Genes Genet. Syst. (2011) **86**, p. 287–294

Positive selection for gains of N-linked glycosylation sites in hemagglutinin during evolution of H3N2 human influenza A virus

Yoshiyuki Suzuki*

Graduate School of Natural Sciences, Nagoya City University, 1 Yamanohata, Mizuho-cho, Mizuho-ku, Nagoya-shi, Aichi-ken 467-8501, Japan

(Received 9 October 2011, accepted 15 November 2011)

The number of N-linked glycosylation sites in the globular head of hemagglutinin (HA) has increased during evolution of H3N2 human influenza A virus. Here natural selection operating on the gains of N-linked glycosylation sites was examined by using the single-site analysis and the single-substitution analysis. In the single-site analysis, positive selection was not inferred at the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, but was detected at antigenic sites. In contrast, in the single-substitution analysis, positive selection was detected for the amino acid substitutions generating N-linked glycosylation sites. The single-site analysis and the single-substitution analysis appeared to be suitable for detecting recurrent and episodic natural selection, respectively. The gains of N-linked glycosylation sites were likely to be positively selected for the function of shielding antigenic sites from immune responses. At the antigenic sites, positive selection appeared to have operated not only on the radical substitution but also on the conservative substitution in terms of the charge of amino acids, suggesting that the antigenic drift is not a by-product of the evolution of receptor binding avidity in HA of human H3N2 virus.

Key words: influenza A virus, N-linked glycosylation site, positive selection, single-site analysis, single-substitution analysis

INTRODUCTION

Influenza A virus is classified as the genus *Influenzavirus A* in the family *Orthomyxoviridae* (Carstens, 2010). The genome of this virus is an eight-segmented and negative-stranded RNA, encoding envelope glycoproteins, matrix proteins, nonstructural proteins, nucleoproteins, and polymerase subunits. The envelope glycoproteins include hemagglutinin (HA) and neuraminidase (NA), with HA existing 4–5 times more abundantly than NA in virions. According to the antigenicity of HA and NA, influenza A virus is classified into subtypes H1–H16 and N1–N9, respectively (World Health Organization, 1980).

Influenza A virus is an etiological agent of influenza (Shope, 1931). Genomic sequence data of this virus are available for the strains circulating since 1918. In the human population, H1N1 virus circulated in 1918–1956, followed by H2N2 virus in 1957–1967. Subsequently, H3H2 virus has been circulating in 1968-present. In addition, H1N1 virus closely related to that observed in early 1950s reappeared in 1977 and circulated thereafter, and another H1N1 virus (A(H1N1)pdm09) emerged in 2009 (World Health Organization, 2011). A(H1N1)pdm09 virus has been circulating as the predominant H1N1 virus in 2009-present.

HA is a homotrimeric type I transmembrane glycoprotein. The HA gene encodes HA0, consisting of a signal peptide (amino acid sites [–16]–[–1] in H3N2 virus), HA1 (1–328), and HA2 (330–550) (Skehel and Wiley, 2000). Signal peptide directs the co-translational transport of HA into the endoplasmic reticulum (ER). HA1 is the sialic acid receptor-binding protein and the major target of humoral immunity. HA2 is an anchor protein to the envelope and mediates fusion of the envelope and the endosomal membrane.

The ectodomain of HA is composed of the globular head (amino acid sites 58–272 of HA1) and the fibrous stem (1–57 and 273–328 of HA1 and 330–514 of HA2) (Wilson et al., 1981). Antigenic sites are distributed in the globular head, constituting 5 epitopes (A–E) in H3N2 virus (Suzuki, 2004b). In the ectodomain, asparagine [N] of the sequon, which is defined as the sequence of N, any amino acid except for proline [P], and serine [S] or threo-

Edited by Fumio Tajima
* Corresponding author. E-mail: yossuzuk@nsc.nagoya-cu.ac.jp

nine [T], mostly serves as an N-linked glycosylation site, where high-mannose or complex oligosaccharide is attached with various forms (Hebert et al., 1997; Schulze, 1997; Daniels et al., 2003; Abe et al., 2004; Blackburne et al., 2008; Das et al., 2010). O-linked glycosylation has not been reported for influenza A virus.

The N-linked glycan is involved in the folding of ectodomain in the ER lumen by binding to lectin chaperones (Hebert et al., 1997; Daniels et al., 2003; Cui et al., 2009). N-linked glycans in the fibrous stem are involved in the fusion activity of HA. In the globular head, N-linked glycosylation sites usually overlap with antigenic sites. N-linked glycans may be involved in shielding of antigenic sites from binding by antibodies (Skehel et al., 1984; Tsuchiya et al., 2002; Das et al., 2010; Wei et al., 2010; Wanzeck et al., 2011) and major histocompatibility complex (Jackson et al., 1994), interference with proteolytic activity of HA, and recognition by collectins for neutralization (Vigerust et al., 2007). In addition, structural complexity of N-linked glycans is positively and negatively correlated with HA-receptor binding specificity and affinity, respectively (Tsuchiya et al., 2002; Wang et al., 2009; de Vries et al., 2010).

N-linked glycosylation sites in the fibrous stem are usually conserved among influenza A viruses (Sun et al., 2011). However, in the globular head of H1N1 virus, the number of N-linked glycosylation sites, which was 1 in 1918, increased up to 6 and is 4 at present (Igarashi et al., 2008; Das et al., 2010; Sun et al., 2011). Although the number of N-linked glycosylation sites remained 1 in 1957–1967 for H2N2 virus (Tsuchiya et al., 2001, 2002; Abe et al., 2004), the number increased from 2 to 6 or 7 in 1968-present for H3N2 virus (Abe et al., 2004).

It was unclear whether the increase in the number of N-linked glycosylation sites observed in human H1N1 and H3N2 viruses was due to neutral evolution (Zhang et al., 2004) or positive selection. In general, the number of N-linked glycosylation sites is negatively correlated with the GC-content, because N, which is included in the sequon, is encoded by GC-poor (AAY) codons and P, which is not included in the sequon, is encoded by GC-rich (CCN) codons (Cui et al., 2009). Although the GC-content of influenza A virus has decreased in humans because of mutation bias (Rabadan et al., 2006) and natural selection against CpG dinucleotides (Greenbaum et al., 2008), the observed increase in the number of N-linked glycosylation sites appeared to exceed the expectation from the decrease in the GC-content (Cui et al., 2009).

In the phylogenetic tree for HA of H3N2 virus, it was observed that the gain/loss ratio of N-linked glycosylation sites was greater in the branches more proximal to the root (i.e., trunk branches > non-trunk interior branches > non-trunk exterior branches) (Cherry et al., 2009). From this observation, positive selection was inferred for gains of N-linked glycosylation sites. However, since the occurrences of gains and losses are interdependent (e.g., N-linked glycosylation sites should be gained in proximal branches to be lost in distal branches), the null hypothesis of the equal gain/loss ratio for proximal and distal branches without positive selection may not hold.

Positive selection was also inferred for the amino acid substitutions generating N-linked glycosylation sites by DEPS, which is the method for detecting directional evolution of protein sequences (Kosakovsky Pond et al., 2008). In this method, asymmetrical substitution rates between pairs of amino acids are identified at each site, assuming that the pattern of substitution is the same at all sites without positive selection, which appears to be unrealistic. For detecting asymmetrical substitution rates, it is required that a number of substitutions occurred between pairs of amino acids at single sites. DEPS has been reported to produce many false-positives (Nozawa et al., 2009).

The purpose of the present study was to examine natural selection operating on the gains of N-linked glycosylation sites in the globular head of HA during evolution of human H3N2 virus by using the single-site analysis and the single-substitution analysis. The human H3N2 virus was analyzed because of the clinical importance as currently circulating in the human population, and of the availability of sequence data containing a sufficiently large amount of genetic variation for the statistical analysis.

MATERIALS AND METHODS

Sequence data A total of 3,206 nucleotide sequences for the entire protein-coding region of HA for human H3N2 virus, excluding laboratory and vaccine strains, were retrieved from the Influenza Virus Resource at the National Center for Biotechnology Information (Bao et al., 2008) as of May 19, 2011. After eliminating sequences for the same strains as others, sequences identical to others, sequences derived from incidental human infections of swine strains, and sequences with minor gaps, ambiguous nucleotides, and premature termination codons, 2,043 sequences were used in the following analysis. A sequence from duck H3N8 virus was added as the outgroup to identify the position of the root for the phylogenetic tree of human sequences. Each sequence consisted of 1,688 nucleotide sites.

Phylogenetic analysis Multiple alignment of 2,044 human and duck sequences was made by using the computer program MAFFT (version 6.853b) (Katoh et al., 2002), which did not contain any gaps. Phylogenetic trees were constructed by the neighbor-joining method (Saitou and Nei, 1987) with the p distance (Nei and Kumar, 2000) and the maximum composite likelihood (MCL) distance (Tamura et al., 2004), which are known to produce reliable phylogenetic trees when a large num-

ber of closely related sequences is analyzed, using MEGA (version 5.05) (Tamura et al., 2011). The nucleotide sequence at each interior node of the phylogenetic tree was inferred by the maximum parsimony method (Fitch, 1971; Hartigan, 1973) using PAML (version 4.4b) (Yang, 2007).

Single-site analysis of natural selection: d_N-d_S test Natural selection operating at the amino acid sequence level can be detected by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) substitutions under the assumption that synonymous mutations are neutral or nearly neutral; the relationships $r_S < r_N$, $r_S > r_N$, and $r_S = r_N$ indicate positive, negative, and no selection, respectively (Kimura, 1977; Hughes and Nei, 1988). Nonsynonymous substitutions may be divided into conservative and radical substitutions according to whether they retain or alter a property of amino acids, respectively. If charge is considered, conservative and radical substitutions may be defined as nonsynonymous substitutions within and between charge categories (Hughes et al., 1990). Arginine [R], histidine [H], and lysine [K] are positively charged; aspartic acid [D] and glutamic acid [E] are negatively charged; and N, P, S, T, alanine [A], cysteine [C], glutamine [Q], glycine [G], isoleucine [I], leucine [L], methionine [M], phenylalanine [F], tryptophan [W], tyrosine [Y], and valine [V] are neutral (Arinaminpathy and Grenfell, 2010). Natural selection operating on conservative and radical substitutions can be detected separately by comparing the rates of these substitutions with r_S, in a similar manner to the comparison of r_S and r_N (Hughes et al., 1990; Suzuki, 2007).

For examining natural selection at single amino acid sites of HA, the numbers of synonymous and nonsynonymous differences and sites were computed at single codon sites for each branch of the phylogenetic tree by comparing the nucleotide sequences at the ancestral and descendant nodes (Suzuki and Gojobori, 1999). Here the transition/transversion rate ratio of nucleotide mutation (κ) was required for computing the numbers of synonymous and nonsynonymous sites. Using the ratio of the transitional/transversional nucleotide diversity at 96 four-fold degenerate sites of 2,043 human sequences, κ was estimated to be 4.057. Therefore, κ was assumed to be 4 in the computation. For each codon site, the numbers of synonymous and nonsynonymous differences were summed and the numbers of synonymous and nonsynonymous sites were averaged with the weight proportional to the branch length over all branches of the phylogenetic tree, to obtain the total numbers of synonymous (c_S) and nonsynonymous (c_N) differences and the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites (Suzuki and Gojobori, 1999; Suzuki, 2004a). Although multiple substitutions were not corrected in this method, the degree of underestimation for c_S and c_N appeared to be negligible for the data set analyzed in the present study, because the branch lengths were generally very small (Saitou, 1989). The total numbers of synonymous (d_S) and nonsynonymous (d_N) substitutions over the phylogenetic tree were computed as c_S/s_S and c_N/s_N, respectively. The r_N/r_S value was estimated as d_N/d_S, and the null hypothesis of no selection ($d_S = d_N$) was tested by computing the probability (p) of obtaining the observed or more biased values for c_S and c_N, which were assumed to follow a binomial distribution with the probabilities of occurrence of synonymous and nonsynonymous substitutions given by $s_S/(s_S + s_N)$ and $s_N/(s_S + s_N)$, respectively.

When the test is conducted for multiple codon sites, it is necessary to correct for multiple testing. In the Bonferroni correction, the family-wise significance level ($\alpha = 0.05$ in the present study) is divided by the number (n) of tests to obtain the corrected significance level (α_c) for individual tests. In this approach, α_c may become unrealistically small when n is large. It should be noted, however, that a number of nucleotide differences ($c_S + c_N$) is required for detecting a significant difference between d_S and d_N at a codon site. For example, if $s_S = 1$ and $s_N = 2$ with $\alpha_c = 0.05$, at least 9 ($c_S = 0$ and $c_N = 9$) and 3 ($c_S = 3$ and $c_N = 0$) nucleotide differences are required for detecting positive and negative selection, respectively (Suzuki, 2008a; Nozawa et al., 2009). In other words, the codon sites with $(c_S + c_N) < 9$ and $(c_S + c_N) < 3$ are not testable for positive and negative selection, respectively, and may be eliminated from the test to reduce n. Since the numbers of nucleotide differences required for detecting positive and negative selection may differ, the tests of positive and negative selection may be conducted separately using different α_c. In the test of positive selection, the probability (p_0) of observing 0 synonymous and $(c_S + c_N)$ nonsynonymous differences or more biased values is computed at each codon site as indicated above. The codon sites are ranked (r) according to p_0 in ascending order, and those with $p_0 < \alpha/r$ are considered to be detectable as positively selected with correction. The test is conducted only for these (n_c) sites; positive selection is inferred when $p < \alpha_c$ ($= \alpha/n_c$) and $d_S < d_N$. Negative selection can be inferred in a similar manner.

Single-site analysis of natural selection: interior-exterior test In the phylogenetic tree of individuals sampled from a population, advantageous and deleterious mutations tend to be accumulated on the branches more proximal and distal to the root, respectively (McDonald and Kreitman, 1991). Therefore, d_N/d_S for proximal branches may be greater and smaller than that for distal branches at the codon sites under positive and negative selection, respectively. Since interior branches are usually more proximal than exterior branches in the phylogenetic tree, positive and negative selection may be inferred when d_N/d_S for the former branches is greater and smaller than that for the latter branches, respectively

(Pybus et al., 2007).

For examining natural selection at single amino acid sites of HA using this approach, c_S, c_N, s_S, and s_N were computed separately for interior (c_{Sint}, c_{Nint}, s_{Sint}, and s_{Nint}) and exterior (c_{Sext}, c_{Next}, s_{Sext}, and s_{Next}) branches. The expected values of c_{Sint}, c_{Nint}, c_{Sext}, and c_{Next} ($E[c_{Sint}]$, $E[c_{Nint}]$, $E[c_{Sext}]$, and $E[c_{Next}]$) were obtained under the null hypothesis of equal d_N/d_S for interior and exterior branches, fixing the total numbers of nucleotide differences for interior and exterior branches. The goodness-of-fit of the null hypothesis was examined by using p for the χ^2 value computed from 4 classes with 1 degree of freedom. The correction for multiple testing can be conducted in a similar manner to the d_N-d_S test. In the test of positive selection, p_0 for the χ^2 value is computed at each codon site under the assumption that the numbers of synonymous and nonsynonymous differences are 0 and $c_{Sint} + c_{Nint}$ for interior branches and $c_{Sext} + c_{Next}$ and 0 for exterior branches, respectively. The codon sites are ranked according to p_0 in ascending order, and those with $p_0 < \alpha/r$ are considered to be detectable as positively selected with correction. The test is conducted only for these sites; positive selection is inferred when $p < \alpha_c$ ($= \alpha/n_c$) with $c_{Nint} > E[c_{Nint}]$. Negative selection can be inferred in a similar manner.

It should be noted that the χ^2 value may be unreliable when the expected value is < 5 for any class (Sokal and Rohlf, 1995). In this case, $c_{Sint} + c_{Next}$ and $c_{Sext} + c_{Nint}$ may be compared with $E[c_{Sint}] + E[c_{Next}]$ and $E[c_{Sext}] + E[c_{Nint}]$, respectively, to obtain p for the χ^2 value computed from 2 classes with 1 degree of freedom. Positive and negative selection may be inferred when $p < \alpha_c$ ($= \alpha/n_c$) with ($c_{Sext} + c_{Nint}$) > ($E[c_{Sext}] + E[c_{Nint}]$) and ($c_{Sint} + c_{Next}$) > ($E[c_{Sint}] + E[c_{Next}]$), respectively, in a similar manner as above.

Single-substitution analysis of natural selection The fitness effect of amino acid substitutions may be advantageous, deleterious, or neutral. Under the assumption that natural selection operating at each amino acid site did not change to any large extent during evolution, which may be the case for human H3N2 virus with relatively short evolutionary history after transmission into the human population, the fitness effect of reverse substitutions may be deleterious, advantageous, and neutral if that of original substitutions was advantageous, deleterious, and neutral, respectively (Bazykin and Kondrashov, 2011). Therefore, natural selection operating on single amino acid substitutions may be inferred by detecting natural selection operating on reverse substitutions; positive, negative, and no selection are inferred for original substitutions when negative, positive, and no selection are detected for reverse substitutions, respectively (two-tailed test). If the fitness effect of original substitutions is assumed to be only advantageous or neutral (McDonald and Kreitman, 1991), positive and no selection are inferred for original substitutions when negative and no selection are detected for reverse substitutions, respectively (one-tailed test). The two-tailed test is adopted in the present study.

For examining natural selection operating on the amino acid substitutions generating N-linked glycosylation sites in HA, natural selection operating on reverse substitutions was inferred by comparing the rate of this substitution (r_{Nrev}) with r_S, focusing on the branches where the ancestral amino acid was the substituted form in the phylogenetic tree. For example, if an amino acid substitution D → N generated an N-linked glycosylation site, r_S and r_{Nrev} (causing N → D) were compared at the codon site using the branches where the ancestral amino acid was N. The numbers of reverse nonsynonymous differences (c_{Nrev}) and sites (s_{Nrev}) were computed in a similar manner to the computation of c_N and s_N. The null hypothesis of no selection for reverse substitution ($d_S = d_{Nrev}$) was tested with correction in a similar manner to the comparison of d_S and d_N. Although the single-substitution analysis is conceptually different from the single-site analysis, they are methodologically similar to each other. It has been shown that the latter analysis is generally conservative and reliable in the computer simulation and real data analysis (Suzuki and Gojobori, 1999; Suzuki, 2004a, 2007). It should be noted that in the single-substitution analysis, positive selection for original substitutions can be detected even when the number of (reverse) nonsynonymous substitutions is 0.

RESULTS

Identification of amino acid substitutions generating N-linked glycosylation sites When the phylogenetic tree was constructed for HA of 2,043 human H3N2 viruses and a duck H3N8 virus with the p distance and the ancestral sequences were inferred at interior nodes, N-linked glycosylation sites were observed in some sequences of the human lineage at amino acid sites 6, 7, 8, 22, 38, 45, 276, 278, 285, 483, and 498 in the fibrous stem and at sites 63, 81, 122, 126, 133, 144, 165, 171, and 246 in the globular head. The amino acid substitutions generating N-linked glycosylation sites were observed at amino acid sites 63 (D → N), 124 (G → S), 126 (D → N and T → N), 133 (D → N), 144 (D → N, I → N, and T → N), 173 (K → T), and 248 (N → S and N → T) in the globular head (Table 1).

Single-site analysis of natural selection The single-site analysis of natural selection using the d_N-d_S test and the interior-exterior test was conducted for the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, as indicated above. No site was inferred as positively or negatively selected by either test (Table 1). Natural selection was not detected even

Table 1. The p values obtained from the single-site analysis and the single-substitution analysis for the amino acid sites and substitutions generating N-linked glycosylation sites using the phylogenetic tree constructed with the p distance

Position	Substitution	Single-site analysis						Single-substitution analysis	
		d_N-d_S test		Interior-exterior test					
				4 classes		2 classes			
		Individual	Combined	Individual	Combined	Individual	Combined	Individual	Combined
63	D → N	*0.0498*[a]		N.A.[b]		0.527		*0.0625*	
124	G → S	0.829		N.A.[b]		*0.957*		*0.625*	
126	D → N	*0.0220*		N.A.[b]		0.733		*0.151*	
	T → N								
133	D → N	*0.0578*		N.A.[b]		*0.278*		*0.267*	
144	D → N	*0.0108*	*0.209*	N.A.[b]	0.244	0.0460	0.291	0.218	*0.0477*
	I → N								
	T → N								
173	K → T	0.516		N.A.[b]		0.505		N.A.[d]	
248	N → S	*0.160*		N.A.[b]		N.A.[b]		*1*	
	N → T								
p_c	Positive	0.0125	0.05	N.A.[c]	0.05	0.00833	0.05	0.0167	0.05
p_c	Negative	*0.00714*	*0.05*	N.A.[c]	*0.05*	*0.00833*	*0.05*	*0.0167*	*0.05*

[a] Values are indicated in plain text and in italic when the configuration of test statistics was in favor of positive and negative selection, respectively.
[b] Not applicable because the expected number of differences was < 5 for some classes.
[c] Not applicable because all individual sites were not testable for natural selection.
[d] Not applicable because no substitution was observed.

when the c_S, c_N, s_S, and s_N values for these sites were combined to increase the sensitivity of the tests. To examine the property of the amino acid sites that are identified as positively selected by these tests, the tests were performed using all sites of HA. Although the interior-exterior test failed to detect natural selection at any site, the d_N-d_S test identified negative selection at many sites and positive selection at sites 53 and 138, which were antigenic sites included in epitopes C and A, respectively. When nonsynonymous substitutions were divided into conservative and radical substitutions according to whether they retain or alter the charge of amino acids, positive selection was inferred to have operated on conservative and radical substitutions at sites 138 and 53 by the d_N-d_S test, respectively.

Single-substitution analysis of natural selection The single-substitution analysis of natural selection was conducted for the amino acid substitutions generating N-linked glycosylation sites, as identified above. Specifically, natural selection was examined for the reverse substitutions, N → D at amino acid site 63, S → G at site 124, N → D and N → T at site 126, N → D at site 133, N → D, N → I, and N → T at site 144, T → K at site 173, and S → N and T → N at site 248, focusing on the branches where the ancestral amino acid was the substituted form in the phylogenetic tree. No selection was inferred for each of the reverse substitutions. However, negative selection was detected when the c_S, c_{Nrev}, s_S, and s_{Nrev} values for these substitutions were combined (Table 1), suggesting that the original amino acid substitutions generating N-linked glycosylation sites were positively selected.

Similar results were obtained when all of the above analyses were repeated using the phylogenetic tree constructed with the MCL distance (data not shown). Although it was possible that the mutation bias affected the results of single-substitution analysis, negative selection for the reverse substitutions was also identified when the pattern of nucleotide substitution obtained from the analysis of PB2 (Rabadan et al., 2006) was used (data not shown).

DISCUSSION

Positive selection for gains of N-linked glycosylation sites In the single-substitution analysis of HA for human H3N2 virus, positive selection was detected for the amino acid substitutions generating N-linked glycosylation sites. However, all of the sites where these substitutions occurred were antigenic sites and some of the substitutions increased the positive charge of HA, which can also be the target of positive selection (Suzuki, 2006; Hensley et al., 2009). Nevertheless, the effect of antigenic variation at these sites may be shielded by the N-linked glycans attached to the newly generated and closely located N-linked glycosylation sites (Skehel et al., 1984; Jackson et al., 1994; Tsuchiya et al., 2002; Das et al., 2010; Wei et al., 2010; Wanzeck et al., 2011), and the effect of increment of positive charge by some of these

substitutions may be cancelled by the increment of negative charge in the sialic acid or sulfuric acid, which may be added to the N-linked glycans (Spiro and Spiro, 2000). Therefore, positive selection appears to have operated for gains of N-linked glycosylation sites.

The N-linked glycans of HA are involved in several functions, such as folding of ectodomain, fusion activity of HA, shielding of antigenic sites, proteolytic activity of HA, recognition by collectins, and receptor binding, as discussed above. However, since most of the gains of N-linked glycosylation sites apparently occurred around the receptor-binding pocket in the three-dimensional structure of HA (Abe et al., 2004; Kobayashi and Suzuki, in preparation), it is likely that positive selection has operated on shielding of antigenic sites or receptor binding. It should be noted that the number of N-linked glycosylation sites has increased only in the influenza A viruses circulating in human. In particular, the number stayed constant in influenza A viruses circulating in swine, which apparently expresses similar distributions of the receptors with sialic acid α2,3-galactose and α2,6-galactose linkages in organs (Nelli et al., 2010; Sriwilaijaroen et al., 2011), but weaker immune responses against influenza A viruses (Nerome et al., 1995) compared with human. These observations suggest that the target of positive selection was not the receptor binding but the shielding of antigenic sites by the gains of N-linked glycosylation sites in human H3N2 virus.

It has been proposed that the gain and loss of N-linked glycosylation sites may not have been involved in antigenic changes of HA in human H3N2 virus, because the gain and loss were not found to be coincided with the transition of antigenic clusters (Smith et al., 2004), as well as the increase in the rate of change in the substitution pattern at amino acid sites (Blackburne et al., 2008). However, the existence of antigenic clusters in human H3N2 virus itself has been questioned (Shih et al., 2007; Suzuki, 2008b; Bhatt et al., 2011). In addition, the antigenic change appeared to occur continuously even within antigenic clusters (Suzuki, 2008b). Therefore, the gains of N-linked glycosylation sites may be involved in antigenic changes of HA in human H3N2 virus.

Properties of single-site analysis and single-substitution analysis In the study of natural selection for HA of human H3N2 virus, the single-site analysis failed to detect positive selection at the amino acid sites where the substitutions generating N-linked glycosylation sites were observed, but identified positive selection at antigenic sites. In contrast, the single-substitution analysis succeeded in detecting positive selection for the amino acid substitutions generating N-linked glycosylation sites. At the antigenic sites, positive selection is considered to operate recurrently, so that the virus can escape from immune responses continuously (Suzuki, 2008b). Many amino acid substitutions may be accumulated at the antigenic sites, and the excess of nonsynonymous substitutions over synonymous substitutions is detected as the signature of positive selection by the single-site analysis. In contrast, positive selection for gains of N-linked glycosylation sites is considered to be directional. Once a substitution generating an N-linked glycosylation site occurs, further substitutions may be suppressed at the site, so that the advantageous effect of shielding antigenic sites is maintained unless natural selection changes. The suppression of further amino acid substitutions, especially that of the reverse substitution, is detected as the signature of positive selection for the original substitution by the single-substitution analysis. These observations indicate that the single-site analysis is suitable for detecting recurrent natural selection (Suzuki, 2010), whereas the single-substitution analysis is suitable for detecting episodic natural selection.

In the single-site analysis, the d_N/d_S test appeared to be more efficient than the interior-exterior test in detecting natural selection. This is partly because in the latter test the c_S and c_N values were divided into c_{Sint}, c_{Sext}, c_{Nint}, and c_{Next} at single codon sites, which may be too small for obtaining statistical significance. It should also be noted that in the interior-exterior test the difference in the d_N/d_S value between interior and exterior branches can occur not only when positive or negative selection operated but also when natural selection was weakened or strengthened during evolution.

Antigenic drift is not a by-product of the evolution of receptor binding avidity In the single-site analysis of HA for human H3N2 virus, the number of amino acid sites identified as positively selected was only 2, which appeared to be relatively small compared with the envelope glycoproteins of other viruses, such as hepatitis C virus (HCV) and human immunodeficiency virus type 1 (HIV-1) (Suzuki and Gojobori, 2001; Yang et al., 2003). This is probably because the antigenic sites, which are usually the targets of positive selection, were shielded from immune responses by N-linked glycans after the gains of N-linked glycosylation sites in HA of human H3N2 virus, although N-linked glycans are also known to be attached to the envelope glycoproteins of HCV and HIV-1 (Zhang et al., 2004). In fact, a reduction in d_N/d_S has been observed at the antigenic sites of HA that are likely to be shielded by N-linked glycans after the gains of N-linked glycosylation sites during evolution of human H3N2 virus (Kobayashi and Suzuki, in preparation).

Although influenza A virus was believed to escape from immune responses by changing the antigenicity gradually through mutations at antigenic sites (antigenic drift) and abruptly through reassortment of genomic segments encoding HA and NA (antigenic shift), it was proposed that the antigenic drift is a by-product of repeated natural

selection for increased and decreased receptor binding avidity of virus in immune and naïve individuals, which is caused by the amino acid substitutions increasing and decreasing the positive charge of HA, respectively (Hensley et al., 2009). The receptor binding avidity of virus was considered to be positively correlated with the positive charge of HA, because the sialic acid receptor and the host cell membrane are negatively charged. According to this hypothesis, it is expected that positive selection has operated only on the radical substitution in terms of the charge of amino acids at antigenic sites. However, in the single-site analysis of HA for human H3N2 virus, positive selection was identified not only on the radical substitution but also on the conservative substitution at antigenic sites, suggesting that the antigenic drift is not a by-product of the evolution of receptor binding avidity of HA, but the evolutionary mechanism of influenza A virus where amino acid substitutions inhibit recognition of antigenic sites by immune responses.

The author thanks Yuki Kobayashi and two anonymous reviewers for valuable comments.

REFERENCES

Abe, Y., Takashita, E., Sugawara, K., Matsuzaki, Y., Muraki, Y., and Hongo, S. (2004) Effect of the addition of oligosaccharides on the biological activities and antigenicity of influenza A/H3N2 virus hemagglutinin. J. Virol. **78**, 9605–9611.

Arinaminpathy, N., and Grenfell, B. (2010) Dynamics of glycoprotein charge in the evolutionary history of human influenza. PLoS One **5**, e15674.

Bao, Y., Bolotov, P., Dernovoy, D., Kiryutin, B., Zaslavsky, L., Tatusova, T., Ostell, J., and Lipman, D. (2008) The Influenza Virus Resource at the National Center for Biotechnology Information. J. Virol. **82**, 596–601.

Bazykin, G. A., and Kondrashov, A. S. (2011) Detecting past positive selection through ongoing negative selection. Genome Biol. Evol. **3**, 1006–1013.

Bhatt, S., Holmes, E. C., and Pybus, O. G. (2011) The genomic rate of molecular adaptation of the human influenza A virus. Mol. Biol. Evol. **28**, 2443–2451.

Blackburne, B. P., Hay, A. J., and Goldstein, R. A. (2008) Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. **4**, e1000058.

Carstens, E. B. (2010) Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2009). Arch. Virol. **155**, 133–146.

Cherry, J. L., Lipman, D. J., Nikolskaya, A., and Wolf, Y. I. (2009) Evolutionary dynamics of N-glycosylation sites of influenza virus hemagglutinin. PLoS Curr. **1**, RRN1001.

Cui, J., Smith, T., Robbins, P. W., and Samuelson, J. (2009) Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. Proc. Natl. Acad. Sci. USA **106**, 13421–13426.

Daniels, R., Kurowski, B., Johnson, A. E., and Hebert, D. N. (2003) N-linked glycans direct the cotranslational folding pathway of *influenza* hemagglutinin. Mol. Cell **11**, 79–90.

Das, S. R., Puigbo, P., Hensley, S. E., Hurt, D. E., Bennink, J. R., and Yewdell, J. W. (2010) Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. PLoS Pathog. **6**, e1001211.

de Vries, R. P., de Vries, E., Bosch, B. J., de Groot, R. J., Rottier, P. J. M., and de Haan, C. A. M. (2010) The influenza A virus hemagglutinin glycosylation state affects receptor-binding specificity. Virology **403**, 17–25.

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

Greenbaum, B. D., Levine, A. J., Bhanot, G., and Rabadan, R. (2008) Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. **4**, e1000079.

Hartigan, J. A. (1973) Minimum mutation fits to a given tree. Biometrics **29**, 53–65.

Hebert, D. N., Zhang, J.-X., Chen, W., Foellmer, B., and Helenius, A. (1997) The number and location of glycans on influenza hemagglutinin determine folding and association with calnexin and calreticulin. J. Cell Biol. **139**, 613–623.

Hensley, S. E., Das, S. R., Bailey, A. L., Schmidt, L. M., Hickman, H. D., Jayaraman, A., Viswanathan, K., Raman, R., Sasisekharan, R., Bennink, J. R., et al. (2009) Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. Science **326**, 734–736.

Hughes, A. L., and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature **335**, 167–170.

Hughes, A. L., Ota, T., and Nei, M. (1990) Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7**, 515–524.

Igarashi, M., Ito, K., Kida, H., and Takada, A. (2008) Genetically destined potentials for *N*-linked glycosylation of influenza virus hemagglutinin. Virology **376**, 323–329.

Jackson, D. C., Drummer, H. E., Urge, L., Otvos, L. Jr., and Brown, L. E. (1994) Glycosylation of a synthetic peptide representing a T-cell determinant of influenza virus hemagglutinin results in loss of recognition by $CD4^+$ T-cell clones. Virology **199**, 422–430.

Katoh, K., Misawa, K, Kuma, K.-i., and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**, 3059–3066.

Kimura, M. (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. Nature **267**, 275–276.

Kosakovsky Pond, S. L., Poon, A. F. Y., Leigh Brown, A. J., and Frost, S. D. W. (2008) A maximum likelihood method for detecting directional evolution in protein sequences and its application to influenza A virus. Mol. Biol. Evol. **25**, 1809–1824.

McDonald, J. H., and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351**, 652–654.

Nei, M., and Kumar, S. (2000) Molecular Evolution and Phylogenetics, pp. 165–186. Oxford University Press, Oxford, New York.

Nelli, R. K., Kuchipudi, S. V., White, G. A., Perez, B. B., Dunham, S. P., and Chang, K.-C. (2010) Comparative distribution of human and avian type sialic acid influenza receptors in the pig. BMC Vet. Res. **6**, 4.

Nerome, K., Kanegae, Y., Shortridge, K. F., Sugita, S., and Ishida, M. (1995) Genetic analysis of porcine H3N2 viruses originating in southern China. J. Gen. Virol. **76**, 613–624.

Nozawa, M., Suzuki, Y., and Nei, M. (2009) Reliabilities of iden-

tifying positive selection by the branch-site and the site-prediction methods. Proc. Natl. Acad. Sci. USA **106**, 6700–6705.

Pybus, O. G., Rambaut, A., Belshaw, R., Freckleton, R. P., Drummond, A. J., and Holmes, E. C. (2007) Phylogenetic evidence for deleterious mutation load in RNA viruses and its contribution to viral evolution. Mol. Biol. Evol. **24**, 845–852.

Rabadan, R., Levine, A. J., and Robins, H. (2006) Comparison of avian and human influenza A viruses reveals a mutational bias on the viral genomes. J. Virol. **80**, 11887–11891.

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 Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**, 406–425.

Schulze, I. T. (1997) Effects of glycosylation on the properties and functions of influenza virus hemagglutinin. J. Infect. Dis. **176**, S24–S28.

Shih, A. C.-C., Hsiao, T.-C., Ho, M.-S., and Li, W.-H. (2007) Simultaneous amino acid substitutions at antigenic sites drive influenza A hemagglutinin evolution. Proc. Natl. Acad. Sci. USA **104**, 6283–6288.

Shope, R. E. (1931) Swine influenza. III. Filtration experiments and etiology. J. Exp. Med. **54**, 373–380.

Skehel, J. J., and Wiley, D. C. (2000) Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. **69**, 531–569.

Skehel, J. J., Stevens, D. J., Daniels, R. S., Douglas, A. R., Knossow, M., Wilson, I. A., and Wiley, D. C. (1984) A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc. Natl. Acad. Sci. USA **81**, 1779–1783.

Smith, D. J., Lapedes, A. S., de Jong, J. C., Bestebroer, T. M., Rimmelzwaan, G. F., Osterhaus, A. D. M. E., and Fouchier, R. A. M. (2004) Mapping the antigenic and genetic evolution of influenza virus. Science **305**, 371–376.

Sokal, R. R., and Rohlf, F. J. (1995) Biometry. 3rd edition, pp. 685–793. W. H. Freeman and Company, New York.

Spiro, M. J., and Spiro, R. G. (2000) Sulfation of the *N*-linked oligosaccharides of influenza virus hemagglutinin: temporal relationships and localization of sulfotransferases. Glycobiology **10**, 1235–1242.

Sriwilaijaroen, N., Kondo, S., Yagi, H., Takemae, N., Saito, T., Hiramatsu, H., Kato, K., and Suzuki, Y. (2011) *N*-glycans from porcine trachea and lung: predominant NeuAcα2-6Gal could be a selective pressure for influenza variants in favor of human-type receptor. PLoS One **6**, e16302.

Sun, S., Wang, Q., Zhao, F., Chen, W., and Li, Z. (2011) Glycosylation site alteration in the evolution of influenza A (H1N1) viruses. PLoS One **6**, e22844.

Suzuki, Y. (2004a) New methods for detecting positive selection at single amino acid sites. J. Mol. Evol. **59**, 11–19.

Suzuki, Y. (2004b) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21**, 2352–2359.

Suzuki, Y. (2006) Natural selection on the influenza virus genome. Mol. Biol. Evol. **23**, 1902–1911.

Suzuki, Y. (2007) Inferring natural selection operating on conservative and radical substitution at single amino acid sites. Genes Genet. Syst. **82**, 341–360.

Suzuki, Y. (2008a) False-positive results obtained from the branch-site test of positive selection. Genes Genet. Syst. **83**, 331–338.

Suzuki, Y. (2008b) Positive selection operates continuously on hemagglutinin during evolution of H3N2 human influenza A virus. Gene **427**, 111–116.

Suzuki, Y. (2010) Statistical methods for detecting natural selection from genomic data. Genes Genet. Syst. **85**, 359–376.

Suzuki, Y., and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16**, 1315–1328.

Suzuki, Y., and Gojobori, T. (2001) Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. Gene **276**, 83–87.

Tamura, K., Nei, M., and Kumar, S. (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA **101**, 11030–11035.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. **28**, 2731–2739.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.-N., and Nakamura, K. (2001) Antigenic structure of the haemagglutinin of human influenza A/H2N2 virus. J. Gen. Virol. **82**, 2475–2484.

Tsuchiya, E., Sugawara, K., Hongo, S., Matsuzaki, Y., Muraki, Y., Li, Z.-N., and Nakamura, K. (2002) Effect of addition of new oligosaccharide chains to the globular head of influenza A/H2N2 virus haemagglutinin on the intracellular transport and biological activities of the molecule. J. Gen. Virol. **83**, 1137–1146.

Vigerust, D. J., Ulett, K. B., Boyd, K. L., Madsen, J., Hawgood, S., and McCullers, J. A. (2007) N-linked glycosylation attenuates H3N2 influenza viruses. J. Virol. **81**, 8593–8600.

Wang, C.-C., Chen, J.-R., Tseng, Y.-C., Hsu, C.-H., Hung, Y.-F., Chen, S.-W., Chen, C.-M., Khoo, K.-H., Cheng, T.-J., Cheng, Y.-S. E., et al. (2009) Glycans on influenza hemagglutinin affect receptor binding and immune response. Proc. Natl. Acad. Sci. USA **106**, 18137–18142.

Wanzeck, K., Boyd, K. L., and McCullers, J. A. (2011) Glycan shielding of the influenza virus hemagglutinin contributes to immunopathology in mice. Am. J. Respir. Crit. Care Med. **183**, 767–773.

Wei, C.-J., Boyington, J. C., Dai, K., Houser, K. V., Pearce, M. B., Kong, W.-P., Yang, Z.-y., Tumpey, T. M., and Nabel, G. J. (2010) Cross-neutralization of 1918 and 2009 influenza viruses: role of glycans in viral evolution and vaccine design. Sci. Transl. Med. **2**, 24ra21.

Wilson, I. A., Skehel, J. J., and Wiley, D. C. (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature **289**, 366–373.

World Health Organization (1980) A revision of the system of nomenclature for influenza viruses: a WHO memorandum. Bull. W. H. O. **58**, 585–591.

World Health Organization (2011) Review of the 2010–2011 winter influenza season, northern hemisphere. Wkly. Epidemiol. Rec. **86**, 222–227.

Yang, W., Bielawski, J. P., and Yang, Z. (2003) Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J. Mol. Evol. **57**, 212–221.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. **24**, 1586–1591.

Zhang, M., Gaschen, B., Blay, W., Foley, B., Haigwood, N., Kuiken, C., and Korber, B. (2004) Tracking global patterns of *N*-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin. Glycobiology **14**, 1229–1246.