Pacific Symposium on Biocomputing 14:492-503 (2009)

COMBINING MUTUAL INFORMATION WITH STRUCTURAL ANALYSIS TO SCREEN FOR FUNCTIONALLY IMPORTANT RESIDUES IN INFLUENZA HEMAGGLUTININ

PETER M. KASSON

Departments of Structural Biology and Chemistry, Stanford University, Stanford, CA 94305 USA

VIJAY S. PANDE

Department of Chemistry, Stanford University, Stanford, CA 94305 USA

Influenza hemagglutinin mediates both cell-surface binding and cell entry by the virus. Mutations to hemagglutinin are thus critical in determining host species specificity and viral infectivity. Previous approaches have primarily considered point mutations and sequence conservation; here we develop a complementary approach using mutual information to examine concerted mutations. For hemagglutinin, several overlapping selective pressures can cause such concerted mutations, including the host immune response, ligand recognition and host specificity, and functional requirements for pH-induced activation and membrane fusion. Using sequence mutual information as a metric, we extracted clusters of concerted mutation sites and analyzed them in the context of crystallographic data. Comparison of influenza isolates from two subtypes—human H3N2 strains and human and avian H5N1 strains—yielded substantial differences in spatial localization of the clustered residues. We hypothesize that the clusters on the globular head of H3N2 hemagglutinin may relate to antibody recognition (as many protective antibodies are known to bind in that region), while the clusters in common to H3N2 and H5N1 hemagglutinin may indicate shared functional roles. We propose that these shared sites may be particularly fruitful for mutagenesis studies in understanding the infectivity of this common human pathogen. The combination of sequence mutual information and structural analysis thus helps generate novel functional hypotheses that would not be apparent via either method alone.

1. Introduction

Influenza virus is a major cause of both seasonal epidemic respiratory disease and periodic high-mortality pandemics. The most significant of these latter events within recent history, the pandemic of 1918-1919, caused approximately 50 million deaths worldwide.[1] Influenza viruses circulate extensively in birds as well as humans and other mammals, and the three major pandemics of the 20th century (1918, 1957, 1968) were all likely due to epizootic transfer from

viruses infecting other species into the human population. This has likely occurred both via adaptation of avian viruses to human hosts and via genetic reassortment between avian or mammalian and human-specific viruses.

More recently, the spread of a highly pathogenic avian influenza virus (HPAI H5N1) and a number of epizootic infections of humans (with a case-fatality rate of approximately 60% [2]) has raised concern of another imminent pandemic. Fortunately, the H5N1 virus has thus far not displayed efficient human-to-human transmission. It has been postulated that the poor human-to-human transmissibility of H5N1 may be due to inefficient viral replication in the upper respiratory epithelium of humans. Since the viral hemagglutinin protein is the primary determinant of both cell entry and antibody-mediated immunity, mutations to the hemagglutinin molecule that increase the efficiency by which human respiratory epithelial cells are infected would be an important permissive factor in human-to-human spread of either an adapted avian H5N1 virus or an avian-human reassortant.

We would therefore like to understand the functional control of influenza hemagglutinin and the means by which the molecule might mutate to alter host range or to evade new therapeutic agents. Informatics-based methods allow computationally efficient screening of the large number of potential hemagglutinin mutations. Such efficiency is required because hemagglutinin is over 500 amino acids in length, yielding a mutation space of $\sim 20^{500}$, and mutations both near and distant from the ligand binding site have been shown experimentally to alter ligand selectivity. We therefore propose a stepwise approach in which informatics-based methods are used to generate an initial set of predictions that can be further refined by a combination of physics-based computational methods and targeted experimental mutagenesis.

Influenza functional regulation differs fundamentally from the canonical systems for which many function-prediction methods were developed. Computational methods that have been used in other systems to predict ligand-binding specificity include shape-based analysis of the ligand-binding pocket [3], analysis of conserved residues [4], evolutionary trace methods [4], and methods that combine phylogenetic and information-theoretic characterizations [5]. For control of hemagglutinin function, particularly ligand specificity switches, experimental characterization of isolates displaying partial specificity switches identified both single point mutations and concerted mutations among several residues[6]. The evolutionary pressure of the host immune response and the frequent recombination events undergone by influenza may also complicate the mutational assumptions of phylogenetic methods, and the lack of crystal structures of a hypothetical human-adapted H5N1 hemagglutinin

challenge shape-based methods. While all of these methods may be helpful in studying influenza function, there clearly exists the opportunity for novel methodology to yield additional insight.

Computational prediction methods are particularly helpful for influenza because systematic experimental screening for functionally important mutants is challenging. Hemagglutinin is heavily glycosylated, and the glycan residues affect ligand binding[7-9]. Furthermore, glycosylation patterns vary according to the cell culture system used to express the hemagglutinin protein. Because of these challenges and biosafety issues associated with live H5N1 virus, experimental studies to date have focused on retrospective analysis of mutations observed in clinical isolates rather than prospective screening[6, 10].

Given the complexity of influenza evolution and the absence of phylogenetically distinct lineages of human H5N1 influenza, a statistical approach that does not utilize a model of evolutionary correlation provides a reasonable alternative. Close co-variation between a pair of residues is suggestive of selective pressure and likely functional linkage. This is distinct from conserved residues, which may be required for conformational stability, and hypervariable residues, which may mutate to achieve immune evasion with no functional consequence. In many classic examples of small-molecule recognition by proteins, key residues involved in ligand recognition can be mutated but only in concert. The classic example of this is the catalytic triad in serine proteases; a point mutation to a key residue can destroy activity, but activity can be rescued by a complementary mutation to another residue. In this spirit, co-evolving residues have previously been used as a means of identifying functional significance in other systems [11], including HIV protease drug resistance [12, 13]. We employ a slightly different approach here; instead of introducing a distance matrix for protein mutation and computing co-variance, we compute pair-wise mutual information between residues in a discrete fashion similar to [14]. We then employ crystallographic data to evaluate the sequence-based predictions.

We have used sequence mutual information to analyze closely-linked mutation sites in hemagglutinin from two influenza subtypes: H3N2 isolates that have circulated in humans since 1968 and H5N1 isolates that circulate primarily in birds and have been responsible for millions of poultry deaths and ~400 human deaths since 1997 [2]. To interpret the mutual information data, we employ a combination of hierarchical clustering and visualization using available crystallographic data. Comparison of the two influenza subtypes in this regard is particularly informative: major differences in the residues identified may reflect the different selective pressures on each, while the

common set of residues identified may reflect common functional roles. Taken together, these results suggest hypotheses regarding functional regulation of hemagglutinin that are testable via future mutagenesis experiments.

2. Methods

2.1. *Sequences and structures*

Sequences were obtained from the NCBI Influenza Virus Resource [15]; all 2103 full-length human H3N2 hemagglutinin sequences and all 1516 full-length human or avian H5N1 hemagglutinin sequences in the database as of July 2008 were used. Multiple-sequence alignment was performed for each influenza subtype separately and for pooled sequences from both subtypes using MUSCLE.[16] Crystal structures of A/Aichi/2/1968 (H3N2) hemagglutinin (1HGG) [17] and A/Viet Nam/1194/2004 (H5N1; 2IBX)[6] were used for visualization and distance calculations. Since the human H5N1 structure does not include a ligand, ligand coordinates from a related avian H5 hemagglutinin[18] were used for distance calculations after structural alignment of the two hemagglunins.

2.2. *Identification of closely-linked residues*

Pairwise mutual information was calculated in a discrete fashion between each residue position in the H3N2 and H5N1 multiple sequence alignments respectively as follows:

$$I(i,j) = H(i) + H(j) - H(i,j) \tag{1}$$

where

$$H(i) = \sum_{a \in A} -p(x_i = a)\log p(x_i = a) \tag{2}$$

and

$$H(i,j) = \sum_{a \in A} \sum_{b \in A} -p(x_i = a, x_j = b)\log p(x_i = a, x_j = b), \tag{3}$$

for sequence positions *i* and *j*, where the variable x_i represents the values of the multiple sequence alignment at position *i*. Mutual information captures nonlinear relationships more efficiently than covariance-based methods; use of a substitution matrix to give varying mismatch probabilities according to the chemical similarity of the amino acids involved may add further sensitivity.

Symmetric uncertainty[19] was used to normalize the pairwise mutual information matrix as follows: $U(i,j)=2*I(i,j) / (I(i,i) + I(j,j))$ for all positions i and j. Single-linkage hierarchical clustering was performed in MATLAB using U as the distance metric and guaranteeing $U(i,i) =1$, all i. The 99.9^{th} percentile of all non-self symmetric uncertainty values was calculated, and the corresponding distance metric was used as a threshold for cluster identification.

2.3. *Cluster visualization*

Each residue in a cluster was mapped to the corresponding residue in the crystal structure for H3N2 or H5N1 hemagglutinin as appropriate. These residues were then visualized using Pymol (DeLano Scientific, Palo Alto CA).

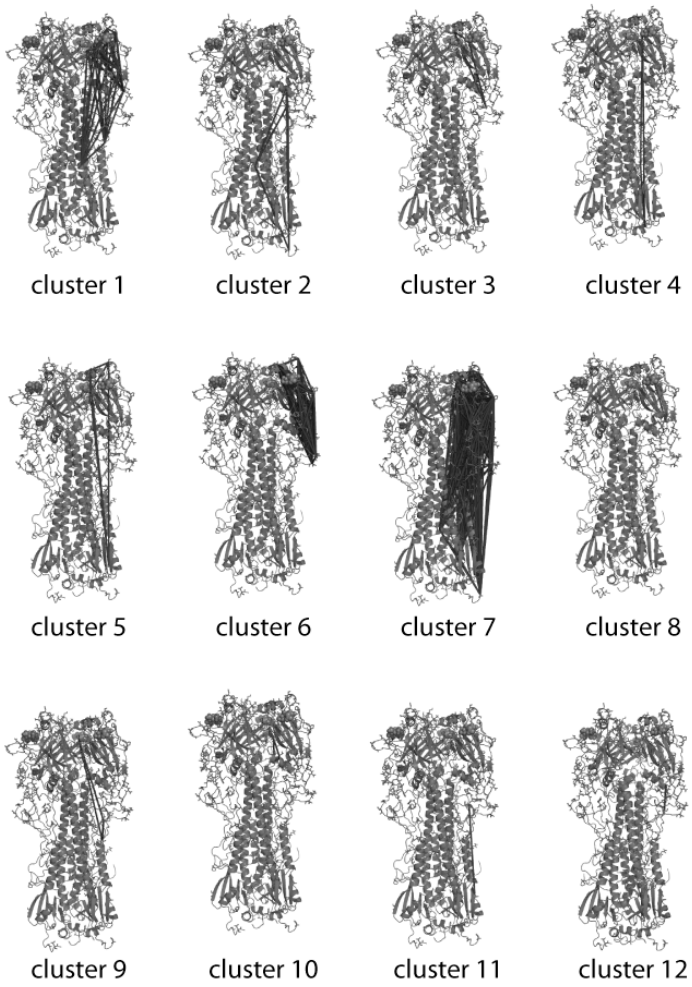

Figure 1. Clusters of closely-linked residues for H3N2 hemagglutinin.
Clusters identified by hierarchical clustering on sequence mutual information are visualized using crystallographic data. Red lines connect α-carbons within each cluster; sialic acid residues of the ligand are visualized via orange spheres.

3. Results

3.1. *Closely-linked residues in H3N2 hemagglutinin*

Closely-linked residues were identified based on hierarchical clustering on normalized sequence mutual information; clusters were selected using a

threshold corresponding to the 99.9th percentile of non-self mutual information; for H3N2 hemagglutinin 75 residues in the mature hemagglutinin molecule were identified within 12 clusters. These clusters are visualized on the crystal structure of H3N2 A/Aichi/68 hemagglutinin in Fig. 1. Clusters fall into three general categories: few residues, short-range interactions (clusters 3, 8, 10, 12); few residues, long-range interactions (clusters 2, 4, 5, 9, 11); many residues primarily localized to the globular head (clusters 1, 6, 7). This last category is intriguing because both the ligand-binding sites and the major antibody-recognition epitopes are located here. In isolation, these clusters are interesting but less informative; we have compared the clustering patterns between two influenza subtypes in order to correlate clustering difference with biological differences between the subtypes.

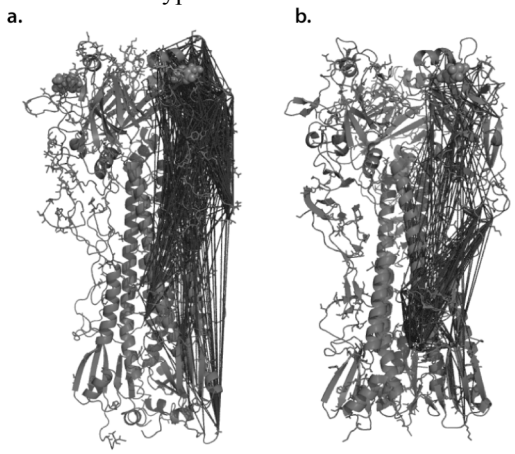

Figure 2. Comparison of closely-linked residues between influenza subtypes.
H3N2 hemagglutinin is rendered in panel (a), H5N1 in panel (b). Clusters are denoted by red lines connecting C-alpha atoms in the A/B monomer. Clusters are obtained via hierarchical clustering on a normalized pairwise mutual information matrix (symmetric uncertainty), with a distance cutoff corresponding to the 99.5th percentile of all non-self interactions in the matrix. The sialic acid of the ligand is shown as orange spheres.

3.2. *Differences between hemagglutinin subtypes*

We have analyzed pooled human and avian H5N1 isolates in the same fashion as the human H3N2 isolates to enable a more informative analysis. Fig. 2 shows all identified clusters for H3N2 and H5N1 hemagglutinin superimposed onto their respective crystal structures. One immediately apparent difference between the two subtypes is the frequency of linked residues on the globular head. This frequency is much greater in human H3N2 isolates than in the

predominantly avian H5N1 isolates. We have quantified this relationship by measuring the distribution of radial distances from the sialic-acid-binding pocket to the residues identified as closely-linked for each subtype and comparing that to the distances for all residues of that subtype. This combined structural and sequence-informatic analysis clearly shows an enrichment of residues in the globular head for H3N2 compared to H5N1. While a large number of H3N2 HA residues have been implicated in antibody recognition, crystal structures of neutralizing antibodies show binding primarily to the globular head[20, 21]. It is also believed that the human immune response plays a major role in driving the evolution of the H3N2 virus [22, 23]. Since clusters of closely linked residues involving the globular head are identified in the human H3N2 isolates but not in the primarily avian H5N1 isolates (where the human immune response is not a major factor), we hypothesize that these clusters may be largely involved in immune escape.

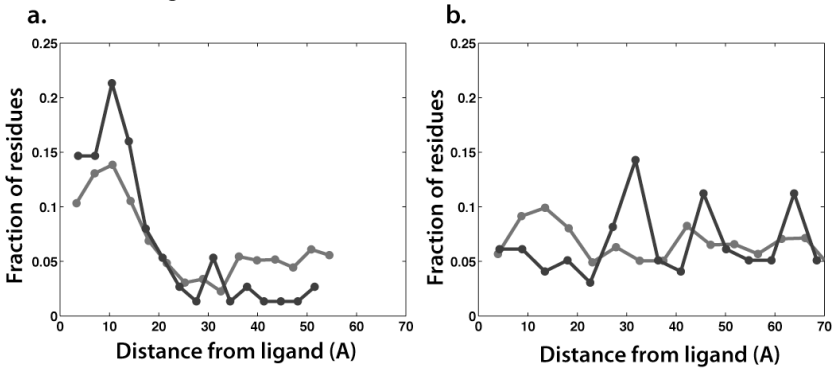

Figure 3. Distance from ligand for closely-linked residues in H3N2 and H5N1 hemagglutinin. Plotted are radial distribution functions for H3N2 hemagglutnin (**a**) and H5N1 hemagglutinin (**b**); closely-linked residues are plotted in red, while all protein residues are plotted in blue.

3.3. *Commonalities between hemagglutinin subtypes*

While human H3N2 isolates differ from H5N1 isolates in the role of the human immune response in driving viral evolution, they share many basic functional characteristics: binding of sialic-acid-terminated glycans, pH-induced conformational activation, and catalysis of membrane fusion. We therefore hypothesize that closely linked residues identified via independent analysis of each of these subtypes may be involved in some of the common functional roles. We have identified 12 residues meeting our detection threshold for both H3N2 and H5N1 isolates; these are listed in Table 1 and mapped to the H3N2 structure

and visualized in Figure 4. These residues are an attractive target for future mutagenesis experiments; we hypothesize that single mutants in these positions may have altered activity in one or more of hemagglutinin's functional roles.

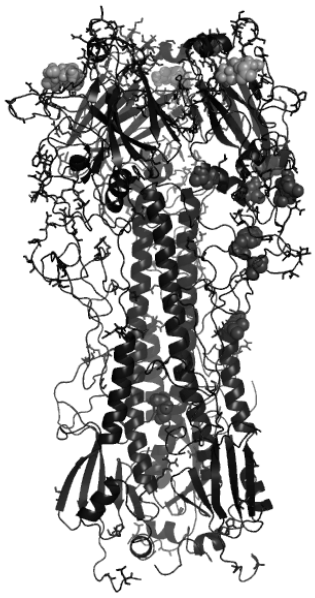

Figure 4. Residues identified in common between influenza subtypes.
Closely-linked residues identified in common between H3N2 and H5N1 subtypes are rendered in red on a structure of H3N2 hemagglutinin. The sialic acid of the ligand is shown as orange spheres.

	Position	A/Aichi/2/68 sequence
HA1	82	Q
HA1	83	T
HA1	106	A
HA1	107	S
HA1	115	S
HA1	164	L
HA1	174	F
HA1	197	Q
HA1	287	S
HA1	302	Y
HA2	2	L
HA2	59	T

Table 1. Closely-linked residues identified via independent analysis of both H3N2 and H5N1 subtypes.

4. Conclusions

We wish better to understand the functional regulation of influenza hemagglutinin, particularly with regard to ligand-binding specificity and catalysis of membrane fusion. Such an understanding will enable better surveillance for host-range changes in influenza and potentially assist in the development of novel small-molecule inhibitors. Using sequence mutual information as a robust metric of co-evolution, we have identified clusters of closely linked residues in both H5N1 and human H3N2 influenza isolates. Structure-based visualization facilitates the interpretation of these clusters in the context of existing functional data. By comparing the clusters for these two subtypes, we hypothesize that the extensive association network on the globular head of H3N2 may be linked to the human immune response, while the residues identified in common between H5N1 and H3N2 may be important to the basic functional roles of hemagglutinin. We suggest that these residues may be useful targets for future mutagenesis experiments. Our combined sequence-informatic and structural analysis thus enables generation of novel functional hypotheses that are not readily apparent via either method alone.

Acknowledgments

The authors thank R. Brandman, K. Branson, and O. Troyanskaya for helpful discussions. This work was supported by a fellowship from the Berry Foundation to P.K.

References

1. N. P. Johnson and J. Mueller, "Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic," *Bull Hist Med*, vol. 76, pp. 105-15, 2002.
2. "World Health Organization Cumulative number of Confirmed Human Cases of Avian Influenza A(H5N1)," 2008.
3. S. Jones and J. M. Thornton, "Searching for functional sites in protein structures," *Curr Opin Chem Biol*, vol. 8, pp. 3-7, 2004.
4. O. Lichtarge, H. R. Bourne, and F. E. Cohen, "An evolutionary trace method defines binding surfaces common to protein families," *J Mol Biol*, vol. 257, pp. 342-58, 1996.
5. G. Casari, C. Sander, and A. Valencia, "A method to predict functional residues in proteins," *Nat Struct Biol*, vol. 2, pp. 171-8, 1995.

6. S. Yamada, Y. Suzuki, T. Suzuki, M. Q. Le, C. A. Nidom, Y. Sakai-Tagawa, Y. Muramoto, M. Ito, M. Kiso, T. Horimoto, K. Shinya, T. Sawada, M. Kiso, T. Usui, T. Murata, Y. Lin, A. Hay, L. F. Haire, D. J. Stevens, R. J. Russell, S. J. Gamblin, J. J. Skehel, and Y. Kawaoka, "Haemagglutinin mutations responsible for the binding of H5N1 influenza A viruses to human-type receptors," *Nature*, vol. 444, pp. 378-82, 2006.
7. V. P. Marinina, A. S. Gambarian, N. V. Bovin, A. B. Tuzikov, A. A. Shilov, B. V. Sinitsyn, and M. N. Matrosovich, "[The effect of losing glycosylation sites near the receptor-binding region on the receptor phenotype of the human influenza virus H1N1]," in *Mol Biol (Mosk)*, vol. 37, 2003, pp. 550-5.
8. A. S. Gambaryan, V. P. Marinina, A. B. Tuzikov, N. V. Bovin, I. A. Rudneva, B. V. Sinitsyn, A. A. Shilov, and M. N. Matrosovich, "Effects of host-dependent glycosylation of hemagglutinin on receptor-binding properties on H1N1 human influenza A virus grown in MDCK cells and in embryonated eggs," in *Virology*, vol. 247, 1998, pp. 170-7.
9. M. Ohuchi, R. Ohuchi, A. Feldmann, and H. D. Klenk, "Regulation of receptor binding affinity of influenza virus hemagglutinin by its carbohydrate moiety," in *J Virol*, vol. 71, 1997, pp. 8377-84.
10. P. Auewarakul, O. Suptawiwat, A. Kongchanagul, C. Sangma, Y. Suzuki, K. Ungchusak, S. Louisirirotchanakul, H. Lerdsamran, P. Pooruk, A. Thitithanyanont, C. Pittayawonganon, C. T. Guo, H. Hiramatsu, W. Jampangern, S. Chunsutthiwat, and P. Puthavathana, "An avian influenza H5N1 virus that binds to a human-type receptor," *J Virol*, vol. 81, pp. 9950-5, 2007.
11. W. R. Atchley, K. R. Wollenberg, W. M. Fitch, W. Terhalle, and A. W. Dress, "Correlations among amino acid sites in bHLH protein domains: an information theoretic analysis," *Mol Biol Evol*, vol. 17, pp. 164-78, 2000.
12. N. G. Hoffman, C. A. Schiffer, and R. Swanstrom, "Covariation of amino acid positions in HIV-1 protease," *Virology*, vol. 314, pp. 536-48, 2003.
13. Y. Liu, E. Eyal, and I. Bahar, "Analysis of correlated mutations in HIV-1 protease using spectral clustering," *Bioinformatics*, vol. 24, pp. 1243-50, 2008.
14. L. C. Martin, G. B. Gloor, S. D. Dunn, and L. M. Wahl, "Using information theory to search for co-evolving residues in proteins," *Bioinformatics*, vol. 21, pp. 4116-24, 2005.
15. Y. Bao, P. Bolotov, D. Dernovoy, B. Kiryutin, L. Zaslavsky, T. Tatusova, J. Ostell, and D. Lipman, "The influenza virus resource at the National Center for Biotechnology Information," *J Virol*, vol. 82, pp. 596-601, 2008.

16. R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, pp. 1792-7, 2004.
17. N. K. Sauter, J. E. Hanson, G. D. Glick, J. H. Brown, R. L. Crowther, S. Park, J. J. Skehel, and D. C. Wiley, "Binding of influenza virus hemagglutinin to analogs of its cell-surface receptor, sialic acid: analysis by proton nuclear magnetic resonance spectroscopy and X-ray crystallography," in *Biochemistry*, vol. 31, 1992, pp. 9609-21.
18. Y. Ha, D. J. Stevens, J. J. Skehel, and D. C. Wiley, "X-ray structures of H5 avian and H9 swine influenza virus hemagglutinins bound to avian and human receptor analogs," *Proc Natl Acad Sci U S A*, vol. 98, pp. 11181-6, 2001.
19. I. H. Witten and E. Frank, *Data mining : practical machine learning tools and techniques*, 2nd ed. Amsterdam ; Boston, MA: Morgan Kaufman, 2005.
20. T. Bizebard, B. Gigant, P. Rigolet, B. Rasmussen, O. Diat, P. Bosecke, S. A. Wharton, J. J. Skehel, and M. Knossow, "Structure of influenza virus haemagglutinin complexed with a neutralizing antibody," *Nature*, vol. 376, pp. 92-4, 1995.
21. D. Fleury, B. Barrere, T. Bizebard, R. S. Daniels, J. J. Skehel, and M. Knossow, "A complex of influenza hemagglutinin with a neutralizing antibody that binds outside the virus receptor binding site," *Nat Struct Biol*, vol. 6, pp. 530-4, 1999.
22. D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier, "Mapping the antigenic and genetic evolution of influenza virus," *Science*, vol. 305, pp. 371-6, 2004.
23. B. P. Blackburne, A. J. Hay, and R. A. Goldstein, "Changing selective pressure during antigenic changes in human influenza H3," *PLoS Pathog*, vol. 4, pp. e1000058, 2008.