Data Mining in Amino Acid Sequences of H3N2 Influenza Viruses Isolated during 1968 to 2006

Kimihito Ito, Manabu Igarashi, Ayato Takada
Hokkaido University Research Center for Zoonosis Control
Sapporo 060-0818, Japan

Abstract

The hemagglutinin (HA) of influenza viruses undergoes antigenic drift to escape from antibody-mediated immune pressure. In order to predict possible structural changes of the HA molecules in future, it is important to understand the patterns of amino acid mutations and structural changes in the past. We performed a retrospective and comprehensive analysis of structural changes in H3 hemagglutinins of human influenza viruses isolated during 1968 to 2006. Amino acid sequence data of more than 2000 strains have been collected from NCBI Influenza virus resources. Information theoretic analysis of the collected sequences revealed a number of simultaneous mutations of amino acids at two or more different positions (correlated mutations). We also calculated the net charge of the HA1 subunit, based on the number of charged amino acid residues, and found that the net charge increased linearly from 1968 to 1984 and, after then, has been saturated. This level of the net charge may be an upper limit for H3 HA to be functional. It is noted that ”correlated mutations” with the conversion of acidic and basic amino acid residues between two different positions were frequently found after 1984, suggesting that these mutations contributed to counterbalancing effect to keep the net charge of the HA . These approaches may open the way to find the direction of future antigenic drift of influenza viruses.

1. Introduction

Influenza A virus causes highly contagious, acute respiratory illness, and continues to be a major cause of morbidity and mortality. The viral strains mutates from year to year, causing the annual epidemic world wide.

The hemagglutinin (HA) is the major surface glycoprotein of influenza viruses and plays an important role in virus entry into host cells. The HA undergoes antigenic drifts which occur by accumulation of a series of amino acid substitutions. Influenza viruses that escape from antibody-mediated immune pressure of human population acquire a new antigenic structure and continuously cause epidemic in the world.

In order to predict possible structural changes of the HA molecules in future, it is important to understand the patterns of antigenic drift of influenza viruses in the past. We have been studying computational methods that include information theoretic analysis to find patterns of amino acids substitutions, molecular modeling to reveal the changes in 3D structure of antigenically different HAs, and molecular simulation to investigate the antigen-antibody interaction.

We performed a retrospective and comprehensive analysis of structural changes in H3 HAs of human influenza viruses isolated during 1968 to 2006. Amino acid sequence data of more than 2000 strains have been collected from NCBI Influenza virus resources and used for the analysis. These approaches may open the way to find the direction of future antigenic drift of influenza viruses.

2. Information Theoretic Analysis of the Past HA Sequences

The entropy and mutual information are used to find simultaneous mutations of amino acids at two or more different positions (correlated mutations). The entropy of an amino acid position represents the uncertainty of the amino acids in the position. The amino acid positions with frequent mutations are expected to have higher entropy. The mutual information between two amino acid positions represents the average reduction in uncertainty about one amino acid position that results from learning the amino acid of another position. The pairs of amino acid positions that tend to be involved in correlated mutations are expected to have higher mutual information values.

Definition 1 Entropy: The entropy of X is defined to be the average Shannon information of an outcome.

$$H(X) = \sum_{x \in Ax} P(x) \log \frac{1}{P(x)}$$

Example 1 Suppose we have 6 protein sequences in which the amino acid of positions 1, 2, 3 are the following.

position 1	D	D	D	D	D	D
position 2	Q	Q	Q	Q	Q	I
position 3	D	D	I	I	N	V

The entropy of position 1,2, and 3 are

$$\begin{aligned} H(X_1) &= 0.00, \\ H(X_2) &= 0.65, \\ H(X_3) &= 1.91, \end{aligned}$$

respectively.

Definition 2 Mutual Information: The mutual information between X and Y is defined to be the average reduction in uncertainty about x by knowing the value of y.

$$I(X;Y) = H(X) - H(X|Y)$$

$$H(X|Y) = \sum_{xy \in AxAy} P(x,y) \log \frac{1}{P(x|y)}$$

Example 2 Suppose we have 6 protein sequences in which the aminos acid of positions 1, 2, 3 are the following.

position 1	Q	Q	Q	Q	Q	Q
position 2	Q	Q	Q	I	I	V
position 3	D	D	D	N	N	K
position 4	Q	I	Q	Q	I	Q

The Mutual Information of among position 1, 2, 3 include

$$\begin{aligned} I(X_1;X_2) &= 0.00, \\ I(X_2;X_3) &= 1.45, \\ I(X_3;X_4) &= 0.12. \end{aligned}$$

3. Correlated Mutations Found by the Analysis

X	Y	$I(X;Y)$
144	156	1.0687064407
156	276	1.0606584462
62	158	1.0568431741
135	262	1.0462450293
156	158	0.9904144166
196	276	0.9548333157
62	276	0.9544171044
62	156	0.9440218619
158	276	0.9313234350
156	196	0.9269716745

Table 1: The top 10 pairs of amino acid positions in HA1 subunit that have high mutual information values.

2183 amino acid sequences of HA1s of human influenza viruses have been collected from NCBI Influenza virus resources. Mutual information for every pair of two amino acid positions in the HA1 subunit is calculated. Table 1 shows the top 10 pairs that have high mutual information values. The analysis founds a number of the correlated mutations in amino acid positions of the HA1, and three of them are shown in Figure 1.

The 3D structure model in Figure 2 depicts the location of amino acid positions that appear in the pairs in Table 1.

The correlated mutation between amino acid position 163 and 248, which have mutual information 0.61 during 1968 to 1989, resulted

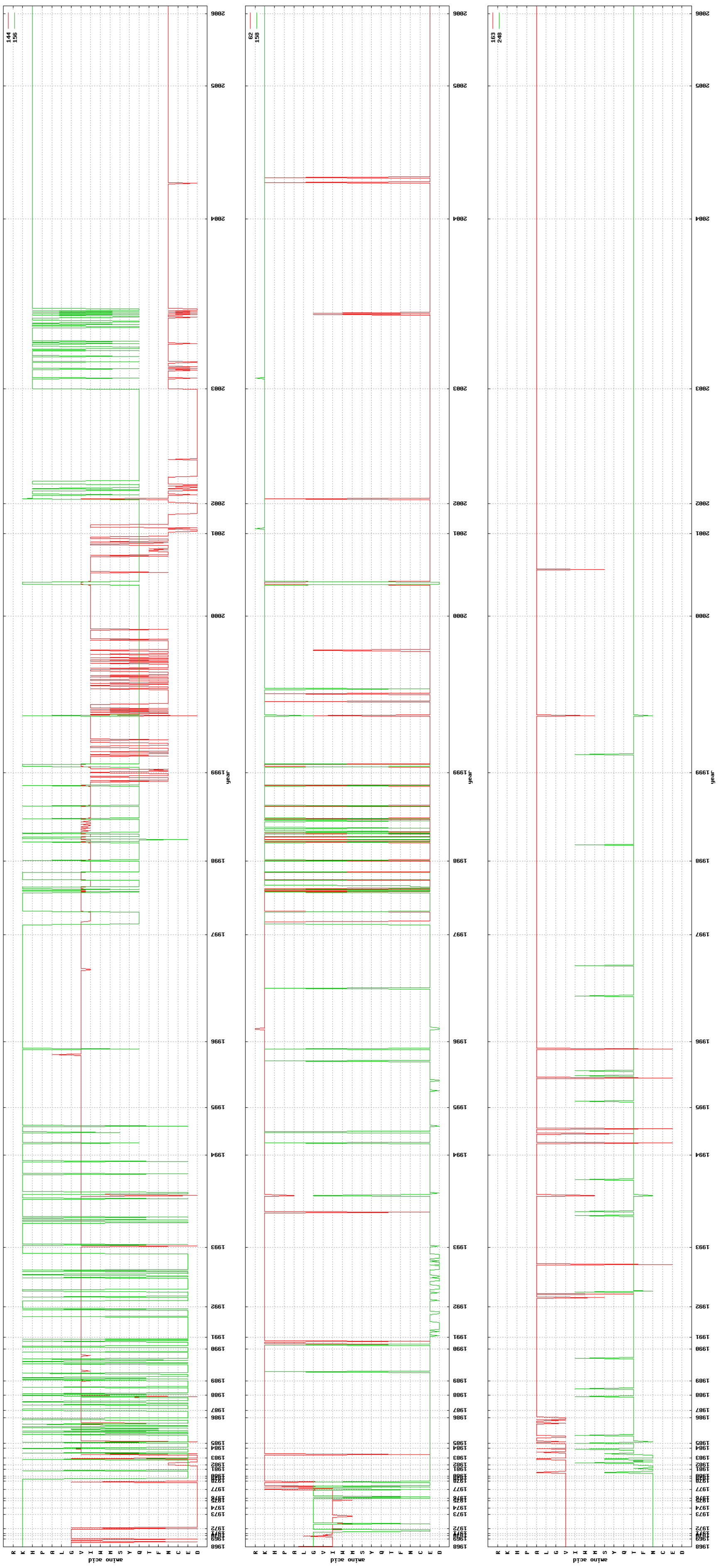


Figure 1: Correlated mutations found by the analysis. Series of amino acid changes in the positions 144-156(left), 62-158(middle), and 163-248(right) are shown.

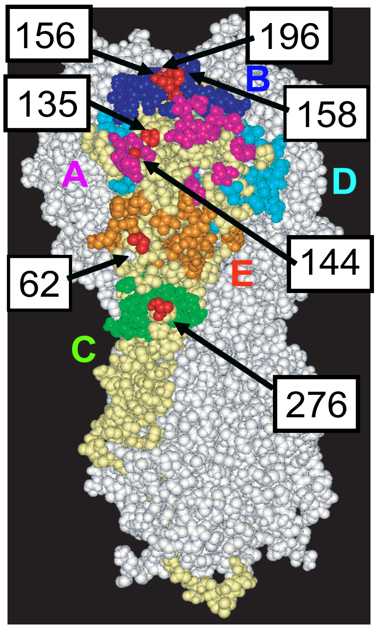


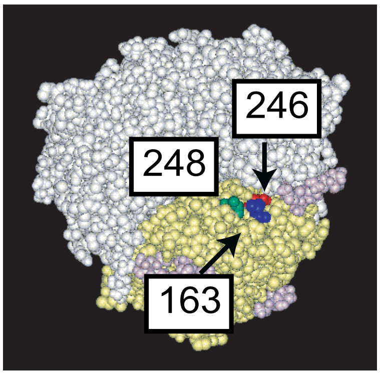
Figure 2: The 3D structure model of an HA molecule. The locations of amino acid positions that are involved in the correlated mutations in Table 1 are highlighted. The commonly used definitions of the five antigenic sites (A-E) are shown in different colors.

from the addition of an N-linked glycosylation site. The acquisition of new oligosaccharide chains is known to be significant for the antigenic drift of HA. Figure 3(a)(b) shows that the oligosaccharide chain that is added at position 246 produced a steric hindrance which likely required the substitution of valine at position 163 with the smaller amino acid alanine.

4. Retrospective Net Charge Analysis of Hemagglutinins

To study the change of the net charge of HAs, which may be associated with anitigenic properties, we have also analyzed the HA1 subunits of influenza viruses isolated from human, swine, and avian hosts.

The net charge at neutral pH are calculated, assuming that glutamic and aspartic acid have -1 charge, and arginine and lysin have +1 charge at this pH. Figure 4 shows that the net charge of HA1s increased linearly from 1968 to 1984, while avian virus HAs have retained their net charge since 1963. This result suggests that the mutations that increased the charge of HA molecule were required for the adaptation of avian virus to human population. After 1984, the increase in the net charge in human virus HA has been saturated. This level of the net charge might be an upper limit for H3 HA to be functional. Correlated mutations that exchange the charges among two or more different amino acid positions were frequently found after 1984. These co-mutations include E82K/K83E(1989), N145K/G135E(1991), E135K&E156K/S133D&K145N&R189S-(1993), D124G/G172D&R197Q(1995), N145K/K135T(1996), E158K&N276K/-K62E&K156Q(1998), D271N/K92T(2000), T92K/N271D(2001), E83K&D144N&-W222R/R50G&N126D&G225D(2003), and D126N/K145N(2004). It suggests that these co-mutations contributed to counter balancing effect to keep the net charge of the HA.



year	163	246	248
1968	V	N	N
1982	V	N	T
1985	A	N	T
1989	A	N	T

glycosylation site

Figure 3: The correlated mutation among amino acids at 163, 246, and 248.(a) The 3D structure model of an HA. The locations of the amino acids at 163, 246, and 248 are indicated. (b)The amino acids in the glycosylation site during 1968 to 1989.

For further investigation, we performed homology modeling to build 3D structure models of HAs of antigenically different influenza viruses (Figure 5). We found that antigenic drifts with frequent substitutions

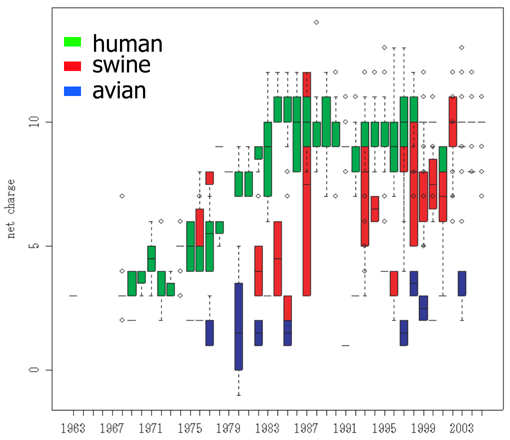


Figure 4: The net charges of H3 HA1 of influenza viruses isolated from human, swine, and avian.

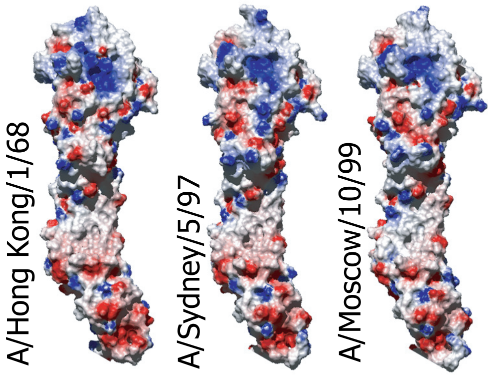


Figure 5: 3D structure model of three antigenically different HAs of human H3N2 viruses. The electrostatic potential on the surface of HAs are shown.

of charged amino acids resulted in the change of the charge distribution on the HA surface.

5. Conclusion

A number of correlated mutations in HAs are found by analyzing large scale sequence data of influenza viruses. A retrospective net charge analysis shows the increase of the net charge in HAs of human H3N2 viruses. It might be required for avian influenza viruses to increase the positive charge in order to adapt to human population. There seems to be an upper bound on possible net charge of hemagglutinin. Charge compensations have been made by co-mutations since 1984 and these co-mutations might contribute to keep the net charge constant in HA molecules.

References

[1] Atchley, W.R., Wollenberg, K.R., Fitch, W.M., Terhalle, W., Dress, A.W.: Correlations Among Amino Acid Sites in bHLH Protein Domains: An Information Theoretic Analysis, Molecular Biology and Evolution 17,pp.164-178 (2000)

[2] Bush, R.M., Bender ,C.A., Subbarao, K., Cox, N.J., Fitch, W.M.: Predicting the evolution of human influenza A, Science, 286(5446),pp.1921-1925.(1999)

[3] Ghedin, E., Sengamalay, N.A., Shumway, M., Zaborsky, J., Feldblyum, T., Subbu, V., Spiro, D.J., Sitz, J., Koo, H., Bolotov, P., Dernovoy, D., Tatusova, T., Bao, Y., St, George, K., Taylor, J., Lipman, D.J., Fraser, C.M., Taubenberger, J.K., Salzberg, S.L.: Large-scale sequencing of human influenza reveals the dynamic nature of viral genome evolution, Nature, 437(7062),pp.1162-1166 (2005)

[4] Martin LC, Gloor GB, Dunn SD, Wahl LM. Using information theory to search for co-evolving residues in proteins. Bioinformatics, 21(22),pp.4116-4124, (2005)