A Coding Theoretical Approach to Predict Sequence Changes in H5N1 Influenza A Virus Hemagglutinin

Keiko Sato, Toshihide Hara and Masanori Ohya
Department of Information Science, Tokyo University of Science, Noda, Japan

Keywords: Receptor Binding Domain, H5N1, Code Structure, Sequence Prediction.

Abstract: The changes in the receptor binding domain of influenza A virus hemagglutinin lead to the appearance of new viral strains that evade the immune system. To prepare the future emergence of potentially dangerous outbreaks caused by divergent influenza strains including human-adapted H5N1 strains, it is imperative that we understand the rule stored in the sequence of the receptor binding domain. Information of life is stored as a sequence of nucleotides, and the sequence composed of four nucleotides seems to be a code. It is important to determine the code structure of the sequences. Once we know the code structure, we can make use of mathematical results concerning coding theory for research in life science. In this study, we applied various codes in coding theory to sequence analysis of the 220 loop in the receptor binding domain of H1, H3, H5 and H7 subtype viruses isolated from humans. Sequence diversity in the 220 loop has been observed even within the same hemagglutinin subtype. However, we found that the code structure of the 220 loop from the same subtype remains unchanged. Our results indicate that the sequences at the 220 loop have the structure of subtype-specific codes. In addition, in view of these finding, we predicted possible amino acid changes in the 220 loop of H5N1 strains that will emerge in the future. Our method will facilitate understanding of the evolutionary patterns of influenza A viruses, and further help the development of new antiviral drugs and vaccines.

1 INTRODUCTION

Influenza A viruses have eight pieces of segmented RNA, which encode 11 proteins (Olsen et al., 2006). The antigenic properties in the two viral surface proteins, hemagglutinin and neuraminidase, are used to classify influenza A viruses into different subtypes. Currently Influenza A viruses circulating among humans are the H1N1 and H3N2 subtypes. Although other subtypes such as H5N1 and H7N9 have not yet gained the ability to spread efficiently from person to person, these virus subtypes have occasionally infected humans.

High-pathogenicity avian H5N1 influenza viruses exhibiting high lethality continue to pose threats to our lives since their emergence in China in 1996. According to the World Health Organization (WHO), there have been 826 human cases with H5N1 influenza infection since 2003, and approximately 53% of the cases have died (as of March 31, 2015). Despite the high mortality, H5N1 viruses have not yet gained the ability to spread efficiently from person to person. However, the outbreaks of H5N1 have been reported among domestic poultry and wild birds in many countries (Durand et al., 2015; Pfeiffer et al., 2011; Yamamoto et al., 2011). In addition, recent studies reported that a reassortant influenza virus containing a hemagglutinin protein from an H5N1 virus with four mutations can be transmitted between ferrets (Imai et al., 2012). The viral surface protein, hemagglutinin mediates binding of the virus to target cells via the host cell receptor, sialic acid (Jiang et al., 2012; Rumschlag-Booms and Rong, 2013). The hemagglutinin of avian influenza viruses preferentially binds sialic acid receptors (α2,3-SA) on epithelial cells in the intestinal tract of birds and in the lower respiratory tract of humans, whereas the hemagglutinin of human influenza viruses preferentially binds another type of sialic acid (α2,6-SA) (Schrauwen and Fouchier, 2014; Yen and Peiris, 2012). The receptor binding domain (RBD) of hemagglutinin, situated at the outer surface on top of the viral spike, is composed of three major structural elements: a 130-loop (residues 134-138), a 190-helix (residues 188-190), and a 220-loop

Sato K., Hara T. and Ohya M.
A Coding Theoretical Approach to Predict Sequence Changes in H5N1 Influenza A Virus Hemagglutinin.
DOI: 10.5220/0005659701590167
In *Proceedings of the 9th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOINFORMATICS 2016)*, pages 159-167
ISBN: 978-989-758-170-0

(residues 221-228) based on H3 numbering (Das et al., 2009; Durand et al., 2015; Jiang et al., 2012; Stevens et al., 2006). It is considered that the mutations in the RBD could affect the receptor binding avidity and specificity of hemagglutinin (Chen et al., 2011; de Vries et al., 2013; de Vries et al., 2014; Schrauwen and Fouchier, 2014). The RBD is the primary target of neutralizing antibodies, which are induced by virus infection or by vaccination with specific antigen (Bright et al., 2003; Chen et al., 2011; Jiang et al., 2012; Khurana et al., 2011; McCullough et al. 2012). However, the mutations in the RBD lead to change in viral immunogenicity and antigenicity (Chen et al., 2011; Xu et al., 2010). Jiang et al. (2012) state that RBD plays a critical role in the elucidation of antiviral immune response and protective immunity. McCullough et al. (2012) also state that a better understanding of mutations in the RBD may be useful in vaccine and drug design effort. To prepare the future emergence of potentially dangerous outbreaks caused by divergent influenza strains including human-adapted H5N1 strains, it is imperative that we understand the rule stored in the sequence of the RBD.

Information of life is stored as a code composed of four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). Therefore, we can consider that the DNA or gene in each organism is a code showing its inherent structure. In protein coding region, each group of three consecutive nucleotides is called a codon, and each codon corresponds to one amino acid. The total number of three nucleotide groups is the third power of 4, which means we have 64 codons. However, only 20 proteinogenic amino acids exist in nature. Moreover, it is supposed that the third nucleotide for a codon will not play an essential role in making of an amino acid. This shows that a gene has redundancy to correct errors to some extent. In other words, it has a structure that is similar to one of an error-correcting/detecting code for the transmission of information. In life-science research, it is important to determine the code structure of the target gene. Once we know the code structure, we can make use of mathematical results concerning coding theory for research in life science. How can the RBD sequences of influenza A viruses be discussed using coding theory? The present study was conducted to find out the code structure of the 220 loop of influenza A viruses, and to predict sequence changes in the 220 loop of H5N1 virus.

2 METHODS

2.1 Sequence Data

We applied artificial codes in coding theory to sequence analysis of the 220 loop in the H1, H3, H5 and H7 RBD. All full-length amino acid and nucleotide sequences of hemaggulutinin from influenza A H1, H3, H5, and H7 subtypes were downloaded from the Influenza Research Database on September 2014. The hemaggulutinin data set consists of 8,941 human sequences from the H1 subtype between 1918 and 2014, 6,013 human sequences from H3 subtype between 1968 and 2014, 230 human sequences from the H5 subtype between 1997 and 2013, and 51 human sequences from H7 subtype between 1996 and 2014. The sequences were aligned using MAFFT (Katoh and Toh, 2008) which can quickly process a large dataset.

2.2 Sequence Analysis of the 220 Loop by Coding Theory

We explain how to encode the nucleotide sequence of the 220 loop to detect the code structure. The method for applying artificial codes to sequence analysis has been described in detail previously (Ohya and Sato, 2000; Sato et al., 2013). Since the Galois Field GF(4) consists of four elements, 0, 1, α and α^2 such that $\alpha^2+\alpha+1=0$, the four nucleotides can be expressed in each of four elements. There are a total of 24 (= 4!) different possible combinations to map the four nucleotides to the four elements in GF(4).

First, an important part of the nucleotide sequence of the 220 loop from an influenza strain, namely the nucleotide sequence excluding the third nucleotide of each codon, is transformed into the information sequence which consists of the elements of GF(4). Next, the information sequence is grouped into blocks and then encoded into code words of an error-correcting/detecting code C. The total length of such a code (code word length) is multiples of 3 and the length of the information symbols (information block length) is multiples of 2. The check symbols in each code word are placed into the corresponding position of the third nucleotide of codon. Then, the encoded sequence, which consists of the set of the code words, is written back to nucleotide sequence. We call it the encoded nucleotide sequence. After that, the encoded nucleotide sequence is converted into amino acid sequence. We call it the encoded amino acid sequence. Finally, the degree of similarity between the amino acid sequence of the

220 loop from the influenza strain and the encoded amino acid sequence described above is computed. We think that if the amino acid sequence of the 220 loop is identical to the encoded amino acid sequence generated by the code C, i.e. the similarity is 100%, then the nucleotide sequence of the 220 loop has the structure of the code C. Therefore, it is possible to find the code structure of the 220 loop by computing the degree of similarity for various artificial codes. Artificial codes used for our study are the so-called linear codes, cyclic codes, Bose-Chaudhuri-Hocquenghem (BCH) codes, self-orthogonal codes and Iwadare codes. Practically, we used 95 types of codes including differences in generator polynomial.

Let X_i $(i = 1, 2, \cdots, 230)$ be 230 amino acid sequences of the 220 loop from the H5 subtype. As described above, we encode the 230 nucleotide sequences of the 220 loop in a code C, and then get the encoded amino acid sequences X_i^C $(i = 1, 2, \cdots, 230)$. Because the 220 loop is composed of 8 amino acid residues, a degree to measure the similarity between X_i and X_i^C is denoted by rate of coincidence (RC) as follows:

$$\mathrm{RC}(X_i, X_i^C) = 1 - a/8 \quad (0 \leq \mathrm{RC}(X_i, X_i^C) \leq 1),$$

where a is the numbers of sites for which two amino acid sequences differ from each other. $\mathrm{RC}(X_i, X_i^C)$=1 means that the similarity between X_i and X_i^C is 100%. If all of the 230 amino acid sequences of the 220 loop from H5 subtype are identical to the encoded amino acid sequences generated by the code C, i.e. $\sum_{i=1}^{230} \mathrm{RC}(X_i, X_i^C)/230 = 1$, then 100% of the 220 loop nucleotide sequences have the structure of the code C.

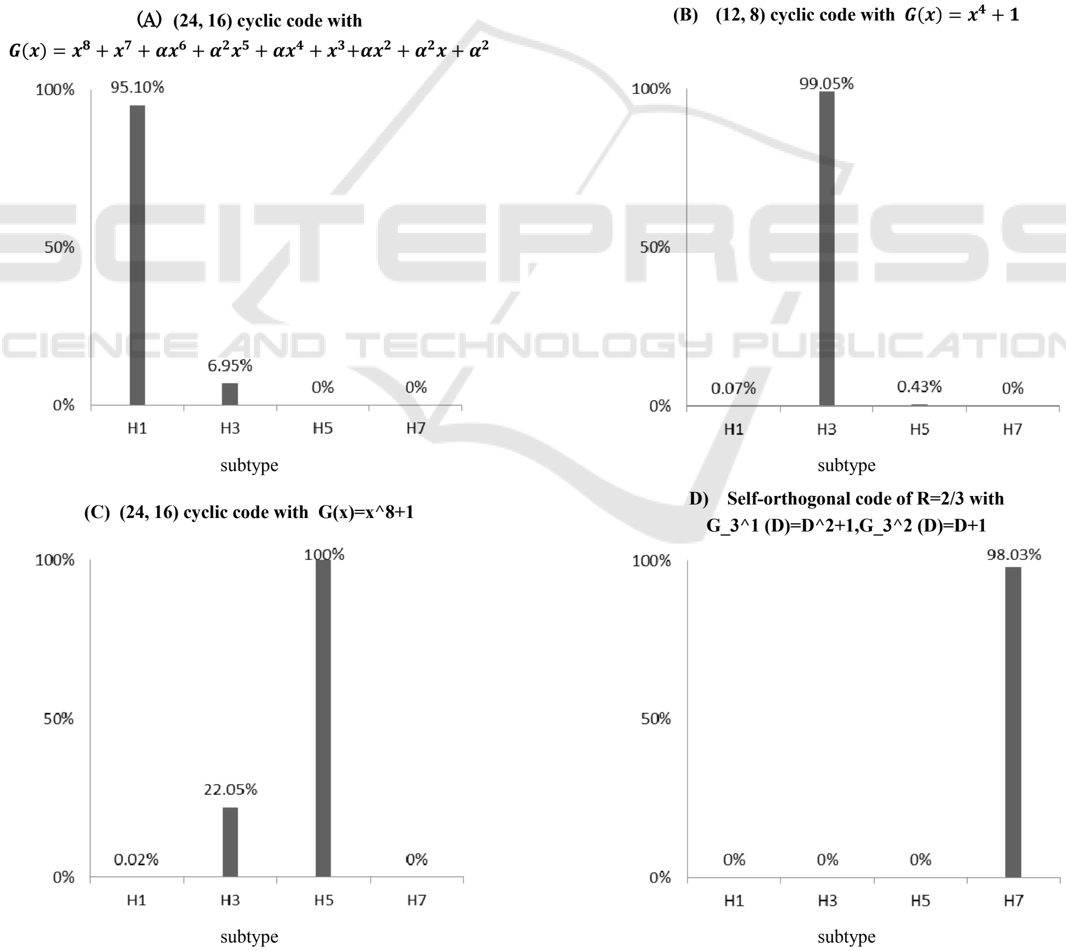

Figure 1: Percentage of the 220 loop nucleotide sequences with the structure of the indicated codes for H1, H3, H5, and H7 subtypes. The percentage was calculated to the second decimal place. G is the generator polynomial of each code. For figures (A) and (B), the correspondence between the four nucleotides and the elements in GF(4) was given as A → 0, C →1, T → α and G → α^2. For figures (C) and (D), C → 0, A →1, G → α and T → α^2.

By using 95 types of codes for each case of the 24 representations of the four nucleotides in the elements of GF(4), we tried to find the code structure of the 220 loop in each of the H1, H3, H5 and H7 subtype viruses in this way.

Once we found the code structure for the 200 loop of influenza A virus by using various artificial codes, we can apply this results to the prediction of amino acid residues in the 220 loop of influenza strains that will emerge in the future. The 220 loop is composed of 8 amino acid residues (24 nucleotides). The 220 loop of the H5N1 viruses isolated since 1997 showed nucleotide changes in 5 positions (the first at codons 221, 222 and 223, and the second at codons 226 and 227) out of 16 positions excluding the third nucleotide position from each of the 8 codons. Therefore we consider the 5 positions as variable positions, while the remaining 11 positions as no variable positions. Given the possibility of any one of four nucleotides at each of the 5 positions, the information sequence is composed of 16 nucleotides as follows, where N stands for any one of four nucleotides: NCNANTAAGGCNANGG. In other words, as for the information sequence of length 16, 1,024 (= $4^5 \times 1^{11}$) patterns are made through combination of these 16 positions. To predict possible amino acid changes in the 220 loop of H5N1 influenza hemagglutinin, each of these information sequences was encoded using the encoding scheme of the code characterizing the 220 loop sequences from H5N1 viruses.

3 RESULTS

3.1 The Code Structure of the 220 Loop of Influenza A Viruses

Figure 1 shows the percentage of the 220 loop nucleotide sequences with the structure of the indicated codes for their respective subtypes. Interestingly, more than 95% (8,504/8,941) of the 220 loop nucleotide sequences of the H1 subtype that infected humans between 1918 and 2014 had the structure of the (24, 16) cyclic code with the generator polynomial $G(x) = x^8 + x^7 + \alpha x^6 + \alpha^2 x^5 + \alpha x^4 + x^3 + \alpha x^2 + \alpha^2 x + \alpha^2$ (Figure 1(A)). Almost all of the 220 loop nucleotide sequences from other subtypes (H3, H5 and H7) did not have that structure. For the H3 subtype that infected humans between 1968 and 2014, more than 99% (5,956/6,013) of the 220 loop nucleotide sequences had the structure of the (12, 8) cyclic code with the generator polynomial $G(x) = x^4 + 1$ (Figure 1(B)).

Table 1: Possible amino acid changes in the 220 loop of H5N1 influenza A strains that will emerge in the future.

Residues 221-228		
TEMNGQNG	SEVKGLNG	PKLNGQNG
TEMNGQSG	SEVKGLTG	PKLNGQSG
TEMNGRTG	SEVNGRNG	PKLKGLIG
TEMKGPIG	SEVNGRSG	PKLKGLTG
TQINGQNG	SEVKGPSG	PKLNGRNG
TQINGQSG	SEVKGPTG	PKLNGRSG
TQIKGLIG	SELNGQNG	PKLKGPIG
TQIKGLTG	SELNGQSG	PKLKGPTG
TQIKGPNG	SELKGLTG	PKVNGQNG
TQIKGPIG	SELNGRNG	PKVNGQSG
TQIKGPTG	SELNGRSG	PKVKGLIG
TQVKGLIG	SELKGPTG	PKVKGLTG
TQVKGLTG	SQINGHIG	PKVNGRNG
TQVKGPIG	AKINGQNG	PKVNGRSG
TQVKGPTG	AKINGQSG	PKVKGPIG
TQLKGLIG	AEMNGHIG	PKVKGPTG
TQLKGLTG	AEMNGHTG	PEMNGQNG
TQLKGPIG	AEMKGLNG	PEMNGHIG
TQLKGPTG	AEMKGLIG	PEMNGQSG
SKINGQNG	AEMKGLSG	PEMNGHTG
SKINGQSG	AEMNGRIG	PEMKGLNG
SKIKGLTG	AEMNGRTG	PEMKGLSG
SKINGRNG	AEMKGPIG	PEMNGRNG
SKINGRSG	AEMKGPTG	PEMNGRSG
SKIKGPTG	AQVKGLIG	PEMKGPNG
SKVNGQNG	AQVKGLTG	PEMKGPSG
SKVNGQSG	AQVKGPIG	PEVNGQNG
SKVKGLSG	AQVKGPTG	PEVNGQSG
SKVKGLTG	AQLKGLIG	PEVKGLIG
SKVNGRNG	AQLKGLTG	PEVKGLTG
SKVNGRSG	AQLKGPIG	PEVNGRNG
SKVKGPTG	AQLKGPTG	PEVNGRSG
SKLNGQNG	PKINGQNG	PEVKGPIG
SKLNGQSG	PKINGHIG	PEVKGPTG
SKLKGLTG	PKINGQSG	PELNGQNG
SKLNGRNG	PKIKGLNG	PELNGQSG
SKLNGRSG	PKIKGLIG	PELKGLIG
SKLKGPTG	PKIKGLSG	PELKGLTG
SEMNGQNG	PKIKGLTG	PELNGRNG
SEMNGQSG	PKINGRNG	PELNGRSG
SEMKGPNG	PKINGRSG	PELKGPIG
SEVNGQNG	PKIKGPIG	PELKGPTG
SEVNGQSG	PKIKGPTG	

Those from other subtypes (H1, H5 and H7) did not have that structure. In addition, we found the code structure characterizing the 220 loop sequences from the H5 and H7 subtypes, respectively. All (230/230) of the nucleotide sequences of the H5 subtype that infected humans between 1997 and 2013 had the structure of the (24, 16) cyclic code with the generator polynomial $G(x) = x^8 + 1$ (Figure 1(C)). For the H7 subtype that infected humans between 1996 and 2014, approximately 98% (50/51) of the nucleotide sequences had the structure of the self-orthogonal code of information rate R=2/3 with the generator polynomial $G_3^1(D) = D^2 + 1$, $G_3^2(D) = D + 1$ (Figure 1(D)). The amino acid sequences of the 220 loop are diverse even within the same subtype (Tables S1-S4). However, surprisingly, the code structure of the 220 loop from the same subtype remains unchanged.

3.2 Future Sequence Changes in H5N1 220 Loop

We found the mutation rules for the 200 loop of influenza A virus hemagglutinin by using various artificial codes in information transmission. As became clear above, the 220loop human sequences from H5N1 strains have preserved the structure of a specific code since the emergence of H5N1 in humans in 1997. In this study of predicting sequences, we used 95 types of codes including differences in generator polynomials on the condition that C, A, G and T of nucleotides correspond to 0, 1, α and α^2 of Galois Field GF(4), respectively. Every 220 loop amino acid sequence belonging to the H5 subtype was identical to the encoded amino acid sequence generated by the (24, 16) cyclic code with generator polynomial $G(x) = x^8 + 1$ (the similarity is 100%) and was not identical to that generated by any of different 65 types of codes (the similarity is 0%).

Table 1 shows possible amino acid changes in the 220 loop of H5N1 influenza strains that will emerge in the future. These are composed of 128 sequences out of the 1,024 encoded amino acid sequences generated by the (24, 16) cyclic code with generator polynomial $G(x) = x^8 + 1$, the rest of which were removed because of overlap with the encoded amino acid sequences generated by the 65 types of codes. The possible changes we predicted are based on the assumption that although sequence diversity in the 220 loop of H5N1 hemagglutinin will be observed even from now on, the code structure will probably not change.

4 CONCLUSIONS

Influenza A H1 and H3 subtypes, which have circulated among humans for nearly 100 years since the pandemic of 1918 and for nearly 50 years since the pandemic of 1968 respectively, continue to change by accumulation of mutations in the hemagglutinin. Similarly, other subtypes such as H5 and H7, which have occasionally caused human infections, change by mutations in the hemagglutinin. These changes, particularly the changes in the RBD of the hemagglutinin, lead to the appearance of new viral strains that evade the immune system. Therefore, it is imperative for us to understand the mutational patterns in the RBD. Sequence diversity in the 220 loop of the RBD, has been observed among different hemagglutinin subtypes, or even within the same subtype. However, the code structure of the 220 loop from the same subtype remains unchanged. Our results indicate that the sequences at the 220 loop have the structure of subtype-specific codes. The first goal of this study was to find out the code structure of the 220 loop of influenza A viruses. We fortunately found the rules of mutations for the loop by using various codes in information transmission. These findings may be very helpful in predicting sequence changes in the 220 loop and may provide clues to the decision of vaccine strain and the development of new antiviral drugs. The 220 loop of the RBD is definitely an attractive target for developing antiviral drugs.

The second goal of this study was to predict sequence changes in the 220 loop of H5N1 virus. Based on the assumption that the code structure of the 220 loop from the same subtype will probably not change even from now on, we predicted possible amino acid changes in the 220 loop of H5N1 influenza strains that will emerge in the future. We cannot deny the possibility that a pandemic H5N1 strain transmissible between humans may not possess the amino acid changes predicted here. Monitoring the molecular changes in hemagglutinin is important for the accurate sequence prediction. However, our method, which determines the code structure of the 220 loop of influenza A virus hemagglutinin, will facilitate understanding of the evolutionary patterns of influenza A viruses, and further help the development of new antiviral drugs and vaccines. Through the generation of mutant viruses possessing hemagglutinin gene with mutations of the 220 loop predicted in our method and the examination of the growth and transmissibility of the mutant viruses in animal models, suitable vaccine candidates will be selected. It is expected that the 220 loop-based influenza vaccines would be effective against divergent influenza strains, including those that may cause pandemics in the future.

REFERENCES

Bright, R.-A. et al., 2003, Impact of glycosylation on the immunogenicity of a DNA-based influenza H5 HA vaccine. *Virology*, 308, 270-278.

Chen, M.-W. et al., 2011, Broadly neutralizing DNA vaccine with specific mutation alters the antigenicity and sugar-binding activities of influenza hemagglutinin, *Proceedings of the National Academy of Sciences,* 108, 3510–3515.

Das, P. et al., 2009, Free energy simulations reveal a double mutant avian H5N1 virus hemagglutinin with altered receptor binding specificity, *J. Comput. Chem.,* 30, 1654–1663.

de Vries, R.-P. et al., 2013, Evolution of the hemagglutinin protein of the new pandemic H1N1 influenza virus: maintaining optimal receptor binding by compensatory substitutions. *J. Virol.*, 87, 13868-13877.

de Vries, R.-P. et al., 2014, Hemagglutinin receptor specificity and Structural Analyses of Respiratory Droplet-transmissible H5N1 Viruses. *J. Virol.*, 88, 768-773.

Durand, L.-O. et al., 2015, Timing of Influenza A(H5N1) in Poultry and Humans and Seasonal Influenza Activity Worldwide, 2004-2013, *Emerg. Infect. Dis.*, 21, 202-208.

Imai, M. et al., 2012, Experimental adaptation of an influenza H5 HA confers respiratory droplet transmission to a reassortant H5 HA/H1N1 virus in ferrets, *Nature*, 486, 420-428.

Jiang, S. et al., 2012, Receptor-binding domains of spike proteins of emerging or re-emerging viruses as targets for development of antiviral vaccines, *Emerging Microbes & Infections,* 1, e13.

Katoh, K. and Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program, *Brief. Bioinform.,* 9, 286–298.

Khurana, S. et al., 2011, Bacterial HA1 vaccine against pandemic H5N1 influenza virus: evidence of oligomerization, hemagglutination, and cross-protective immunity in ferrets, *J. Virol.,* 85, 1246–1256.

McCullough, C. et al., 2012, Characterization of influenza hemagglutinin interactions with receptor by NMR, *PloS one,* 7, e33958.

Ohya, M. and Sato, K., 2000, Use of information theory to study genome sequences. *Reports on Mathematical Physics,* 46, 419–428.

Olsen, B. et al., 2006, Global patterns of influenza A virus in wild birds. *Science,* 312, 384–388.

Pfeiffer, D.-U. et al., 2011, Implications of global and regional patterns of highly pathogenic avian influenza virus H5N1 clades for risk management, *Vet. J.*, 190, 309-316.

Rumschlag-Booms, E. and Rong, L., 2013, Influenza a virus entry: implications in virulence and future therapeutics, *Advances in virology,* 2013, 121924.

Sato, K. et al., 2013, The code structure of the p53 DNA-binding domain and the prognosis of breast cancer patients. *Bioinformatics,* 29, 2822–2825.

Schrauwen, E.-J., 2014, Host adaptation and transmission of influenza A viruses in mammals. *Emerg. Microbes Infect.*, 3, e9.

Stevens, J. et al., 2006, Structure and receptor specificity of the hemagglutinin from an H5N1 influenza. *Science*, 312, 404-410.

Xu, Q. et al., 2010, Influenza H1N1 A/Solomon Island/3/06 virus receptor binding specificity correlates with virus pathogenicity, antigenicity, and immunogenicity in ferrets. *J. Virol.*, 84, 4936-4945.

Yamamoto, N. et al., 2011, Characterization of a non-pathogenic H5N1 influenza virus isolated from a migratory duck flying from Siberia in Hokkaido, Japan, in October 2009. *Virol. J.*, 8, 65.

Yen, H.-L. and Peiris, J.-S., 2012, Virology: bird flu in mammals, *Nature*, 486, 332-333.

APPENDIX

Table S1: Amino acid sequence diversity in the 220 loop of human H1 hemagglutinin.

H1 strain	Residues 221-228
AF117241|A/South_Carolina/1/18|H1N1	PKVRDQAG
CY010788|A/WSN/1933_TS61|H1N1	PKVKDQHG
U08904|A/WS/1933|H1N1	PKVRDQPG
DQ508905|A/Wilson_Smith/1933|H1N1	PKVRDQHG
U08903|A/NWS/1933|H1N1	PKVRNQPG
CY040170|A/Puerto_Rico/8_SV14/1934|H1N1	PKVKGQAG
CY146857|A/Puerto_Rico/8_SV40/1934|H1N1	PKVKDQAG
CY147326|A/BH/JY2/1935|H1N1	PKVRDQTG
CY020445|A/Henry/1936|H1N1	PEVRDQAG
CY013271|A/Hickox/1940|H1N1	PKVRGQAG
CY045772|A/Melbourne/1/1946|H1N1	PEVKDQAG
CY077768|A/Netherlands/002P1/1951|H1N1	PKVRNQAG
CY009340|A/Malaysia/54|H1N1	PKVRGQPG
CY008988|A/Denver/1/1957|H1N1	PKVRDQSG
CY125862|A/Kw/1/1957|H1N1	PKVRGQSG
CY021717|A/California/10/1978|H1N1	PKVRGQEG
CY028724|A/California/45/1978|H1N1	PKVRDQEG
CY020173|A/Lackland/7/1978|H1N1	PKVRDQKG
CY017203|A/Memphis/23/1983|H1N1	PKVRNQEG
CY104862|A/TayNguyen/TN182/2006|H1N1	PKVRDQGG
EU100724|A/Solomon_Islands/03/2006|H1N1	PKVRDREG
EU199338|A/Texas/06/2007|H1N1	PKVRBQEG
CY027779|A/Kentucky/UR06_0339/2007|H1N1	PKVREQEG
CY118091|A/Malaysia/1794173/2007|H1N1	LKVRDQEG
EU516017|A/Hawaii/31/2007|H1N1	PKIRDQEG
CY073960|A/Mexico/UASLP_009/2008|H1N1	PKLRDQDG
GU367325|A/Novgorod/01/2009|H1N1	PKVREREG
CY049076|A/Singapore/ON141/2009|H1N1	PKVGDQEG
CY051455|A/Wisconsin/629_S0339/2009|H1N1	TKVRDQEG
CY095906|A/Zhejiang/8/2009|H1N1	PKVRDQER
CY054606|A/Thailand/THB0405/2009|H1N1	PRVRDQEG
CY122835|A/Singapore/GP2242/2009|H1N1	PQVRDQEG
CY075897|A/Blore/NIV1196/2009|H1N1	PKMRGKEG
KC781609|A/California/33/2009|H1N1	PKMRDQEG
CY083399|A/Great_Lakes/WRAIR1664P/2009|H1N1	PKVKEQEG
KC782207|A/South_Carolina/18/2009|H1N1	PKVKDQEG
KC781375|A/Oregon/35/2009|H1N1	HKVRDQEG
CY095955|A/Zhejiang/86/2009|H1N1	PKVRDQEA
CY057254|A/New_York/5186/2009|H1N1	PKVMDQEG
CY069114|A/Madrid/INS296/2009|H1N1	PKVRAQEG
HM581919|A/Iran/15583/2009|H1N1	PKVRDRQG
KF411180|A/Qingdao/FF85/2009|H1N1	PKVRDSEG
CY067632|A/Qingdao/66/2010|H1N1	PKVRDQEW
CY092952|A/Chile/15/2010|H1N1	PKLRDQEG
CY079544|A/Switzerland/5165/2010|H1N1	PKVREQAG
JQ796827|A/Zhejiang/HZ19/2011|H1N1	PIVRDQEG
JQ396238|A/Kenya/145/2011|H1N1	PKGRDQEG
KF451900|A/Kenya/262/2013|H1N1	PKVKEQDG
KM013710|A/Shiraz/87/2013|H1N1	PKVRDHEG
KJ645782|A/Gainesville/03/2014|H1N1	PKVRSQEG

All groups of identical sequences in the 220 loop sequences from H1 subtype that infected humans between 1918 and 2014 were represented by the oldest sequence in the group.

Table S2: Amino acid sequence diversity in the 220 loop of human H3 hemagglutinin.

H3 strain	Residues 221-228
CY011120|A/Northern_Territory/60/1968|H3N2	PWVRGLSS
V01103|A/NT/60/68/29c|H3N2	PWVRGQSS
AB284320|A/Aichi/2/1968|H3N2	PWVGGLSS
CY033529|A/Hong_Kong/1_9_MA21_3/1968|H3N2	PWIRGLSS
CY112249|A/Hong_Kong/1/1968|H3N2	PWVRGMSS
CY112297|A/Bilthoven/6022/1972|H3N2	PWVRGPSS
CY113957|A/Akita/4/1993|H3N2	PWVRGQPS
CY113981|A/Lyon/672/1993|H3N2	PWVRGLPS
CY114149|A/Hong_Kong/56/1994|H3N2	PWVRGISS
CY118426|A/Malaysia/07831/1995|H3N2	PWVRGVSS
CY009676|A/New_York/576/1997|H3N2	PWIRGVSS
CY121424|A/California/32/1999|H3N2	HWVRGVSS
CY001397|A/New_York/156/2000|H3N2	PWERGVSS
EU856922|A/Hong_Kong/CUHK22072/2000|H3N2	PWVRDVSS
EU856918|A/Hong_Kong/CUHK21932/2001|H3N2	PWIRDVSS
EU856946|A/Hong_Kong/CUHK24749/2001|H3N2	PRVRDVSS
DQ415319|A/TW/872/02|H3N2	HRVRDVSS
CY112933|A/Fujian/411/2002|H3N2	PRVRGVSS
CY003096|A/New_York/403/2002|H3N2	PWGRGVSS
CY007843|A/Canterbury/14/2002|H3N2	PWARGVSS
EU857019|A/Hong_Kong/CUHK50200/2002|H3N2	PRIRDVSS
EU103747|A/Denmark/87/2003|H3N2	PRVRDVPS
EF568926|A/Thailand/Siriraj_02/2003|H3N2	PRVRDIPS
AY531033|A/Wyoming/3/03|H3N2	PRVRDISS
EU857094|A/Hong_Kong/CUHK83422/2003|H3N2	LRVRDVPS
DQ249261|A/Taiwan/30005/2004|H3N2	PRVRHIPS
CY105310|A/HaNoi/HN30147/2004|H3N2	TRVRDVPS
CY013517|A/Wellington/58/2004|H3N2	SRVRDIPS
CY002064|A/New_York/392/2004|H3N2	PRIRDVPS
CY163648|A/Wisconsin/67/2005|H3N2	PRIRNIPS
EU283414|A/Hiroshima/52/2005|H3N2	PRVRNIPS
CY016595|A/South_Australia/18/2005|H3N2	LRVRDIPS
CY016028|A/Western_Australia/74/2005|H3N2	PRIRDIPS
KJ855363|A/Mexico/DIF29/2006|H3N2	LRVRNIPS
CY020357|A/New_York/923/2006|H3N2	PRVRBIPS
EU716471|A/Texas/03/2008|H3N2	HRVRNIPS
CY037543|A/California/UR07_0053/2008|H3N2	PRIKNIPS
FJ179354|A/Minnesota/14/2008|H3N2	PKVRNIPS
GQ385889|A/New_Hampshire/01/2009|H3N2	PRVREIPS
CY050125|A/Qingdao/1329/2009|H3N2	PRVGNIPS
CY091837|A/Guangdong/322/2010|H3N2	TRVRNIPS
JX946754|A/Qingdao/FF184/2010|H3N2	PRLRNIPS
KC882891|A/District_Of_Columbia/02/2010|H3N2	ARVRNIPS
KC882953|A/Minnesota/04/2011|H3N2	SRVRNIPS
CY162984|A/Peru/PER345/2011|H3N2	PRVRNVPS
KC892741|A/New_Jersey/08/2011|H3N2	LRIRNIPS
KC892638|A/California/34/2011|H3N2	PRIRBIPS
KJ942608|A/Hawaii/22/2012|H3N2	PRIRNSPS
KF598718|A/British_Columbia/004/2012|H3N2	HRIRNIPS
KC892959|A/Hawaii/02/2012|H3N2	TRIRNIPS
CY134996|A/Texas/JMM_37/2012|H3N2	PRIRNVPS
KF789696|A/Maine/05/2012|H3N2	PRIRNNPS
KF790228|A/Hawaii/30/2012|H3N2	SRIRNIPS
CY141264|A/Texas/3249/2013|H3N2	PRIRSIPS
KF789872|A/Hawaii/02/2013|H3N2	LRIRDIPS
KM064043|A/Texas/14/2014|H3N2	HRIRDIPS

All groups of identical sequences in the 220 loop sequences from H3 subtype that infected humans between 1968 and 2014 were represented by the oldest sequence in the group.

Table S3: Amino acid sequence diversity in the 220 loop of human H5 hemagglutinin.

H5 strain	Residues 221-228
GU052142|A/Hong_Kong/485/1997|H5N1	PKVNGQSG
GU052089|A/Hong_Kong/378.1/2001|H5N1	SKVNGQSG
AB212054|A/Hong_Kong/213/2003|H5N1	SKVNGQNG
EF107522|A/Thailand/1_KAN_1A_/2004|H5N1	SEVNGQSG
EF456802|A/Viet_Nam/JPHN30321/2005|H5N1	SKINGQSG
DQ371929|A/Anhui/2/2005|H5N1	SKVNGRSG
KF918470|A/Cambodia/X0810301/2013|H5N1	SKVKGLSG

All groups of identical sequences in the 220 loop sequences from H5 subtype that infected humans between 1997 and 2013 were represented by the oldest sequence in the group.

Table S4: Amino acid sequence diversity in the 220 loop of human H7 hemagglutinin.

H7 strain	Residues 221-228
GU053110|A/England/AV877/1996|H7N7	PQVNGQSG
CY181569|A/Anhui/DEWH72_08/2013|H7N9	PQVNGLSG
KF018039|A/Taiwan/1/2013|H7N9	PQVNGPSG
KC853766|A/Hangzhou/1/2013|H7N9	PQVNGISG
KF609511|A/Shanghai/JS01/2013|H7N9	TQVNGQSG

All groups of identical sequences in the 220 loop sequences from H7 subtype that infected humans between 1996 and 2014 were represented by the oldest sequence in the group.