Journal of Virology

# Evidence for N-Glycan Shielding of Antigenic Sites during Evolution of Human Influenza A Virus Hemagglutinin

**Yuki Kobayashi and Yoshiyuki Suzuki**
***J. Virol.* 2012, 86(7):3446. DOI: 10.1128/JVI.06147-11.**
**Published Ahead of Print 18 January 2012.**

JVI
Journals.ASM.org

# Evidence for N-Glycan Shielding of Antigenic Sites during Evolution of Human Influenza A Virus Hemagglutinin

**Yuki Kobayashi and Yoshiyuki Suzuki**

Graduate School of Natural Sciences, Nagoya City University, Nagoya City, Aichi, Japan

**After the emergence of influenza A viruses in the human population, the number of N-glycosylation sites (NGS) in the globular head region of hemagglutinin (HA) has increased continuously for several decades. It has been speculated that the addition of NGS to the globular head region of HA has conferred selective advantages to the virus by preventing the binding of antibodies (Ab) to antigenic sites (AS). Here, the effect of N-glycosylation on the binding of Ab to AS in human influenza A virus subtype H3N2 (A/H3N2) was examined by inferring natural selection at AS and other sites (NAS) that are located close to and distantly from the NGS in the three-dimensional structure of HA through a comparison of the rates of synonymous ($d_S$) and nonsynonymous ($d_N$) substitutions. When positions 63, 122, 126, 133, 144, and 246 in the globular head region of HA were non-NGS, the $d_N/d_S$ was >1 and positive selection was detected at the AS located near these positions. However, the $d_N/d_S$ value decreased and the evidence of positive selection disappeared when these positions became NGS. In contrast, $d_N/d_S$ at the AS distantly located from the positions mentioned above and at the NAS of any location were generally <1 and did not decrease when these positions changed from non-NGS to NGS. These results suggest that the attachment of N-glycans to the NGS in the globular head region of HA prevented the binding of Ab to AS in the evolutionary history of human A/H3N2 virus.**

Influenza virus is a member of the family *Orthomyxoviridae*, which contains a segmented and negative-stranded RNA genome in an enveloped virion. This virus is classified into types A to C, and influenza A virus has caused four global pandemics in humans during the last hundred years (18, 22). Influenza A virus is further classified into subtypes based on the genetic and antigenic characteristics of hemagglutinin (HA) and neuraminidase (NA), which constitute envelope glycoproteins. To date, 16 HA (H1 to H16) and 9 NA (N1 to N9) subtypes have been identified, and influenza A viruses bearing H1N1 and H3N2 subtypes (A/H1N1 and A/H3N2 viruses) are currently cocirculating in humans.

Influenza A virus escapes from host immune responses by changing the antigenicity of HA and NA both gradually (antigenic drift) and abruptly (antigenic shift) (16). The HA consists of signal peptide, HA1 (amino acid positions 1 to 328 in the H3 numbering system), and HA2 (amino acid positions 330 to 550), and it forms trimers in virions (30). The ectodomain of an HA trimer can be divided into globular head and stem regions (30). The globular head region is composed of a part of HA1 (amino acid positions 58 to 272), while the stem region is composed of the other parts of HA1 and HA2. Antigenic sites (AS), which are targets of antibodies (Ab), constitute five epitopes (A to E) in A/H3N2 viruses. Epitopes A, B, D, and E reside in the globular head region, while epitope C is in the stem region of HA1 (24, 29).

Although the escape from host immune responses occurs through changes in amino acids at AS that are recognized by Ab (14, 18, 23), it has been proposed that the attachment of an oligosaccharide to the N-glycosylation site (NGS), which is the Asn residue of the sequon (Asn-Xaa-Ser/Thr, where Xaa is any amino acid except for Pro), in the globular head region also contributes to the escape (1, 21, 27, 28). This is based on observations in experimental studies that some N-glycans attached to the globular head region interfere with the binding of Ab to AS by masking the surface of HA (1, 21, 27, 28).

In the evolutionary studies of A/H1N1 and A/H3N2 viruses, it has been observed that the number of NGS in the globular head region has increased continuously for several decades after the emergence of these viruses in the human population (4, 6–8, 11, 32). It was speculated that the addition of NGS to the globular head region conferred selective advantages to the viruses by preventing the binding of Ab to AS. However, this hypothesis has been questioned, because the gain of NGS in the globular head region of HA did not appear to influence the amino acid variation at AS, and no correlation was observed between the transitions of antigenic clusters and gains or losses of NGS during the evolution of human A/H3N2 virus (4, 8).

If N-glycans attached to the globular head region contribute to preventing the binding of Ab to AS, it is expected that natural selection for amino acid substitutions at AS is weakened (strengthened) when AS that are not covered (covered) by N-glycans become covered (not covered) through gains (losses) of NGS. Natural selection for amino acid substitutions may be identified from a comparison of the rate of nonsynonymous substitution ($d_N$) to that of synonymous substitution ($d_S$). Positive, negative, and no selection are inferred when the $d_N/d_S$ value is >1, <1, and 1, respectively. It has been observed that $d_N/d_S$ at AS is often >1, whereas that at other sites (NAS) is usually <1 in HA of human A/H1N1 and A/H3N2 viruses (20, 23, 24, 31).

The purpose of the present study was to examine the effect of gains and losses of NGS on $d_N/d_S$ at AS and NAS that are covered and not covered by N-glycans during the evolution of human A/H3N2 virus.

**Received** 26 August 2011 **Accepted** 6 January 2012
**Published ahead of print** 18 January 2012
Address correspondence to Y. Suzuki, yossuzuk@nsc.nagoya-cu.ac.jp.
Supplemental material for this article may be found at http://jvi.asm.org/.

doi:10.1128/JVI.06147-11

## MATERIALS AND METHODS

**Sequences.** All (2,158) complete HA-coding nucleotide sequences of A/H3N2 viruses isolated from humans in the period between 1968 and 2010 were retrieved from the Influenza Virus Resource (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) on 15 June 2010. After removing the sequences containing undetermined nucleotides and minor gaps, 1,903 HA sequences were used for the following analysis.

**Phylogenetic tree.** A multiple alignment of HA sequences was made using the computer program MAFFT (12). The optimum model of nucleotide substitution for these sequences was selected using MODELTEST (17) with PAUP (version 4.0). The general time-reversible model with a gamma distribution for the rate of heterogeneity among sites (G) and invariable sites (I) (GTR+G+I) (gamma shape parameter = 1.3549, proportion of invariable sites = 0.2696, and transition/transversion rate ratio [$\kappa$] = 10) and the transversional model with G and I (TVM+G+I) (gamma shape parameter = 1.3543, proportion of invariable sites = 0.2695, and $\kappa$ = 9) were selected as the most appropriate models by the hierarchical likelihood ratio test (hLRT) and Akaike information criterion (AIC), respectively.

Since TVM is a special case of GTR, a phylogenetic tree of human A/H3N2 virus was constructed using 1,903 complete HA-coding nucleotide sequences by the neighbor-joining (NJ) method with the evolutionary distances measured assuming GTR+G+I using PAUP. The HA sequence of A/duck/Hong Kong/7/1975 (accession no. CY006026) was used as the outgroup to root the phylogenetic tree. After constructing the phylogenetic tree, the branch lengths were reestimated by the maximum likelihood (ML) method with GTR+G+I.

**Reconstruction of ancestral HA nucleotide sequences.** For each interior node of the phylogenetic tree, the ancestral HA nucleotide sequence was reconstructed by the maximum parsimony method using PAML (9, 10). In the HA amino acid sequences obtained by translating the HA nucleotide sequences at the interior and exterior nodes, the Asn residue of the sequon was considered an NGS. For each of the amino acid positions where an NGS was observed at some nodes of the phylogenetic tree, the state of N-glycosylation (whether the amino acid position was an NGS or a non-NGS) was inferred for each branch by referring to the state of the nodes located at the ends of the branch. Each node of the branch was identified as ancestral or descendant because the phylogenetic tree was rooted. For each branch, the amino acid position was inferred to be an NGS (non-NGS) when the position was an NGS (non-NGS) at both nodes. A gain (loss) of an NGS was inferred to have occurred at the branch when the position was a non-NGS (NGS) at the ancestral node and an NGS (non-NGS) at the descendant node.

**Prediction of 3D structures for N-glycans.** The three-dimensional (3D) structures for the variants of N-glycans were constructed using SWEET (5). The structures of sugar chains were inferred using formulae for N-glycans that were reported to be attached to the HA of A/H1N1 viruses (15) (see Table S1 in the supplemental material). The distance between the NGS and each amino acid position in the 3D structure of HA was computed using the structure of trimeric HA for A/Memphis/102/72 (Protein Data Bank [PDB] no. 3HMG). Amino acid positions located within 10 or 15 Å from the NGS were assumed to be covered by N-glycans, whereas those located farther than 10 or 15 Å away were assumed not to be covered by N-glycans (see Results). The AS (NAS) of HA1 located within 10 and 15 Å from NGS are denoted $AS_{10c}$ ($NAS_{10c}$) and $AS_{15c}$ ($NAS_{15c}$), respectively, and the AS (NAS) of HA1 located farther than 10 and 15 Å away are denoted $AS_{10uc}$ ($NAS_{10uc}$) and $AS_{15uc}$ ($NAS_{15uc}$), respectively. When an amino acid site was located within 10 or 15 Å of more than one NGS, that site was categorized as $AS_{10c}$, $AS_{15c}$, $NAS_{10c}$, or $NAS_{15c}$ of an NGS that was generated first during evolution. The amino acid sites that were possibly covered by any NGS during evolution were eliminated from $AS_{10uc}$, $AS_{15uc}$, $NAS_{10uc}$, and $NAS_{15uc}$ of all NGS, resulting in the same set of amino acid sites included in $AS_{10uc}$, $AS_{15uc}$, $NAS_{10uc}$, and $NAS_{15uc}$ for all NGS. The amino acid positions constituting sequons during any period of evolution were eliminated from all categories.

**Estimation of $d_N/d_S$ at AS and NAS.** For each amino acid position that was found to be an NGS during some period in the evolutionary history of human A/H3N2 virus, $d_S$ and $d_N$ at $AS_{10c}$, $AS_{15c}$, $AS_{10uc}$, $AS_{15uc}$, $NAS_{10c}$, $NAS_{15c}$, $NAS_{10uc}$, and $NAS_{15uc}$ were estimated for the branches where the position was an NGS and a non-NGS separately. When an amino acid site in $AS_{10c}$, $AS_{15c}$, $NAS_{10c}$, and $NAS_{15c}$ of a particular position was considered to possibly be covered by an N-glycan attached to another position on some branches where the former position was a non-NGS, these branches were eliminated from the computation of $d_N$ and $d_S$ for the non-NGS branches of the former position. In addition, three branches connecting HA sequences derived from human infections of triple-reassortant A/H3N2 viruses were excluded from all computations of $d_S$ and $d_N$ (Fig. 1) (2).

The $d_S$ and $d_N$ values were computed as *Sd*/*S* and *Nd*/*N*, respectively, where *Sd* and *Nd* denote the numbers of synonymous and nonsynonymous differences summed over the branches of sites included in each category, and *S* and *N* denote the numbers of synonymous and nonsynonymous sites averaged over the branches of sites included in each category with the weight proportional to the branch length. The numbers of synonymous and nonsynonymous differences as well as the numbers of synonymous and nonsynonymous sites for each branch of individual codon sites were obtained by comparing the sequences at the ancestral and descendant nodes using the modified Nei-Gojobori method with $\kappa$ = 10. Since positions 165 and 285 appeared to be NGS throughout the evolutionary history of human A/H3N2 virus, the analysis of $d_S$ and $d_N$ was not conducted with these positions.

**Statistical test.** The effect of gains and losses of NGS on $d_N/d_S$ was examined at AS and NAS that can be covered and not covered by N-glycans during evolution. At $AS_{10c}$, $AS_{15c}$, $AS_{10uc}$, $AS_{15uc}$, $NAS_{10c}$, $NAS_{15c}$, $NAS_{10uc}$, and $NAS_{15uc}$ of amino acid positions that were found to be NGS during some evolutionary period, the average $d_N/d_S$ value during the evolutionary history of A/H3N2 virus was estimated using all branches of the phylogenetic tree irrespective of whether the position under consideration was an NGS or a non-NGS. The average $d_N/d_S$ value was used to obtain the expected values of *Sd* and *Nd* for the branches where the position was an NGS and a non-NGS separately. The null hypothesis of equal $d_N/d_S$ for NGS and non-NGS branches was tested by the chi-square test with one degree of freedom using the expected and observed values of *Sd* and *Nd* for both branches.

Natural selection operating at $AS_{10c}$, $AS_{15c}$, $AS_{10uc}$, $AS_{15uc}$, $NAS_{10c}$, $NAS_{15c}$, $NAS_{10uc}$, and $NAS_{15uc}$ was examined for the branches where the position under consideration was an NGS and a non-NGS separately. The proportions of *S* and *N* were used to obtain the expected values of *Sd* and *Nd* (25), and the null hypothesis of selective neutrality was tested by the chi-square test with one degree of freedom. Positive and negative selection were inferred when the observed values of *Nd* and *Sd* were significantly greater than the expected values of *Nd* and *Sd*, respectively.

## RESULTS

**Size of N-glycans attached to HA.** Two types of N-glycans, high-mannose and complex oligosaccharides, are known to be attached, in various forms, to HA (13, 15). To infer the amino acid positions possibly covered by N-glycans in the 3D structure of HA, the maximum distance between sugar molecules within N-glycans was measured using formulae for N-glycans that were reported to be attached to HA (15). The maximum distance obtained was 16.78 to 34.63 Å (average, 25.4 Å) (see Table S1 in the supplemental material). If we assume that the shape of N-glycans is approximately spherical, the positions located within 8 to 17 Å of an NGS are considered to be covered by N-glycans. Therefore, the amino acid positions located within 10 or 15 Å of an NGS were assumed to be covered by N-glycans, whereas those located farther than 10 and 15 Å away were assumed to not be covered in the present study. The AS (NAS) located within 10 and 15 Å of an NGS are

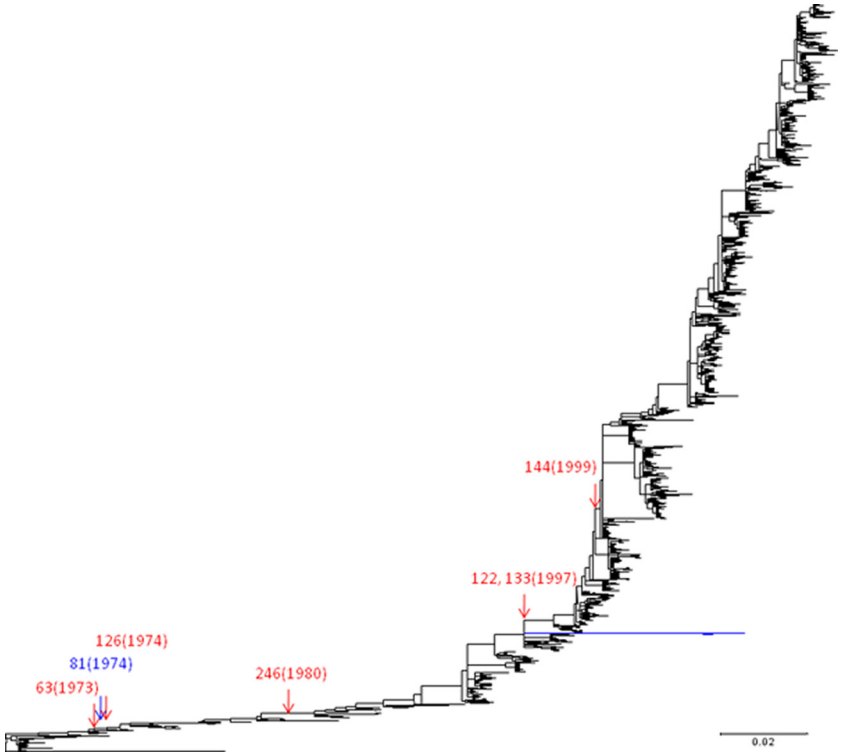


**FIG 1** Gains and losses of NGS in the globular head region of HA in the evolutionary history of human A/H3N2 virus. The phylogenetic tree was constructed for 1,904 HA-coding nucleotide sequences by the NJ method with GTR+G+I. The red and blue arrows indicate the trunk branches where gains and losses of NGS in the globular head region occurred, respectively, with the amino acid positions of NGS indicated. The numbers in parentheses indicate the earliest (for gains) or most recent (for losses) year when the virus with the NGS was isolated. The long branches colored in blue were excluded from the computation of $d_S$ and $d_N$, because the HA sequences at the ends were derived from human infections of triple-reassortant A/H3N2 viruses.

denoted $AS_{10c}$ ($NAS_{10c}$) and $AS_{15c}$ ($NAS_{15c}$), respectively, and the AS (NAS) located farther than 10 or 15 Å away are denoted $AS_{10uc}$ ($NAS_{10uc}$) and $AS_{15uc}$ ($NAS_{15uc}$), respectively (see Materials and Methods for details).

**Gains and losses of NGS in HA of human A/H3N2 virus.** A phylogenetic tree was constructed for 1,903 complete HA-coding nucleotide sequences of A/H3N2 viruses isolated from humans in the period between 1968 and 2010 (Fig. 1), and the ancestral nucleotide sequence was inferred at each interior node. When the nucleotide sequences at the interior and exterior nodes were translated into amino acid sequences, some of them contained NGS at amino acid positions 63, 81, 122, 126, 133, 144, 165, and 246 in the globular head region and at positions 6, 7, 8, 22, 38, 45, 276, 285, 483, and 498 in the stem region (Fig. 2). N-glycans appeared to cover AS when they were attached to positions 45, 63, 81, 122, 126, 133, 144, 165, 246, 276, and 285 (Table 1).

The occurrences of gains and losses of NGS that may cover AS in the evolutionary history of human A/H3N2 virus was inferred from the analysis of nucleotide sequences at the interior and exterior nodes of the phylogenetic tree. When A/H3N2 virus emerged in the human population in 1968, only positions 81 and 165 were NGS in the globular head region of HA (19). The NGS at position 165 has been maintained throughout the evolutionary history of the virus. However, position 81 became a non-NGS in 1974 (Fig. 1). The NGS generated at positions 63, 122, 126, 133, 144, and 246 have been retained during evolution. In the stem region of HA, position 285 has been maintained as an NGS throughout the evo-

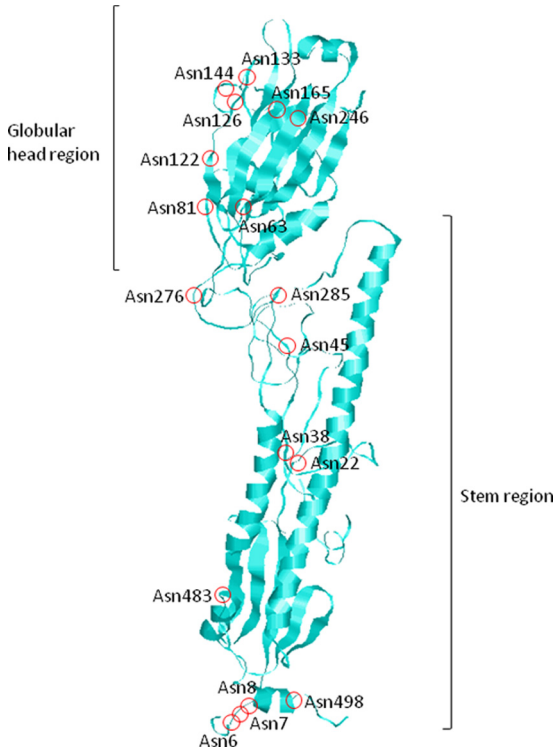


**FIG 2** Amino acid positions of NGS in the 3D structure of human A/H3N2 virus HA (PDB no. 3HMG) identified in the present study. RasMol (3) was used to visualize the 3D structure.

**TABLE 1** Amino acid sites possibly covered by N-glycan in the 3D structure of trimeric HA

| Antigenic site[a] and NGS[b] | Amino acid site (epitope: amino acid position[s]) possibly covered by N-glycan[c] |
|---|---|
| $AS_{10c}$ | |
| 45 | C: **44**, 45, 46, 47, 297, **312** |
| 63 | D: **96**; E: **62**, 63, **75**, 78, **91**, **92**, **94** |
| 81 | A: **150**; D: **117**, **121**; E: **78**, **80**, 81, 82, 83 |
| 122 | A: 122, 124, 150; D: 121 |
| 126 | A: 124, 126, **130**, **132**, **168**; B: 128, **129**, 164, 165; D: 167 |
| 133 | A: **131**, 132, 133, 135, 146, **152**; B: **155** |
| 144 | A: 135, **137**, **138**, **140**, **142**, **143**, 144, 145, 146 |
| 246 | B: 163, 164, 165, **186**; D: **201**, 203, **212**, **216**, **217**, 218, 219, 244, 246, 247, 248 |
| 276 | C: 50, 51, **53**, **54**, **275**, 276, 278 |
| $AS_{15c}$ | |
| 45 | C: 44, 45, 46, 47, 48, **294**, 297, 299, 300, 305, **307**, 308, 309, **310**, **311**, **312** |
| 63 | D: **96**; E: 59, **62**, 63, 67, **75**, 78, 80, 82, 87, 88, 91, **92**, **94** |
| 81 | A: 122, **150**, D: **117**, **121**, **177**, **179**; E: **57**, **59**, **67**, **78**, **80**, 81, 82, 83, **87**, **260**, **261** |
| 122 | A: 122, 124, 126, 132, 133, 150, 152, 168; D: 121, 167, **170**, **172**, **174**, **176**, 177, 179; E: 80, 81 |
| 126 | A: 122, 124, 126, 130, **131**, **132**, 133, **152**, 168; B: 128, 129, **155**, **157**, 163, 164, 165; D: 167, 244, 246, 247 |
| 133 | A: 122, 124, 126, 130, 131, 132, 133, 135, **137**, **140**, **142**, 144, 145, 146, 150, 152; B: 128, 129, 155, **156**, 157, **158**, **194** |
| 144 | A: 133, 135, 137, **138**, 140, 142, **143**, 144, 145, 146; D: 96, **226**; E: 75 |
| 246 | A: 126, 130, 168; B: 128, 129, 163, 164, 165, 186, 187, 188, **190**, **197**; D: 167, **182**, 201, 203, **209**, 212, **213**, **214**, **215**, 216, 217, 218, 219, 227, 228, 229, 242, 244, 246, 247, 248 |
| 276 | C: 48, 50, 51, **53**, **54**, 273, 275, 276, 278, 279, **280**; E: 57, 59 |

[a] $AS_{10c}$ and $AS_{15c}$ denote the antigenic sites located within 10 and 15 Å from NGS, respectively.
[b] Amino acid positions of NGS analyzed in the present study.
[c] Epitopes and amino acid positions possibly covered by N-glycans attached to the indicated amino acid positions. Amino acid positions used for the computation of $d_N/d_S$ are denoted in boldface. Amino acid positions with underlines indicate those constituting sequons during some period in the evolutionary history of human A/H3N2 virus. The numbering system of amino acid sites follows the H3 numbering system with reference to A/Memphis/102/72 (PDB no. 3HMG).

**TABLE 2** $d_N/d_S$ values estimated at antigenic sites possibly covered by N-glycan

| Antigenic site,[a] NGS class,[b] and NGS[c] | $d_N/d_S$[d] | |
|---|---|---|
| | Non-NGS | NGS |
| $AS_{10c}$ | | |
| Maintained | | |
| 63 | NA | 1.26 |
| 122 | NA | NA |
| 126 | 0.73 | **0.11** |
| 133 | 0.43 | 0.44 |
| 144 | **2.25** | 1.54 |
| 246 | 1.01 | 0.82 |
| Avg | 1.30 | 0.82 |
| Lost | | |
| 45 | 0.95 | NA |
| 81 | **0.42** | 0.50 |
| 276 | 1.52 | NA |
| Avg | 1.07 | 0.94 |
| $AS_{15c}$ | | |
| Maintained | | |
| 63 | NA | 1.47 |
| 122 | 0.40 | 0.39 |
| 126 | 0.99 | 0.66 |
| 133 | 2.20 | 1.21 |
| 144* | 1.93 | 0.74 |
| 246 | 0.97 | **0.33** |
| Avg** | 1.42 | **0.73** |
| Lost | | |
| 45 | **0.36** | NA |
| 81 | **0.54** | 0.23 |
| 276 | 1.75 | NA |
| Avg | **0.70** | 0.45 |

[a] $AS_{10c}$ and $AS_{15c}$ denote the antigenic sites located within 10 and 15 Å from NGS, respectively.
[b] NGS class of the indicated amino acid positions. Maintained indicates that the amino acid positions were maintained as NGS after the gains of NGS during evolution. Lost indicates that the NGS at the amino acid positions were lost during evolution.
[c] Amino acid positions of NGS analyzed in the present study. An asterisk is attached to the amino acid position when the difference in $d_N/d_S$ values between the non-NGS and NGS branches was statistically significant (*, $P < 0.05$; **, $P < 0.01$).
[d] The $d_N/d_S$ value at AS when the indicated amino acid positions were non-NGS and NGS. The $d_N/d_S$ values are indicated in boldface when positive selection ($d_N/d_S > 1$) or negative selection ($d_N/d_S < 1$) was detected ($P < 0.05$). NA indicates that $d_N/d_S$ was incalculable because $d_S = 0$.

lutionary history of the virus. The NGS at position 276 reverted to a non-NGS 3 years after its generation in 1993. At position 45, the gains and losses of NGS occurred sporadically since 1997.

**Effect of N-glycosylation on natural selection at AS and NAS.** N-glycans appeared to cover AS when they were attached to positions 45, 63, 81, 122, 126, 133, 144, 165, 246, 276, and 285, as indicated above. For each of these amino acid positions, the effect of N-glycosylation on natural selection at AS was examined by computing $d_N/d_S$ at $AS_{10c}$, $AS_{15c}$, $AS_{10uc}$, and $AS_{15uc}$ for the branches where the amino acid position under consideration was an NGS and a non-NGS separately. When these positions were non-NGS, $d_N/d_S$ values at $AS_{10c}$ and $AS_{15c}$ were estimated to be 0.42 to 2.25 and 0.36 to 2.20, respectively (Table 2). (Note that $d_N/d_S$ was incalculable for some cases because $d_S = 0$.) Positive selection was detected at $AS_{10c}$ of position 144 ($P = 0.03$ by $\chi^2$ test). When the positions mentioned above became NGS, $d_N/d_S$ at $AS_{10c}$ and $AS_{15c}$ generally decreased to 0.11 to 1.54 and 0.23 to 1.47, respectively, and no positive selection was detected. A reduction of $d_N/d_S$ was observed at $AS_{10c}$ of 3 out of 5 positions and at $AS_{15c}$ of 6 out of all 6 positions where the comparison of $d_N/d_S$ between the NGS and non-NGS branches could be made, and the latter observation was statistically significant ($P = 0.03$ by binomial test). Furthermore, the reduction of $d_N/d_S$ at $AS_{15c}$ of position 144 was statistically significant ($P < 0.05$ by $\chi^2$ test).

The NGS analyzed above may be classified into two groups according to whether they were maintained (positions 63, 122, 126, 133, 144, and 246) or lost (positions 45, 81, and 276) during the evolution of human A/H3N2 virus. When the average $d_N/d_S$ values at $AS_{10c}$ and $AS_{15c}$ for the positions included in each class was computed for the NGS and non-NGS branches separately, it was observed that the $d_N/d_S$ value for the NGS branches generally was smaller than that for the non-NGS branches (Table 2). In particular, the difference was statistically significant at $AS_{15c}$ of the class of positions maintained as NGS during evolution ($P < 0.01$ by $\chi^2$ test).

**TABLE 3** $d_N/d_S$ values estimated at antigenic sites possibly not covered by N-glycan

| Antigenic site,[a] NGS class,[b] and NGS[c] | $d_N/d_S$[d] | |
|---|---|---|
| | Non-NGS | NGS |
| $AS_{10uc}$ | | |
| Maintained | | |
| 63 | **0.45** | **0.56** |
| 122 | **0.62** | **0.52** |
| 126 | **0.52** | **0.56** |
| 133 | **0.60** | **0.52** |
| 144 | **0.60** | **0.51** |
| 246 | **0.56** | **0.55** |
| Avg | **0.59** | **0.54** |
| Lost | | |
| 45 | **0.56** | **0.32** |
| 81 | **0.56** | **0.49** |
| 276 | **0.54** | 0.73 |
| Avg | **0.55** | **0.57** |
| $AS_{15uc}$ | | |
| Maintained | | |
| 63 | **0.39** | **0.73** |
| 81 | **0.73** | 0.47 |
| 122 | 0.74 | **0.66** |
| 126 | 0.72 | **0.70** |
| 133 | 0.75 | **0.67** |
| 144 | **0.65** | 0.76 |
| 246 | **0.53** | 0.73 |
| Avg | **0.67** | **0.71** |
| Lost | | |
| 45 | **0.69** | NA |
| 81 | **0.73** | 0.47 |
| 276* | **0.67** | 2.48 |
| Avg | **0.70** | 0.91 |

[a] $AS_{10uc}$ and $AS_{15uc}$ denote the antigenic sites located farther than 10 and 15 Å away from NGS, respectively.
[b] NGS class of the indicated amino acid positions. Maintained indicates that the amino acid positions were maintained as NGS after the gains of NGS during evolution. Lost indicates that the NGS at the amino acid positions were lost during evolution.
[c] Amino acid positions of NGS analyzed in the present study. An asterisk is attached to the amino acid position when the difference in $d_N/d_S$ between the non-NGS and NGS branches was statistically significant (*, $P < 0.05$).
[d] The $d_N/d_S$ value at AS when the indicated amino acid positions were non-NGS and NGS. The $d_N/d_S$ values are indicated in boldface when negative selection ($d_N/d_S < 1$) was detected ($P < 0.05$). NA indicates that $d_N/d_S$ was incalculable because $d_S = 0$.

For $AS_{10uc}$ and $AS_{15uc}$, the $d_N/d_S$ values were estimated to be 0.45 to 0.62 and 0.39 to 0.75, respectively, when the positions 45, 63, 81, 122, 126, 133, 144, 246, and 276 were non-NGS (Table 3). Negative selection was detected in most cases, and no positive selection was detected at all. When these positions became NGS, the $d_N/d_S$ values at $AS_{10uc}$ and $AS_{15uc}$ did not change to any large extent (they were 0.32 to 0.73 and 0.47 to 0.76, respectively), except for $AS_{15uc}$ of position 276, where an increase in $d_N/d_S$ (from 0.67 to 2.48) was observed. Negative selection again was detected in most cases.

It should be noted that the $d_N/d_S$ value at AS may be influenced by natural selection operating not only on the specific function of AS but also on the HA as a whole. To eliminate the possibility that the reduction of $d_N/d_S$ at $AS_{10c}$ and $AS_{15c}$ observed above reflected a change in functional constraints on the entire region of HA, e.g., due to a change in the effective population size of human A/H3N2 virus, $d_N/d_S$ was also computed for $NAS_{10c}$, $NAS_{15c}$, $NAS_{10uc}$, and $NAS_{15uc}$.The $d_N/d_S$ values obtained at NAS generally were much smaller than those obtained at AS, and negative selection was detected in most cases whether the amino acid positions 45, 63, 81, 122, 126, 133, 144, 246, and 276 were non-NGS or NGS (see Tables S2 and S3 in the supplemental material). Although a significant difference in $d_N/d_S$ was observed between NGS and non-NGS branches in some cases, $d_N/d_S$ was usually greater in the former branches than in the latter, which was the opposite of the pattern observed at $AS_{10c}$ and $AS_{15c}$.

## DISCUSSION

After the emergence of A/H3N2 virus in the human population in 1968, NGS were generated at amino acid positions 63, 122, 126, 133, 144, and 246 in the globular head region of HA. Since the glycosidic linkage of oligosaccharides may be flexible, N-glycans may have multiple conformations in solution (26). In addition, the structure of N-glycans may vary among NGS in HA (13, 15). For these reasons, two different distances from NGS, 10 and 15 Å, were used as the possible ranges of amino acid positions covered by N-glycans in the present study. When positions 63, 122, 126, 133, 144, and 246 were non-NGS, the $d_N/d_S$ value was $>1$ on average for both $AS_{10c}$ and $AS_{15c}$ and positive selection was detected at $AS_{10c}$ of position 144, suggesting that $AS_{10c}$ and $AS_{15c}$ of these positions contained target sites of Ab. However, the $d_N/d_S$ values at $AS_{10c}$ and $AS_{15c}$ decreased, and the evidence of positive selection disappeared when the positions mentioned above became NGS. In contrast, the $d_N/d_S$ values at $AS_{10uc}$ and $AS_{15uc}$ generally were $<1$ and did not change to any large extent whether the positions mentioned above were NGS or non-NGS. In addition, the $d_N/d_S$ values at $NAS_{10c}$, $NAS_{15c}$, $NAS_{10uc}$, and $NAS_{15uc}$ often were greater when these positions were NGS than when they were non-NGS, which was the opposite of the pattern observed at $AS_{10c}$ and $AS_{15c}$. These results indicate that natural selection for amino acid substitutions at $AS_{10c}$ and $AS_{15c}$ was weakened due to the gain of NGS in the globular head region of HA, supporting the hypothesis that the N-glycans attached to the NGS in the globular head region of HA prevented the binding of Ab to AS in the evolutionary history of human A/H3N2 virus.

The results obtained in the present study may contradict those obtained in a previous study, where a reduction of amino acid variability was not observed after gains of NGS in the globular head region at ±5 amino acid sites from the NGS in the linear sequence of HA for human A/H3N2 virus (8). There are several differences between the present and previous studies that may have caused conflicting results. First, in the present study, amino acid sites possibly covered by N-glycans were identified using the distance from NGS in the 3D structure of trimeric HA, taking into account the size of N-glycans, whereas in the previous study, the ±5 amino acid sites from NGS were assumed to be covered by N-glycans. In addition, the AS and NAS were analyzed separately in the present study, whereas they were mixed in the previous study. It is therefore likely that the effect of N-glycans on natural selection for amino acid substitutions at AS could be detected more efficiently and specifically in the present study than in the previous study. Second, in the present study, the change in natural selection for amino acid substitutions was measured using $d_N/d_S$, whereas the number of different amino acids was used in the previous study. Since the amino acid variability is more strongly influenced by the mutation rate than the $d_N/d_S$ value, it is likely that

the change in natural selection was more reliably identified in the present study than in the previous study.

The N-glycans attached to NGS at positions 45 and 276 in the stem region of HA appeared to have covered AS in the evolutionary history of human A/H3N2 virus. However, the NGS at these positions have been generated only sporadically and have not been maintained for a long evolutionary period. It is known that natural selection for amino acid substitutions operates strongly around the receptor binding pocket (RBP), which is located in the globular head region of HA, because Ab may interfere with viral entry into the host cell more efficiently when they bind to the amino acid positions closely located to the RBP (24, 29). Indeed, positive selection was detected at $AS_{10c}$ of amino acid position 144, which was located close to the RBP, and the NGS generated in the globular head region appeared to be maintained longer than those generated in the stem region during evolution. It is therefore conceivable that N-glycans attached to the globular head region of HA were advantageous to the virus by preventing the neutralization by Ab, whereas those attached to the stem region were less advantageous.

Furthermore, the NGS at amino acid position 81, which was located in the globular head region, also was lost during evolution. This event apparently occurred after the generation of NGS at position 63, which was located adjacent to position 81 in the 3D structure of HA. The $d_N/d_S$ values at $AS_{10c}$ and $AS_{15c}$ of position 63 were incalculable when this position was a non-NGS. However, experimental studies have shown that N-glycosylations at positions 63 and 81 both block the binding of Ab to AS (1, 21). Since some of the AS that were likely to be covered by N-glycans attached to positions 63 and 81 overlapped (29), and since the NGS at position 63 has been maintained after its generation, it is speculated that the generation of NGS at position 63 compensated for the advantageous effect of N-glycosylation at position 81 by shielding the overlapping amino acid sites from the access of Ab, which may have allowed for the loss of the NGS at position 81 after the generation of the NGS at position 63.

## ACKNOWLEDGMENTS

We thank two anonymous reviewers for valuable comments.

Y.K. was supported by a JSPS Research Fellowship for Young Scientists (21-4991).

## REFERENCES

1. **Abe Y, et al.** 2004. Effect of the addition of oligosaccharides on the biological activities and antigenicity of influenza A/H3N2 virus hemagglutinin. J. Virol. **78**:9605–9611.
2. **Bastien N, et al.** 2009. Parotitis in a child infected with triple-reassortant influenza A virus in Canada in 2007. J. Clin. Microbiol. **47**:1896–1898.
3. **Bernstein HJ.** 2000. Recent changes to RasMol, recombining the variants. Trends Biochem. Sci. **25**:453–455.
4. **Blackburne BP, Hay AJ, Goldstein RA.** 2008. Changing selective pressure during antigenic changes in human influenza H3. PLoS Pathog. **4**:e1000058.
5. **Bohne A, Lang E, von der Lieth CW.** 1999. SWEET-WWW-based rapid 3D construction of oligo- and polysaccharides. Bioinformatics **15**:767–768.
6. **Cherry JL, Lipman DJ, Nikolskaya A, Wolf YI.** 2009. Evolutionary dynamics of N-glycosylation sites of influenza virus hemagglutinin. PLoS Curr. **1**:RRN1001.
7. **Cui J, Smith T, Robbins PW, Samuelson J.** 2009. Darwinian selection for sites of Asn-linked glycosylation in phylogenetically disparate eukaryotes and viruses. Proc. Natl. Acad. Sci. U. S. A. **106**:13421–13426.
8. **Das SR, et al.** 2010. Glycosylation focuses sequence variation in the influenza A virus H1 hemagglutinin globular domain. PLoS Pathog. **6**:e1001211.
9. **Fitch WM.** 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.
10. **Hartigan J.** 1973. Minimum evolution fits to a given tree. Biometrics **29**:53–65.
11. **Igarashi M, Ito K, Kida H, Takada A.** 2008. Genetically destined potentials for N-linked glycosylation of influenza virus hemagglutinin. Virology **376**:323–329.
12. **Katoh K, Misawa K, Kuma K, Miyata T.** 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30**:3059–3066.
13. **Keil W, et al.** 1985. Carbohydrates of influenza virus. Structural elucidation of the individual glycans of the FPV hemagglutinin by two-dimensional 1H n.m.r. and methylation analysis. EMBO J. **4**:2711–2720.
14. **McHardy AC, Adams B.** 2009. The role of genomics in tracking the evolution of influenza A virus. PLoS Pathog. **5**:e1000566.
15. **Mir-Shekari SY, Ashford DA, Harvey DJ, Dwek RA, Schulze IT.** 1997. The glycosylation of the influenza A virus hemagglutinin by mammalian cells. A site-specific study. J. Biol. Chem. **272**:4027–4036.
16. **Nelson MI, Holmes EC.** 2007. The evolution of epidemic influenza. Nat. Rev. Genet. **8**:196–205.
17. **Posada D, Crandall KA.** 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics **14**:817–818.
18. **Russell CA, et al.** 2008. The global circulation of seasonal influenza A (H3N2) viruses. Science **320**:340–346.
19. **Seidel W, et al.** 1991. Intraepidemic variants of influenza virus H3 hemagglutinin differing in the number of carbohydrate side chains. Arch. Virol. **120**:289–296.
20. **Shen J, Ma J, Wang Q.** 2009. Evolutionary trends of A(H1N1) influenza virus hemagglutinin since 1918. PLoS One **4**:e7789.
21. **Skehel JJ, et al.** 1984. A carbohydrate side chain on hemagglutinins of Hong Kong influenza viruses inhibits recognition by a monoclonal antibody. Proc. Natl. Acad. Sci. U. S. A. **81**:1779–1783.
22. **Smith GJ, et al.** 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature **459**:1122–1125.
23. **Suzuki Y.** 2006. Natural selection on the influenza virus genome. Mol. Biol. Evol. **23**:1902–1911.
24. **Suzuki Y.** 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol. Biol. Evol. **21**:2352–2359.
25. **Suzuki Y, Gojobori T.** 1999. A method for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **16**:1315–1328.
26. **Varki A, et al.** 1999. Essentials of glycobiology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
27. **Wang CC, et al.** 2009. Glycans on influenza hemagglutinin affect receptor binding and immune response. Proc. Natl. Acad. Sci. U. S. A. **106**:18137–18142.
28. **Wang W, et al.** 2010. Glycosylation at 158N of the hemagglutinin protein and receptor binding specificity synergistically affect the antigenicity and immunogenicity of a live attenuated H5N1 A/Vietnam/1203/2004 vaccine virus in ferrets. J. Virol. **84**:6570–6577.
29. **Wiley DC, Wilson IA, and Skehel JJ.** 1981. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature **289**:373–378.
30. **Wilson IA, Skehel JJ, Wiley DC.** 1981. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 A resolution. Nature **289**:366–373.
31. **Wolf YI, Viboud C, Holmes EC, Koonin EV, Lipman DJ.** 2006. Long intervals of stasis punctuated by bursts of positive selection in the seasonal evolution of influenza A virus. Biol. Direct. **1**:34.
32. **Zhang M, et al.** 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.