JOURNAL OF PURE AND APPLIED MICROBIOLOGY, June 2013. Vol. 7(2), p. 1251-1254

Forecasting Influenza Epidemic based on Mutation Probability of HA/NA Protein

Jie Gao*, Ling Zhang and Peixuan Jin

School of Science, Jiangnan University, Wuxi - 214 122, China.

(Received: 30 March 2013; accepted: 22 May 2013)

The HA/NA protein sequences of the influenza virus from 1914 to 2012 are studied and the early-warning signal indicator values (the differences between predictable and unpredictable portions) for influenza pandemic outbreak are worked out through the calculation of the actual and predicted frequencies of amino-acid pairs. It is concluded that the differences between predictable and unpredictable portions for the pandemic years are significantly smaller than those for the adjacent years. Furthermore, it is also verified that the mutations of influenza virus HA/NA protein will cause influenza pandemic.

Key words: Influenza epidemic, Mutation probability, HA/NA protein sequence.

Influenza virus, a member of the Orthomyxoviridae family, is an enveloped RNA virus with a genome in the form of eight segments, and moreover HA and NA are the major of membrane-spanning proteins of influenza virus.[1-2] Because of the particular structure of the virus genome and function of the viral proteins, antigenic shift and antigenic drift easily change which make the great variation of antigenicity. It brings immense difficulty to the prevention of the outbreak of influenza virus[3-5].

On the other hand, proteins are essential elements of life[6-7]. Many experts and scholars researched the protein. Yu *et al.,* proposed the new chaos game representation and correlation analysis of protein sequences based on the detailed HP model[8-9]. Zhu *et al.,* researched the structural analysis of protein sequences based on the quasi-amino acids code[10]. Gao *et al.,* put forward the CGR-walk model based on the CGR coordinates for the protein sequences and a long-memory ARFIMA(p,d,q) model was introduced into the protein sequence analysis.[11] Liu *et al*., proposed the method of forecasting bases for DNA sequences of influenza virus[12]. Ren *et al.,* used DNA sequences of influenza virus to study the large intrinsic information contained in the flu outbreak, thus it can provide pandemic early-warning signal[13]. These have enlightenment and guidance on my study.

Yan *et al*., found computational mutation based on the random principle of nature. It overcame the limitation of bioinformatics and computational biology, where the letters and measures are not subject to amino-acid position and composition, the sequence length, neighboring amino acids, etc. It used amino-acid pair predictability to quantify a whole protein, which provides living measures to quantitatively analyze protein. Protein mutations particularly narrowed the difference between actual and predicted frequencies in amino-acid pairs[14-16].

This paper analyzes the HA/NA protein sequences of the influenza virus from 1914 to 2012, and works out the early-warning signal indicator values through the calculation of the actual and predicted frequencies of amino-acid pairs. We

* To whom all correspondence should be addressed.
Tel: 0510-85913885; Fax: 0510-85910227;
E-mail: ezhun6669@sina.com

conclude that the differences between predictable and unpredictable portions for the pandemic years are significantly smaller than those for the adjacent years. So we can gain the abnormal information to establish a warning system prior to the outbreak of the flu pandemic. That is to say, we can construct the early warning signals for outbreak of influenza pandemic, and make full preparation prior to a pandemic.

MATERIALS AND METHODS

For scientificity and universality, and in view of the collection and integrity of the data, we select the HA/NA protein sequences of the influenza virus from 1914 to 2012 for which the homology is relatively high (data from the NCBI website at: www.ncbi.nlm.nih.gov/).

As we compare the bioinformatics with our reading, we also initially consider a protein as an unknown text. For analyzing an unknown text, we can use counted frequencies to compare with the natural frequencies of elements. We convert the protein sequence into a single number with the amino-acid pair calculated frequencies and actual frequencies.

For example, there are 155 amino acids in the CAA28414 HA protein sequence. The first and second amino acids can be counted as an amino-acid pair, the second and third as another amino-acid pair, the third and fourth, until the 154th and 155th, thus there are totally 154 amino-acid pairs. The CAA28414 HA protein sequence has 6 alanines (A) and 15 glutamic acids (E): we can predict the appearance of amino-acid pair AE, which must appear one time (6 /155x15 /154x 154= 0.58), and it indeed appears 1 time, thus AE appearance is predictable. By contrast, we can predict the appearance of amino-acid pair EE in this protein: it must appear once (15 /155x14 / 154x154=1.35); actually there is 3 EE, so the appearance of EE is unpredictable.

In this manner, all amino-acid pairs in the HA/NA protein sequences can be classified as predictable and unpredictable. For this CAA28414 HA protein sequence, its predictable and unpredictable potions are 16.62% and 83.38%. Their difference is 66.76%.

There is a mutation of the CAA28414 HA protein sequence at position 104, Which V mutates into D. Although this mutation is related to a single amino acid, its predictable and unpredictable portion becomes 18.71% and 81.29%.Their difference is 62.58%.

In this way, each protein has a unique number, either predictable or unpredictable portion, to distinguish itself from others. Moreover, protein mutations narrow the difference between predictable and unpredictable portions in amino-acid pairs.

RESULTS AND DISCUSSION

Through the above analysis and methods, we calculate predictable portions of the HA/NA protein sequences of the influenza virus from 1914 to 2012 (Table 1). We first analyze the HA protein sequences of the influenza virus. From table 1, a comparison of the HA sequences from 1914 to 2012 indicates that the difference between the predictable and unpredictable portions for the pandemic years are obviously smaller than those for the adjacent years (the numbers in bold in Table 1). We know that protein mutations narrow the difference between actual and predicted frequencies in amino-acid pairs in Section 2. So we also verify that influenza virus protein mutations cause influenza pandemic. Secondly we study the NA protein sequences using the same method as HA protein sequences, and find the result of NA sequences is strikingly similar to HA sequences. As we know, pandemics are the most dramatic presentation of influenza. There have been three times in the twentieth century: the 1918 H1N1 pandemic, the 1957 H2N2 pandemic, and the 1968 H3N2 pandemic. Especially 2009 influenza virus A/H1N1 pandemic, the World Health Organization raised its alert level to its sixth grade.

In view of above analysis, we can draw table 1 HA/NA sequences (Fig.1 and Fig.2), and observe that the predictable portion (or the difference between predictable and unpredictable portions) can be regarded as an early-warning indicator value of influenza pandemic. Obviously, the predictable portions of influenza pandemic years are bigger than those of the adjacent years, i.e. the difference between predictable and unpredictable portions for the pandemic years are significantly smaller than those for the adjacent years.

Table 1. HA/NA protein sequences and predictable portions

year	HA protein sequence	Predictable portions (%)	NA protein sequence	Predictable portions (%)
1914-1915	CAA28414 ABS54296	16.62 20.28	EHR60808 ADZ64539	24.00 15.34
1916-1917	CAA42701 AAM22278	17.60 20.00	EHR60810 EHR60811	23.47 21.35
1918-1919	AAD17218 AAO65769	23.21 25.16	AAD17218 EIF48153	23.21 29.71
1953-1954	AAY19498 ABO52280	18.59 17.35	ADT78910 ABO52283	23.84 23.24
1955-1956	CAA34904 ACV41548	19.33 18.47	EIO42164 ACZ45145	18.16 20.90
1957-1958	ADT79119 ADT78898	18.36 19.09	AEM53459 ABP49440	22.75 23.93
1959-1960	ABR75306 ADT78961	24.78 19.25	EKA55505 AAO46219	26.60 23.72
1961-1962	AEZ68721 ADT78887	20.11 19.07	EIA87509 EHU20909	15.76 15.26
1963-1964	ACV49589 AAA43090	19.61 19.07	ADX06985 AAO46228	24.44 23.93
1965-1966	ACS93986 AEZ68718	21.97 17.94	AEA04315 AEA04411	23.72 22.01
1967-1968	AEZ68715 AAA43247	17.94 21.57	ABO44060 AEA04346	23.29 17.27
1969-1970	AFE18703 AFG71909	24.84 18.58	ACD85388 EHW99905	24.09 14.92
1971-1972	AEM60147 ACS93984	19.68 19.20	AEM60150 ACS93999	20.95 23.62
1973-1974	ACS93982 ABD66296	18.05 19.14	AFG98899 ACS68364	22.44 22.39
1975-1976	ACF25478 AFJ13080	17.58 20.86	ACZ45105 AFN11833	22.01 23.29
1977-1978	ACD88505 ABB19011	18.89 20.18	ABI84569 AEK65243	22.55 25.54
1979-1980	ABB19551 ACD88516	18.23 20.71	ABB19680 AFJ13163	23.79 23.50
1981-1982	ABO52731 AFM72032	19.36 21.06	AFJ12896 AFN11829	22.01 22.86
1983-1984	ABB88132 ABB20444	22.50 22.18	ABB88133 ACL79914	23.16 19.64
1985-1986	ABB17681 ACI41038	19.79 17.23	AFJ12962 ABB17739	22.16 23.67
1987-1988	ACF25679 ABI84816	19.75 17.58	AET74772 ABJ09132	23.09 23.52
1989-1990	ABB87281 ACI26604	17.23 20.18	AEB40996 AET74377	21.75 23.94
1991-1992	ACE78941 AFO64835	18.12 18.05	AFG99946 ABW97494	22.86 22.80
1993-1994	ADY70686 AET75956	18.76 20.53	AET75442 ACE81885	22.81 21.54
1995-1996	ACI89788 ABB19382	21.30 18.05	ACF25055 AFJ74451	23.25 23.50
1997-1998	AFJ78298 AET76470	17.55 17.83	AFJ78301 AFA51697	22.80 23.93
1999-2000	ACI90075 ACF25019	22.32 22.20	ACF33698 AF398422_1	21.75 23.93
2001-2002	AET77035 AEF59357	18.76 19.47	ACA47518 ACA47716	22.90 23.66
2003-2004	ABB87195 AEX33629	20.78 18.76	ACA47991 ADB07954	24.65 25.23
2005-2006	ABH03489 ABI30876	18.67 19.47	ABZ91692 ADB26166	22.77 23.26
2007-2008	ACH95480 AEN68955	21.82 20.63	AEH42714 ACF17951	17.46 22.73
2009-2010	ADD62041 AEN94754	29.54 19.79	ACT20894 AET50892	26.64 22.99
2011-2012	AEX91696 AFF59151	19.43 21.16	BAL14219 AFF59153	19.08 25.89

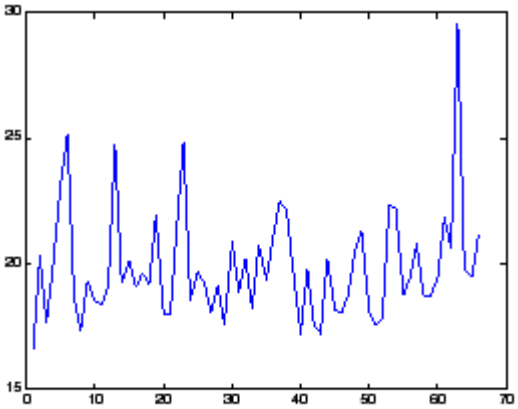

Fig. 1. The predictable percentage of HA sequence from 1914 to 2012

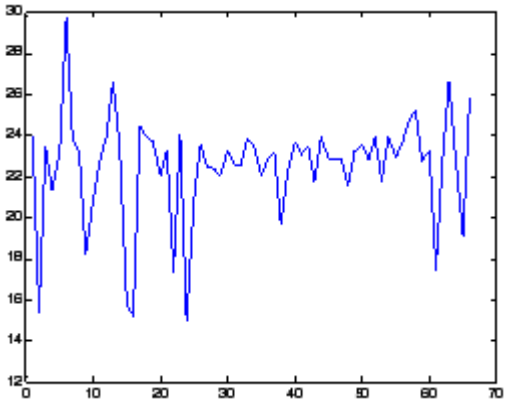

Fig. 2. The predictable percentage of NA sequence from 1914 to 2012

ACKNOWLEDGMENTS

The project was supported by the National Natural Science Foundation of China (Grant No. 11271163) and the Fundamental Research Funds for the Central Universities (Grant No. JUSRP21117).

REFERENCES

1. Zhang W, Wang C Y, Gao Y W, Xia C Z. Molecular biology research progress of influenza virus. *Chinese Journal of comparative medicine*. 2010; **20**: 74-79.(in Chinese)
2. Morens D., Folkers G., Fauci A. The challenge of emerging and re-emerging infectious disease. *Nature*. 2004; **430**: 242-249.
3. Zhu F, Wang H, Fang Hÿ*et al*.,. A Novel influenza A (H1N1) vaccine in various age groups. *New Engl Med*. 2009; **361**: 2414-2423ÿ
4. Taubenberger J.K., Reid A.H. and Fanning G. The 1918 Influenza Virus: a killer comes into view. *Virology*. 2000; **274**: 241-245.
5. David M. M, Anthony S. F. The 1918 influenza pandemic: insights for the 21st century. *The Journal of Infectious Diseases.* 2007; **195**: 1018-1028.
6. Jiang F, Li N. Protein structural codes and nucleation sites for protein folding. *Chin. Phys. B.* 2007; **16**: 392-404.
7. Chen Y, Wang J, Wang W. The preferences of orientations between the pairs of amino acids. *Chin. Phys.B.* 2007;**16**: 868-875.
8. Yu Z G, Anh V, Lau K S. Fractal analysis of measure representation of large proteins based on the detailed HP model. *Phys. A.* 2004; **337**: 171-184.
9. Yu Z G, Anh V, Lau K S. Chaos game representation of protein sequences based on the detailed HP model and their multifractal and correlation analyses. *J. Theo. Biol.* 2004; **226**: 341-348.
10. Zhu P, Tang X Q, Xu Z Y. The structural analysis of protein sequences based on the quasi-amino acids code. *Chin. Phys. B.* 2009;**18**: 363-369.
11. Gao J, Jiang L L, Xu Z Y. Chaos game representation walk model for the protein sequences. *Chin. Phys. B.* 2009; **18**: 4571-4579.
12. Liu J, Gao J. Long-memory ARFIMA model for DNA sequences of influenza A virus. *Acta Phys. Sin.* 2011; **60**: 702-707. (in Chinese)
13. Ren D, Gao J. Early-warning signals for an outbreak of the influenza pandemic. *Chin. Phys. B.* 2011; **20**: 370-376.
14. Yan S M, Wu G. Creation and Application of Computational Mutation. *Journal of Guangxi Academy of Sciences.* 2010; **26**: 130-139.
15. Yan S M, Wu G. Mutation Pattern of Amino-Acid Pairs in Human Henmoglobin β -Chain. *Journal of Guangxi Academy of Sciences.* 2011; **27**: 29-33.
16. Yan S M, Zuo W P, Zhu Q X, Huang Y Y, Pan L X, Wu G. Statistical Evidence for Cross-Species Infection and Cross-Subtype Mutation in Matrix Protein 2 Family of Influenza A Virus. *Journal of Guangxi Academy of Sciences.* 2010; **26**:45-51.