Bioinformatics Advance Access published August 25, 2009

Sequence analysis

Using a Mutual Information Based Site Transition Network to Map the Genetic Evolution of Influenza A/H3N2 Virus

Zhen Xia[1], Gulei Jin[1] , Jun Zhu[1] and Ruhong Zhou[1,2,3,*]

[1]Institute of Bioinformatics, Zhejiang University, Hangzhou 310027, P R China

[2]Computational Biology Center, IBM Thomas J Watson Research Center, Yorktown Heights, NY 10598

[3]Department of Chemistry, Columbia University, New York, NY 10027

Associate Editor: Prof. John Quackenbush

ABSTRACT

Motivation: Mapping the antigenic and genetic evolution pathways of influenza A is of critical importance in the vaccine development and drug design of influenza virus. In this paper, we have analyzed more than 4,000 A/H3N2 hemagglutinin (HA) sequences from 1968 to 2008 to model the evolutionary path of the influenza virus, which allows us to predict its future potential drifts with specific mutations.

Results: The mutual information (MI) method was used to design a site transition network (STN) for each amino acid site in the A/H3N2 hemagglutinin sequence. The STN network indicates that most of the dynamic interactions are positioned around the epitopes and the RBD regions, with strong preferences in both the mutation sites and amino acid types being mutated to. The network also shows that antigenic changes accumulate over time, with occasional large changes due to multiple co-occurring mutations at antigenic sites. Furthermore, the cluster analysis by subdividing the STN into several subnetworks reveals a more detailed view about the features of the antigenic change: The characteristic inner sites and the connecting inter-subnetwork sites are both responsible for the drifts. A novel 5-step prediction algorithm based on the STN shows a reasonable accuracy in reproducing historical HA mutations. For example, our method can reproduce the 2003-2004 A/H3N2 mutations with ~70% accuracy. The method also predicts seven possible mutations for the next antigenic drift in the coming 2009-2010 season. The site transition network approach also agrees well with the phylogenetic tree and antigenic maps based on HA inhibition assays.

Availability: All code and data are available at http://ibi.zju.edu.cn/birdflu/

Contact: ruhongz@us.ibm.com

Supplementary Information: http://ibi.zju.edu.cn/birdflu/

Key Words: H3N2 evolution, site transition network, mutual information, antigenic drift, phylogentic tree, co-evolution

1 INTRODUCTION

With a wide geographic distribution and a rapid evolution rate, the influenza virus is one of the most emergent and fatal diseases of human and poultry. It causes an estimated 500,000 deaths worldwide in humans and tens of millions in avians every year (Cox and Subbarao, 2000; Hilleman, 2002; Horimoto and Kawaoka, 2005). The main influenza antigens targeted by our immune system are the viral surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA) (Horimoto and Kawaoka, 2005). Based on the antigenic distinctions of the HA and NA proteins, the influenza A viruses have been subdivided into 16 HA and 9 NA subtypes, respectively (Fouchier, et al., 2005; Webster, et al., 1992). The HA1 domain of the HA is the primary protein component of vaccines to provide protective immunity to influenza A virus infection. The accumulated substitutions at the antibody sites of HA1, called antigenic drifts, are the main factors causing the influenza virus to escape human immunity in A/H3N2 and other subtypes (Webster, et al., 1982). Therefore, a number of approaches have been applied to understand the antigenic evolution of HA (Huelsenbeck and Dyer, 2004; Nielsen and Yang, 1998; Suzuki, 2004; Yang, et al., 2000; Yang and Swanson, 2002; Zhou, et al., 2008). These methods effectively identified several single mutation sites that are under the pressure of positive selection (also known as Darwinian selection, a process by which new advantageous genetic variants sweep a population) (Bush, et al., 1999; Fitch, et al., 1997; Plotkin and Dushoff, 2003; Suzuki and Gojobori, 1999; Zhou, et al., 2008). Recently, in a pioneering work by Smith et al. the antigenic evolution of HA1 (A/H3N2) has been mapped using the hemagglutination inhibition (HI) assays, which provides a direct link between a viral genotype and an inferred phenotype (Smith, et al., 2004). Later studies suggest that multiple mutations at antigenic sites cumulatively enhance the antigenic drift of influenza virus A/H3N2 (Du, et al., 2008; Shih, et al., 2007). By analyzing 2,248 A/H3N2 HA1 amino acid sequences, Shih et al. found that positive selections are ongoing most of the time (i.e., not sporadic), and multiple mutations at antigenic sites cumulatively enhance antigenic drift (Shih, et al., 2007). Similarly, by examining the nucleotides over the entire genome of human A/H3N2 viruses, Du et al. have shown that the co-occurring nucleotide modules apparently underpin the dynamics of human A/H3N2 evolution (Du, et al., 2008). These studies show that antigenic drift might be enhanced by simultaneous multi-site mutations in addition to the accumulation of single site mutations (Shih, et al., 2007). Given that the influenza A virus is under rapid mutations, with substitution rates estimated to be 5.7×10^{-3} substitutions per site per year for HA1 domain (Chen and Holmes, 2006), it is non-trivial to predict future mutations and make an efficient influenza vaccine before a potential variant causes epidemics or even pandemics. To make the situation worse, the evolutionary dynamics of influenza A virus can be shaped by a complex interplay between the aforementioned rapid mutations, frequent

*To whom correspondence should be addressed: ruhongz@us.ibm.com

reassortments, widespread gene flows, natural selections, functional interactions among gene segments, and global epidemiological dynamics. In particular, the frequent reassortments can cause sporadic jumps in antigenic space (or antigenic shifts) (Holmes, et al., 2005; Nelson, et al., 2006; Rambaut, et al., 2008). Whether new vaccines could effectively fight against future strains of influenza A virus still remains a great worldwide concern. Therefore, a better understanding of the evolutional direction of the influenza virus is critical for subsequent development of effective vaccines against future strains (Bush, et al., 1999; Koelle, et al., 2006; van Nimwegen, 2006). In this paper, we focus on antigenic drifts due to rapid mutations in order to understand the interplay between various mutations and propose a predictive model to describe the evolution trajectory based on available HA amino acid sequences (for analysis on the genome-wide reassortments, see excellent studies by Holmes and coworkers (Holmes, et al., 2005; Nelson, et al., 2006; Rambaut, et al., 2008)). Because of the large amount of available sequences and longer history of A/H3N2, the A/H3N2 subtype was chosen for this study instead of the more recent and fatal A/H5N1. However, the methodology developed here should be equally applicable to A/H5N1, once more sequence data becomes available. The varieties and time sequences of mutations (site substitutions) in HA were analyzed, especially for the sites of HA1. The mutual information (MI) method was used to calculate the MI score, or correlation, between any two residue sites, thus generating a MI matrix for all pairs of sites. Then a site-site transitional relevance network, named the Site Transition Network (STN), was built based on the time sequenced MI matrices (Butte and Kohane, 2000; Butte, et al., 2000; Margolin, et al., 2006). The STN network maps the genetic evolution history of the virus, which reveals the underlying mechanism of antigenic drifts. We then applied the clustering analysis to subdivide the MI matrix into clusters, which shows various groups of sites with highly correlated and co-occurring mutations. These clusters also reveal the hidden secrets of the antigenic changes – both characteristic inner sites (sites within a cluster or subnetwork) and connecting inter-cluster (or inter-subnetwork) sites are responsible for the antigenic drifts. These results suggest that the influenza A/H3N2 evolution is often enhanced by simultaneous multi-site co-mutations. Finally, we developed a 5-step prediction algorithm to forecast the potential future A/H3N2 mutations in 2009~2010 season. Our current prediction strategy might shed light in identifying the trends in the HA sequence evolution, and provide guidelines for future vaccine development.

2 METHODS AND MATERIALS

2.1 Network inference algorithm: MI

We used Mutual Information (MI) method to calculate the correlation (or a measure for co-mutation) between any two residue sites (Butte and Kohane, 2000; Butte, et al., 2000; Faith, et al., 2007; Margolin, et al., 2006). Mutual Information value for a pair of discrete variables, x and y (mutation of sites), can be defined as:

$$I(x, y) = S(x) + S(y) - S(x,y) \qquad (1)$$

where $S(t)$ is the entropy of an arbitrary variable t. Entropy for a discrete variable is defined as the average of the log probability of its states:

$$s(t) = -\langle \log p(t_i) \rangle = -\sum_i p(t_i) \log p(t_i) \qquad (2)$$

where $p(t_i)$ is the probability associated with each discrete state. In this case, it is the probability for each residue site to have a mutation at a particular year. We can compute the MI value for each pair of amino acid sites, and thus obtain a MI matrix for HA1 at any given year (see below). The evolution of this MI matrix forms the evolution network. For statistical significance, we also compute the P values using Monte Carlo simulations with one million iterations by defining a null-hypothesis model, in which each pair of existing sites is randomly shuffled.

2.2 Data collection and preparation

All sequences of the HA1 of A/H3N2 were downloaded from NCBI's Influenza Virus Resource up to November 1, 2008 (Bao Y., 2008). Sequences without the record of the year and sequences with a length less than the full length of HA1 (312 residues) were removed. A total of 4,064 sequences were obtained after the above cleaning. Then, all these sequences were aligned by the Muscle program (Edgar, 2004). A small number of sequences are a few residues longer than the conventional length, and these insertions after alignment are removed to be consistent with other sequences and previous literature (with the remaining residues mapped back to the original 312-residue numbering).

All the remaining sequences from the above procedure were divided into 41 bins, with each bin representing sequences from a particular year (41 years for sequences from 1968 to 2008). Considering the large imbalance in the sequence numbers for each year (fewer sequences in earlier years and more sequences in recent years), a sampling with replacement method was used to avoid the overwhelming weight from those recent high-yield years. Ten sequences from each year were randomly selected to form one input sequence sample in each calculation. A total of 2,000 different samples were chosen to ensure that enough statistics were obtained. The output MI matrices from these samples are then averaged (arithmetic average) and normalized (with mean 0 and standard deviation of 1) to generate the final "MI Matrix".

2.3 Predicting future site mutations

With the above mutual information matrix (MI Matrix) we can design a predictive model to identify potential mutations in upcoming seasons, which is based on the fact that there is a strong preference in both the mutation sites and amino acid types being mutated to (more below). A total of five steps are involved in this site mutation prediction. A general example is provided in Section 3.4.1 to illustrate each step in more detail. Before we start, we first define the year that we want to predict as the "Target Year" N, and the years N-1 and N-2 as "Induction Years".

The following are the steps involved. Step 1: Calculate all sites that are under positive selection in HA1 before year N. Here, the "positive selection site" is defined as a site that has been mutated between successive years and then remains fixed in the population for at least 1 year, similar to the definition used by Shih et al.'s (Shih, et al., 2007). Step 2: Find the sites that just mutated in any of the "Induction Years" and also belonged to the positive selection sites. Such sites are considered as the initial state of the present network. Step 3: Use all of the available sequences before year "N" as data source to construct the sequence input sample file described above, and calculate the MI matrix. Step 4: Since the MI matrix quantifies the interaction between any two sites by a mutual information score, for each site X in HA1, we sum up the scores between the site X and all those sites found in Step 2 (i.e., newly mutated positive selection sites in Induction Years). The sites with high MI scores are chosen as

predicted sites. Step 5: Find the most probable amino acid type for each predicted mutation site from Step 4 by searching the historical amino acid type database for each site. Historical data suggest that there is a strong preference for each residue site to have some specific amino acid types (see Results section below). We therefore use the most probable amino acid type other than the current one as the final mutated type. The time evolution of this "mutation or co-mutation trajectory" from the dynamic MI matrices is called the Site-Transition-Network (STN) in the following (see Figure 1 for one example).

3 RESULTS AND DISCUSSION

3.1 Site Transition Network for Influenza A/H3N2 HA1

We calculated the two dimensional MI Matrix for each year using the above procedure and the time sequence of this MI Matrix forms the Site Transition Network (STN). Before showing a full scale 312-sites STN (which seems not necessary, more below), we identified the positive selection sites first. A total of 63 positive selection sites, as defined in the Methods section, are found from 1968 to 2008, consistent with previous results obtained by Shih et al. using a somewhat smaller data set (about half the size of ours) (Shih, et al., 2007). Most of these positive selection sites are within the RBD (receptor binding domain) and other antigenic regions (more later), which are responsible for the most antigenic transitions in the evolution history of A/H3N2 (Shih, et al., 2007). The other sites are mostly the conserved sites, which host the basic skeleton structure of HA1 but have little effect on the antigenic transitions. They have no or little mutual information (interaction) with other sites (i.e., isolated nodes in MI Matrix or STN if plotted). Therefore, in the following we only show a smaller, 63-site (positive selection sites only) MI matrix or STN network for simplicity, unless otherwise explicitly stated. Figure 1 shows one such Site Transition Network. Each node represents a mutation site, and each branch represents the interaction between a pair of mutation sites if its normalized MI score is larger than 0.5 (this threshold is arbitrary, it is chosen to avoid too many branches in the graph for clarity. Any nodes with no branches are removed from the graph). As shown in Figure 1, the network displays a clear trajectory on how the various HA1 sites shape the antigenic transition during the influenza evolution history. It should be noted that the MI Matrix is by itself 2-dimensional (2-D), therefore, with the inclusion of time series the STN becomes a 3-dimensional (3-D) plot if all information to be shown. In order to show key features of the HA1 evolution clearly, we have plotted a 2-D STN with "inter-linked" MI Matrices in Figure 1 (i.e., for positive selection sites in any given year, link to sites from previous years whenever possible), which covers all important antigenic transition periods. Therefore, some sites might occur more than once in Figure 1 (e.g. sites 2, 50, etc). The clusters of sites with the same or similar colors represent that they were substituted in relatively close years, and have more interactions and closer relationships than other sites.

To evaluate the effectiveness of the network, we compared it with the phylogenetic tree (PT) of the HA1 sequences calculated by the program MEGA4 (Tamura, et al., 2007). Figure 2 shows the phylogenetic tree of the A/H3N2 HA1. We found a rough match on the influenza evolutional history between the STN network and the phylogenetic tree. When compared both under the time evolution, the STN network can identify, most of the time, the specific sites among the neighboring groups in the phylogenetic tree. In addition, the connections between two antigenic groups in the STN network can also explain the overlaps in years between neighboring groups in the phylogenetic tree (represented by different color in Figure 2). Larger genetic distance in phylogenetic tree usually indicates a larger difference in antigenic space. For example, the larger genetic distance in the phylogenetic tree between 93-95 and 95-97 groups (genetic distance 0.009, colored grass green in Figure 2) are consistent with many substitutions at sites 135, 145, 226, 262, 172, 197 and 278 in our STN network (shown in a broader color range in Figure 1 near BE92 to WU95 antigenic change and also shown in Table 1). In addition, the genetic distance (0.0019) between 97-99 and 99-00 groups is found to be smaller than the distance (0.012) between 00-02 and 02-03 groups, which is also consistent with substitutions found at sites of 202 and 225 for 99-00 and sites 25, 57, 186 and 189 for 02-03 in our STN network (near SY97-FU02 antigenic transition in Figure 1). However, not all the cases are closely mapped between the phylogenetic tree and the STN network. For example, the EN72-VI75-TX77 successive antigenic transitions show different behaviors: in the phylogenetic tree: The TX77 strain came directly from EN72 but not VI75, while STN network indicates that the TX77 strain came from both EN72 (indicated by substitutions of sites 53, 137, 188) and TX77 (indicated by substitutions of sites 63, 83, 145, 189).

We also compared the STN network with the "antigenic maps" from Smith and coworkers in a recent study (Smith, et al., 2004). The "antigenic map" utilizes the serum HI (hemagglutination inhibition) assays to measure the cross-immunity "distance" between each strain of the influenza to every other strain in an "antigenic space" plot. The distance between any two viral strains represents the cross-immunity diversity of corresponding strains. We found that both the STN network and the antigenic map indicate many similar characteristic features. For example, the site mutations 122, 144, 155, 188, and 207 from the antigenic map responsible for the HK68-EN72 antigenic transition showed the propinquity in the STN network. However, the site transition network shows more connections among different "simultaneous mutation" groups than the antigenic map. It seems that the antigenic drift in A/H3N2 occurs more smoothly at sequence level, indicating that the mutation on antigenic site of HA happens all the time and some positive substitutions might result in a partial structural change in the antigenic regions. Simultaneous occurrence of several such substitutions and structural changes in a specific group of these sites may gain enough power to induce an antigenic change that cumulatively enhances the antigenic drift.

3.2 Cluster analysis revealing co-evolving sites responsible for antigenic drift

To further understand the site transition network and the relationship among multiple co-mutating sites, we performed a cluster analysis on the reduced MI Matrix consisting of 63 positive selection sites obtained from all sequence data up to

2008. The hierarchical cluster analysis method was used to cluster the sites of HA1, with program dChip originally designed for DNA microchip array analysis (Li and Wong, 2003) (http://www.biostat.harvard.edu/complab/dchip/

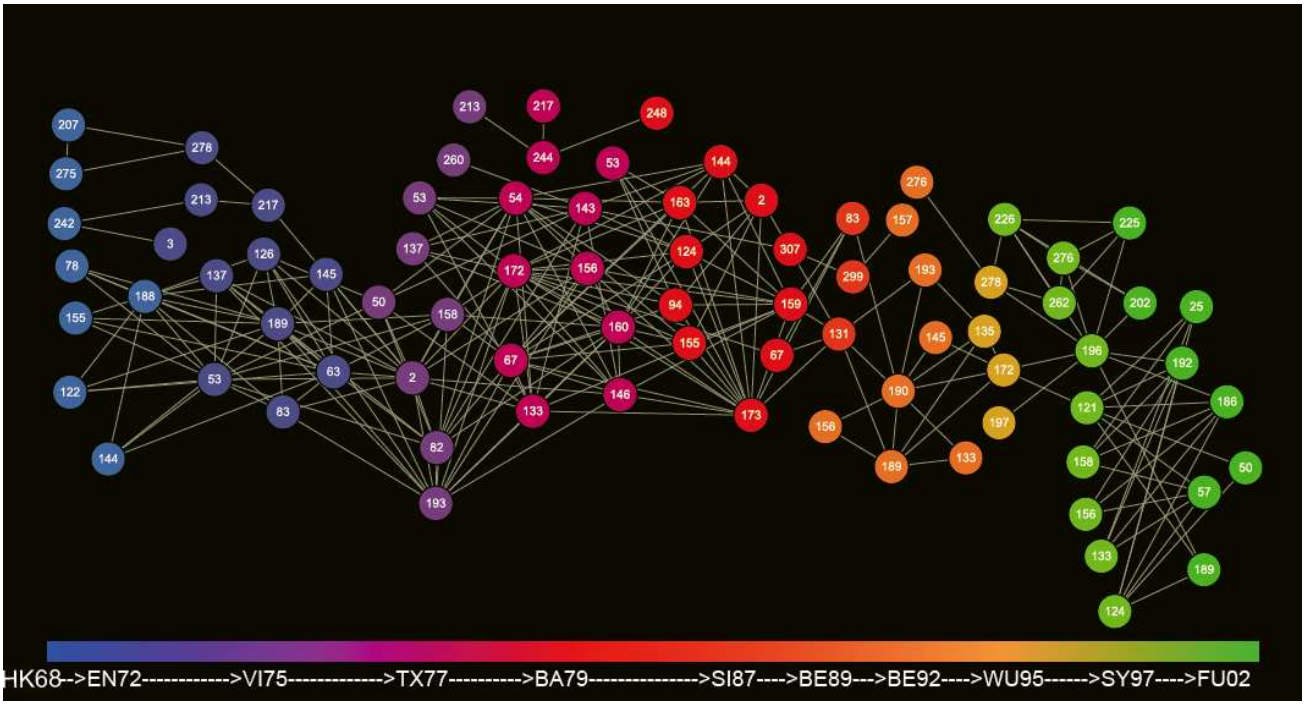

Fig. 1: The site transition network (STN) from year 1968 to 2008. A total of 63 positive selection sites out of 312 were found, which were then chosen to plot the network (other sites were omitted since they are mostly conserved with little interactions with other nodes and will be seen as isolated nodes in the network if plotted). Each node represents a site with its residue number marked on top, and each branch represents the interaction between a pair of mutation sites if its normalized MI score (normalized with a mean 0 and standard deviation of 1) in Step 4 is larger than 0.5 (see text for more details). Sites substituted in relatively close years were drawn in the proximate color. The nodes with same color mean that they might mutate simultaneously to induce an antigenic change.

Table 1: Amino acid sites responsible for co-occurring mutations. A total of 5 clusters were obtained from the cluster analysis, with each cluster representing roughly one or several antigenic transitions. The sites shown in red color in column 2 and 3 are the identical sites obtained form both the real historical data and our clustering analysis (and the ones in gray are missed one).

Antigenic Changes	Observed Mutant Sites in HA1 During the Antigenic Transitions	Sites in different Groups in Cluster Analysis
HK68-EN72	78 122 188 207 242 144 155 275	78 122 188 207 242 276
EN72-VI75	53 137 213 145 189 217 278	2 53 137 213 260 244 50
VI75-TX77	2 137 213 260 50 82 158 193	
TX77-BK79	54 133 143 156 160 172 197 217 53 146 162 244	54 133 143 156 160 172 197 217 121
BK79-SI87	124 155 189	124 135 145 157 189 190 196 226 262 276
SI87-BE89	145	
BE89-BE92	145 157 189 190 276 156	
BE92-WU95	135 145 226 262 172 197 278	
WU95-SY97	276 196 226 62 156 158	
SY97-FU02	25 57 75 83 131 142 144 155 186 222 225 227 50 156 159 189 202	25 57 75 83 131 142 144 155 186 222 225 227 63 82 94 126 192 202 299

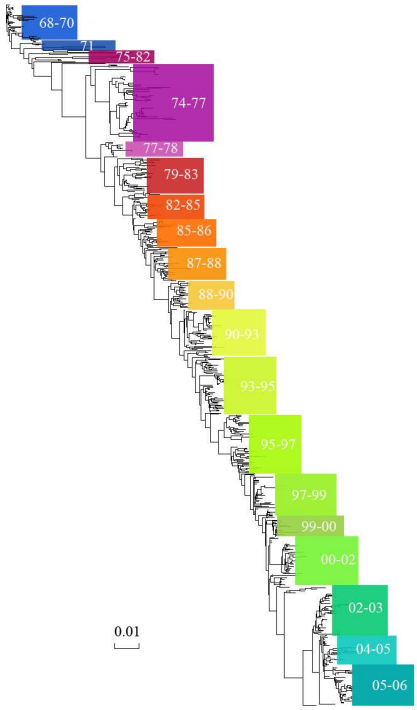

Fig. 2: Phylogenetic tree of the HA1 sequences. The size of the square is arbitrary, and is color-coded according to season and tree topology. Multiple trees were built using neighbor-joining (NJ) algorithm and with program MEGA version 4 (Tamura, Dudley, Nei, and Kumar 2007).

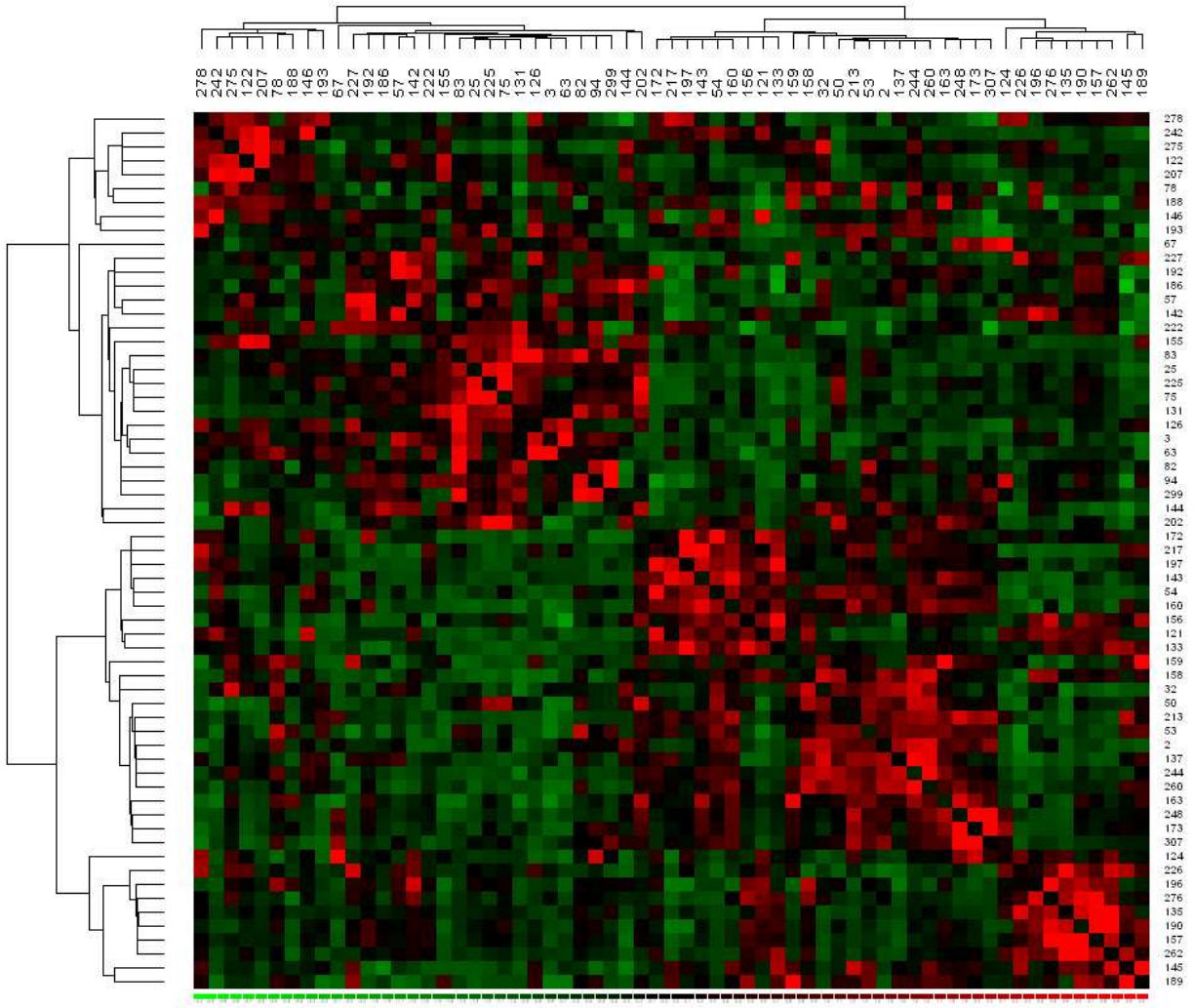

Fig. 3: A/H3N2 site clusters based on agglomerative hierarchical cluster analysis. The numbers in the axes represent amino acid sites of the HA1. The red color indicates the MI score above the mean and the green color below the mean. The red clusters represent the multiple co-occurring or co-evolving residue sites. This figure also shows that there are two different types of sites involved in antigenic transitions, characteristic sites and connection sites (see text for more details).

/clustering.htm). Here, each site mimics a "gene", and each row of the normalized MI scores with other sites represents the "gene expression data" (i.e., the feature vector). The idea behind this is that if any two sites *i* and *j* (in addition to their own MI score which is analyzed in above STN in Figure 1) have high MI scores with some other common site *k* or sites {*k1, k2, k3,…*}, they will show high correlations and thus be clustered together with those common sites. These clustered or grouped sites represent the highly-correlated co-occurring mutations in HA1. Thus, this provides another angle to examine the simultaneous multi-site mutations in antigenic drifts. The "distance" (similarity metric) between any two sites *i* and *j* is defined as (1-r), with r the Pearson correlation coefficient between the *i*th and *j*th feature vectors (array of MI scores). The "centroid-linkage" (distance between the centroids of two clusters) is used for cluster merging when combining rows and columns (Li and Wong, 2003). Figure 3 shows the final clusters for the HA1 sites. Each cluster represents a co-occurring or co-evolving substitution group, and such co-occurring mutations are believed to be the trigger that leads to one specific antigenic change. The antigenic maps (Smith, et al., 2004) also classify the strains of influenza A/H3N2 into several clusters to indicate the corresponding relationship among the strains between different time periods. Interestingly, our clusters also revealed similar antigenic drifts in the A/H3N2 evolutional history. For example, the mutations of sites 122, 207 and 188 in HA1 responsible for the HK68-EN72 antigenic drift were clustered together. Moreover, the sites 172, 217, 197, 143, 54, 160, 156, 133 were also clustered together that caused the TX77-BK79 drift. At the same time, we found one cluster (including site 124, 196, 190, 262, 145, 189, 276 colored in orange in Table 1) comprising 5 antigenic transitions from BK79 to SY97. For example, site 145 mutated 4 times during BK79 to SY97 (BK79-SI87, BE89-BE92, BE92-WU95 and WU95-SY97). Site 124 mutated 3 times and site 189, 262, 276 mutated twice. The high frequencies of mutations in these sites show strong connections (interactions) among related antigenic groups. More results are shown in Table 1. These cluster analyses suggest that, even without the expensive serum HI assays for corresponding strains, the sequence data alone can also uncover most of the antigenic drifts and evolutionary

characteristics of influenza A/H3N2. These findings imply that a cluster of co-evolving sites on HA1 might have caused the protein surface to have large enough structural changes at antigenic sites that can induce the antigenic drift to a new strain and thus the escape of immunity.

Furthermore, these sites can be classified into two types through the cluster analysis: one type for sites that mostly connect with other sites in the same cluster (the "inner cluster sites", or "characteristic sites"), and the other type for sites that have less connections with the inner cluster but more connections with neighboring clusters (named as "connection sites"). One natural question becomes "Which type of sites, if exists, is more important for triggering an antigenic change?" Since the "connection sites" have more interactions with other clusters, it seems logical that such sites might first induce the influenza escaping of the immunity, and then the "characteristic sites" help settle down the escape activity in a new antigenic strain. However, after careful examination of the entire network of HA sites, we found that the problem is much more complicated, and it is hard to draw an unambiguous conclusion to the above hypothesis. There are evidences showing that both the "characteristic sites" and "connection sites" can maintain a very low mutation frequency before they become fixed in one antigenic group. Sometimes, the "connection sites" and "characteristic sites" played a similar role in an antigenic transition event. For example, in the antigenic change of BK79 to SI87, the mutations of "connection sites" T155H and Q189R first appeared in year 1968 and kept in a low frequency and never fixed until the "characteristic site" mutation G124D happened around 1984. However, in the antigenic change of BE89 to BE92, "characteristic site" mutations (S133D, E156K and E190D) first appeared in 1968. In addition, another group of "characteristic site" mutations (K145N and T262N) appeared in 1975. All of these 5 "characteristic sites" (133, 156, 190, 145, 262) remained in low mutation frequency during the period before "connection sites" mutated (S157L, R189S and T272N) in 1989. After that, a new antigenic change of BE89 to BE92 happened. These studies indicate that both the mutations of connection sites and characteristic sites are determining factors for influenza A to transfer form one antigenic cluster to another, which help A/H3N2 to consistently escape immunity and vaccine.

3.3 Strong selection pressures on certain sites

Similarly, a strong preference in mutation sites can be found by calculating the level of conservation (the mutation frequency) for each site of HA1. We used program Consurf (Glaser, et al., 2003; Landau, et al., 2005) to map the residue conservation levels on HA 3D X-ray crystal structure (PDB ID: 2HMG) (Weis, et al., 1990). The conservation scores calculated from the 4,064 HA1 sequences are mapped onto the HA1 structure, as shown in Figure 4, with most variable residues colored in cyan and most conserved residues colored in magenta. These results indicate that the HA1 mutations have a strong preference for the positive selection sites. As shown in the figure, sites with high mutation rates are concentrated in RBD and five known antigenic regions (A, B, C, D and E) on the HA1 protein surface. On the contrary, the inner protein residues are mostly well conserved. This finding indicates that HA is holding its basic skeleton as an "anchor" to maintain its overall structure, while generating rapid mutations in RBD and antigenic regions (both characteristic and connection sites) to drive the influenza to adapt new hosts and vaccines. These findings are consistent with the X-ray crystal structures of hemagglutinin, where RBD and epitope regions show larger B-factors while most of the inner residues show very rigid structures (Gamblin, et al., 2004; Ha, et al., 2003; Stevens, et al., 2006).

Further analysis showed that HA1 also has preferences for particular amino acid types in addition to the positions. We found that a particular site almost always wants to mutate to some specific amino acid types. A detailed analysis of all 312 sites' amino acid (AA) types revealed that the average number of amino acid types is only ~2.36 AA/site (ranging from 1 to 8 AA/site). Therefore, the average probability for a site to mutate to a new amino acid that did not appear in the history is quite low. For instance, in year 2005 and 2006, the probability is only ~0.248% and

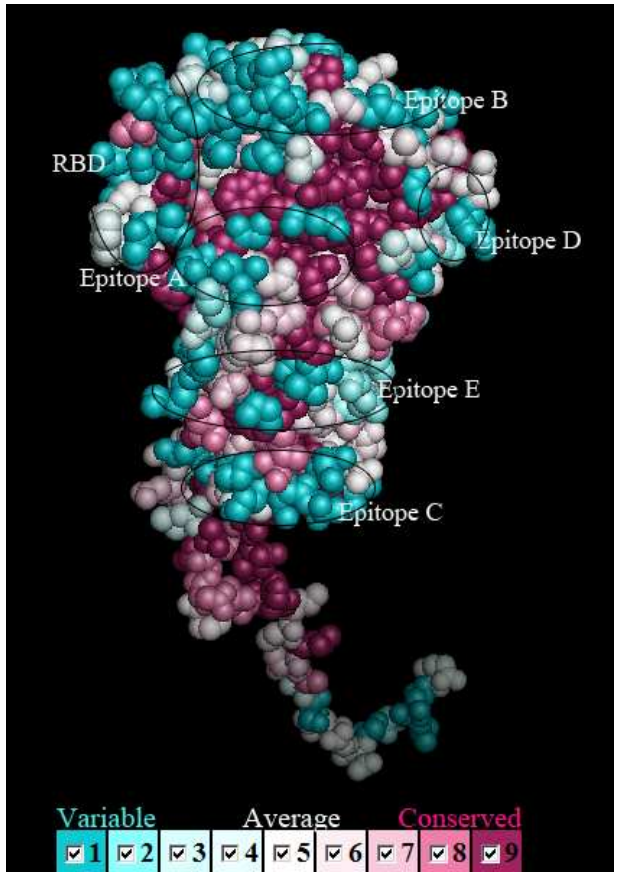

Fig. 4: Identification of the conserved and variable regions on HA1 protein surface, by mapping the conservation score calculated from the 4,064 A/H3N2 HA1 sequences onto the HA1 structure (Glaser, et al., 2003; Landau, et al., 2005). The level of amino acid site's conservation is represented with score 1 to 9, and colored from cyan (most variable) to magenta (most conserved). This figure shows that most of the mutations happen in the RBD and antigenic regions on the protein surface, while the main "skeleton" structure underneath the protein surface is largely unaltered. These results were calculated with The ConSurf Web Server (http://consurf.tau.ac.il/index.html) and rendered by PyMol (DeLano, 2008).

~0.08%, respectively. This result suggests that the mutation in HA1 has preference in the selection of amino acid types – sites prefer to choose the amino acids that have already appeared in the history (i.e., those can survive the natural selection). In addition, the selection trend for each specific site from 1968 to 2008 was calculated. We subdivided the 20 amino acids into 3 subclasses. Class 1 are non-polar residues (including A, F, G, I, L, M, P, V, W), Class 2 are polar residues (including C, N, Q, S, T, Y, H), and Class 3 are charged residues (D, E, K, R). The trend results indicate that in general HA1 sites have a mild tendency in substituting within the same amino acid subclass, which means the mutations have a greater chance to stay within the same class of residues to keep their structural and functional characteristics at that specific position (Weis, et al., 1988). Of course, there are exceptions -- some mutations occurred from one residue subclass to another, like sites 133, 135, 137, 155, 225 and 227 in the RBD region (at least substituted once from one residue subclass to another). It is interesting to note that, in some extreme cases, such cross-class mutations can cause critical conformational changes that can induce the switch of influenza virus binding specificity. For example, in the recent A/H5N1 strains, Wilson's group has shown that the Q226L & G228S double mutation in H5N1 (A/Vietnam/1203/2004) could cause the virus stride over the barrier between avian and human (Stevens, et al., 2006). In other words, these two cross-class mutations might change the binding specificity from avian (α2,3-linkage sialic acids) to human (α2,6-linkage sialic acids). Similarly, Zhou and coworkers predicted another double mutation, V135S & A138S (also cross-class) to be candidates for such a binding specificity switch from avian to human (Das, et al., 2009). Furthermore, we found that more than 99.2% sequences of A/H3N2 selected residue Ser in position 228 (while 99.9% sequences of A/H5N1 chose residue Gly in position 228 up to now). This result indicates that site 228 has very strong preference to choose Ser as its fixed site to preserve the binding specificity of A/H3N2. The historical strong preference for mutation site positions as well as amino acid types makes it possible for us to predict the HA1 mutations.

3.4 Network guided prediction of the A/H3N2 evolution

The Site Transition Network, as shown in Figure 1, shows that the antigenic drifts can be enhanced by cumulative multi-site co-occurring substitutions at the epitope regions of HA protein surface. The co-evolutional mutations give us an opportunity to use present network to predict the future mutations, which might induce the next antigenic change to a new strain. As shown above, when the new mutations are generated, their site positions and amino acid types are not random but have several strong preferences. These preferences are deducted from the statistical analysis of the historical sequence data from 1968 to 2008. Obviously, the more abundant the sequence data (and longer history) is, the higher the prediction accuracy is. In this section, we first validate the prediction method by using previous years' data, such as the year 2003-2004, as an example, and then predict the possible mutations in the upcoming season of 2009-2010.

3.4.1 Validation with the known historical mutation data

A validation test was done to evaluate our 5-step predictive method based on the STN. Since extensive studies have been done with the FU02 antigenic change, we chose the 2003-2004 season as an example to illustrate our method using sequence data from 1968 to 2002. Following the steps detailed in Section 2.3, we first identified 63 positive selection sites in HA1 from 1968 to 2002. All of these positive selection sites are on the protein surface of HA1, with most of them from antigenic epitopes (57 of 63). Then, we determined all the new mutation sites that appeared in Induction Years 2001~2002. We found 18 positive selection sites mutated in these two years 2001~2002 (Step 2). These sites were residues 57, 62, 121, 124, 133, 137, 142, 156, 158, 172, 190, 192, 193, 196, 197, 226, 262 and 276. Then a new MI matrix was constructed using all sequence data from 1968 to 2002 by MI method in Step 3. For each site in HA1 (Step 4), a sum score is obtained by adding the MI scores of this site with all the other 18 sites identified in Step 2. We found 14 sites (2, 50, 54, 75, 78, 94, 131, 135, 155, 156, 157, 202, 213 and 222) with the MI score above the threshold, i.e., at least one interaction link in the network. These sites were thus chosen as the candidates for possible mutation sites in year 2003-2004.

Table 2A: The prediction results of the year 2003-2004 season for the validation of the method based on the site transition network. Compared to the real sequence data from 2003-2004, 7 out of 11 sites were successfully predicted.

Sites mutation in 2003-2004	R50G	H75Q	E83K	A131T	V144N	H155T	Q156H	S186G	V202I	W222R	G225D
Sites predicted successfully	√	√		√		√	√		√	√	

Table 2B: The statistical results for the site mutation predictions from 1999 to 2008. The accuracy rate in line 2 is the number of successful predicted sites vs. the number of truly mutated sites. The mark "N/A" in year 06-07 is due the zero sites fixed in that year.

Year	99-00	00-01	01-02	02-03	03-04	04-05	05-06	06-07	07-08	Average
Accuracy rate:	4/7	3/7	7/9	8/9	5/6	7/10	3/4	N/A	2/3	39/55
Accuracy Percentage	57.1%	42.9%	77.8%	88.9%	83.3%	70.0%	75.0%	N/A	66.7%	70.9%± 13.8%

Finally, in Step 5, we assigned each predicted site the most probable amino acid type according to their historical frequency. In reality, a total of 10 mutations occurred in 2003-2004, and 7 of these 10 mutations (mutation R50G, H75Q, A131T, H155T, Q156H, V202I and W222R) are included in our prediction. This means that our STN based method has a ~70% agreement rate in the prediction (sensitivity=0.70, positive predictive value PPV=0.50). Detailed results are summarized in Table 2A.

For further validation, we also performed predictions for the site mutations in every year from 1999 to 2008 (see Table 2B). We found the accuracy of predicted results was fairly stable, around 70%, which means that the network guided method is a fairly reliable tool to predict the antigenic drifts due to rapid mutations. In some cases, a higher accuracy was obtained, for example, ~89% for the 2002~2003 season. Table 2B summarizes the statistical results (sensitivity=0.71 $\pm$ 0.14, PPV=0.65$\pm$0.12).

It should be noted that in the above approach, the sites in the Induction Years are not repeated/included in the next season's predictions, only those new sites generated from STN are included (i.e., those with high mutual information scores with the positive selection sites in Induction Years). The underlying assumption for this approach is the following: All of the positive selection sites did not mutate twice within the two year time window. In fact, new strains from antigenic drifts almost never happened in less than two years period so far. If one site mutates twice within two years, it means that the site can not fix ("site fixation") during that time period. None of the 63 positive selection sites have mutated twice within the two year time window in the past.

3.4.2 Identifying possible mutations which might become dominant in year 2009-2010

Following same procedure for the above validation predictions, we used all the sequence data available up to 2008 to predict the mutations in year 2009-2010. Similarly, in the "Induction Years," four site mutations (G50E, K140I, S193F and D225N) were identified during 2007-2008 season. Our method then predicted 7 possible site mutations, G5E, N6S, I58V, A106T, T128A, R142k, and V166M, for the next antigenic drift, which may appear at the earliest in year 2009-2010. These predictions, of course, need to be validated by actual antigenic drifts in the coming seasons.

4 CONCLUSION

In this paper, more than 4,000 influenza A/H3N2 sequences from 1968 to 2008 were analyzed to model the influenza A virus evolutional behavior. The mutual information (MI) method was used to draw correlation between co-occurring mutations and then design a site transition network (STN) for virus evolutionary path. The STN network indicates that the dynamic interactions between different sites are mainly near the epitopes and the RBD regions of HA1 protein. The network also shows that antigenic changes accumulate over time, with occasional large changes due to multiple co-occurring mutations at antigenic sites. Furthermore, the cluster analysis by subdividing the entire MI Matrix or network into several clusters reveals a more detailed view about the correspondence between simultaneous multi-site mutations and characteristics of the antigenic drift. These results from our STN network approach agree well with those from the phylogenetic tree (Fig. 2) and "antigenic map" (Smith, et al., 2004) analyses. The current approach provides a novel bridge in connecting the sequence data with hemagglutination inhibition (HI) assay data, and suggests that the A/H3N2 virus is constantly under significant selection pressure, and mutates rapidly but relatively smoothly at sequence and structure levels. Meanwhile, the occasional co-occurring multi-site mutations cumulatively enhance the antigenic drifts in generating new strains.

The strong preferences in mutation sites, as well as in the types of final amino acids after mutation, provide us a promising way to predict the future site mutations. A novel 5-step STN network based method was then designed to predict the future mutations using mutual information between different sites. The validation tests show that the method can reproduce the known mutations with previous years' data at about 70% accuracy. We then predicted 7 possible mutations for the next antigenic drift, which may appear at the earliest in the coming 2009-2010 season.

Finally, it should be noted that multiple factors can influence the influenza evolution in reality, such as location, seasons, new vaccine pressure, other selection pressure, and so on. The influenza epidemics outbreak in different locations around the world, and such discontinuities in geography weaken the time connection between two successive antigenic transitions, in which our method is based on. Recently, large collections of New York State and New Zealand data were used to study the genomic and epidemiological dynamics of human influenza A virus, and a sink-source model was suggested to describe the influenza viral ecology (Rambaut, et al., 2008). Such work indicates that the location diversity should be taken into account in future prediction methods, once more geographic based data becomes available.

ACKNOWLEDGEMENTS

We would like to thank Isidore Rigoutsos, Alice McHardy, Payel Das, Laxmi Parida and Ajay Royyuru for many useful discussions. We also thank Ian Wilson, Jim Paulson, and Peter Palese for helpful comments in the beginning of this project.

REFERENCES

Bao Y., P.B., D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman. (2008) The influenza virus resource at the National Center for Biotechnology Information, *J. Virol.*, **82(2)**, 596-601.

Bush, R.M., Bender, C.A., Subbarao, K., Cox, N.J. and Fitch, W.M. (1999) Predicting the evolution of human influenza A, *Science*, **286**, 1921-1925.

Bush, R.M., Fitch, W.M., Bender, C.A. and Cox, N.J. (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A, *Molecular biology and evolution*, **16**, 1457-1465.

Butte, A.J. and Kohane, I.S. (2000) Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements, *Pacific Symposium on Biocomputing*, 418-429.

Butte, A.J., Tamayo, P., Slonim, D., Golub, T.R. and Kohane, I.S. (2000) Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks, *Proc Natl Acad Sci U S A*, **97**, 12182-12186.

Chen, R. and Holmes, E.C. (2006) Avian influenza virus exhibits rapid evolutionary dynamics, *Molecular biology and evolution*, **23**, 2336-2341.

Cox, N.J. and Subbarao, K. (2000) Global epidemiology of influenza: past and present, *Annu Rev Med*, **51**, 407-421.

Das, P., Li, J., Royyuru, A.K. and Zhou, R. (2009) Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity, *J Comput Chem*.

DeLano, W.L. (2008) The PyMOL Molecular Graphics System, **DeLano Scientific**, San Carlos, CA, USA.

Du, X., Wang, Z., Wu, A., Song, L., Cao, Y., Hang, H. and Jiang, T. (2008) Networks of genomic co-occurrence capture characteristics of human influenza A (H3N2) evolution, *Genome research*, **18**, 178-187.

Edgar, R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity, *BMC bioinformatics*, **5**, 113.

Faith, J.J., Hayete, B., Thaden, J.T., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J.J. and Gardner, T.S. (2007) Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles, *PLoS biology*, **5**, e8.

Fitch, W.M., Bush, R.M., Bender, C.A. and Cox, N.J. (1997) Long term trends in the evolution of H(3) HA1 human influenza type A, *Proceedings of the National Academy of Sciences of the United States of America*, **94**, 7712-7718.

Fouchier, R.A., Munster, V., Wallensten, A., Bestebroer, T.M., Herfst, S., Smith, D., Rimmelzwaan, G.F., Olsen, B. and Osterhaus, A.D. (2005) Characterization of a novel influenza A virus hemagglutinin subtype (H16) obtained from black-headed gulls, *J Virol*, **79**, 2814-2822.

Gamblin, S.J., Haire, L.F., Russell, R.J., Stevens, D.J., Xiao, B., Ha, Y., Vasisht, N., Steinhauer, D.A., Daniels, R.S., Elliot, A., Wiley, D.C. and Skehel, J.J. (2004) The structure and receptor binding properties of the 1918 influenza hemagglutinin, *Science*, **303**, 1838-1842.

Glaser, F., Pupko, T., Paz, I., Bell, R.E., Bechor-Shental, D., Martz, E. and Ben-Tal, N. (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information, *Bioinformatics (Oxford, England)*, **19**, 163-164.

Ha, Y., Stevens, D.J., Skehel, J.J. and Wiley, D.C. (2003) X-ray structure of the hemagglutinin of a potential H3 avian progenitor of the 1968 Hong Kong pandemic influenza virus, *Virology*, **309**, 209-218.

Hilleman, M.R. (2002) Realities and enigmas of human viral influenza: pathogenesis, epidemiology and control, *Vaccine*, **20**, 3068-3087.

Holmes, E.C., Ghedin, E., Miller, N., Taylor, J., Bao, Y., St George, K., Grenfell, B.T., Salzberg, S.L., Fraser, C.M., Lipman, D.J. and Taubenberger, J.K. (2005) Whole-genome analysis of human influenza A virus reveals multiple persistent lineages and reassortment among recent H3N2 viruses, *PLoS Biol*, **3**, e300.

Horimoto, T. and Kawaoka, Y. (2005) Influenza: lessons from past pandemics, warnings from current incidents, *Nat Rev Microbiol*, **3**, 591-600.

Huelsenbeck, J.P. and Dyer, K.A. (2004) Bayesian estimation of positively selected sites, *J Mol Evol*, **58**, 661-672.

Koelle, K., Cobey, S., Grenfell, B. and Pascual, M. (2006) Epochal evolution shapes the phylodynamics of interpandemic influenza A (H3N2) in humans, *Science*, **314**, 1898-1903.

Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. and Ben-Tal, N. (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures, *Nucleic acids research*, **33**, W299-302.

Li, C. and Wong, W.H. (2003) *DNA-Chip Analyzer (dChip).* Springer, New York.

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R. and Califano, A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC bioinformatics*, **7 Suppl 1**, S7.

Nelson, M.I., Simonsen, L., Viboud, C., Miller, M.A., Taylor, J., George, K.S., Griesemer, S.B., Ghedin, E., Sengamalay, N.A., Spiro, D.J., Volkov, I., Grenfell, B.T., Lipman, D.J., Taubenberger, J.K. and Holmes, E.C. (2006) Stochastic processes are key determinants of short-term evolution in influenza a virus, *PLoS Pathog*, **2**, e125.

Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, **148**, 929-936.

Plotkin, J.B. and Dushoff, J. (2003) Codon bias and frequency-dependent selection on the hemagglutinin epitopes of influenza A virus, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 7152-7157.

Rambaut, A., Pybus, O.G., Nelson, M.I., Viboud, C., Taubenberger, J.K. and Holmes, E.C. (2008) The genomic and epidemiological dynamics of human influenza A virus, *Nature*, **453**, 615-U612.

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

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

Stevens, J., Blixt, O., Tumpey, T.M., Taubenberger, J.K., Paulson, J.C. and Wilson, I.A. (2006) Structure and receptor specificity of the hemagglutinin from an H5N1 influenza virus, *Science*, **312**, 404-410.

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

Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites, *Molecular biology and evolution*, **16**, 1315-1328.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0, *Molecular biology and evolution*, **24**, 1596-1599.

van Nimwegen, E. (2006) Epidemiology. Influenza escapes immunity along neutral networks, *Science*, **314**, 1884-1886.

Webster, R.G., Bean, W.J., Gorman, O.T., Chambers, T.M. and Kawaoka, Y. (1992) Evolution and ecology of influenza A viruses, *Microbiol Rev*, **56**, 152-179.

Webster, R.G., Laver, W.G., Air, G.M. and Schild, G.C. (1982) Molecular mechanisms of variation in influenza viruses, *Nature*, **296**, 115-121.

Weis, W., Brown, J.H., Cusack, S., Paulson, J.C., Skehel, J.J. and Wiley, D.C. (1988) Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid, *Nature*, **333**, 426-431.

Weis, W.I., Brunger, A.T., Skehel, J.J. and Wiley, D.C. (1990) Refinement of the influenza virus hemagglutinin by simulated annealing, *Journal of molecular biology*, **212**, 737-761.

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

Yang, Z. and Swanson, W.J. (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes, *Mol Biol Evol*, **19**, 49-57.

Zhou, R., Das, P. and Royyuru, A.K. (2008) Single Mutation Induced H3N2 Hemagglutinin Antibody Neutralization: A Free Energy Perturbation Study, *J. Phys. Chem. B*, **112**, 15813–15820.