

Examining the Probability of the Critical Mutation of H5N8 by Comparing with H7N9 and H5N1 Using Apriori Algorithm and Support Vector Machine

Dae Young Kim, Hye-Jun Kim, Junhyeok Bae, and Taeseon Yoon

Abstract—In January, 2014, the outbreak of H5N8 in South Korea started in one duck farm. The representative outbreak of H5N8 is to turkey 1983 Ireland and to duck 2010 China. Obviously, people were concerned about whether this Influenza A virus is highly pathogenic or human transmissible. In this research to identify the probability of H5N8's pathogenic rate, we will investigate its chance to have Cytokine Storm, a deadly attribute of Influenza A virus, by seeking similarity in glycoprotein amino acid sequence with H5N1, which has the same hemagglutinin subtype, using Support Vector Machine. In addition, to identify H5N8's human transmissible possibility, we will compare the its year-on-year glycoprotein amino acid mutating trend with H7N9, which was previously known not to be transmissible to human but mutated to infect human, using Apriori Algorithm.

Index Terms—Apriori algorithm, H5N8, H5N1, H7N9, influenza A Virus, support vector machine.

I. INTRODUCTION

January 17th 2014, the onset of H5N8 virus was found in a duck farm in Gochang-gun, Jeollabuk-do, South Korea and about 200 million 3 thousand fowls were killed [1]. The outbreak of highly pathogenic influenza A virus subtype H5N8 was reported in turkey 1983 Ireland and duck 2010 China but hasn't been reported to infect humans. However, only a year ago, H7N9 virus infected 49 people and killed 11 of them in only two weeks after its first outbreak April 2013 in China and until November 4th 2013, it infected 134 people and killed 45 people overall [2]. Also in Taiwan, it was reported that H7N9 had infected 2 people by December 31 2013 and had killed the patients by January 20th 2014 [3], [4]. The initial victims of H7N9 in China were people engaged in poultry farms. So, it is assumed that they have got infected through ducks and chickens because it is not yet verified that H7N9 can spread through humans. However it is a remarkable phenomenon that the human-non-infectious H7N9 turned out to be contagious to human after all. Recently, various types of Influenza A virus such as H7N9 and H5N8 are found in East Asia. In this situation, we are going to weigh the possibility of H5N8 being able to infect human by comparing its mutating tendency with H7N9's using amino acid sequence data [5]. In addition, we are going to compare it with H5N1, influenza A virus with the same

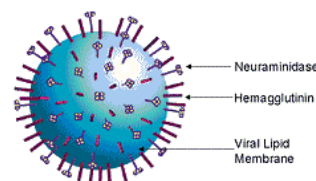
hemagglutinin subtype, to weigh the possibility whether H5N1's risky attribute, Cytokine Storm, could occur in H5N8.

II. INFLUENZA A VIRUS SUBTYPES

A. The Glycoprotein of Influenza; HA (Hemagglutinin) and NA (Neuraminidase)

There are glycoproteins on the Surface of influenza A virus. These proteins is classified into 11 types (HA, NA, M1, M2, NS1, NS2 (NEP), NP, PA, PB1, PB1-F2, and PB2) and each of them have distinguishing abilities. The two representative proteins that contribute to viral infection process and multiplied viruses' releasing process are HA (Hemagglutinin) and NA (Neuraminidase) (see Fig. 1).

External Structure of Influenza Virus



Major Surface Antigens

Fig. 1. The External Structure of Influenza Virus and two major glycoproteins ; hemagglutinin and neuraminidase (Kuby J. In: Immunology, 2nd ed. W.H. Freeman and Co., NewYork, 1991, 487).

The infection process begins as the HA protein combines with sialic acid in epithelial cell, the viral acceptor, and clings to the host cell. Then, the virus reproduces in the host cell. The NA protein severs the combination made by HA and contributes to the escaping process of the reproduced viruses.

In this manner, the HA protein and NA protein plays a significant role in the viral infection so all of the Influenza A subtype viruses are coined by using letter "H" from the HA protein and "N" from the NA protein. There are various subtypes in this HA and NA protein and accordingly, there are also a multitude of subtypes of Influenza A virus. Recently, new subtypes were found in bats, so overall 18 types of HA and 11 types of NA are currently known [6], [7].

B. H5N8 and H5N1

As mentioned in the Introduction, no human infected case was reported in the outbreak of H5N8 in 1983 Ireland, and in 2010 China, and even in the recent outbreak of H5N8 in South Korea. On the contrary, it is quite different for H5N1, which was previously called avian flu, that numerous cases of

Manuscript received March 11, 2014; revised May 15, 2014.

Dae Young Kim, Hye-Jun Kim, Junhyeok Bae, and Taeseon Yoon are with the Natural Science Department, Hankuk Academy of Foreign Studies, South Korea (e-mail : dae8177@naver.com).

infection in human were reported. Since 2003, human infected case of H5N1 was reported in Vietnam and China and about 570 people were infected and 330 of them died of the infection up till now.

Especially, H5N1 triggers cytokine storm which resulted in 1918 global flu pandemic [8]. Cytokine storm is a hyper immunization resulted from over releasing of cytokine, an immune substance in human. (see Fig. 2) As macrophage provides virus particles to T cells, they will be activated for cytokine emission in terms of immune system. This damages the lung or induces MODS (Multiple Organ Dysfunction Syndrome), causing high fatality [9]. H5N8's Cytokine Storm triggering ability is not yet verified, but since it has the same HA protein subtype as H5N1, identifying the similarity of amino acid sequence between them will be helpful in anticipating H5N8's mutation tendency of having deadly features.

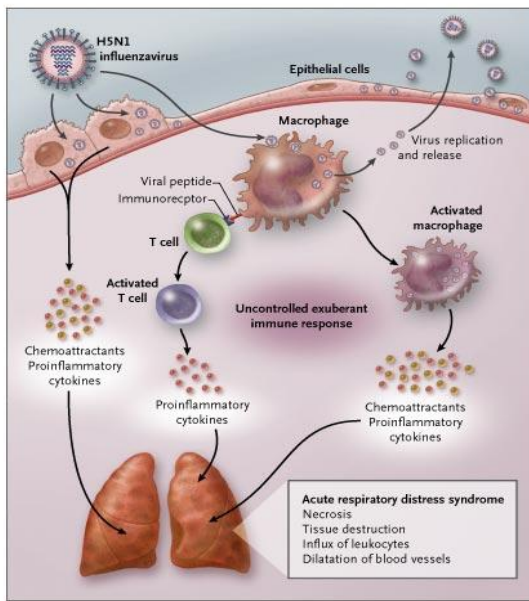


Fig. 2. The Mechanism of Cytokine Storm evoked by Influenza Virus (Osterholm. New England Journal of Medicine, 352 (18): 1839, Fig. 3. May 5, 2005.)

C. H5N8 and H7N9

About one year before the outbreak of H5N8 in South Korea, human-non-infectious H7N9 was reported to infect human for the first time in China and the infected patients tends to be increasing. Although H5N8 hasn't made infection in human yet, it may mutate like H7N9 and be lethal to human. Such being the case, we are going to investigate the possibility of mutation in H5N8 by comparing its amino acid sequence with H7N9.

III. EXPERIMENTS

A. Algorithms

The similarity of hemagglutinin and neuraminidase, glycoprotein of H5N8 and H5N1, will be measured by using Support Vector Machine. Support Vector Machine is a data classifying algorithm, which first sets up a function by going through a series of training process and then locating certain

pattern between two data groups therefore identifying their similarity [10]-[12]. In this research, we used Normal function and Kernel function as a vector function in each window size 5, 7, and 9. Kernel function was used for mapping process in support vector machine to distribute non linear data and we designated Linear Polynomial Function, Radial Basis Function (R.B.F.), and Quadratic Polynomial Function as Kernel function. In addition, to find out the year-on-year mutation tendency, Apriori algorithm was used. This experiment was also done in the order of window size 5, 7 and 9 and then compared with one another.

B. Data

The data used to compare the similarity of the glycoprotein amino acid sequence between H5N8 and H5N1 are the amino acid sequence of hemagglutinin and neuraminidase from H5N1 infected chicken from Hong Kong in 2001 and H5N8 infected duck from China Jiangsu in 2010.

The data used to investigate the year-on-year mutation tendency of H5N8 and H7N9 are the amino acid sequence of hemagglutinin and neuraminidase from H5N8 infected duck from Ireland in 1983, Jiangsu in 2010 and H7N9 infected human from Shanghai in 2013.

All of these datasets were obtained from NCBI (National Center for Biotechnology Information) in FASTA form [5].

IV. RESULTS

A. Similarity between H5N8 and H5N1 Examined by SVM

First of all, the datasets of amino acid sequences of hemagglutinin and neuraminidase were divided by linear separation using normal function and non-linear separation using kernel function. The accuracy values from this classification done by each functions are shown in the Tables. (See Table I, II, and III)

TABLE I: THE ACCURACY VALUES OF HEMAGGLUTININ AND NEURAMINIDASE FOR EACH FUNCTION IN WINDOW SIZE 5

	Normal Function	Linear Polynomial Function	R.B.F.	Quadratic Polynomial Function
Hemagglutinin	50.00%	65.00%	60.00%	40.00%
Neuraminidase	50.00%	55.00%	75.00%	70.00%

TABLE II: THE ACCURACY VALUES OF HEMAGGLUTININ AND NEURAMINIDASE FOR EACH FUNCTION IN WINDOW SIZE 7

	Normal Function	Linear Polynomial Function	R.B.F.	Quadratic Polynomial Function
Hemagglutinin	62.50%	68.75%	56.25%	81.25%
Neuraminidase	43.75%	43.75%	93.75%	62.50%

TABLE III: THE ACCURACY VALUES OF HEMAGGLUTININ AND NEURAMINIDASE FOR EACH FUNCTION IN WINDOW SIZE 9

	Normal Function	Linear Polynomial Function	R.B.F.	Quadratic Polynomial Function
Hemagglutinin	42.86%	42.86%	50.00%	57.14%
Neuraminidase	50.00%	66.67%	91.67%	91.67%

B. The Comparison of Mutation for Decades of H5N8 (1983-2010) and H7N9 (1988-2009-2013)

For precise interpretation, the output data with accuracy ($0 \leq a \leq 1$) under 0.8 were neglected. Fig. 3, Fig. 5, and Fig. 7 are the amino acid sequence of hemagglutinin in H5N8 and H7N9 at window size 5, 7, and 9 respectively. Fig. 4, Fig. 6, and Fig. 8 are the amino acid sequence of neuraminidase in H5N8 and H7N9 also at window size 5, 7, and 9 respectively.

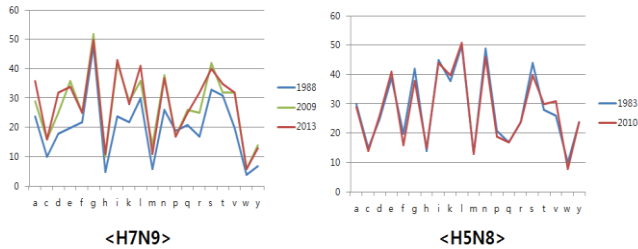


Fig. 3. The graph of amino acid frequency of hemagglutinin in window size 5.

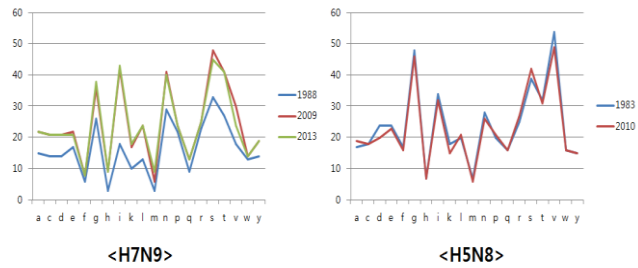


Fig. 4. The graph of amino acid frequency of neuraminidase in window size 5.

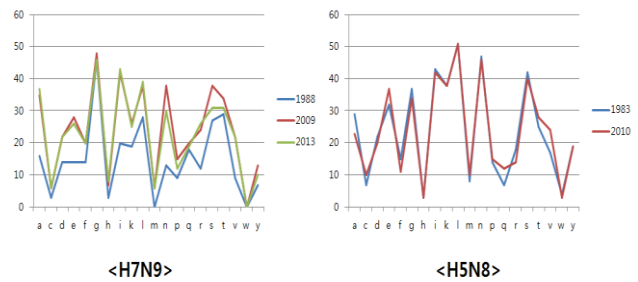


Fig. 5. The graph of amino acid frequency of hemagglutinin in window size 7.

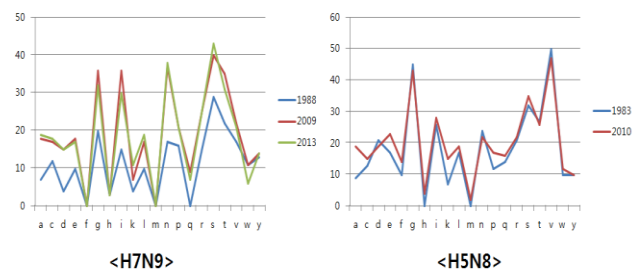


Fig. 6. The graph of amino acid frequency of neuraminidase in window size 7.

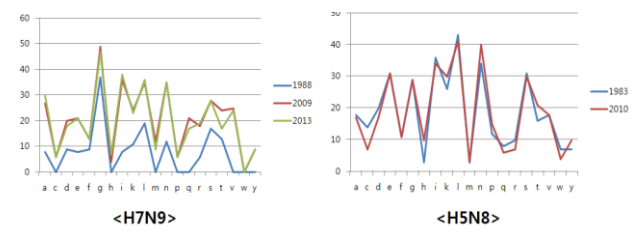


Fig. 7. The graph of amino acid frequency of hemagglutinin in window size 9.

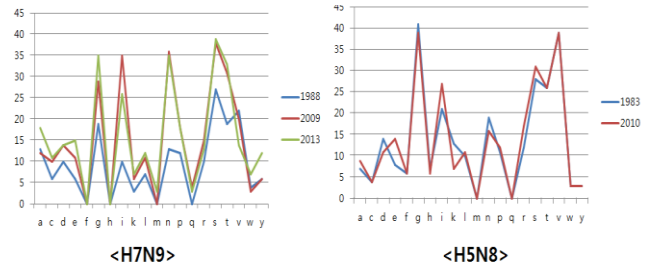


Fig. 8. The graph of amino acid frequency of neuraminidase in window size 9.

Through all window sizes, H5N8 hasn't mutated as dramatic as H7N9 from 1980s to 21st century.

V. DISCUSSION

A. Similarity between H5N8 and H5N1

See the amino acid dataset of hemagglutinin and neuraminidase in H5N8 which was classified by SVM in the condition of window size 9, where the result is more precise compared to other window sizes (see Table III). It showed out that the accuracy of hemagglutinin was higher than neuraminidase. In other words, the classification of neuraminidase in H5N8 and H5N1 was better done than hemagglutinin and this means that the similarity of hemagglutinin in H5N8 and H5N1 is much higher. The high similarity of hemagglutinin in H5N1 and H5N8 can be explained by the same hemagglutinin subtype; H5. But since the similarity of neuraminidase is low, we cannot conclude that the H5N8 of 2010 and H5N1 are alike. Therefore, it is viewed that the H5N1's feature of cytokine storm will not take place in H5N8 but this needs further prudent research to be confirmed.

B. Comparison of the Mutational Attribution of H5N8 and H7N9

H7N9 was not infectious to humans in 1988 to 2009. But, as the amino acid sequence of hemagglutinin and neuraminidase in H7N9 changed dramatically, it became infectious to humans starting from 2013. Especially, despite the same host, fowls, the sequence shows a great change. (See Fig. 3 ~ Fig. 8). As seen in the case of H7N9, the sufficient condition for the influenza to be infectious to human is to have a considerable change in the amino acid sequence even for the same host. However, compared to the mutation in H7N9, there was only a little change in the sequence between the 1983 H5N8 and 2010 H5N8. Thus, we concluded that the possibility of H5N8 to be infectious to humans is comparatively low.

REFERENCES

- [1] R. Herriman. (Jan. 26, 2014). South Korea H5N8 avian flu update, virus spreads to chicken farm. [Online]. Available: <http://www.theglobaldispatch.com/south-korea-h5n8-avian-flu-update-virus-spreads-to-chicken-farm-43052/>
- [2] National Health and Family Planning Commission of the People's Republic of China. China Reports 3rd H7N9 Case in November. [Online]. Available: http://www.npfp.gov.cn/data/201311/t20131129_408006.html
- [3] C.-F. Chen and S. C. Chang. Taiwan reports 2nd case of imported H7N9 since April. [Online]. Available: <http://focustaiwan.tw/news/asoc/201312310041.aspx>

- [4] Centers for Disease Control, R.O.C. (Taiwan). Taiwan's second imported H7N9 case died of severe pneumonia with septic shock on January 20, 2014. [Online]. Available: <http://www.cdc.gov.tw/english/info.aspx?treeid=BC2D4E89B154059B&nowtreeid=EE0A2987CFBA3222&tid=BBFB90143BD5A724>
- [5] National Center for Biotechnological Information. [Online]. Available: <http://www.ncbi.nlm.nih.gov/>
- [6] S. Tong *et al.*, "A distinct lineage of influenza a virus from bats," in *Proc. the National Academy of Sciences of the United States of America*, 2012, vol. 109, no. 11, pp. 4269-4274.
- [7] S. Tong *et al.*, "New world bats harbor diverse influenza A viruses," *PLoS pathogens*, vol. 9, no. 10, pp. e1003657, 2013.
- [8] T. Watanabe, S. Watanabe, K. Shinya, J. H. Kim, M. Hatta, and Y. S. Kawaoka, "Viral RNA polymerase complex promotes optimal growth of 1918 virus in the lower respiratory tract of ferrets," *PNAS*, 2009.
- [9] M. T. Osterholm, "Preparing for the Next Pandemic," *N Engl J Med* 2005, vol. 352, pp. 1839-1842, May 5, 2005.
- [10] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, June 1998.
- [11] C. Setubal and J. Medianis, *Introduction to Computational Molecular Biology*, 1997.
- [12] G. Gordon, *Support Vector Machines and Kernel Methods*, 2004.



Dae Young Kim was born in Daegu, South Korea in 1996. He is now in Hankuk Academy of Foreign Studies. He is interested in influenza virus research and published a paper about H5N1 hemagglutinin sequence analysis with artificial neural network on International Conference on Engineering and Applied Science. He also submitted a paper about comparison of several influenza a viruses' glycoprotein and accepted to BDM 14. He is now majoring bioinformatics, and proteomics for molecular modeling in the Academy.



Hye-Jun Kim was born in 1996 in South Korea, and is currently studying in Hankuk Academy of Foreign Studies, majoring in biology and studying bioinformatics. She is researching influenza virus and published a paper about H5N1 hemagglutinin sequence analysis with artificial neural network on International Conference on Engineering and Applied Science. She also submitted a paper about comparison of influenza a viruses' glycoprotein and accepted to BDM 14.



Bae Jun Hyeok was born in Seoul, South Korea in 1996. He is now in Hankuk Academy of Foreign Studies. He is interested in influenza virus research and published a paper about H5N1 hemagglutinin sequence analysis with artificial neural network on International Conference on Engineering and Applied Science. He also submitted a paper about comparison of several influenza a viruses' glycoprotein and accepted to PAKDD 2014. He completed the Computer Science Progress of Ajou University Science Genius Education. He is now majoring in computer science and algorithm in the Academy.



Taeseon Yoon was born in Seoul, Korea in 1972. He was a Ph.D. Candidate in computer education from the Korea University, Seoul, Korea, in 2003. From 1998 to 2003, he was with an EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea, as a lecturer and Ansan University as an adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a computer science and statistics teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.