral receptors is affected by amino acid changes in antigenic sites of the *HA1* gene product (38). This implies that even in antigenic sites of the *HA1* gene, the fraction of nondeleterious mutations at nonsynonymous sites is not so large and that nonsynonymous substitutions occur at restricted sites. If this were the case, the number of nonsynonymous substitutions between distantly related sequences would be underestimated by methods that assume the uniform substitution rate among nonsynonymous sites. Thus, it is expected that the values of the  $\bar{d}_N/\bar{d}_S$  ratio between (sub)groups are smaller than those within the (sub)groups. As a result, it is possible that even though positive selection operates more strongly in the internodal branches between (sub)groups than in the branches within the (sub)groups, the difference between the values of  $\bar{d}_S$  and  $\bar{d}_N$  is statistically significant within the (sub)groups but not between the (sub)groups. Therefore, the result that we could not find any convincing evidence for the arguments of Fitch *et al.* (2) may be attributed to the methods used. This is because all of the methods used assume the uniform substitution rate among nonsynonymous sites.

Although it is now unclear which of the above interpretations of results *i-iii* is correct, it is likely that positive selection operates on antigenic sites of the *HA1* gene of human influenza A viruses (H1 subtype). Probably, some of the amino acid changes in antigenic sites of the *HA1* gene product are advantageous to the escape of human influenza A viruses from host immune systems. Thus, the present paper clearly shows the advantageous evolution of viruses at the molecular level by refined statistical methods of molecular evolutionary studies. To clarify in more detail the evolution of the *HA1* gene of influenza A viruses, much more sequence data are required. In particular, nucleotide sequences of influenza A viruses not only for human but also for animals should be analyzed. Moreover, it is of interest to examine whether the neuraminidase (sialidase) gene shows the same pattern of  $\bar{d}_S < \bar{d}_N$ , since it is known that the product of this gene is also recognized by host immune systems.

We thank Drs. T. Ohta and F. Tajima for their helpful suggestions and comments during the course of this study. We are also grateful to Dr. T. Ohta, who provided us with computing facilities.

1. Gojobori, T., Moriyama, E. N. & Kimura, M. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 10015–10018.

2. Fitch, W. M., Leiter, J. M. E., Li, X. & Palese, P. (1991) *Proc. Natl. Acad. Sci. USA* **88**, 4270–4274.
3. Sugita, S., Yoshioka, Y., Itamura, S., Kanegae, Y., Oguchi, K., Gojobori, T., Nerome, K. & Oya, A. (1991) *J. Mol. Evol.* **32**, 16–23.
4. Takahata, N. (1990) in *Population Biology of Genes and Molecules*, eds. Takahata, N. & Crow, J. F. (Baifukan, Tokyo), pp. 267–286.
5. Kimura, M. (1968) *Nature (London)* **217**, 624–626.
6. Miyata, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, UK).
7. Kimura, M. & Ohta, T. (1974) *Proc. Natl. Acad. Sci. USA* **71**, 2848–2852.
8. Kimura, M. (1977) *Nature (London)* **267**, 275–276.
9. Felsenstein, J. (1978) *Syst. Zool.* **27**, 401–410.
10. Saitou, N. & Imanishi, T. (1989) *Mol. Biol. Evol.* **6**, 514–525.
11. Nei, M. & Gojobori, T. (1986) *Mol. Biol. Evol.* **3**, 418–426.
12. Miyata, T. & Yasunaga, T. (1980) *J. Mol. Evol.* **16**, 23–36.
13. Li, W.-H., Wu, C.-I. & Luo, C.-C. (1985) *Mol. Biol. Evol.* **2**, 150–174.
14. Pamilo, P. & Bianchi, N. O. (1993) *Mol. Biol. Evol.* **10**, 271–281.
15. Li, W.-H. (1993) *J. Mol. Evol.* **36**, 96–99.
16. Kimura, M. (1980) *J. Mol. Evol.* **16**, 111–120.
17. Caton, A., Brownlee, C. G., Yewdell, J. W. & Gerhard, W. (1982) *Cell* **31**, 417–427.
18. Nei, M. & Li, W.-H. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 5269–5273.
19. Hiti, A. L., Davis, A. R. & Nayak, D. P. (1981) *Virology* **111**, 113–124.
20. Winter, G., Fields, S. & Brownlee, G. G. (1981) *Nature (London)* **292**, 72–75.
21. Nakajima, S., Nakajima, K. & Kendal, A. P. (1983) *Virology* **131**, 116–127.
22. Raymond, F. L., Caton, A. J., Cox, N. J., Kendal, A. P. & Brownlee, G. G. (1983) *Nucleic Acids Res.* **11**, 7191–7203.
23. Beklemishev, A. B., Blinov, V. M., Vassilendo, S. K., Golovin, S. Y., Gutorov, V. V., Karginov, V. A., Mamaev, L. V., Mikryukov, N. N., Netesov, S. V., Petrenko, V. A., Petrov, N. A. & Sandakhchiev, L. S. (1984) *Bioorg. Khim.* **10**, 1535–1543.
24. Beklemishev, A. B., Blynov, V. M., Vassilenko, S. K., Golovin, S. Y., Karginov, V. A. & Mamayev, L. V. (1986) *Bioorg. Khim.* **12**, 375–381.
25. Concannon, P., Cummings, I. W. & Salser, W. A. (1984) *J. Virol.* **49**, 276–278.
26. Robertson, J. S. (1987) *J. Gen. Virol.* **68**, 1205–1208.
27. Cox, N. J., Black, R. A. & Kendal, A. P. (1989) *J. Gen. Virol.* **70**, 299–313.
28. Rajakumar, A., Swierkosz, E. M. & Schulze, I. T. (1990) *Proc. Natl. Acad. Sci. USA* **87**, 4154–4158.
29. Rocha, E., Cox, N. J., Black, R. A., Harmon, M. W., Harrison, C. J. & Kendal, A. P. (1991) *J. Virol.* **65**, 2340–2350.
30. Hayashida, H., Toh, H., Kikuno, R. & Miyata, T. (1985) *Mol. Biol. Evol.* **2**, 289–303.
31. Saitou, N. & Nei, M. (1986) *Mol. Biol. Evol.* **3**, 57–74.
32. Nei, M. & Jin, L. (1989) *Mol. Biol. Evol.* **6**, 240–300.
33. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro, H. N. (Academic, New York), pp. 21–132.
34. Takahata, N. & Tajima, F. (1991) *Mol. Biol. Evol.* **8**, 494–502.
35. Efron, B. (1982) *The Jackknife, the Bootstrap and Other Sampling Plans* (Soc. Ind. Appl. Math., Philadelphia).
36. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.
37. Tajima, F. (1993) *Mol. Biol. Evol.* **10**, 677–688.
38. Suzuki, Y., Kato, H., Naeve, C. W. & Webster, R. G. (1989) *J. Virol.* **63**, 4298–4302.
39. Ina, Y. (1994) *J. Mol. Evol.*, in press.