Three-Dimensional Window Analysis for Detecting Positive Selection at Structural Regions of Proteins

Yoshiyuki Suzuki

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima-shi, Shizuoka-ken, Japan

Detection of natural selection operating at the amino acid sequence level is important in the study of molecular evolution. Single-site analysis and one-dimensional window analysis can be used to detect selection when the biological functions of amino acid sites are unknown. Single-site analysis is useful when selection operates more or less constantly over evolutionary time, but less so when selection operates temporarily. One-dimensional window analysis is more sensitive than single-site analysis when the functions of amino acid sites in close proximity in the linear sequence are similar, although this is not always the case. Here I present a three-dimensional window analysis method for detecting selection given the three-dimensional structure of the protein of interest. In the three-dimensional structure, the window is defined as the sphere centered on the α-carbon of an amino acid site. The window size is the radius of the sphere. The sites whose α-carbons are included in the window are grouped for the neutrality test. The window is moved within the three-dimensional structure by sequentially moving the central site along the primary amino acid sequence. To detect positive selection, it may also be useful to group the surface-exposed sites in the window separately. Three-dimensional window analysis appears not only to be more sensitive than single-site analysis and one-dimensional window analysis but also to provide similar specificity for inferring positive selection in the analyses of the hemagglutinin and neuraminidase genes of human influenza A viruses. This method, however, may fail to detect selection when it operates only on a particular site, in which case single-site analysis may be preferred, although a large number of sequences is required.

Introduction

In the study of molecular evolution, it is important to detect natural selection operating at the amino acid sequence level. Selection can be inferred by comparing the rates of synonymous (r_S) and nonsynonymous (r_N) substitutions; $r_S < r_N$ as an indication of positive selection, whereas $r_S > r_N$ suggests negative selection (Hughes and Nei 1988, 1989). In the protein molecule, different amino acid sites usually perform different biological functions, where the type and strength of selection may vary. Parsimony (Suzuki and Gojobori 1999), likelihood (Suzuki 2004), and Bayesian (Yang et al. 2000) methods have been developed for inferring selection at single amino acid sites (single-site analysis). However, these methods appear to have problems. In the Bayesian method, the neutrality test is not conducted at individual sites independently, and the false-positive rate for determining the presence of positive selection is often unduly high (Suzuki and Nei 2002, 2004). The neutrality test is conducted at individual sites independently in the parsimony and likelihood methods. However, a large number of nucleotide substitutions is required to have accumulated at each codon site to detect significant differences between r_S and r_N. These methods are, therefore, useful for detecting selection when it operates more or less constantly over evolutionary time, but less useful when selection operates temporarily, although this appears to be the case for most biological innovations.

When the number of nucleotide substitutions that have accumulated at each codon site is small, several sites may be grouped to increase the total number of nucleotide substitutions. Although it may be difficult to identify selected sites exactly, the neutrality test should be more sensitive than with single-site analysis. There have been two approaches for grouping sites. In the first approach, the amino acid sites involved in similar functions are grouped because the type and strength of selection is likely to be similar for these sites (Hughes and Nei 1988, 1989). This approach is applicable when the functions of amino acid sites are known, but this is not the case for most proteins of interest. In the second, a sliding window is used to detect selection (one-dimensional window analysis) (Clark and Kao 1991). It is not necessary to understand the functions of amino acid sites in this approach. However, the sites included in a window are not necessarily involved in similar functions, because functions are determined by the three-dimensional structure rather than by the linear sequence. It is possible that the type and strength of selection is different among the sites in a window. In this case, the sensitivity of neutrality test should be low (Endo, Ikeo, and Gojobori 1996).

The purpose of this paper is to solve these problems, at least partially, by developing a method of three-dimensional window analysis for detecting selection in structural regions of proteins. It is shown that three-dimensional window analysis not only is more sensitive than single-site analysis and one-dimensional window analysis but also provides as much specificity as these methods for inferring positive selection in the analyses of the hemagglutinin (*HA*) and neuraminidase (*NA*) genes of human influenza A viruses.

Materials and Methods

Three-Dimensional Window Analysis

This method is based on the assumption that the amino acid sites located in close proximity in the three-dimensional structure are more likely to be involved in similar functions than those located in close proximity in the linear sequence. The rationale behind this assumption is the fact that functions of amino acid sites are determined by the three-dimensional structure rather than by the linear

Key words: Three-dimensional structure, window analysis, positive selection, human influenza A virus, hemagglutinin, neuraminidase.

E-mail: yossuzuk@lab.nig.ac.jp.

Mol. Biol. Evol. 21(12):2352–2359. 2004
doi:10.1093/molbev/msh249
Advance Access publication September 8, 2004

sequence. In the presence of the coordinate data for the protein of interest, the three-dimensional window in the three-dimensional structure is defined as a sphere centered on the α-carbon of an amino acid site. The window size is the radius of the sphere. The amino acid sites whose α-carbons are included in the window are grouped for the neutrality test. Here, for simplicity only the α-carbons are used. In addition, most of the data available in the Protein Data Bank (PDB) consist of α-carbon coordinates. Note that some sites may be included more than once in a window when the protein molecule forms homomultimers. In such cases, each site may be counted only once to make the neutrality test unbiased. The three-dimensional window is moved within the three-dimensional structure by sliding the central site progressively along the primary sequence from the N-terminus towards the C-terminus.

When the coordinate data for the side chains are available, the amino acid sites exposed on the protein surface in the three-dimensional window may be identified and grouped separately to allow for detection of positive selection, which is known to operate on the exposed sites in many cases (Hughes 1999). The exposed sites can be identified using relative solvent accessibility, which is defined as the solvent accessibility of a given site divided by the maximum solvent accessibility of the corresponding amino acid (Kabsch and Sander 1983). The maximum accessibilities for 20 amino acids are available in Rost and Sander (1994), and the accessibility of a given site can be computed by DSSP (version DsspCMBI-April-2000) (Kabsch and Sander 1983) using the coordinate data. An amino acid site is judged exposed when the relative accessibility is greater than 16% (Rost and Sander 1994).

The neutrality test is conducted for the grouped sites by extending the parsimony and likelihood methods for detecting selection at single amino acid sites. In the parsimony method of single-site analysis, the average numbers of synonymous (s_S) and nonsynonymous (s_N) sites as well as the total numbers of synonymous (c_S) and nonsynonymous (c_N) substitutions for a given phylogenetic tree are computed at each codon site of the sequences. Positive and negative selection are inferred if c_N and c_S are significantly larger than the expected values, respectively. In the parsimony method of three-dimensional window analysis, the s_S and s_N values are averaged and the c_S and c_N values are summed over the grouped sites to obtain S_S, S_N, C_S, and C_N. The neutrality test is conducted similarly to the single-site analysis, where s_S, s_N, c_S, and c_N are replaced with S_S, S_N, C_S, and C_N, respectively. In the likelihood method of single-site analysis, the r_N/r_S value is estimated at each codon site by the maximum-likelihood method. Positive and negative selection are inferred if $r_N/r_S > 1$ and $r_N/r_S < 1$, respectively, and the likelihood value is significantly larger than that obtained under the assumption $r_N/r_S = 1$. In the likelihood method of three-dimensional window analysis, the r_N/r_S value for the grouped sites is estimated under the assumption that the r_N/r_S values are the same for these sites. The neutrality test is conducted by comparing the likelihood value with that obtained under the assumption $r_N/r_S = 1$ for these sites.

In this paper, three-dimensional window analysis uses the parsimony method. One of the possible problems of the parsimony method is that the number of nucleotide substitutions may be underestimated when the sequence divergence is large, because this method does not correct for multiple substitutions. However, the probability of multiple substitutions appears to be small in the following two data sets because of the small sequence divergence as indicated below (Saitou 1989).

Analysis of the *HA* Genes of Human Influenza A Viruses

Influenza A viruses are etiological agents of influenza. HA is an envelope glycoprotein of 566 amino acids and forms homotrimers in the virion. This protein is classified into 15 subtypes (H1 to H15) according to the antigenicity that is determined mainly by five epitopes (A, B, C, D, and E). The amino acid positions included in each epitope are as follows: positions 122, 124, 126, 130 to 133, 135, 137, 138, 140, 142 to 146, 150, 152, and 168 in epitope A; positions 128, 129, 155 to 160, 163 to 165, 186 to 190, 192 to 194, and 196 to 198 in epitope B; positions 44 to 48, 50, 51, 53, 54, 273, 275, 276, 278 to 280, 294, 297, 299, 300, 304, 305, and 307 to 312 in epitope C; positions 96, 102, 103, 117, 121, 167, 170 to 177, 179, 182, 201, 203, 207 to 209, 212 to 219, 226 to 230, 238, 240, 242, 244, and 246 to 248 in epitope D; and positions 57, 59, 62, 63, 67, 75, 78, 80 to 83, 86 to 88, 91, 92, 94, 109, 260 to 262, and 265 in epitope E (Wiley, Wilson, and Skehel 1981; Macken et al. 2001; Wright and Webster 2001). Positive selection has been identified at many of these positions in H3 HA from human influenza A viruses (Fitch et al. 1997; Bush et al. 1999; Suzuki and Gojobori 1999). In addition, influenza viruses are usually passaged in embryonated chicken eggs before isolation. It has been reported that 22 amino acid sites (egg-selection sites) of the H3 HA of human influenza A viruses are positively selected during the passage (Bush et al. 2000). These sites are positions 111, 126, 137, 138, 144, 145, 155, 156, 158, 159, 185, 186, 193, 194, 199, 219, 226, 229, 246, 248, 276, and 290. Therefore, the sensitivity and specificity of single-site analysis, one-dimensional window analysis, and three-dimensional window analysis, were examined using this gene. It was hypothesized that positive selection should be predicted for the windows related to epitopes and egg-selection sites, whereas negative selection should be inferred at other windows.

All nucleotide sequences of the H3 *HA* genes of human influenza A viruses were extracted from the international nucleotide sequence database (DDBJ release 56). The sequences, including ambiguous nucleotides and minor gaps, were removed. In addition, I eliminated all sequences except for one when multiple sequences had been determined for the same strain. The accession numbers of the remaining 24 sequences were AB019354 to AB019357, AF017270, AF348176, AF363502 to AF363504, AF382318, AF382320, AF382322, AF382324, AJ252129, AJ252131, AJ289703, AY035591, AY271794, J02132, J02135, M55059, U26830, U97740, and X05907.

A multiple alignment of these sequences was made using ClustalW (Thompson, Higgins, and Gibson 1994). The phylogenetic tree was constructed by the neighbor-joining method (Saitou and Nei 1987) with the number of synonymous substitutions (Nei and Gojobori 1986). The total and average branch lengths of the phylogenetic tree were 0.539 and 0.0120, respectively. Single-site analysis was conducted with ADAPTSITE (version 1.3) (Suzuki, Gojobori, and Nei 2001). The two-parameter model (Kimura 1980) of nucleotide mutation was used for estimating s_S and s_N. The transition/transversion ratio (R) was estimated as the ratio of the transitional to transversional nucleotide diversities at the fourfold degenerate site and was determined to be $R = 3.09$. In one-dimensional window analysis, the window was defined as a certain number (window size) of continuous amino acid sites in the linear sequence, and all sites included in the window were grouped. The neutrality test was conducted in a way similar to three-dimensional window analysis. The window was progressively moved from the N-terminus towards the C-terminus of the sequence. For three-dimensional window analysis, the coordinate data for H3 HA from strain A/Hong Kong/1/68 (H3N2) with the accession number 1KEN in the PDB was used (Barbey-Martin et al. 2002). The HA of this strain is one of the best-characterized HAs among all H3 strains. The significance level was 5% for all neutrality tests.

Analysis of the *NA* Genes of Human Influenza A Viruses

NA is also an envelope glycoprotein of 469 amino acids, and forms homotetramers in the virion. This protein is classified into nine subtypes (N1 to N9). Several epitopes have been identified in N2 NA of human influenza A viruses. One of the best-characterized epitopes is called NC–41, which consists of positions 326 to 330, 343 to 347, 366 to 372, 399 to 403, and 430 to 435 (Colman et al. 1987; Wright and Webster 2001). There appears to be no indication of positive selection for any epitope of NA in the literature. Single-site analysis, one-dimensional window analysis, and three-dimensional window analysis, were used to examine the gene for indications of positive selection.

All nucleotide sequences of the N2 *NA* genes of human influenza A viruses were extracted from the DDBJ (release 56). The sequences, including ambiguous nucleotides and minor gaps, were removed. In addition, I eliminated all sequences except for one when multiple sequences had been determined for the same strain. The accession numbers of the remaining 120 sequences were AB126623, AF038-260 to AF038265, AF382329, AF382332, AF386761, AF386763, AF503463 to AF503469, AF503471, AF5-33730, AF533731, AF533733, AF533735, AF533738, AF533742 to AF533745, AF533747 to AF533750, AJ2-91403, AJ307599 to AJ307606, AJ307609, AJ307611, AJ307612, AJ307614 to AJ307621, AJ307623, AJ307624, AJ307626, AJ307627, AJ307629, AJ316063, AJ457931 to AJ457938, AJ457940, AJ457942 to AJ457946, AJ457956, AJ457957, AJ457960, AJ457962 to AJ457966, AJ489846 to AJ489849, AY271795, D10164, K01150, U42632 to U42637, U42770 to U42780, U43417 to U43424, U43426, U51245 to U51247, and U71140 to U71143.

These sequences were analyzed in a way similar to the analysis of the *HA* genes. The total and average branch lengths of the phylogenetic tree were 1.04 and 0.00438, respectively. The R value was estimated as 1.53. The coordinate data for N2 NA was available for the strain A/Tokyo/3/67 (H2N2) with the accession number 1NN2 in the PDB (Varghese and Colman 1991). The NA of this strain is one of the best-characterized NAs among all N2 strains.

Results

Positive Selection for the *HA* Genes

In this study, the 24 sequences of the *HA* genes of human influenza A viruses represent a relatively small data set, whereas the 120 sequences of the *NA* genes represent a relatively large data set. Although both positive and negative selection can be detected by single-site analysis, one-dimensional window analysis, and three-dimensional window analysis, the performance for inferring positive selection is emphasized in this paper because it is related to biological innovation and, consequently, is more interesting than negative selection.

Positive selection was not inferred for any site of HA by single-site analysis. Note that positive selection has been inferred for many sites by previous analyses using many (> 100), but partial, sequences of the same gene, as mentioned above. It is likely that the number of sequences analyzed in this study was so small that the number of nucleotide substitutions that had accumulated at each codon site was insufficient to detect significant differences between r_S and r_N.

One-dimensional window analysis was used to increase the number of nucleotide substitutions for the neutrality test. Various window sizes, ranging from 2 to 20 were examined. However, positive selection was not inferred for any window of any size.

In three-dimensional window analysis, window sizes ranging from 1 Å to 10 Å were examined to make the average number of sites in a window compatible with that in one-dimensional window analysis (≤ 20) (table 1). The number of amino acid sites in a window increased on average as the window size increased, and its distribution was roughly unimodal for each window size (Supplementary Material online). The analyses with window sizes of 1 Å, 2 Å, and 3 Å were the same as single-site analysis because only one site was included in every window. One window each was inferred as positively selected when the window size was 6 Å, 7 Å, and 8 Å. The positively selected window of 6 Å consisted of five sites that were included in epitope A (three sites were also involved in egg selection) and two sites that were involved neither in any epitope nor in egg selection. However, the latter sites, positions 136 and 139, were invariable codon sites. Note that invariable codon sites do not contribute very much to the neutrality test because $c_S = c_N = 0$ at these sites. This window was, therefore, considered to be involved in epitope A. The positively selected window of 7 Å was apparently obtained by adding position 140, which was included in epitope A,

Table 1
Positively Selected Windows in Three-Dimensional Window Analysis Where Exposed and Buried Sites for H3 HA of Human Influenza A Viruses Were Not Distinguished

Window Size (Å)	Average Number of Sites[a]	Positively Selected Window[b]
1	1.00	None
2	1.00	None
3	1.00	None
4	3.04	None
5	3.69	None
6	6.23	(**135**, [136], **137***, [**138***], [139], **145***, **146**)
7	8.96	(**135**, [136], **137***, [**138***], [139], [**140**], **145***, **146**)
8	11.18	(**135**, [136], **137***, [**138***], [139], [**140**], **144***, **145***, **146**)
9	14.67	None
10	19.37	None

[a] The average number of amino acid sites in the three-dimensional window of a given size.

[b] The amino acid positions in positively selected window are indicated within parentheses for each window. The positions included in epitope A are bold-faced. The positions involved in egg selection are indicated with asterisks and the invariable codon positions are indicated within brackets.

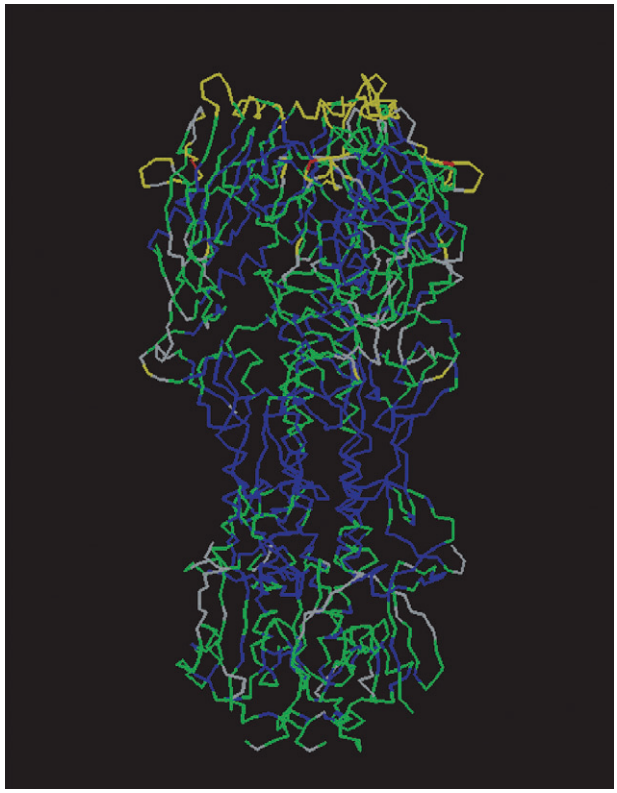

FIG. 1.—Selection profile of the H3 HA of human influenza A viruses inferred by three-dimensional window analysis without distinguishing the exposed and buried sites (window size = 8 Å). The coordinate data for the HA homotrimer of the strain A/Hong Kong/1/68 (H3N2) was obtained from the PDB (accession number 1KEN). Only the backbone structure was visualized by RASMOL (version 2.6) (Sayle and Milner-White 1995). The central amino acid sites of the three-dimensional windows where $r_S < r_N$ (significant; positively selected), $r_S < r_N$ (not significant), $r_S = r_N$, $r_S > r_N$ (not significant), and $r_S > r_N$ (significant; negatively selected) are colored red, yellow, grey, green, and blue, respectively.

to the positively selected window of 6 Å. However, because this position was also an invariable codon site, the positively selected window of 7 Å was indistinguishable from that of 6 Å from the statistical viewpoint, and the former window could be reduced to the latter. The positively selected window of 8 Å was apparently obtained by adding position 144, which was a variable codon site and involved in epitope A (and also in egg selection), to the positively selected window of 7 Å. This window was, therefore, also involved in epitope A (fig. 1).

Because the coordinate data for the side chains were available, the sites exposed on the protein surface in each window were identified and grouped separately for the neutrality test. To make the average number of exposed sites in a window 20 or less, window sizes ranging from 1 Å to 14 Å were examined (table 2). Positive selection was inferred for 1, 1, 4, 10, 11, 9, 7, and 10 windows for window sizes of 6 Å, 8 Å, 9 Å, 10 Å, 11 Å, 12 Å, 13 Å, and 14 Å, respectively. (When the positively selected windows inferred for different window sizes consisted of the same set of amino acid sites, the window was shown only for the smaller window size in table 2.) The positively selected window of 6 Å consisted of three sites, which were all included in epitope A (two sites were also involved in egg selection). Note here that the minimum number of exposed sites in a window required for inferring positive selection (related to epitope A) was three, which was smaller than that (seven) required when the exposed and buried sites were not distinguished (table 1). The positively selected window of 8 Å was apparently obtained by adding an invariable codon site (position 140) and a variable codon site (position 144) to the positively selected window of 6 Å. Because both sites were included in epitope A, this window was also considered to be involved in epitope A. Three of the positively selected windows of 9 Å and one of 11 Å were apparently obtained by adding the sites included in epitope A and invariable codon sites, as well as small numbers of sites involved in epitope D (and egg selection), to the positively selected window of 6 Å, suggesting that they were also largely involved in epitope A. All other positively selected windows of any size, except for two windows of 14 Å, consisted almost exclusively of the sites included in epitope B, containing positions 189, 190, 192, 193, and 198 in common, which were all included in epitope B. These windows were, therefore, considered to be involved in epitope B. Note that the positively selected window related to epitope B was not inferred when the exposed and buried sites were not distinguished (table 1). Two positively selected windows of 14 Å consisted of a mixture of positions included in epitopes A and B, indicating that this window size was too large to distinguish these epitopes. Note that the sites that were included in the positively selected window but involved neither in any epitope nor in egg selection were all invariable codon sites, except for position 134, which was adjacent to the sites (positions 133 and 135) included in epitope A, suggesting that all positively selected windows were involved in epitopes and egg selection.

Table 2:
Positively Selected Windows in Three-Dimensional Window Analysis of Exposed Sites for H3 HA of Human Influenza A Viruses

Window Size (Å)	Average Number of Sites[a]	Positively Selected Window[b]
1	1.00	None
2	1.00	None
3	1.01	None
4	1.85	None
5	2.07	None
6	2.93	(**135**, **137***, **145***)
7	3.96	None
8	4.86	(**135**, **137***, [**140**], **144***, **145***)
9	6.21	([*96*], **135**, **137***, [**140**], **145***, [224], [225], *226**)
		(**135**, **137***, [**140**], **144***, **145***, [224], [225], *226**)
		(**135**, **137***, [**140**], **141**, **144***, **145***, [224], [225], *226**)
		([187], 188, 189, 190, [192], 193*, 198, [199*])
10	8.16	(156*, 157, 159*, 160, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(156*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(**131**, 155*, 156*, 157, 159*, 160, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(155*, 156*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(155*, 156*, 157, 158*, 159*, 160, 189 190, [192], 193*, 194*, 196, 197, 198)
		(156*, 157, 159*, 160, [162], 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 157, 159*, 160, [162], 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(156*, 157, 158*, 159*, 160, [162], 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 157, 158*, 159*, 160, [162], 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
11	10.64	([95], [*96*], [100], [101], **135**, **137***, [**140**], **145***, [224], [225], *226**)
		(156*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(156*, 158*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(155*, 156*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 158*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198)
		(156*, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(**131**, 156*, 157, 158*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(156*, 157, 158*, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 157, 158*, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(**131**, 155*, 156*, 157, 158*, 159*, 160, [187], 188, 189, 190, [192] 193*, 194*, 196, 197, 198, [199*])
12	13.23	(155*, 156*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], [*216*])
		(156*, 157, 158*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(155*, 156*, 157, 158*, 159*, 160, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
13	16.22	(156*, 158*, [162], [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*, [*216*])
		(156*, 157, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*)
		(155*, 156*, 157, 159*, 160, [162], 163, 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*)
		(155*, 156*, 157, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*)
14	19.62	(128, 129, **131**, [**132**], **133**, 134, **135**, 155*, 156*, 157, 158*, 159*, 160, [162], 163, 189 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(129, **131**, [**132**], **133**, 134, **135**, 155*, 156*, 157, 158*, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*])
		(156*, 157, 158*, 159*, 160, [162], [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], [*216*])
		(156*, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*, [*216*])
		(155*, 156*, 157, 158*, 159*, 160, [162], [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], [*216*])
		(155*, 156*, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*, [*216*])
		(156*, 157, 158*, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*, [*216*])
		(155*, 156*, 157, 158*, 159*, 160, [162], 163, [187], 188, 189, 190, [192], 193*, 194*, 196, 197, 198, [199*], *214*, [*216*])

[a] The average number of amino acid sites in the three-dimensional window of a given size.

[b] The amino acid positions in positively selected window are indicated within parentheses for each window. The positions included in epitopes A, B, and D are bold-faced, underlined, and italicized, respectively. The positions involved in egg selection are indicated with asterisks and the invariable codon positions are indicated within brackets. When the positively selected windows inferred for different window sizes consisted of the same set of amino acid sites, the window was shown only for the smaller window size.

Positive Selection for the *NA* Genes

In single-site analysis of NA, positive selection was inferred at position 267, where the c_S and c_N values were estimated as 0 and 11, respectively. Unfortunately, the biological function of this position did not appear to be known.

In one-dimensional window analysis, however, positive selection was not inferred at any window with sizes ranging from 2 to 20, despite the fact that the number of nucleotide substitutions was increased for the neutrality test compared with single-site analysis. Note that all windows containing position 267 were not inferred as

positively selected. This is apparently because the adjacent sites of position 267 in the linear sequence were negatively selected, and the effect of positive selection at position 267 was diluted when adjacent sites were added for the neutrality test. In fact, the c_S/c_N values at positions 266 and 268 were 1/0 and 2/1, respectively, indicating $r_S > r_N$ at both sites, although not significant.

Window sizes ranging from 1 Å to 10 Å in three-dimensional window analysis were examined to make the average number of sites included in a window 20 or less. The analyses with the window sizes of 1 Å, 2 Å, and 3 Å were essentially the same as single-site analysis because the number of sites included in a window was equal or very close to unity (1.00 for 1 Å and 2 Å and 1.01 for 3 Å). Positive selection was not inferred for any window of the sizes ranging from 4 Å to 10 Å, except for a window of 10 Å. This window consisted of positions **370**, **403**, [427], [428], [429], [**430**], **431**, **432**, [**433**], [**434**], **435**, and [439], where the sites that were included in NC–41 are bold-faced and the invariable codon sites are indicated within brackets. Most of these sites were included in NC–41. In addition, all sites that were not included in NC–41 were invariable codon sites. These results suggest that this window is involved in NC–41. Note that all windows including position 267 were not inferred as positively selected, apparently because the adjacent sites of position 267 in the three-dimensional structure were negatively selected. In fact, the c_S/c_N values at positions 265, 266, 268, and 269, which were in the closest proximity to position 267, were 3/4, 1/0, 2/1, and 0/0, respectively, indicating $r_S \geq r_N$ at all sites, although not significant.

Because the coordinate data for the side chains were available, the sites exposed on the protein surface in each window were identified and grouped separately for the neutrality test. To make the average number of exposed sites in a window 20 or less, window sizes ranging from 1 Å to 15 Å were examined. Positive selection was not inferred for any window of any size, except for one window of 10 Å. This window consisted of positions **370**, [**430**], **431**, **432**, [**433**], [**434**], and **435** (with the same notations as above), which were all included in NC–41. Note that these positions were the subset of positions in the positively selected window inferred without distinguishing the exposed and buried sites, as indicated above. Note also that all windows containing position 267 were not inferred as positively selected, although this position was judged as exposed.

Discussion

In three-dimensional window analysis of the *HA* gene, which represents a relatively small data set, positive selection was inferred for the windows related to epitopes A and B. These results appear to be reliable because these epitopes have been well characterized as positively selected, as mentioned above. In contrast, positive selection was not inferred in single-site analysis and one-dimensional window analysis. Three-dimensional window analysis, therefore, appears to be more sensitive than single-site analysis and one-dimensional window analysis for detecting positive selection. In three-dimensional window analysis, a smaller number of sites were required for inferring positive selection (related to epitope A) when only the exposed sites were grouped than when the exposed and buried sites were not distinguished. In addition, positive selection related to epitope B was inferred only when the exposed sites were grouped. These results were apparently obtained because the amino acid sites involved in similar functions were grouped more efficiently using only the exposed sites than using both the exposed and the buried sites, which, in turn, was more efficient than one-dimensional window analysis. Nevertheless, the specificity of these methods did not appear to be different, because false-positive results did not appear to be obtained in any analysis.

Similar results were obtained in the analysis of the *NA* gene, which represents a relatively large data set. No window was inferred as positively selected in one-dimensional window analysis, but one window related to NC–41 was inferred as positively selected in three-dimensional window analysis. These results appear to be reliable because the monoclonal antibody against NC–41 is known to reduce the fitness of influenza A viruses, although it is unclear whether NC–41 is involved in neutralization (Wright and Webster 2001). It is interesting that position 267 was inferred as positively selected only by single-site analysis, apparently because the adjacent sites of this position were negatively selected both in the linear sequence and in the three-dimensional structure. Although the biological function of position 267 does not appear to be known, it may be worth examination because the parsimony method of single-site analysis is known to be conservative (Suzuki and Nei 2002).

Three-dimensional window analysis, however, appears to have some problems. First, the three-dimensional structure of at least one of the sequences analyzed has to be known. This problem, however, may be solved to some extent as the number of three-dimensional structures increases exponentially in the PDB. Second, it is implicitly assumed that the three-dimensional structures of all sequences analyzed are the same as that of the reference sequence for which the coordinate data are available. Strictly speaking, this assumption should almost always be violated because the three-dimensional structures of proteins should not be exactly the same if one amino acid site is different. However, the overall structures of proteins are known to be quite stable, even if the sequence divergence is large (Branden and Tooze 1999). Therefore, the assumption may be acceptable as long as closely related sequences are analyzed and relatively large window sizes are used. Third, the most appropriate window size for detecting selection may be different according to the proteins analyzed. For example, the number of amino acid sites that contact with an antibody in general ranges from 15 to 22 (Klein and Horejsi 1997), suggesting that the window sizes examined in this study are appropriate. However, such information is usually not available, and various window sizes may be examined for each data set (Fares et al. 2002). Fourth, even when positive selection is inferred, it may be difficult to identify selected sites exactly, as mentioned above. For example, in the analysis of the *HA* genes, the positively selected window of 7 Å was obtained

by adding an invariable codon site to the positively selected window of 6 Å (table 1). Similar examples can also be found in table 2. In these cases, the additional sites did not contribute to the neutrality test but were included in the positively selected windows because the amino acid sites around them were positively selected. Therefore, the invariable codon sites may be eliminated when the functions of amino acid sites in the positively selected windows are examined. At any rate, it is important to recognize that selection inferred by the methods examined in this study remains tentative until it is confirmed by experiments. This is partly because multiple neutrality tests are conducted with a 5% significance level (two-tailed) and positive selection is expected to be inferred for 2.5% of all windows, even if the null hypothesis is true. The significance level may be lowered by using the Bonferroni correction (Sokal and Rohlf 1995), but it does not appear to be suitable in the postgenomic era, where a large number of tests (e.g., for each codon site of the entire coding region) may be conducted based on limited amounts of data (e.g., the number of nucleotide substitutions accumulated at each codon site). Under such conditions, the corrected significance level becomes unrealistically low. It may, therefore, be important not to try to draw conclusions only from the statistical analysis but to confirm the statistical predictions by experiments, because the uncorrected significance level is applicable as long as a particular window is concerned.

In conclusion, three-dimensional window analysis appears to be more sensitive than one-dimensional window analysis for detecting positive selection. Three-dimensional window analysis also appears to be more sensitive than single-site analysis, especially when the number of sequences analyzed is relatively small and positive selection operates over the structural regions of proteins. This method, however, may fail to detect selection when it operates only on a particular site, in which case single-site analysis may be more accurate, although a large number of sequences is required.

The computer programs for the parsimony and likelihood methods of three-dimensional window analysis will be implemented in ADAPTSITE.

Acknowledgments

The author thanks Masatoshi Nei and three anonymous reviewers for their valuable comments on the earlier version of the manuscript. I am also grateful to Satoshi Fukuchi and Akira Kinjo for their valuable suggestions on the analysis of the three-dimensional structure of proteins. I am indebted to Katsuhisa Nakajima and an anonymous reviewer for their valuable information on the amino acid positions included in epitopes of H3 HA of human influenza A viruses.

Literature Cited

Barbey-Martin, C., B. Gigant, T. Bizebard, L. J. Calder, S. A. Wharton, J. J. Skehel, and M. Knossow. 2002. An antibody that prevents the hemagglutinin low pH fusogenic transition. Virology **294**:70–74.

Branden, C., and J. Tooze. 1999. Introduction to protein structure. 2nd edition. Garland Publishing, Inc., New York.

Bush, R. M., W. M. Fitch, C. A. Bender, and N. J. Cox. 1999. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol. Biol. Evol. **16**:1457–1465.

Bush, R. M., C. B. Smith, N. J. Cox, and W. M. Fitch. 2000. Effects of passage history and sampling bias on phylogenetic reconstruction of human influenza A evolution. Proc. Natl. Acad. Sci. USA **97**:6974–6980.

Clark, A. G., and T.-H. Kao. 1991. Excess nonsynonymous substitution at shared polymorphic sites among self-incompatibility alleles of Solanaceae. Proc. Natl. Acad. Sci. USA **88**:9823–9827.

Colman, P. M., W. G. Laver, J. N. Varghese, A. T. Baker, P. A. Tulloch, G. M. Air, and R. G. Webster. 1987. Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. Nature **326**:358–363.

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

Fares, M. A., S. F. Elena, J. Ortiz, A. Moya, and E. Barrio. 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. J. Mol. Evol. **55**:509–521.

Fitch, W. M., R. M. Bush, C. A. Bender, and N. J. Cox. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. Proc. Natl. Acad. Sci. USA **94**:7712–7718.

Hughes, A. L. 1999. Adaptive evolution of genes and genomes. Oxford University Press, New York.

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

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

Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**:2577–2637.

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

Klein, J., and V. Horejsi. 1997. Immunology. 2nd edition. Blackwell Science Ltd, Oxford, UK.

Macken, C., H. Lu, J. Goodman, and L. Boykin. 2001. The value of a database in surveillance and vaccine selection. Pp 103–106 *in* A. D. M. E. Osterhaus, N. Cox, and A. W. Hampson, eds. Options for the control of influenza IV. Elsevier Science, Amsterdam.

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

Rost, B., and C. Sander. 1994. Conservation and prediction of solvent accessibility in protein families. Proteins **20**: 216–226.

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

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

Sayle, R., and E. J. Milner-White. 1995. RASMOL: biomolecular graphics for all. Trends Biochem. Sci. **20**:374.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry. 3rd edition. W. H. Freeman, New York.

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

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

Suzuki, Y., T. Gojobori, and M. Nei. 2001. ADAPTSITE: detecting natural selection at single amino acid sites. Bioinformatics **17**:660–661.

Suzuki, Y., and M. Nei. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol. Biol. Evol. **19**:1865–1869.

———. 2004. False-positive selection identified by ML-based methods: examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. **21**:914–921.

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

Varghese, J. N., and P. M. Colman. 1991. Three-dimensional structure of the neuraminidase of influenza virus A/Tokyo/3/67 at 2.2 Å resolution. J. Mol. Biol. **221**:473–486.

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

Wright, P. F., and R. G. Webster. 2001. Orthomyxoviruses. Pp. 1533–1579 *in* D. M. Knipe, P. M. Howley, D. E. Griffin, R. A. Lamb, M. A. Martin, B. Roizman, and S. E. Straus, eds. Fields virology. 4th edition. Lippincott Williams & Wilkins, Philadelphia.

Yang, Z., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Michele Vendruscolo, Associate Editor

Accepted August 30, 2004