Vaccine 30 (2012) 6327–6337

Contents lists available at SciVerse ScienceDirect

Vaccine

journal homepage: www.elsevier.com/locate/vaccine

Antigenic sites of H1N1 influenza virus hemagglutinin revealed by natural isolates and inhibition assays

Jhang-Wei Huang[a], Wei-Fan Lin[a], Jinn-Moon Yang[a,b,*]

[a] Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, 30050, Taiwan
[b] Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan

ARTICLE INFO

Article history:
Received 9 February 2012
Received in revised form 16 May 2012
Accepted 30 July 2012
Available online 8 August 2012

Keywords:
Influenza vaccine
H1N1 virus
Antigenic site
Hemagglutinin
Antigenic drift

ABSTRACT

The antigenic sites of hemagglutinin (HA) are crucial for understanding antigenic drift and vaccine strain selection for influenza viruses. In 1982, 32 epitope residues (called laboratory epitope residues) were proposed for antigenic sites of H1N1 HA based on the monoclonal antibody-selected variants. Interestingly, these laboratory epitope residues only cover 28% (23/83) mutation positions for 9 H1N1 vaccine strain comparisons (from 1977 to 2009). Here, we propose the entropy and likelihood ratio to model amino acid diversity and antigenic variant score for inferring 41 H1N1 HA epitope residues (called natural epitope residues) with statistically significant scores according to 1572 HA sequences and 197 pairs of HA sequences with hemagglutination inhibition (HI) assays of natural isolates. By combining both natural and laboratory epitope residues, we identified 62 (11 overlapped) residues clustered into five antigenic sites (i.e., A–E) which are highly correlated to the antigenic sites of H3N2 HA. Our method recognizes sites A, B and C as critical sites for escaping from neutralizing antibodies in H1N1 virus. Experimental results show that the accuracies of our models are 81.2% and 82.2% using 41 and 62 epitope residues, respectively, for predicting antigenic variants on 197 paring HA sequences. In addition, our model can detect the emergence of epidemic strains and reflect the genetic diversity and antigenic variant between the vaccine and circulating strains. Finally, our model is theoretically consistent with the evolution rates of H3N2 and H1N1 viruses and is often consistent to WHO vaccine strain selections. We believe that our models and the inferred antigenic sites of HA are useful for understanding the antigenic drift and evolution of influenza A H1N1 virus.

1. Introduction

The H1N1 virus emerged in 1918 and caused about 20–50 million deaths [1]. In 1957, the H1N1 virus disappeared and replaced by H2N2 virus that caused the Asian influenza pandemic [2]. Twenty years later, H1N1 virus reappeared in 1977 and began to circulate worldwide [3]. To design vaccines for the immune system to recognize glycoprotein protein hemagglutinin (HA) of influenza viruses is an efficient way for prevention and control of influenza virus infection. The HA consists of HA1 and HA2 domains and HA1 is the immunogenic part. Here, we focused on HA1 sequences and structures. Accumulated and continual mutations on the HA generate antigenic variants (called antigenic drift), which invade immune system, from circulating viruses. An antigenic site (epitope) is defined as a region of HA protein involving in antibody binding [4,5]. The residue positions located on the antigenic sites are defined as epitope residues. For influenza virus, to identify antigenic sites of HA is the basis for understanding the antigenic drift and vaccine development [5–8].

Previous studies recognized four antigenic sites (A–D) of HA of influenza A/HongKong/68 (H3N2) based on monoclonal antibody-selected variations in two H3N2 epidemic strains (A/Memphis/102/72 and A/Victoria/3/75) [5]. These four sites were then extended into five sites by considering laboratory selected variants and natural isolates [6,7]. Bush et al. [9] summarized 131 H3N2 epitope residues on these five sites [6,7] from previous works. Currently, many methods identifying the antigenic sites of HA highly focused on amino acid mutations, such as hamming distance [10], positively selected codons [11], Shannon entropy [12–15], and a variant Simpson index [16,17], using HA sequences. For H1N1 virus, 32 residues (called laboratory epitope residues) locating on four antigenic sites (Sa, Sb, Ca and Cb) of HA were proposed based on monoclonal antibody-selected variants of influenza A/PR/8/34 [18]. Few studies extended these four sites by analyzing the HA sequences of H1N1 epidemic strains from 1918 and 1934 [19]. Recently, Deem and Pan identified

* Corresponding author at: Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, 30050, Taiwan.
Tel.: +886 3 5712121x56942; fax: +886 3 5729288.
E-mail address: moon@faculty.nctu.edu.tw (J.-M. Yang).

0264-410X/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.vaccine.2012.07.079

161 epitope residues of H1N1 HA by using an entropy-based method and mapping antigenic sites of H3N2 HA onto the antigenic sites of H1N1 HA [12]. These quantitative analyses of the genetic data have revealed the insights of the evolution of influenza viruses. The hemagglutination inhibition (HI) assay is one of the main methods to define the antigenic variant for the influenza surveillance [20,21]. For H3N2 virus, several works have mapped the antigenic and genetic evolution based on HI assays [14,20] and HA sequences. However, the relationship between amino acid changes and antigenic variants remains unclear for H1N1 virus.

To address these issues, we proposed a method to explore antigenic sites of H1N1 virus HA by analyzing both 1572 HA sequences and 197 pairs of HA sequences with HI assays from natural isolates ranging from 1918 to 2009. We identified 41 epitope residues (called natural epitope residues) with significantly high amino acid diversity and antigenic variant scores. According to 41 natural and 32 laboratory epitope residues, we identified 62 residues clustered into five antigenic sites which are highly correlated to the HA antigenic sites of H3N2 virus. Experimental results show that these epitope residues and our models are able to reflect antigenic variants of a given pair HA sequences which are often a vaccine strain and a circulating strain. We also found that our models can reflect the epitope evolution and be consistent to WER vaccine strain selection for 38 seasons (from 1977 to 2010). These results demonstrate that our models are useful for understanding the antigenic drift and selections of WHO influenza A H1N1 vaccine strains.

2. Materials and methods

2.1. Data sets

The HA sequences of human H1N1 virus were downloaded from National Center for Biotechnology Information [22] on June 23, 2011. After removing the sequences with the length shorter than 981 nucleotides or redundant sequences, we finally yielded 1572 seasonal (from 1918 to 2009, called SEQ1572) and 2190 pandemic 2009 HA sequences. Here, we used the set SEQ1572 for identifying epitope residues and antigenic sites of H1N1 HA. These sequences were aligned by using MUSCLE [23] and assigned into 38 influenza seasons for studying the antigenic drifts and WHO vaccine updates.

To quantify the antigenic drift of H1N1, we collected the second dataset (called HIA197), including 197 pairs of HA sequences with HI assays from Weekly Epidemiological Record (WER), from WHO collaborating center and publications (Table S1 in supporting materials). The HI assay, describing whether one (e.g. circulating) strain is recognized by an antibody, is often used for the identification of antigenic variants in current global influenza surveillance system [21]. For each HI assay, we collected the pair HA sequences (327 amino acid residues) and the respective HI assay value (considered as "antigenic distance"). The antigenic distance between strains A and B is defined as $\log_2(HI^{AA}/HI^{AB})$ [24], where HI^{AA} is homologous titer (antisera raised against A) and HI^{AB} is heterologous titer. Among these 197 pairs of HA sequences, 128 pairs with antigenic distance ≥ 2 are considered as "antigenic variants" and the other 69 pairs (antigenic distance < 2) are considered as "similar viruses". For example, the antigenic distance of the pair A/Chile/1/83 and A/Singapore/6/86 is 7 ($\log_2(1280/10)$) and this pair is considered as "antigenic variants". In addition, to compare H1N1 virus with H3N2 virus, we collected 3331 H3N2 HA sequences (called SEQ3331) and 343 HI assays (225 antigenic variants and 118 similar viruses) of H3N2 virus from our previous work [15].

2.2. Amino acid diversity and antigenic variant score

Here, we measured the antigenic drift value of a residue position by computing its amino acid diversity and antigenic variant scores (Fig. 1). We first used Shannon entropy [12–15] to measure the amino acid diversity of a position using the set SEQ1572. Shannon entropy of the residue position i ($i = 1$–327) is defined as $H(i) = -\sum_{T=1}^{20} P(A_i = T)\log(P(A_i = T))$, where $P(A_i = T)$ is the probability of the position i with amino acid type T. The entropy values of 327 HA positions were then normalized into the range from 0 to 1 by dividing each entropy value by the maximum value of entropy over all residues. The position i with high entropy value is more frequently mutated than the one of the position j with low entropy value.

The likelihood ratio (LR), a statistical index, is commonly used in interpretation of laboratory and diagnostic tests [25–27]. Here, the LR(i) was utilized to measure the antigenic variant score of a residue position i and is defined as $\mathrm{LR}(i) = [(TP_i + K_v)/(V + K_v)]/[(FP_i + K_s)/(S + K_s)]$. According to the data set HIA197, TP_i is the number of antigenic variants with mutated position i and V is the total number of antigenic variants (here, V is 128); FP_i is the mutation number of similar viruses on the position i and S is the total number of similar viruses (here, S is 69); The pseudocounts K_v is set to $\sqrt{V}(k_v = 11.3)$ and K_s is set to $\sqrt{S}(k_s = 8.3)$ for zero probability [28] based on various tests. For example, for the position 43, the numbers of *TP* and *FP* are 46 and 4, respectively, among 197 pairs of HA sequences. Therefore, its $LR(i = 43)$ is 2.58.The mutation on the position i with high LR can be a high probability for antigenic variant than this of the mutation on the position j with low LR.

To determine the threshold of the entropy value for identifying high amino acid diversity positions, we used the Z-score to measure the significance of each position (Table S2 in supporting materials). The standard deviation ($\sigma = 0.106$) and mean ($\mu = 0.185$) were calculated from the Shannon entropy of all 327 HA positions. The Z-score of a residue position i ($i = 1$–327) is defined as $Z_i = (H(i) - \mu)/\sigma$, where $H(i)$ is the Shannon entropy of the position i. The Z-score threshold of $H(i)$ is set to 0.5. We also decided the threshold of LR value for identifying high antigenic variant score positions and the threshold is set to 1. LR value greater than 1 indicates that there is more related to antigenic variant than expected by chance. We selected 41 natural epitope residues of the HA based on the thresholds of both LR and entropy values (Table 1).

2.3. Antigenic sites of H1N1 HA

To identify HA antigenic sites, interacting with infectivity-neutralizing antibodies to the escape of immune system, is the basis for studying the antigenic drift and vaccine development [9,14,20,29]. The antibody recognition is highly correlated to the conformation changes on the antigenic sites of HA. We identified the antigenic sites of H1N1 by aligning H1N1 and H3N2 HA sequences and then assigned each H1N1 epitope residue into one of five antigenic sites based on 131 H3N2 HA epitope residues (Table S3 in supporting materials). After H1N1 and H3N2 HA sequences were aligned [30], each H1N1 epitope residue was assigned into the same or the nearest antigenic site according to its aligned H3N2 residue. For example, the H1N1 HA residue 271, which locates on laboratory antigenic site Ca, was aligned to H3N2 HA residue 273 locating on the antigenic site C. Therefore, the residue 271 of H1N1 HA was assigned into antigenic site C (Table 1). Furthermore, the H1N1 HA residue 222, aligned to the residue 225 on H3N2 HA, was assigned into antigenic site D because the nearest antigenic site of the residue 225 is located in the antigenic site D. Based on this rule, we can assign the H1N1 laboratory antigenic

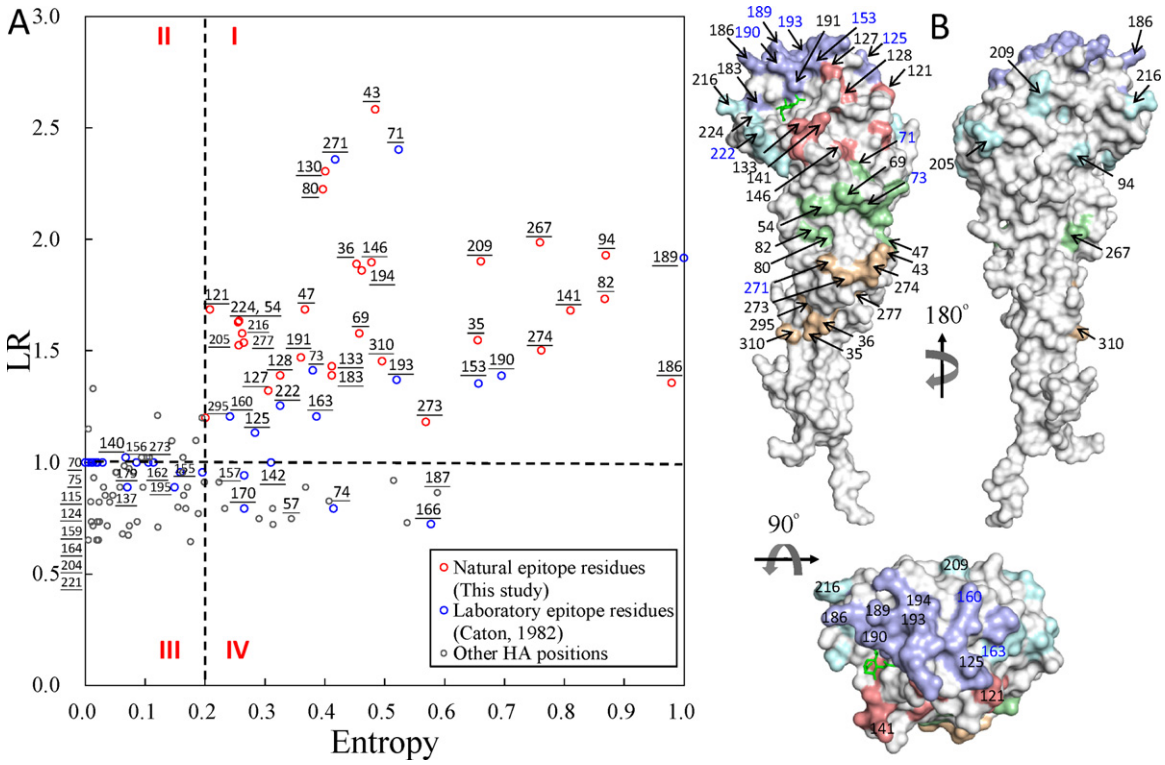

Fig. 1. The relationship between entropy and LR values of HA residues.
(A) HA residues, including laboratory (blue circle) and natural (red circle) epitope residues, are divided into four areas based on entropy and likelihood ratio (LR) values. The residues located in the area I (e.g. 43, 189 and 94) with both high entropy and LR values are selected as natural epitope residues. (B) The structural locations of five epitopes, including A (red), B (blue), C (wheat), D (cyan) and E (green), 41 natural epitope residues, and sialic acid (green) based on HA structure (PDB code 1RVX). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

sites Sa, Sb, Ca and Cb into antigenic sites B + D, B, A + C + D and A + E of H3N2 HA, respectively.

2.4. Variant ratio for vaccine update

To measure WHO vaccine update, we proposed the variant ratio ($VR(y)$) to quantify the vaccine efficiency on the year y. The $VR(y)$ is defined as $VR(y) = V_y/N_y$, where N_y is total number of circulating strains in the year y and V_y is the number of circulating strains which are "antigenic variants" against the vaccine strain in the year y. Here, we considered an influenza vaccine should be updated when the $VR(y) \geq 0.5$ for the year y. The high VR value in the year y mean that the vaccine should be updated due to the vaccine is invalid for the most of circulating strains in this year.

3. Results and discussions

3.1. Antigenic site residues of H1N1 HA

Fig. 1 shows the relationship between entropy and LR values of 118 positions on HA by excluding 209 positions that have no mutations in 197 pairs of HA sequences in set HIA197. Here, the entropy and LR values are used to measure the amino acid diversity and antigenic variant scores of a residue position on HA for the antigenic site identification and antigenic drift of H1N1 virus. A residue position with high entropy means that this position is frequently mutated in the seasonal HA sequences based on set SEQ1572. A residue position with high LR implies that its mutation highly induces antigenic variants. The position 43, locating on the antigenic site C of HA, is the first rank in LR and its LR value is 2.58. Among 197 pairs of HA sequences (HIA197), the position 43 mutates on 50 pairs and 46 of them are antigenic variants (Table 1). We observed that the other positions (e.g., positions 271, 130 and 189) with high LR values have similar behavior, that is, the mutation on these positions should have the high probability to induce antigenic variants. According to the values of entropy (amino acid diversity) and LR (antigenic variant score), 327 HA residues can be classified into four groups (Fig. 1A). In the area I, 41 residues with both high entropy and LR values are considered as "natural epitope residues" (Table 1). Interestingly, only 11 of these 41 residues are overlapped with 32 "laboratory epitope residues" proposed by Caton et al. [18] (Table 1). These 11 epitope residues are consistently located in the protein surface (Table 1 and Fig. 1B) and often highly induce antigenic variants when the circulating strains mutate on these 11 positions. Among these 11 positions, the mutation on position 222 has been associated with severe disease and fatalities for pandemic 2009 H1N1 virus [31] (Fig. 2). Furthermore, the natural and laboratory epitope residues cover 77% (64/83) and 28% (23/83) mutation positions, respectively, among 9 H1N1 vaccine strain comparisons (from 1977 to 2009) (Table 2). The laboratory epitope residues are unable to reflect the evolution (or mutations) of the circulating strains. The low consistence between laboratory and natural epitope residues reflects that they should suffer different population of antibodies.

The positions in the area IV have high amino acid diversity (entropy) values and low antigenic variant (LR) scores. Their LR values < 1 indicate that the mutations on these residues are less preference to cause antigenic variant. For example, the residue 187 mutates on 81 pairs and 31 of them are similar viruses for the set HIA197. Its entropy is high (0.59) but its LR value is low (0.87). This residue was identified as positive selected codon [32]. Conversely, our method, considering both sequences and HI assays, is able to reduce the ill effect of genetic-based methods which only consider

Table 1
The summary of 41 natural epitope residues in H1N1 HA.

Residue position	Entropy	LR	TP[a]	FP[b]	Surface	H1N1 HA epitope		H3N2 HA epitope[d]
						Natural epitope	Laboratory epitope[c]	
121	0.21	1.69	20	2	+[e]	A[f]		
127	0.31	1.32	18	4	+	A		A
128	0.33	1.39	27	7		A		A
130	0.40	2.31	44	5		A		
133	0.41	1.43	23	5		A		A
141	0.81	1.68	29	5	+	A		A
146	0.48	1.90	41	7	+	A		
125	0.28	1.14	20	7	+	B	Sa	B
153	0.66	1.35	26	7	+	B	Sb	B
160	0.24	1.21	22	7	+	B	Sa	B
183	0.41	1.39	22	5		B		B
186	0.98	1.36	53	18	+	B		B
189	1.00	1.92	45	8	+	B	Sb	B
190	0.70	1.39	32	9	+	B	Sb	B
191	0.36	1.47	16	2	+	B		B
193	0.52	1.37	24	6	+	B	Sb	B
194	0.46	1.86	40	7	+	B		B
35	0.66	1.55	23	4	+	C		C
36	0.45	1.89	34	5	+	C		C
43	0.48	2.58	46	4	+	C		C
271	0.42	2.36	24	0	+	C	Ca	C
273	0.57	1.18	32	12	+	C		C
274	0.76	1.50	22	4	+	C		C
277	0.27	1.54	20	3		C		C
295	0.20	1.20	11	2		C		C
310	0.50	1.45	34	9	+	C		C
94	0.87	1.93	28	3	+	D		
163	0.39	1.21	22	7	+	D	Sa	
205	0.26	1.52	17	2	+	D		D
209	0.66	1.90	24	2	+	D		D
216	0.26	1.58	18	2	+	D		D
222	0.33	1.26	55	21	+	D	Ca	
224	0.26	1.63	22	3	+	D		D
47	0.37	1.69	20	2	+	E		
54	0.26	1.63	16	1	+	E		E
69	0.46	1.58	18	2	+	E		
71	0.52	2.40	29	1	+	E	Cb	
73	0.38	1.41	20	4	+	E	Cb	E
80	0.40	2.23	22	0		E		E
82	0.87	1.73	24	3	+	E		
267	0.76	1.99	22	1	+	E		

[a] The number of "antigenic variants" for the position *i* mutated among 197 HA pair sequences.
[b] The number of "similar viruses" for the position *i* mutated among 197 HA pair sequences.
[c] H1N1 HA epitopes identified by Caton et al. [18].
[d] H3N2 HA epitopes defined by Wiley and Skehel [6], Wilson and Cox [7].
[e] The position *i* is the surface residue based on HA structure (PDB code 1RVX [34]).
[f] These five antigenic sites (A–E) are identified by aligning H1N1 and H3N2 HA sequences based on 131 H3N2 HA epitope residues.

the amino acid diversity. Furthermore, the positions in the area II are infrequently mutated and their LR values are high. Finally, the positions in the area III have low LR and entropy values. Here, we discarded these residues in the areas II, III and IV for modeling the antigenic drift and evolution of influenza A H1N1 virus. Please note that some mutations may reflect a neutral polymorphism exist in circulating viral HAs [9,33].

We identified 62 epitope residues by considering both 41 natural and 32 laboratory epitope residues (Table S2 in supporting materials). These 62 residues and the antigenic sites on H1N1 HA protein can be mapped into the five antigenic sites A–E in H3N2 HA protein by using the sequence alignment [30]. Fig. 1B indicates the five epitopes, A (red), B (blue), C (wheat), D (cyan) and E (green), and the epitope residues based on HA structure (PDB code 1RVX [34]). The LR values and structural locations of these 62 epitope residues are shown (Fig. S1 in supporting materials). According to the HA structure (PDB code 1RVX), 35 of 41 natural epitope residues locate on the protein surface. Among 32 laboratory epitope residues, only one residue (residue 271) [18,19] locates on the epitope C. Interestingly, eight residue positions (i.e., 35, 36, 43, 273, 274, 277, 295 and 310) of natural epitope residues are mapped into epitope C.

Recently, Deem and Pan mapped the known H3N2 antigenic sites into H1N1 HA and then extended these sites by using an entropy method to identify additional 31 residues under selective immune pressure for H1N1 virus [12]. Finally, they identified 161 epitope residues and proposed a sequence-based antigenic distance measure, $p_{epitope}$, to predict vaccine effectiveness for H1N1 virus [35]. Among these 161 epitope residues, 40 residues located in the area I (Fig. 1) were totally consistent with our 41 natural epitope residues except the residue 130. For the 197 pairs of HA sequences in set HIA197, the Pearson correlation (0.61) between antigenic distance and $p_{epitope}$ of considering these 40 residues in the area I is much better than the one (0.41) of using total 161 epitope residues (Fig. S2 in supporting materials). In addition, using nine pairs of vaccine strains and dominant circulating strains in seven flu seasons in the Northern hemisphere [35], the Pearson correlations between vaccine effectiveness and number of different amino acids on our natural epitope residues and 40 residues in area I are consistently high (~0.80 in

Table 2
Mutated positions of 9 H1N1 vaccine strain comparisons.

Vaccine A	Vaccine B	HD[a]	Mutated residues							
			Natural epitopes (ratio)[b]	Laboratory epitopes (ratio)[c]	Natural epitope residues					Others
					A	B	C	D	E	
A/USSR/90/77	A/Brazil/11/78	7	4 (4/7, 57%)	0 (0/7, 0%)	128		295	216, 224		56, 253, 258
A/Brazil/11/78	A/Chile/1/83	9	8 (89%)	1 (11%)	121, 128, 130		36, 43, 277	205, **222**[d]		135
A/Chile/1/83	A/Singapore/6/86	14	13 (93%)	6 (43%)	127, 141	**125**, **153**, 186, **189**, **190**, **193**, 194	36, 43	**222**	54	135
A/Singapore/6/86	A/Bayern/7/95	11	2 (18%)	5 (45%)		**153**, **160**				70, 155, 221, 57, 87, 98, 167, 171, 287
A/Bayern/7/95	A/Beijing/262/95	11	11 (100%)	4 (36%)	130, 146	186	43, **271**, 273	**163**, **222**	47, **71**, 80	
A/Beijing/262/95	A/NewCaledonia/20/99	11	10 (91%)	3 (27%)	133	**153**, 183, 186, 191	273, 310	**163**, **222**	69	187
A/NewCaledonia/20/99	A/SolomonIslands/3/2006	12	9 (75%)	2 (17%)	128, 141, 146	194		94, 209	**73**, 82, 267	166, 187,252
A/SolomonIslands/3/2006	A/Brisbane/59/2007	8	7 (88%)	2 (25%)	128, 146	**189**, 194	35, 274		**73**	187
Number of mutations		83	64/83 (77%)	23/83 (23%)	14	17	13	11	9	

Bold values are overlapped positions between laboratory and natural epitope residues.

[a] Hamming distance (HD) of a pair sequences.

[b] Natural covered mutation ratio (n/M), M is the number of mutations of a par vaccines and n is the number of these M mutations covered by 41 natural epitope residues.

[c] Laboratory covered mutation ratio (l/M), M is the number of mutations of a par vaccines and l is the number of these M mutations covered by 32 laboratory epitope residues.

[d] Overlapped positions between laboratory and natural epitope residues.

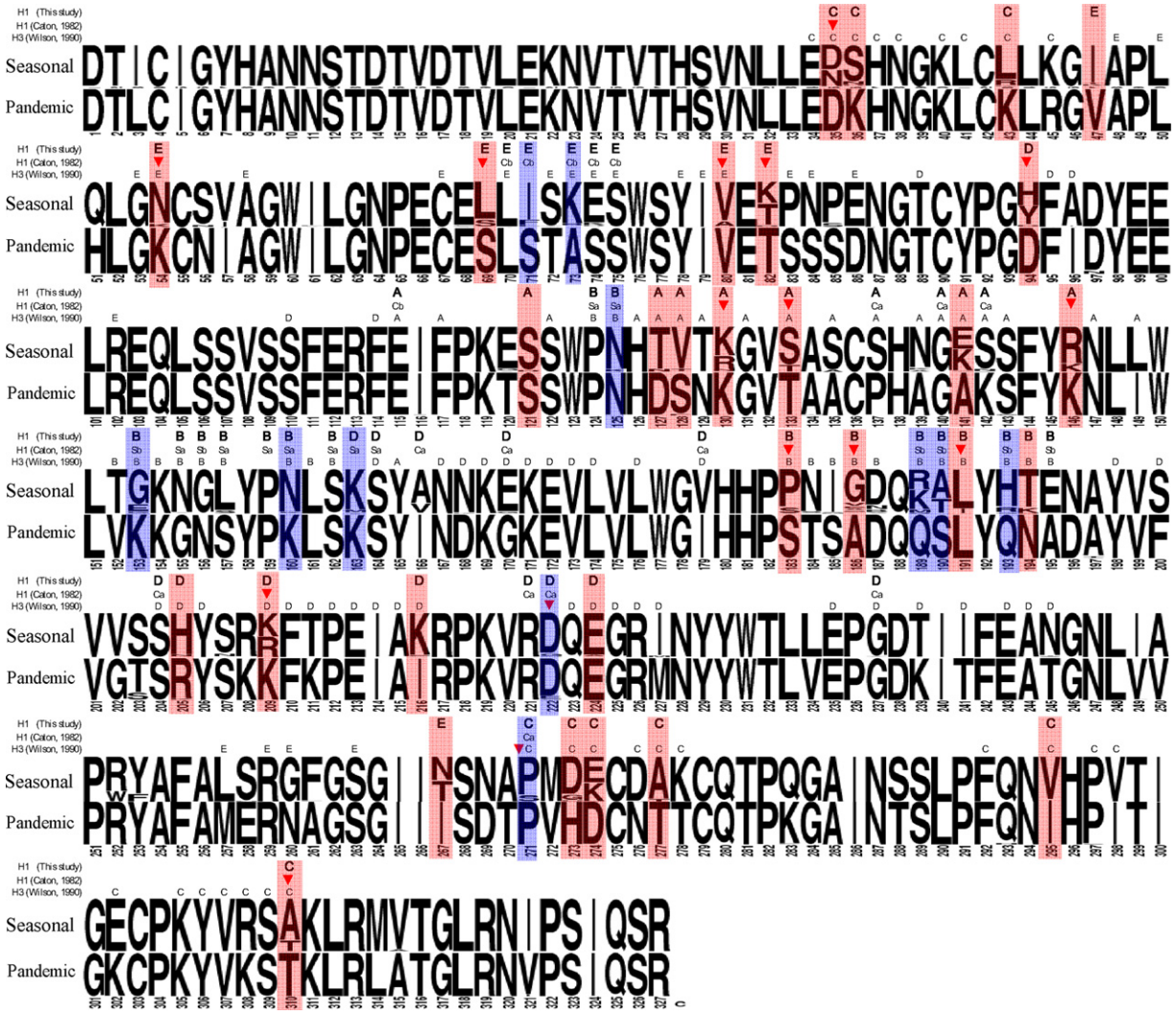

Fig. 2. WebLogo of H1N1 HA residues. The WebLogo of 327 HA amino acid residues of H1N1 virus. The WebLogo of 41 natural epitope residues (red or blue) and 11 overlapped residues (blue) of laboratory and natural epitope residues bases on 1572 H1N1 seasonal and 2190 pandemic 2009 HA sequences. The residues (e.g. positions 35 and 69) sharing at least one amino acid type between seasonal and pandemic viruses denote the red triangles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

Fig. S3 in supporting materials). These results suggest that epitope residues located in the area I are often play the key role for the antigenic distance and vaccine effectiveness.

3.2. Comparisons of natural and laboratory epitopes

Due to the overlap ratio between natural and laboratory epitope residues is low, the comparison between these two antigenic sites is the interesting assessment of the H1N1 HA general antigenicity. Several differences between natural and laboratory epitope residues were observed as follows. First, natural epitope residues are statistically computed from the natural isolate HA sequences, evolved 33 years from 1977 to 2009 with 8 WHO vaccine strain updates. Conversely, the laboratory antigenic sites were derived from 34 laboratory mutant viruses from one virus (i.e. A/PR/8/34 virus) by artificial mutations. Therefore, the screening antibodies for natural epitope residues are naturally similar to human immune system [36] and the antibodies for laboratory epitopes are artificial generated. The laboratory epitope residues and natural epitope residues cover 28% and 77% mutation positions among 9 H1N1 vaccine strain comparisons, respectively (Table 2). Second, the natural epitope residues, which were derived from natural isolates including epidemic strains, are able to reveal the requirements of the emergence of a novel epidemic virus and help the vaccine strains selection. On the other hand, 11 of 32 laboratory epitope residues seldom change and their entropy *Z*-scores are less than 0 (Table S2 in supporting materials) based on SEQ1572 set. Besides, the residue 115 in the laboratory antigenic site Cb was completely conserved. Furthermore, the antibody repertoire and variant viruses for natural epitope identification are more diverse than those of laboratory epitopes. The natural epitopes were identified from 1572 natural isolates that screened by human immune systems [37]. In the other hand, the laboratory epitopes were generated from 34 laboratory mutant viruses that screened by 58 monoclonal antibodies [18]. Finally the numbers of epitope residues in the antigenic site C are 9 and 1 for natural and laboratory epitopes, respectively (Table S2 in supporting materials). The antigenic site C is the most distant site from the receptor-binding site of HA. Besides, 9 of the 11 laboratory epitope Ca residues have no equivalent residue in natural epitopes (except residues 222 and 271). The epitope Ca shows a cleft spanning the monomer-monomer boundary in the HA trimer. The low entropy values of the

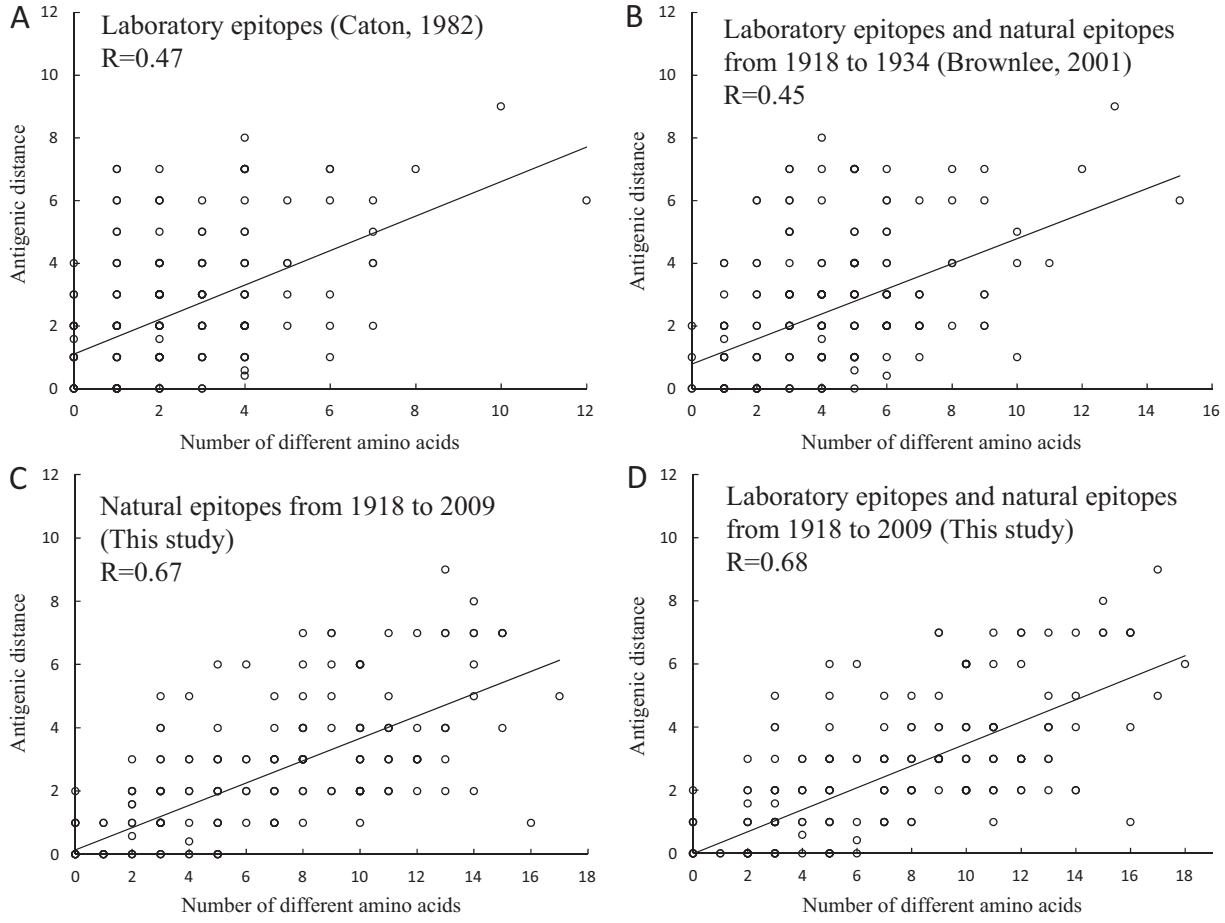

Fig. 3. The correlation between antigenic distance and number of different amino acids. (A) 32 laboratory epitope residues. (B) 52 epitope residues proposed by Brownlee [19]. (C) 41 natural and (D) 62 epitope residues in this study.

epitope Ca residues mean that the residues locating the monomer-monomer interfaces seldom change in natural isolates.

In summary, the origins of screening antibodies and variant viruses are different between natural and laboratory epitopes. 11 of the 32 laboratory epitope residues observed from monoclonal antibody-selected variants seldom mutate in natural isolates. To consider natural isolates, which evolve longer time and are more diverse than laboratory mutant viruses, should reveal a more complete antigenic map of HA1 of H1N1 virus than the one of considering laboratory mutant viruses.

3.3. Epidemic strains and sequence-based antigenic distance

An influenza vaccine strain is often updated to the representative strain of circulating viruses [33] when this vaccine do not match the circulating strains. For the vaccine strain selection, the antigenic and genetic distances are considered as the crucial requirements to understand the emergence of an epidemic strain. The antigenic distance derived from ferret HI assay is widely used to characterize the antigenic properties of circulating strains [21]. However, the HI assay is labor-intensive and time-consuming. Based on HA sequences, we derived the relationships between antigenic distances and number of different amino acids using four computational models with different epitope residues (Fig. 3). The Pearson correlations of these four models are 0.47 (32 laboratory epitope residues), 0.45 (53 epitope laboratory and natural residues from 1918 to 1934 [18]), 0.67 (41 natural epitope residues), and 0.68 (62 laboratory and natural epitope residues).

Nine different WHO H1N1 vaccine strains (after 1977) and their mutation positions between vaccines and circulation strains are listed in Table 2. The laboratory covered mutation ratio (l/M), M is the number of mutations of a pair vaccines and l is the number of these M mutations covered by 32 laboratory epitope residues, is low ($\sim$0.25) on average. Conversely, the natural covered mutation ratio (n/M), n is the number of these M mutations covered by 41 natural epitope residues, is significantly higher ($\sim$0.75) than the l/M value. For example, natural and laboratory epitope residues cover 8 and 1 positions among 9 mutations between the vaccine strains A/Brazil/11/78 and A/Chile/1/83. These results suggested that natural isolates and natural epitope residues play a critical role for identifying the emergence of a novel epidemic strain.

3.4. Changed epitopes and antigenic variants

We used changed-epitope models to measure the antigenic drift for vaccine strain selection and vaccine update (Fig. 4). The antibody recognition of HA is often related to conformation changes on antigenic sites; therefore, to quantify a changed epitope escaping from neutralizing antibodies is the basis to study the antigenic drift [9,14,20,29]. Here, we define the changed epitope if at least two epitope residues mutate (e.g. 41 natural and 32 laboratory epitope residues) on this epitope. The five epitopes of H1N1 HA are defined according to five H3N2 epitopes and their epitope residues proposed by Wilson and Cox [7]. For each epitope of H1N1 HA, we identified its epitope residues by aligning H1N1 and H3N2 HA sequences according to laboratory and natural epitope residues. Here, we predicted a pair HA sequences as “antigenic variants”

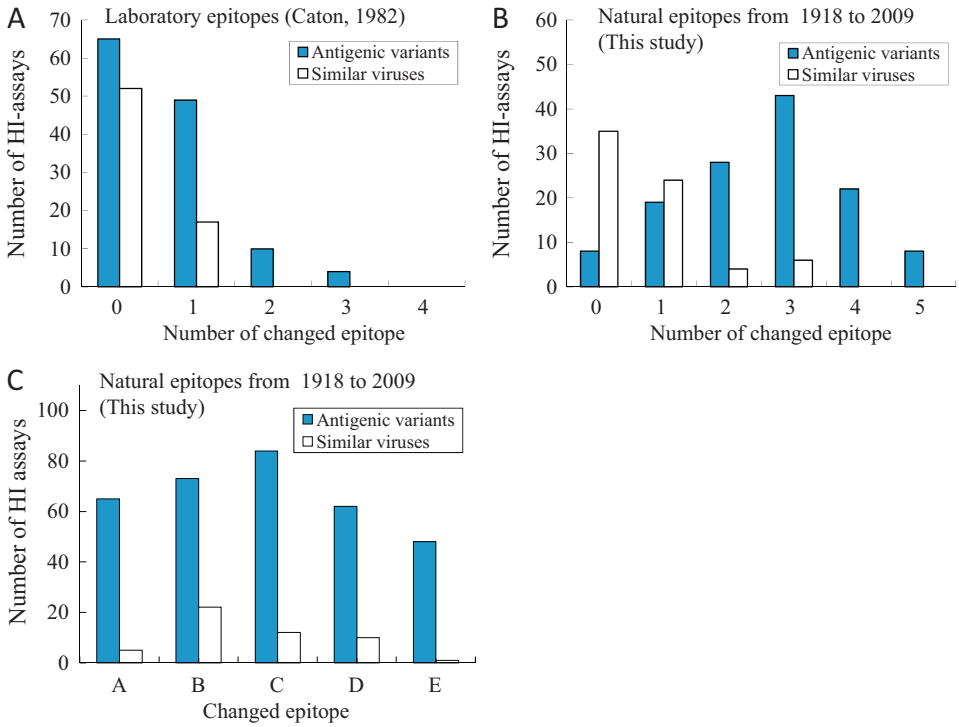

Fig. 4. The relationships between changed epitopes and antigenic variants based on 32 laboratory and 41 natural epitope residues. (A) and (B) are the distributions between the numbers of the changed epitopes and "antigenic variants" using laboratory and natural epitope residues using the set HIA197. Among 128 "antigenic variants" pairs, the numbers of changed epitopes ≤ 1 are 114 and 28 pairs using laboratory and natural epitope residues, respectively. (C) The relationship between antigenic variants and changed epitopes on A–E uses 41 natural epitope residues.

when at least two changed epitopes among five epitopes [15]; conversely, a pair of HA sequences were predicted as "similar viruses".

For 197 pairs in the set HIA197, the accuracies of our models for predicting antigenic variants are 81.2% (160/197, including 101 antigenic variants and 59 similar viruses) (Fig. 4B) and 82.2% using 41 and 62 epitope residues, respectively. Among 128 "antigenic variants" pairs in the set HIA197, 65 (50.8%) and 114 (89.1%) HA pairs have zero and one changed epitopes, respectively (Fig. 4A), based on 32 laboratory epitope residues. Conversely, 8 (6%) and 27 (21.2%) HA pairs have zero and one changed epitopes, respectively (Fig. 4B), by using 41 natural epitope residues. For example, for A/Bayern/7/95 and A/New Caledonia/20/99 (Table S4 in supporting materials), 15 mutation positions induced zero (on laboratory epitopes) and 4 changed epitopes (on natural epitopes). For natural epitope residues, these four changed epitopes include A (mutations on 130, 133 and 146), B (mutations on 153, 183 and 191), C (mutations on 43, 271, 273 and 310), and E (mutations on 47, 69, 71 and 80). These results show that laboratory epitope residues are incomplete for describing the antigenic variants; conversely, natural epitope residues are often able to model changed epitopes and antigenic variants for escaping from the antibody recognition.

Next, we observed the relationships between changed epitopes and antigenic variants for 41 natural epitope residues (Fig. 4C). Using 41 natural epitope residues, the numbers of five changed epitopes for A, B, C, D and E are 65, 73, 84, 62 and 48, respectively, among 128 "antigenic variants". For H1N1 and H3N2 influenza viruses, many studies showed that epitopes A and B near to the receptor binding site play a key role for neutralizing antibodies and antigenic drifts [8,33,38,39]. Interestingly, the epitope C is the most frequently changed among these five epitopes for H1N1 viruses in set HI197.

To observe the role of the epitope C of H1N1 and H3N2 viruses, we analyzed the genetic and antigenic data for the epitope C. Based on the sets SEQ1572 (H1N1) and SEQ3331 (H3N2), we computed the entropy and LR values of epitope residues (Fig. S4 in supporting materials). For each epitope with n epitope residues, we used the mean ($u_{\text{entropy}} = \sum_{i=1}^{n} H(i)/n$) of entropy values and the geometric mean ($u_{LR} = \sqrt[n]{\prod_{i=1}^{n} LR(i)}$) of LR values to measure the amino acid diversity and antigenic variant scores, respectively. For H1N1 virus, the u_{entropy} and u_{LR} of the epitopes A, B and C are consistently high (Fig. S4A and S4B). Conversely, for H3N2 virus, u_{entropy} and u_{LR} of the epitope C are significantly lower than the ones of epitopes A and B (Fig. S4C and S4D). These results suggest that epitopes A, B and C play a key role for neutralizing antibodies and antigenic variants for H1N1 viruses.

3.5. Vaccine update

WHO surveillance network detects the emergence and spread of antigenic variants of H1N1 circulating strains, which is the basis to update the influenza vaccine [21]. Here, we considered an emerging antigenic variant according to emergence of antigenic variants of WER strain, which was the dominant strain in an influenza season. For an influenza season, we applied changed-epitope model using 41 and 62 natural epitope residues to measure changed epitopes and *VR* for detecting the emergence of antigenic variants (Fig. 5). During 38 influenza seasons (1572 seasonal and 2190 pandemic HA sequences in 1977–2010), our model yielded high *VR* values when the vaccines were updated according to WER strains (Fig. 5A). For the 78–79, 82–83, 95–97 and 07–08 seasons, the *VR* values are high (≥0.7 means that the vaccine is invalid for 70% circulating strains in the season); conversely, the *VR* values are low (≤0.4 means that the vaccine is valid for 60% circulating strains in the season) for the 76–78, 86–90, and 00–06 seasons. In addition, for the 82–83 season, 85% (*VR* = 0.85) circulating strains are changed on three epitopes A, C, and D.

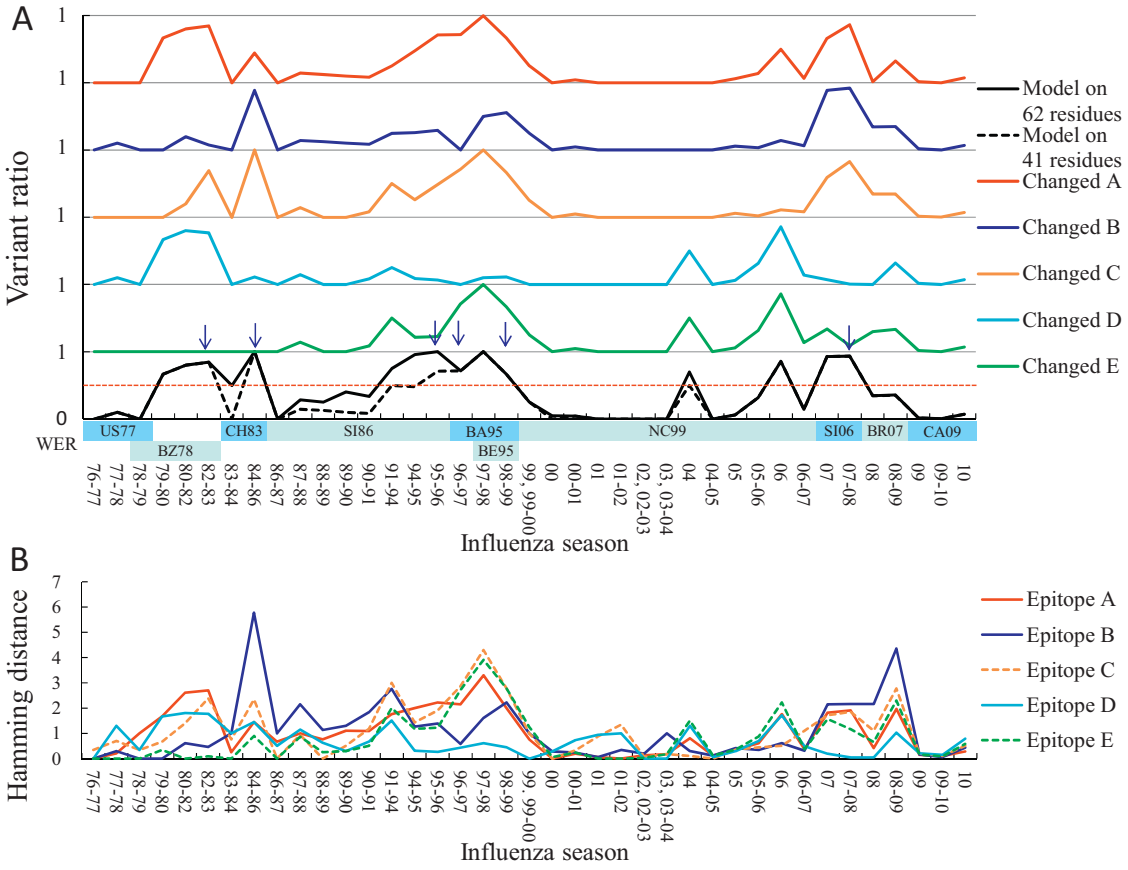

Fig. 5. The epitope evolution and antigenic drift for 38 influenza seasons.
(A) The distributions of variant ratios of 1572 HA sequences for 38 seasons (1976–2010) on five epitopes using 41 natural and 62 epitope residues. Our models are often consistent to WER vaccine strain selection. 6 seasons with emerging variants and followed by the update of WER strain in the next season are labeled (blue arrow). (B) The hamming distances of five epitopes on 1572 HA sequences from 1976 to 2010 based on 41 natural epitope residues.

To observe the antigenic site evolution of H1N1 HA, we calculated the hamming distance (HD) on five epitopes based on 41 natural epitope residues (Fig. 5B). For 9 seasons with WER strain updates, the average HDs of epitopes A, B, C, D and E were 1.66, 1.95, 2.07, 0.80 and 1.28 respectively. The epitope C has the largest HD among these five epitopes. Moreover, previous works showed that the changes in epitopes A and B are often associated with the antigenic drift and have high neutralization efficiencies for the H3N2 virus [33]. Based on this study, we found that the low neutralizing efficiency epitope C should play a key role for the antigenic variants and antigenic drift for H1N1 virus, which may explain the reasons why H1N1 virus causes lower mortality than H3N2 virus. One of the possible explanations is that the change on the epitope C is able to reduce the binding of low neutralizing antibodies and increase the overall viral neutralization [8].

3.6. Evolution rate

The evolution rates of HA1 domains of H1N1 and H3N2 viruses are compared in order to understand the selection pressure. The fixation rates for HA1 domains of H1N1 and H3N2 viruses between 1980 and 2000 were 1.8×10^{-3} and 3.7×10^{-3} nucleotide substitution/site/year, respectively [36]. The 131 epitope residues of H3N2 HA1 were proposed by considering both the monoclonal antibody-selected variants [5] and natural isolates collected from 1968 to 1997 [9]. These 131 epitopes were commonly used to describe antigenic drifts and the changes on antigenic site [8,10,15,20]. This study followed this strategy for antigenic site identification by considering both antibody-selected variants and natural isolates.

Based on the fixation rates of HA1 for H1N1 (1.8×10^{-3}) and H3N2 (3.7×10^{-3}) viruses, we observed the relationship between the fixation rates and the numbers of epitope residues on HA1 domains. Interestingly, the ratio (0.48, 1.8/3.7) of the fixation rates is very closed to the ratio (0.47, 62/131) of the numbers of epitope residues for H1N1 and H3N2 viruses, where 62 is the number of epitope residues identified by our method considering both natural and laboratory epitope residues (Table S2 in supporting materials). This result implies that our model can estimate the evolution rate of HA protein for H1N1 viruses. The comparisons of antigenic sites of H1N1 and H3N2 viruses may provide new insights for understanding evolution of influenza viruses although the isolation years of H1N1 (mainly from 1977–2009) and H3N2 (1968–1997) viruses are different.

3.7. Pandemic H1N1 2009

In 2009, the H1N1 pandemic virus that originated from swine influenza virus presented new threats to public health worldwide. Over five thousands of HA sequences have been deposited in the public database [22]. For these HA sequences, most of the HA positions are conserved due to the short collection period (i.e. from 2009 to 2011). Therefore, the antigenic drift and antigenic structures are poor understood for the pandemic 2009 virus.

The seasonal H1N1 influenza viruses provide valuable opportunity to understand the evolution and antigenic drift for H1N1 pandemic virus. Recently, some studies showed the possibility of applying the evolution trends on 32 laboratory epitope residues

of seasonal influenza viruses to predict the antigenic structure of the pandemic 2009 virus [40]. Here, we applied our models to understand the antigenic evolution of pandemic 2009 virus. Our 62 selected epitope residues can be divided into three groups according to the comparison between seasonal and pandemic viruses (Fig. 2). The first group includes 15 positions (i.e., 70, 75, 115, 121, 124, 125, 140, 159, 162, 163, 164, 204, 221, 224 and 237) which share the same residue types for seasonal and pandemic viruses. The second group includes 16 positions (35, 54, 69, 80, 82, 94, 130, 133, 146, 183, 186, 191, 209, 222, 271 and 310) that share at least one residue type (labeled with red triangle in Fig. 2) between these two type viruses. For example, the position 69 has two residue types (i.e. S and L) in seasonal virus and the amino acid S appeared in pandemic 2009 virus. For the seasonal viruses, the position 69 was dominated by amino acid S from 1977 to 1997 and then dominated by amino acid L from 1998 to 2009. This evolution implies that the amino acid L may dominate position 69 for pandemic 2009 virus. We observed the other 15 positions that have similar behaviors. The third group includes 31 positions which have completely different residue types between seasonal and 2009 pandemic viruses. For these 16 positions of the second group, our models can provide useful amino acid evolution of the HA protein for pandemic 2009 virus.

4. Conclusions

This study demonstrates our method is robust and feasible for exploring the antigenic sites of H1N1 HA using HA sequences and HI assays of natural isolates. We inferred 41 natural epitope residues which have statistically significant amino acid diversity and antigenic variant score among 327 positions. We identified 62 epitope residues and five antigenic sites of HA by considering both 41 natural and 32 laboratory epitope residues. Experimental results show that our models and these 62 epitope residues can measure the genetic diversity and antigenic variant to detect the emergence of epidemic strains for 1572 HA sequences in 38 seasons. In addition, our model proposes epitopes A, B and C should be critical for escaping from neutralizing antibodies in H1N1 virus and vaccine development. Our model is consistent with theoretical evolution rates of H1N1 viruses. We believed that our method is useful for studying influenza A virus evolution and antigenic drifts.

Acknowledgments

This paper was supported by National Science Council, partial supports of Ministry of Education and National Health Research Institutes (NHRI-EX100-10009PI). This paper is also particularly supported by “Aim for the Top University Plan” of the National Chiao Tung University and Ministry of Education. J.-M. Yang also thanks Core Facility for Protein Structural Analysis supported by National Core Facility Program for Biotechnology.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.vaccine.2012.07.079.

References

[1] Johnson NP, Mueller J. Updating the accounts: global mortality of the 1918–1920 Spanish influenza pandemic. Bull Hist Med 2002;76(1):105–15.
[2] Kawaoka Y, Krauss S, Webster RG. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. J Virol 1989;63(11):4603–8.
[3] Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of influenza A viruses. Microbiol Rev 1992;56(1):152–79.
[4] Atassi MZ, Smith JA. A proposal for the nomenclature of antigenic sites in peptides and proteins. Immunochemistry 1978;15(8):609–10.
[5] Wiley DC, Wilson IA, Skehel JJ. Structural Identification of the antibody-binding sites of Hong-Kong influenza hemagglutinin and their involvement in antigenic variation. Nature 1981;289(5796):373–8.
[6] Wiley DC, Skehel JJ. The structure and function of the hemagglutinin membrane glycoprotein of influenza-virus. Annu Rev Biochem 1987;56:365–94.
[7] Wilson IA, Cox NJ. Structural basis of Immune recognition of influenza-virus hemagglutinin. Annu Rev Immunol 1990;8:737–71.
[8] Ndifon W, Wingreen NS, Levin SA. Differential neutralization efficiency of hemagglutinin epitopes, antibody interference, and the design of influenza vaccines. Proc Natl Acad Sci USA 2009;106(21):8701–6.
[9] Bush RM, Fitch WM, Bender CA, Cox NJ. Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 1999;16(11):1457–65.
[10] Plotkin JB, Dushoff J, Levin SA. Hemagglutinin sequence clusters and the antigenic evolution of influenza A virus. Proc Natl Acad Sci USA 2002;99(9):6263–8.
[11] Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM. Predicting the evolution of human influenza A. Science 1999;286(5446):1921–5.
[12] Deem MW, Pan KY. The epitope regions of H1-subtype influenza A, with application to vaccine efficacy. Protein Eng Des Sel 2009;22(9):543–6.
[13] Pan K, Deem MW. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. J R Soc Interface 2011;8(64):1644–53.
[14] Huang JW, King CC, Yang JM. Co-evolution positions and rules for antigenic variants of human influenza A/H3N2 viruses. BMC Bioinformatics 2009;10(Suppl. 1):S41.
[15] Huang JW, Yang JM. Changed epitopes drive the antigenic drift for influenza A (H3N2) viruses. BMC Bioinformatics 2011;12.
[16] Plotkin JB, Dushoff J. Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus. Proc Natl Acad Sci USA 2003;100(12):7152–7.
[17] Li J, Wang Y, Liang Y, Ni B, Wan Y, Liao Z, et al. Fine antigenic variation within H5N1 influenza virus hemagglutinin's antigenic sites defined by yeast cell surface display. Eur J Immunol 2009;39(12):3498–510.
[18] Caton AJ, Brownlee GG, Yewdell JW, Gerhard W. The antigenic structure of the influenza-virus A/PR/8/34 hemagglutinin (H1 subtype). Cell 1982;31(2):417–27.
[19] Brownlee GG, Fodor E. The predicted antigenicity of the haemagglutinin of the 1918 Spanish influenza pandemic suggests an avian origin. Philos Trans R Soc Lond B Biol Sci 2001;356(1416):1871–6.
[20] Smith DJ, Lapedes AS, de Jong JC, Bestebroer TM, Rimmelzwaan GF, Osterhaus AD, et al. Mapping the antigenic and genetic evolution of influenza virus. Science 2004;305(5682):371–6.
[21] Russell CA, Jones TC, Barr IG, Cox NJ, Garten RJ, Gregory V, et al. Influenza vaccine strain selection and recent studies on the global migration of seasonal influenza viruses. Vaccine 2008;26. pp. D31–D4.
[22] Bao YM, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus resource at the national center for biotechnology information. J Virol 2008;82(2):596–601.
[23] Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 2004;32(5):1792–7.
[24] Webster R, Cox NJ, Stohr K. WHO manual on animal influenza diagnosis and surveillance. WHO/CDS/CSR/NCS/20025 2002;Rev.1.
[25] Dujardin B, Vandenende J, Vangompel A, Unger JP, Vanderstuyft P. Likelihood ratios – a real improvement for clinical decision-making. Eur J Epidemiol 1994;10(1):29–36.
[26] Grimes DA, Schulz KF. Refining clinical diagnosis with likelihood ratios. Lancet 2005;365(9469):1500–5.
[27] Jaccard J, Sheng D. A comparison of 6 methods for assessing the importance of perceived consequences in behavioral decisions – applications from attitude research. J Exp Soc Psychol 1984;20(1):1–28.
[28] Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 1993;262:208–14.
[29] Lee MS, Chen JS. Predicting antigenic variants of influenza A/H3N2 viruses. Emerg Infect Dis 2004;10(8):1385–90.
[30] Winter G, Fields S, Brownlee GG. Nucleotide sequence of the haemagglutinin gene of a human influenza virus H1 subtype. Nature 1981;292(5818):72–5.
[31] Chutinimitkul S, Herfst S, Steel J, Lowen AC, Ye J, van Riel D, et al. Virulence-associated substitution D222G in the hemagglutinin of 2009 pandemic influenza A(H1N1) virus affects receptor binding. J Virol 2010;84(22):11802–13.
[32] Li W, Shi W, Qiao H, Ho SY, Luo A, Zhang Y, et al. Positive selection on hemagglutinin and neuraminidase genes of H1N1 influenza viruses. Virol J 2011;8:183.
[33] Cox NJ, Bender CA. The molecular epidemiology of influenza viruses. Semin Virol 1995;6:359–70.
[34] Gamblin SJ, Haire LF, Russell RJ, Stevens DJ, Xiao B, Ha Y, et al. The structure and receptor binding properties of the 1918 influenza hemagglutinin. Science 2004;303(5665):1838–42.
[35] Pan K, Subieta KC, Deem MW. A novel sequence-based antigenic distance measure for H1N1, with application to vaccine effectiveness and the selection of vaccine strains. Protein Eng Des Sel 2011;24(3):291–9.

[36] Ferguson NM, Galvani AP, Bush RM. Ecological and immunological determinants of influenza evolution. Nature 2003;422(6930):428–33.
[37] Fanning LJ, Connor AM, Wu GE. Development of the immunoglobulin repertoire. Clin Immunol Immunopathol 1996;79(1):1–14.
[38] Barbey-Martin C, Gigant B, Bizebard T, Calder LJ, Wharton SA, Skehel JJ, et al. An antibody that prevents the hemagglutinin low pH fusogenic transition. Virology 2002;294(1):70–4.
[39] Xu R, Ekiert DC, Krause JC, Hai R, Crowe JE, Wilson IA. Structural basis of preexisting immunity to the 2009 H1N1 pandemic influenza virus. Science 2010;328(5976):357–60.
[40] Igarashi M, Ito K, Yoshida R, Tomabechi D, Kida H, Takada A. Predicting the antigenic structure of the pandemic (H1N1) 2009 influenza virus hemagglutinin. PLoS One 2010;5(1):e8553.