Available online at www.sciencedirect.com
SCIENCE DIRECT®
Biochemical and Biophysical Research Communications 333 (2005) 70–78
BBRC
www.elsevier.com/locate/ybbrc

Timing of mutation in hemagglutinins from influenza A virus by means of unpredictable portion of amino-acid pair and fast Fourier transform

Guang Wu *, Shaomin Yan

Computational Mutation Project, DreamSciTech Consulting, 301, Building 12, Nanyou A-zone, Jiannan Road, Shenzhen, Guangdong Province CN-518054, China

Received 22 April 2005
Available online 26 May 2005

Abstract

In this study, we calculate the unpredictable portion of amino-acid pairs, which has been developed by us over the last several years, of 1201 hemagglutinins from influenza A viruses dated from 1918 to 2004 in order to compare them with respect to subtypes, species, and years. After noticing the fluctuations of unpredictable portion along the time course, we use the fast Fourier transform to find the mutation periodicity of hemagglutinins. Then we estimate our position at the current cycle of hemagglutinin evolutionary process to determine how many years remain before the next outbreak of influenza and bird flu. Finally, we use the trend line and channel to outlook the hemagglutinins for the next half a century. As our study covers almost all the full-length amino-acid sequences of hemagglutinins from various influenza A viruses, the conclusion will be valid for years until the number of hemagglutinins in protein databank will be significantly increased.

Keywords: Fast Fourier transform; Hemagglutinin; Influenza A virus; Mutation; Periodicity

Due to the unpredictable mutation in influenza A viruses, time and again we are warned of the possible occurrence of pandemics of influenza and the possible outbreak of bird flu. Although the future occurrence of a pandemic, against which the humans have little or no immunity, is unpredictable as to timing or identity of the prospective agent, we know that the influenza A subtypes are frequently different in pandemics and epidemics, for example, the pandemics of 1918, 1957, and 1968 are attributed to H1N1, H2N2, and H3 subtypes [1–3]. This at least suggests that the mutations in influenza A viruses might not occur continually, or the mutations with devastating effects to humans might not occur continually, although we would expect to see the outbreak of influenza or bird flu every year with different intensity.

If, for example, we can arrange the hemagglutinins from influenza A viruses along the time course as each hemagglutinin presents a sample during its evolutionary process, we may find some patterns which help us predict the possible occurrence of mutation in future. However, such an analysis requires us to use a single value to represent a hemagglutinin. Over the last several years, we have developed two computational mutation methods to analyze the protein primary structure [4], of which our first method classifies a protein as randomly predictable and unpredictable portions of amino-acid pairs. We can therefore use either predictable portion or unpredictable one to quantitatively present a protein, then we can use it as a measure to compare different proteins among a protein family [5–7], compare the mutation effect on proteins, for instance, the effect of mutations is generally to reduce the unpredictable portion of amino-acid pairs [8–15], and compare the

* Corresponding author. Fax: +86 755 2528 8156.
E-mail address: hongguanglishibahao@yahoo.com (G. Wu).

0006-291X/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2005.05.094

mutation trend in proteins, for example, the larger is the unpredictable portion, the more sensitive to mutations is a protein [8–15]. Since this measure is sensitive to mutation, we can also use it as a measure to present the mutation process along the time course.

Currently, more than 1000 hemagglutinins of influenza A viruses dated from 1918 to 2004 are available in the protein databank, which offers the possibility of using our computational method to observe the evolutionary process of hemagglutinins and study the pattern of hemagglutinin along the time course.

Materials and methods

The amino-acid sequences of 1201 hemagglutinins from influenza A viruses dated from 1918 to 2004 are obtained from the Medline Protein databank and analyzed according to their subtypes, species, and years (Table 1). The detailed calculations and its rationales have already been published in a number of our previous studies [4], which are based upon the simple permutation principle [16]. As we know that an amino-acid pair in a protein is composed of any 20 kinds of amino acids, so theoretically there are 400 possible types of amino-acid pairs. In terms of amino-acid pairs, distinguishing protein is different either in the numbers of possible types of amino-acid pairs or in the frequency of each type, or both.

Table 1
Hemagglutinins of influenza A viruses analyzed in this study

Subtypes	Species	Number	Years
H0	Human	2	1933–1934
H1	Avian	14	1976–1999
	Human	58	1918–2001
	Swine	64	1930–2002
H2	Avian	19	1961–1998
	Human	21	1957–1968
H3	Avian	43	1974–2003
	Human	85	1968–2002
	Equine	47	1963–1994
	Seal	2	1992
	Swine	48	1977–2002
H4	Avian	33	1956–2000
	Seal	3	1982
	Swine	2	1999
H5	Avian	244	1959–2004
	Human	24	1997–2004
	Others	8	2001–2004
H6	Avian	75	1972–2004
H7	Avian	145	1934–2004
	Human	9	1934–2003
	Equine	22	1956–1977
	Seal	1	1980
H8	Avian	6	1968–1994
H9	Avian	143	1966–2003
	Human	7	1998–1999
	Swine	23	1998–2003
H10	Avian	3	1949
	Minks	3	1984
H11	Avian	3	1956
H12	Avian	2	1976
H13	Avian	8	1977–1984
	Whale	3	1984
H14	Avian	3	1982
H15	Avian	2	1979, 1983
Unidentified	Avian	5	1963–2000
	Equine	2	1971, 1976
	Human	14	1994–1998
	Swine	5	1986–1998
Total		1201	1918–2004

An example for calculations is as follow: a hemagglutinin from chicken H5N1 virus in 2004 (Accession No. AAS87580) is composed of 448 amino acids, thus there are 447 amino-acid pairs. Of 400 possible types, 152 are absent and 248 are present: 115 types appear once, 88 twice, 25 three, 19 four, and 1 five times, respectively.

Randomly predictable present type of amino-acid pair with predictable frequency. There are 41 asparagines (N) and 18 phenylalanines (F) in AAS87580 hemagglutinin, and the frequency of amino-acid pair "NF" is 2 ($41/448 \times 18/447 \times 447 = 1.647$), that is, the "NF" would appear twice in AAS87580 hemagglutinin. Actually we find 2 "NF"s, so the actual frequency of "NF" is 2. In this case, both the presence of the type "NF" and its frequency are predictable, and the difference between its actual and predicted values is 0.

Randomly predictable present type of amino-acid pair with unpredictable frequency. There are 31 glutamic acids (E) in AAS87580 hemagglutinin and the frequency of random presence of amino-acid pair "FE" is 1 ($18/448 \times 31/447 \times 447 = 1.246$), i.e., there would be one "FE" in AAS87580 hemagglutinin. But actually the "FE" appears four times, so the presence of "FE" is predictable, but its frequency is unpredictable, and the difference between its actual and predicted values is 3. This is also the case that the actual frequency of "FE" is larger than its predicted one. Another case is that the actual frequency is smaller than the predicted frequency, for example, the predicted frequency of "NN" is 4 ($41/448 \times 40/447 \times 447 = 3.661$), while its actual frequency is two.

Randomly predictable absent amino-acid pairs. There are 10 methionines (M) in AAS87580 hemagglutinin and the frequency of random presence of "MM" is 0 ($10/448 \times 9/447 \times 447 = 0.201$), i.e., "MM" would not appear in AAS87580 hemagglutinin, which is true in real situation. Thus, the absence of "MM" is predictable.

Randomly unpredictable absent amino-acid pairs. There are 27 isoleucines (I) in AAS87580 hemagglutinin and the frequency of random presence of "EI" is 2 ($31/448 \times 27/447 \times 447 = 1.868$), i.e., there would be two "EI" in AAS87580 hemagglutinin. However, there is no "EI" in AAS87580 hemagglutinin, therefore the absence of "EI" is unpredictable.

Predictable and unpredictable portions of amino-acid pairs. After the above calculations, the amino-acid pairs in a protein can be classified into predictable and unpredictable portions with respect to types and frequencies, and the sum of both predictable and unpredictable portions is 100%. Of the amino-acid pairs in AAS87580 hemagglutinin, the unpredictable portion is 60% and 68% regarding types and frequencies, respectively.

Type mutation and frequency mutation. As there are 400 types of theoretically possible amino-acid pairs and we use the 100% to classify them as predicable and unpredictable types, 0.25% thus represents one of the 400 types, so 0.25% change indicates that one of the 400 types mutates to an unpredictable type from a predictable one or vice versa. This is the type mutation. However, the situation related to the frequency is more complicated because the length of an amino-acid sequence differs in different hemagglutinins. Based on the data studied herein, about 0.18% change can be regarded as a modification in an amino-acid pair, since the average number of amino-acid pairs of 1201 hemagglutinins is 552. This is the frequency mutation.

Determination of periodicity with the fast Fourier transform. At the beginning of this study, we had no intention to use the fast Fourier transform (FFT) to determine the periodicity in hemagglutinins along the time course. After noticing the fluctuations of unpredictable portion of amino-acid pairs along the time course, we used the fast Fourier transform to analyze the evolutionary process of hemagglutinins along the time course because one of the important applications of Fourier

analysis is to determine the periodicities in a chaotic fluctuating dataset. The MatLab (MathWorks, 1984–2001) is used to perform the fast Fourier transform.

Statistics. With respect to the actual and predicted values in a single protein, the statistical inference is carried out as follows. Generally, each of 20 kinds of amino acids has the chance of 1/20 ($p = 0.05$) to repeat once, and a type of amino-acid pair has the chance of 1/400 ($p = 0.0025$) to repeat once. In case of AAS87580 hemagglutinin, there are 41 "N"s, the most abundant amino acid, and 9 "W"s, the least amino acid. If the first amino acid is "N," then the chance of the second amino acid being "N" is 40/447 ($p = 0.089 > 0.05$). If the first amino acid is "W," then the chance of the second amino acid being "W" is 8/447 ($p = 0.018 < 0.05$). Accordingly, the chance of the first amino-acid pair being "NN" is $41/448 \times 40/447$ ($p = 0.008 < 0.01$), and the chance of the second amino-acid pair being "NN" is $39/446 \times 38/445$ ($p = 0.007 < 0.01$).

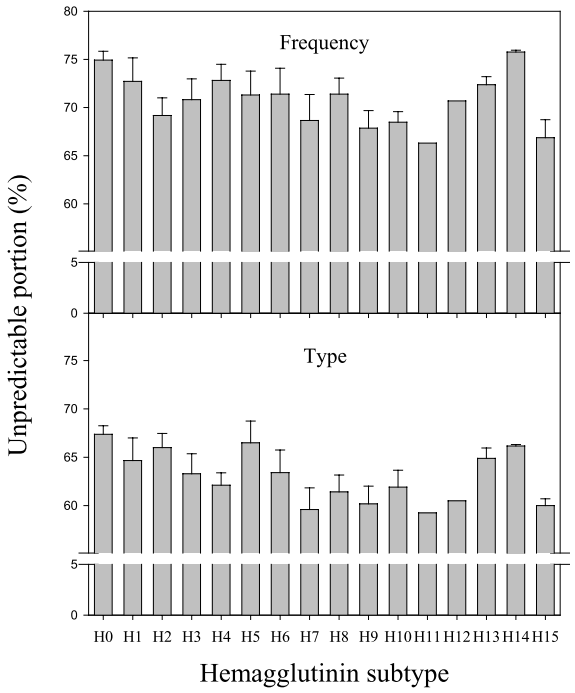

Fig. 1. Unpredictable portion of amino-acid pairs in different subtype hemagglutinins from influenza A viruses. The data are presented as means ± SD.

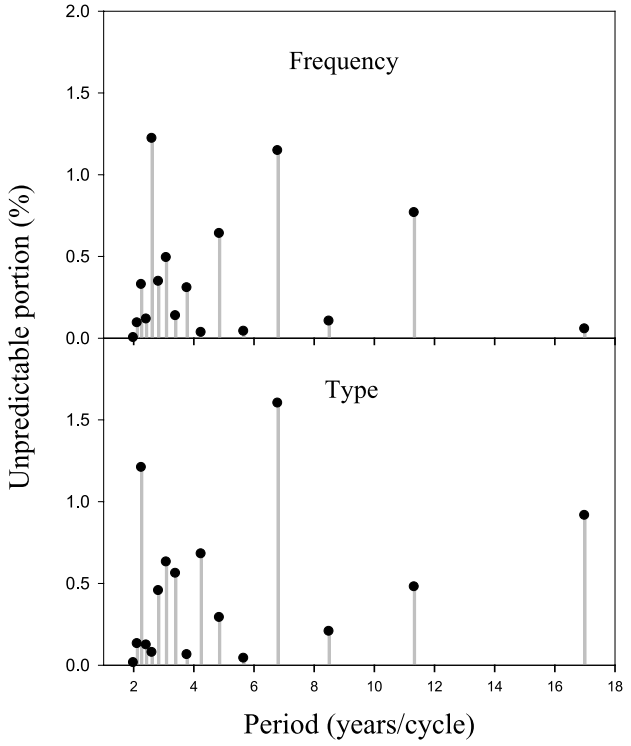

Fig. 3. Periodicity of fluctuated unpredictable portion in influenza A virus hemagglutinins over the last 35 years.

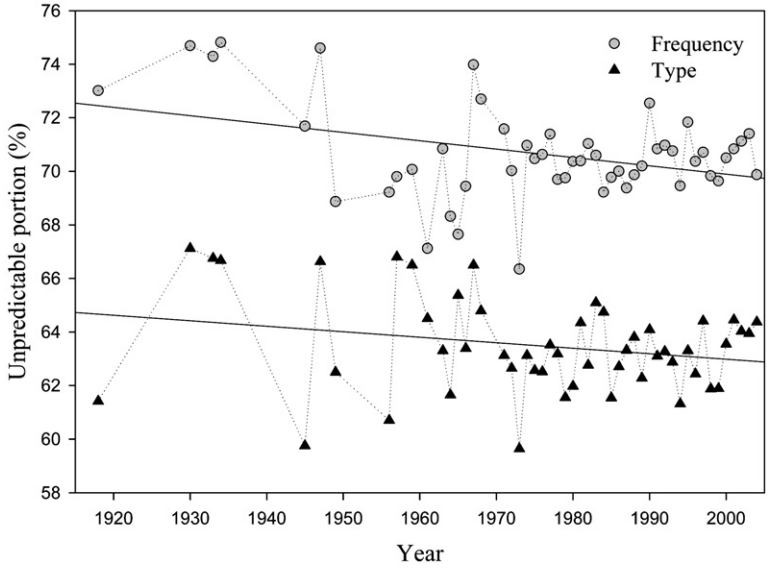

Fig. 2. Time trend of unpredictable portion of amino-acid pairs in 1201 influenza A virus hemagglutinins along the time course.

For the least amino acids "W," the chance of the first amino-acid pair being "WW" is $9/448 \times 8/447$ ($p = 0.0004 < 0.001$), and the chance of the second amino-acid pair being "WW" is $7/446 \times 6/445$ ($p = 0.0002 < 0.001$). For that reason, the probability is less than 0.05 if the difference between the actual and predicted values is greater than or equal to one.

With respect to the comparisons among proteins, the statistical inference is conducted as follows. All the data are examined by the Kolmogorov–Smirnov test to determine their distribution property. For the normal distribution, the data are presented as means ± SD. For the non-normal distribution, the data are presented as median with interquartile range. The outlier is detected according to Healy's method [17]. The one-way ANOVA and the Friedman ANOVA rank tests are used for parametric and non-parametric tests, respectively, followed by comparison tests. The SigmaStat for Windows (SPSS, 1992–2003) is used to operate all the statistical tests, and the $p < 0.05$ is considered statistically significant.

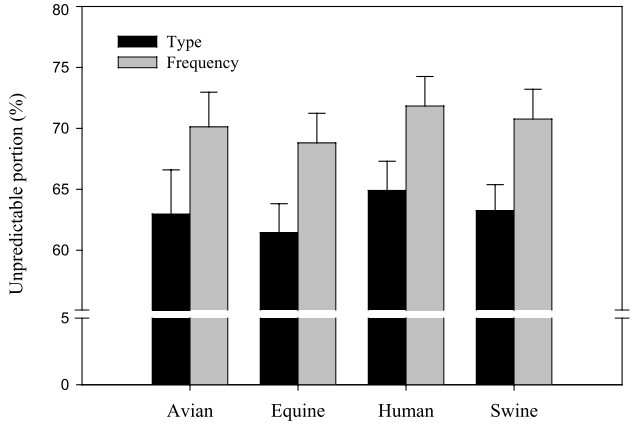

Fig. 4. Unpredictable portion of amino-acid pairs in hemagglutinins of influenza A viruses from different species. The data are presented as means ± SD. Statistically significant difference can be found between any two groups at $p < 0.001$ level, except for the data between avian and swine.

Results and discussion

After the calculations, the amino-acid pairs in a hemagglutinin are classified into randomly predictable and unpredictable portions, of which the sum is 100%, so we only use the unpredictable portion as the measure in this study. Since the calculations are based on almost all the full-length amino-acid sequences of hemagglutinins from various influenza A viruses, the conclusion obtained from these calculations will be valid for years until the number of hemagglutinins in the protein databank will be significantly increased.

Subtype difference

Fig. 1 shows the unpredictable portions of amino-acid pairs in different subtype hemagglutinins from influenza A viruses. Our previous studies show that the larger the unpredictable portion is, the more sensitive to mutation the protein is [8–15]. Although each hemagglutinin subtype has different sensitivity to mutation, a slight trend can be seen in Fig. 1, for example, the unpredictable type decreases from H0 to H4, from H5 to H7, and increases from H11 to H14. Actually we are particularly interested in H0, H1, H2, H3, H5, H7, and H9 because they are linked to previous human infection. An intriguing feature is that the unpredictable types in H5 hemagglutinins are statistically larger than others, except for H0 and H14. Suggestive is that the H5 hemagglutinins are more sensitive to mutation [8–15].

Time trend of all 1201 hemagglutinins

Fig. 2 displays the unpredictable type and frequency of amino-acid pairs of 1201 hemagglutinins from various

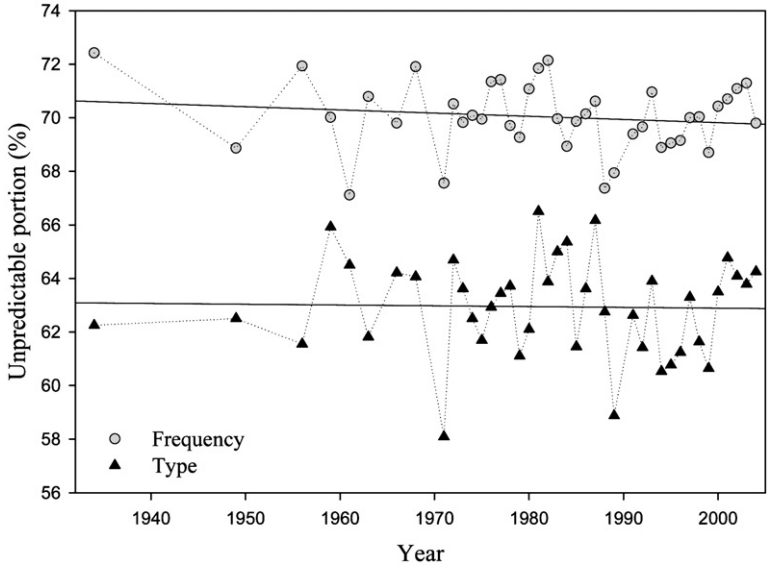

Fig. 5. Time trend of unpredictable portion of amino-acid pairs in avian hemagglutinins from influenza A viruses along the time course.

influenza A viruses over almost a century. Each symbol presents the mean value of unpredictable portion of all hemagglutinins in the given year. In these hemagglutinins, the average values for unpredictable frequency and type are $70.64\% \pm 2.82\%$ (mean $\pm$ SD) and $63.28\% \pm 3.31\%$. Naturally we know that the number of mutations changes from time to time as indicated by dotted lines in Fig. 2 although the efficiency of dotted lines is from 1971 as no data are available in missing years.

In Fig. 2, two regressed lines indicate the time trend, that is, the unpredictable portion of hemagglutinins is decreasing, and the structure of hemagglutinin is becoming more stable and less sensitive to mutation along the

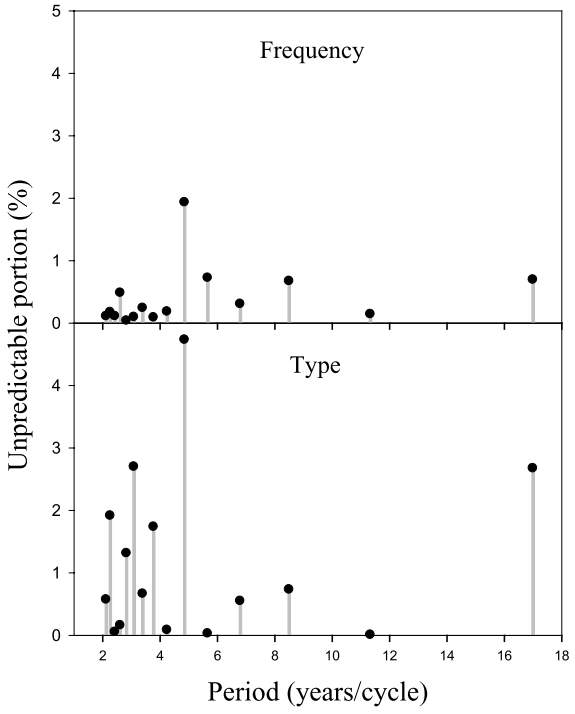

Fig. 6. Periodicity of fluctuated unpredictable portion in avian hemagglutinins over the last 35 years.

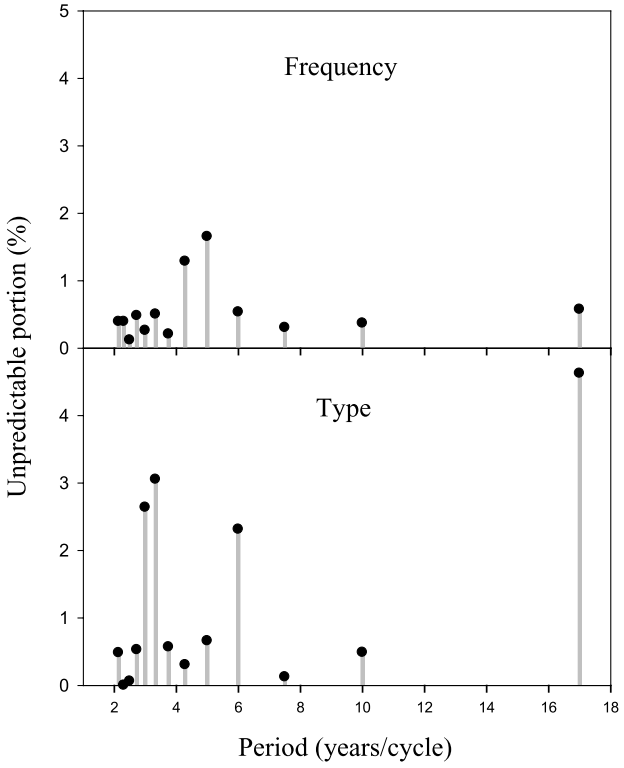

Fig. 8. Periodicity of fluctuated unpredictable portion in human hemagglutinins over the last 35 years.

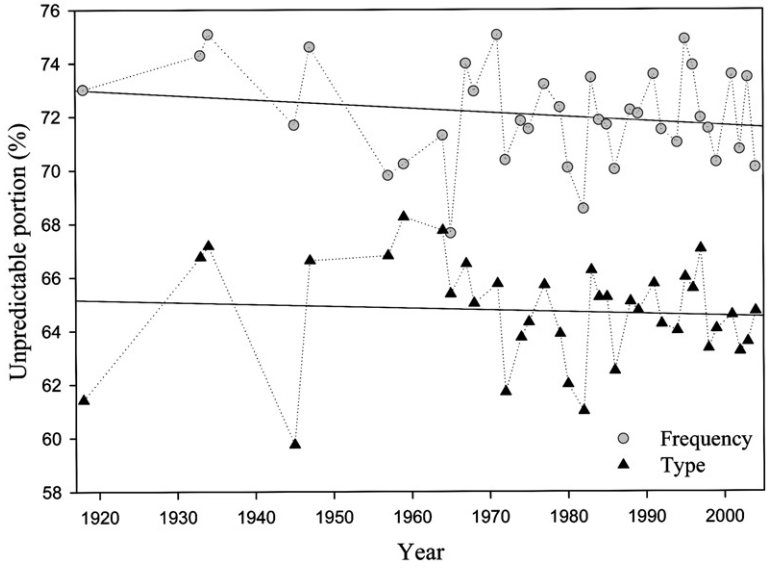

Fig. 7. Time trend of unpredictable portion of amino-acid pairs in human hemagglutinins from influenza A viruses along the time course.

time. However, any spike in fluctuated unpredictable portion suggests that the hemagglutinin may become more sensitive to mutation at some time point leading to the outbreak of influenza and bird flu. To time the point at which the hemagglutinin becomes more sensitive to mutation, we need at first to find the possible periodicity in this fluctuated unpredictable portion from fluctuating dotted lines in Fig. 2.

Periodicity over the last 35 years

Due to data limitation, the unpredictable portions over the last 35 years have been analyzed using the fast Fourier transform because the data have been obtained without any missing year. As can be seen in Fig. 3, the fluctuating dotted lines over the last 35 years in Fig. 2 have many periodicities. For instance, the first prominent peak at both type and frequency is a cycle of about 2.5 years with heights of 1.21% and 1.22%, respectively. Under Materials and methods, we define the type and frequency mutation, in this case the type mutation is 5 (1.21%/0.25% = 4.84) and the frequency mutation 7 (1.22%/0.18% = 6.78). The interpretation of this cycle is that the fluctuation in the unpredictable portion has a periodicity of 2.5 years, which may lead to 5 type mutations and 7 frequency mutations.

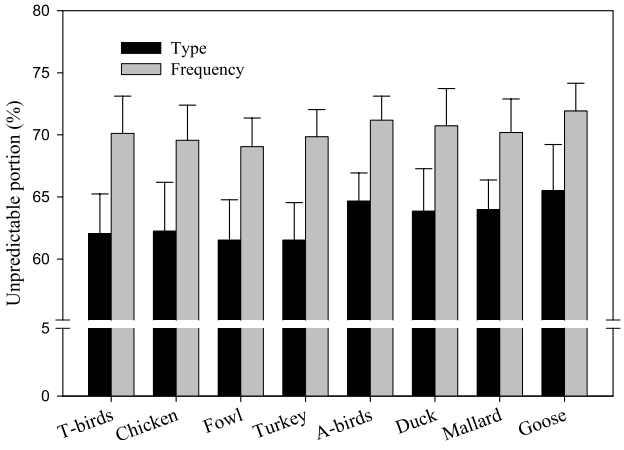

Fig. 9. Unpredictable portion of amino-acid pairs in influenza A virus hemagglutinins of terraneous (T) and aquatic (A) avian isolated from 1934 to 2004. The data are presented as means ± SD.

Species difference

Fig. 4 illustrates the species difference in the unpredictable portion of amino-acid pairs. It can be seen that the unpredictable portions are larger in human, smaller in equine, while intermediate in avian and swine. These large-scale calculations suggest that human hemagglutinins are more sensitive to mutations among these four species.

Time trend and periodicity of avian hemagglutinins

Fig. 5 demonstrates the time trend of unpredictable portion of amino-acid pairs in avian hemagglutinins from influenza A viruses along the time course. Avian have a large reservoir for influenza A viruses, from which new strains mutate frequently [18]. Although the data in avian cover a shorter range than the data in Fig. 3, a slight time trend can still be found in the unpredictable frequency, which is becoming less sensitive to mutations along the time course.

Fig. 6 exhibits the periodicity of mutation in avian hemagglutinins from influenza A viruses. The major peak occurs at 6.8 years at both type and frequency,

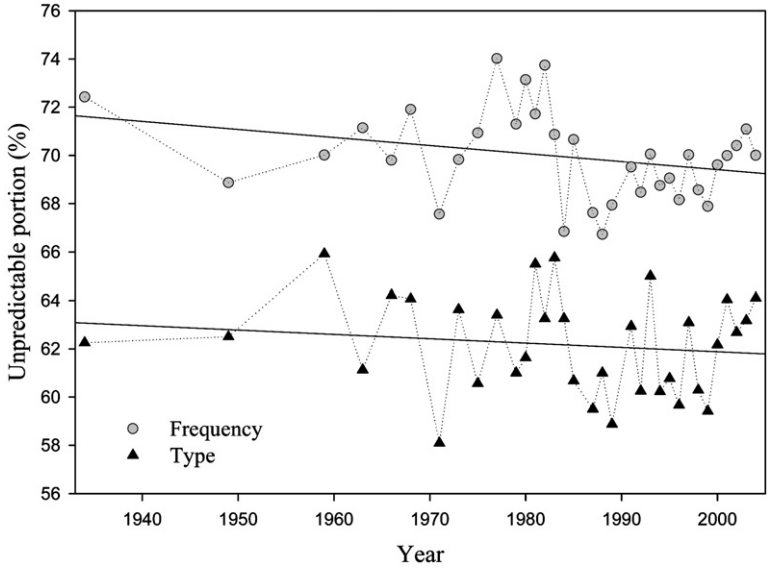

Fig. 10. Time trend of unpredictable portion of amino-acid pairs in hemagglutinins from terraneous avian influenza A viruses along the time course.

but the fluctuation is higher in type than in frequency. Therefore, there would be more type mutations than frequency mutations regarding this periodicity.

Time trend and periodicity of human hemagglutinins

The time trend in human hemagglutinins is more clear from 1918 to 2004 as indicated by two regressed lines (Fig. 7), therefore the human hemagglutinins are less and less sensitive to mutations. This could be a piece of reasoning why the pandemics are less frequent nowadays. Still we notice that the spike in type is narrowing, which means that we have been living in a period of relatively stable human hemagglutinin since 1997 (the last spike), which marked the H5N1 influenza virus spread directly from poultry to humans in Hong Kong and caused the deaths of 6 of 18 persons [19,20]. Four major periodicities can be found in human hemagglutinins after fast Fourier transform (Fig. 8).

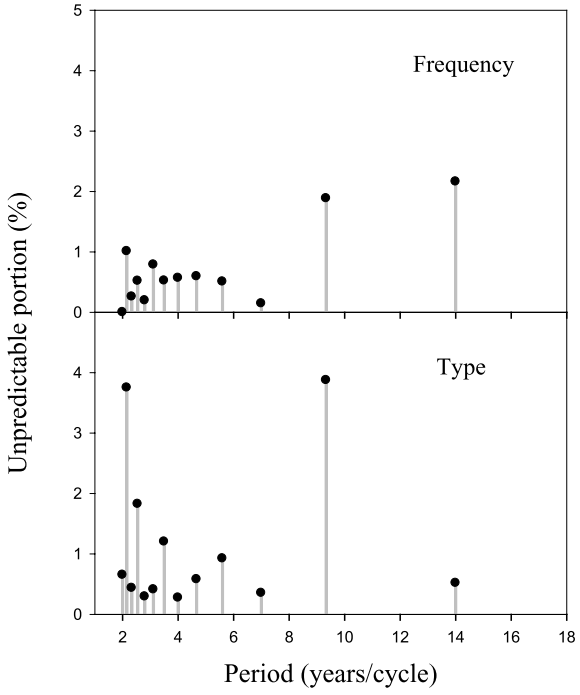

Fig. 11. Periodicity of fluctuated unpredictable portion in terraneous avian hemagglutinins over the last 35 years.

Difference between terraneous and aquatic avian

Furthermore, we estimate the difference in hemagglutinins between terraneous and aquatic avian (Fig. 9). In general, there are more unpredictable types in aquatic avian than in terraneous avian ($p < 0.001$). Fig. 10 demonstrates that the time trend in terraneous avian is similar to what we observe in Figs. 2, 5, and 7. The periodicity in terraneous avian hemagglutinins can be found in Fig. 11.

By clear contrast, the time trend in aquatic avian hemagglutinins is completely different from others (the regressed lines in Fig. 12). Their main periods are 2.0 and 4.3 in frequencies, 4.3 and 5.2 years in types (Fig. 13). In comparison, the terraneous avian has more type mutations with a shorter period whereas the aquatic avian shows more frequency mutations with a shorter period. These results suggest that the terraneous avian hemagglutinins mutate more frequently than the aquatic avian hemagglutinins.

Where are we at the current cycle of hemagglutinin evolutionary process

The unpredictable portion along the time course and its Fourier analysis clearly demonstrate that mutations

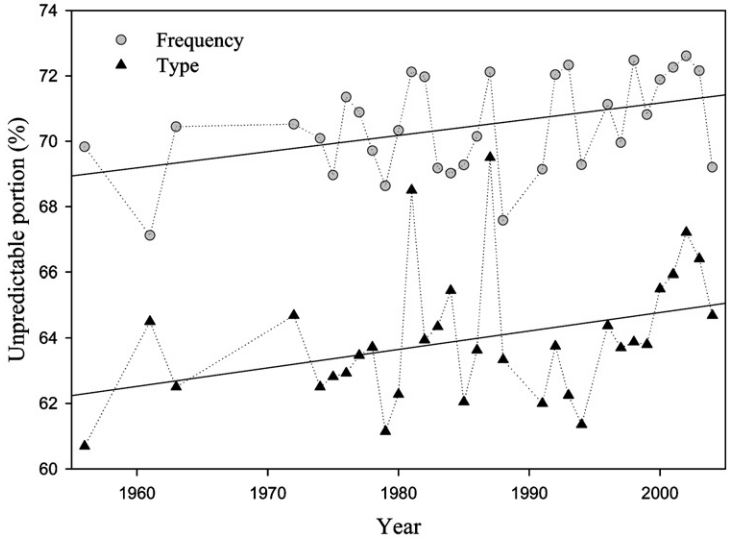

Fig. 12. Time trend of unpredictable portion of amino-acid pairs in hemagglutinins from aquatic avian influenza A viruses along the time course.

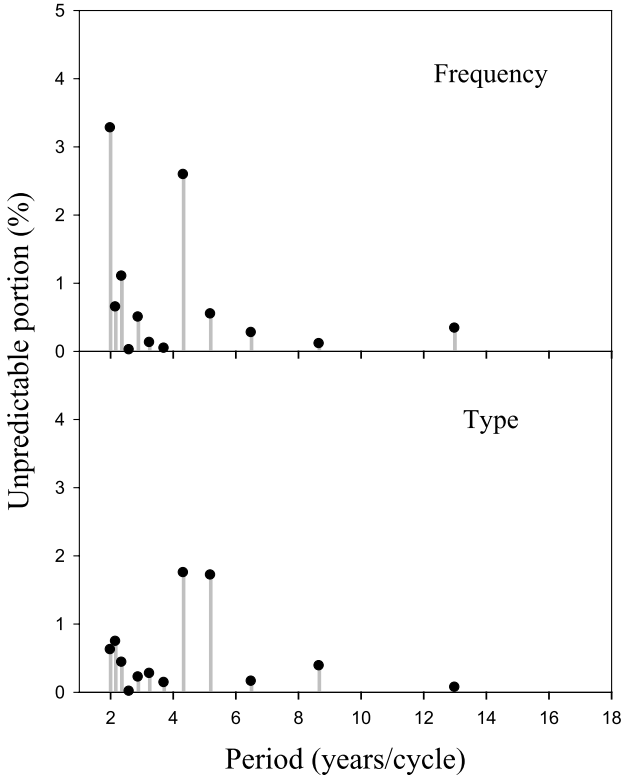

Fig. 13. Periodicity of fluctuated unpredictable portion in aquatic avian hemagglutinins over the last 35 years.

in hemagglutinins have periodicities. Based on these analyses, we can estimate our position at the current cycle of hemagglutinin evolutionary process to determine how many years remain before the next outbreak of influenza and bird flu.

Fig. 14 indicates that the unpredictable portion from 1999 has finished 6/7 of its cycle (pink dashed-dotted line). Compared with five historical lines, our current path is more likely to go downside for the last year of this cycle, then we still have another year of going downside in the next cycle before a spike. Fig. 15 shows the same situation using the data only from human hemagglutinins. As our previous studies clearly demonstrate that the smaller the unpredictable portion is, the less sensitive to mutation the protein is [5–15], therefore we would not expect to see the severe mutations in the next two years.

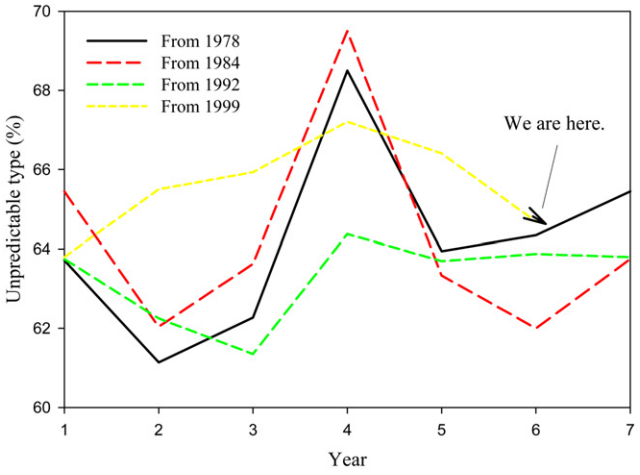

Fig. 15. Our position at the current cycle of human hemagglutinin evolutionary process.

Outlook of hemagglutinin for the next half a century

In the figures of time trend, we notice that the unpredictable portion and fluctuation are becoming smaller along the time course. The trend line and channel suggest the future of hemagglutinins (Fig. 16). In

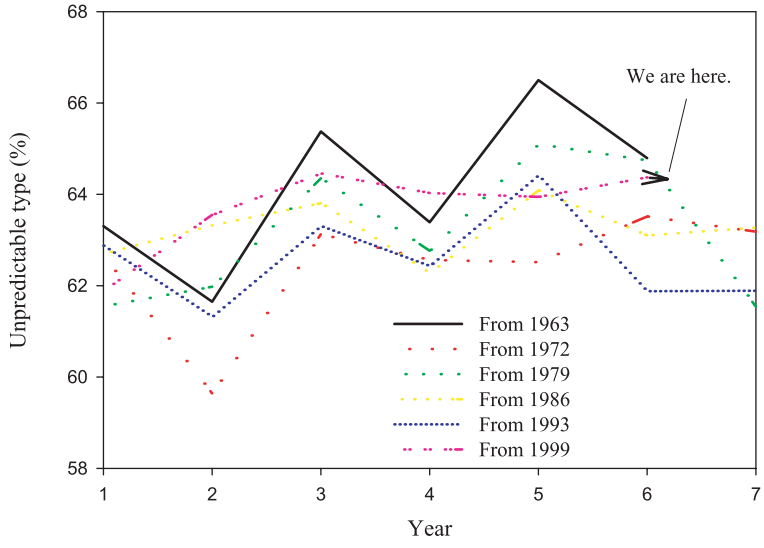

Fig. 14. Our position at the current cycle of hemagglutinin evolutionary process. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

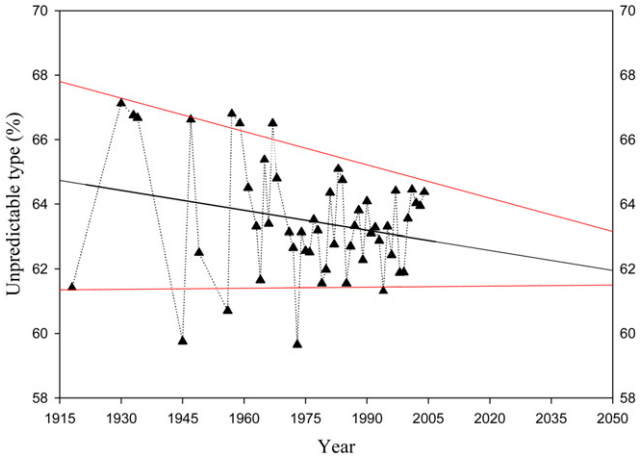

Fig. 16. Outlook of hemagglutinins for the next half a century.

general, we would expect that the unpredictable portion and fluctuation are continuing smaller along trend line and channel although some fluctuations may go out of the band. This outlook can be explained by our method. An unpredictable amino-acid pair should be deliberately constructed for a certain purpose because it requires more time and energy for construction, and a protein needs only absolutely necessary unpredictable amino-acid pairs. During the evolutionary process, nature is trying to minimize the unpredictable portion through mutations which may lead to new unpredictable amino-acid pairs, and the new mutations introduced result in the fluctuations as shown in this study.

Note added in proof

With the time going on, the sentence "We are here." in Figs. 14 and 15 and related text in Results and discussion are referenced to 2004.

References

[1] J.S. Oxford, Influenza A pandemics of the 20th century with special reference to 1918: virology, pathology and epidemiology, Rev. Med. Virol. 10 (2000) 119–133.
[2] R.J. Connor, Y. Kawaoka, R.G. Webster, J.C. Paulson, Receptor specificity in human, avian, and equine H2 and H3 influenza virus isolates, Virology 205 (1994) 17–23.
[3] R.G. Webster, W.J. Bean, O.T. Gorman, T.M. Chambers, Y. Kawaoka, Evolution and ecology of influenza A viruses, Microbiol. Rev. 56 (1992) 152–179.
[4] G. Wu, S.M. Yan, Randomness in the primary structure of protein: methods and implications, Mol. Biol. Today 3 (2002) 55–69.
[5] G. Wu, S. Yan, Fate of 130 hemagglutinins from different influenza A viruses, Biochem. Biophys. Res. Commun. 317 (2004) 917–924.
[6] G. Wu, S. Yan, Reasoning of spike glycoproteins being more vulnerable to mutations among 158 coronavirus proteins from different species, J. Mol. Model. 11 (2005) 8–16.
[7] G. Wu, S. Yan, Prediction of amino acid pairs sensitive to mutations in the spike protein from SARS related coronavirus, Peptides 24 (2003) 1837–1845.
[8] G. Wu, S.M. Yan, Estimation of amino acid pairs sensitive to variants in human phenylalanine hydroxylase protein by means of a random approach, Peptides 23 (2002) 2085–2090.
[9] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human low-density lipoprotein receptor precursor by means of a random approach, J. Biochem. Mol. Biol. Biophys. 6 (2002) 401–406.
[10] G. Wu, S.M. Yan, Determination of amino acid pairs in human haemoglobulin α-chain sensitive to variants by means of a random approach, Comp. Clin. Pathol. 12 (2003) 21–25.
[11] G. Wu, S. Yan, Analysis of amino acid pairs sensitive to variants in human collagen α 5 (IV) chain precursor by means of a random approach, Peptides 24 (2003) 347–352.
[12] G. Wu, S. Yan, Determination of amino acid pairs in human p53 protein sensitive to mutations/variants by means of a random approach, J. Mol. Model. 9 (2003) 337–341.
[13] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human coagulation factor IX precursor by means of a random approach, J. Biomed. Sci. 10 (2003) 451–454.
[14] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human copper-transporting ATPase 2, Biochem. Biophys. Res. Commun. 319 (2004) 27–31.
[15] G. Wu, S. Yan, Determination of amino acid pairs sensitive to variants in human Bruton's tyrosine kinase by means of a random approach, Mol. Simul. 29 (2003) 249–254.
[16] W. Feller, An Introduction to Probability Theory and Its Applications, third ed., John Wiley and Sons, New York, 1968, vol. I, pp. 38–42.
[17] M.J.R. Healy, Outliers in clinical chemistry quality control schemes, Clin. Chem. 25 (1979) 675–677.
[18] S.J. Baigent, J.W. McCauley, Influenza type A in humans, mammals and birds: determinants of virus virulence, host-range and interspecies transmission, Bioessays 25 (2003) 657–671.
[19] K.Y. Yuen, P.K.S. Chan, M. Peiris, D.N.C. Tsang, T.L. Que, K.F. Shortridge, P.T. Cheung, W.K. To, E.T.F. Ho, R. Sung, A.F.B. Cheng, members of the H5N1 Study Group, Clinical features and rapid viral diagnosis of human disease associated with avian influenza A H5N1 virus, Lancet 351 (1998) 467–471.
[20] E.C.J. Claas, A.D.M.E. Osterhaus, R. Van Beck, J.C. de Jong, G.F. Rimmelzwaan, D.A. Senne, S. Krauss, K.F. Shortridge, R.G. Webster, Human influenza A (H5N1) virus related to a highly pathogenic avian influenza virus, Lancet 251 (1998) 472–477.