Analyzing Patterns of Various Avian Influenza Virus by Decision Tree

Seung Jae Lim, Cheolho Heo, Yunsik Hwang, and Taeseon Yoon

***Abstract*—Influenza A virus, well-known for their dangerousness and diversity, take effect once again in recent years. In this case, the cause of this unfavorable situation is H5N8 that is one kind of mutants of influenza A virus. In the past, this virus show miserable effect in Ireland. Also, H5N1 and H1N1, the most common mutants of influenza A virus, have considerable hazard properties. Especially, there is a report that H1N1 took lives of many people and this was called as Swine Flu. Since these three corrupt virus have the most notorious characteristic, we will compare them and find some patterns about their behavior trend. In this experiment, we will try to get much more accurate results decision tree. Lastly, to improve our study, we make use of DNA Sequence Database and proven papers from National Central for Biotechnology Information (NCBI). We take our study's meaning on being able to treat and dealing with virus more easily by finding similarities and particular pattern of these dangerous virus.**

***Index Terms*—Influenza A virus, H1N1, H5N1, H5N8, decision tree.**

I. Introduction

A. Avian Influenza Virus

In our research, we focus on Avian Influenza Virus (AI Virus). This virus, the main cause of avian influenza, give rise to acute viral disease [1]. Furthermore, sometimes, they affect to people extensively. In case of human infection, they show really high amount of fatality since they have very high contagious [2]. When these virus go into human body and infect them, it shows lots of respiratory system symptom such as cough and dyspnea, whole human body symptom such as fever, chills and myalgia and symptoms related with central nerve such headache and consciousness degradation. What was worse, in former days, there was a warning of WHO about H5N1, the most noted AI virus. They described H5N1 is more dangerous than SARS virus [3]-[5].

Although there are some antiviral agents, oseltamivir, amantadine and rimantadine, these virus have plenty of kinds of mutants. They have more rate of arising mutant million times than other virus. For example, H1N1 is known for one of the most dangerous AI virus mutant. In 2009, this virus, called novel swine-origin influenza A, make 14,000 mortality. Including H5N1 and H1N1, Ai virus usually show high rate of mortality when people get them into their internal body [6].

However, nowadays, one kind of new mutant has been found in Korea. This mutant is called as H5N8. Because we are aware of riskiness of them, we focus our object of this research on making a comparison and pattern recognizing H1N1, H5N1 and H5N8 employing decision tree algorithm.

TABLE I: Species of Avian Influenza a Virus Strains [7]

HA subtype designation	NA subtype deignation	Avian influenza A viruses
H1	N1	A/duck/Alberta/35/76(H1N1)
H1	N8	A/duck/Alberta/97/77(H1N8)
H2	N9	A/duck/Germany/1/72(H2N9)
H3	N8	A/duck/Ukraine/63(H3N8)
H3	N8	A/duck/England/62(H3N8)
H3	N2	A/turkey/England/69(H3N2)
H4	N6	A/duck/Czechoslovakia/56(H4N6)
H4	N3	A/duck/Alberta/300/77(H4N3)
H5	N3	A/tern/South Africa/300/77(H4N3)
H5	N4	A/Ethiopia/300/77(H6N6)
H5	N9	A/turkey/Ontario/7732/66(H5N9)
H5	N1	A/chick/Scotland/59(H5N1)
H6	N2	A/turkey/Massachusetts/3740/65(H6N2)
H6	N8	A/turkey/Canada/63(H6N8)
H6	N5	A/shearwater/Australia/72(H6N5)
H6	N6	A/jyotichinara/Ehiopia/73(H6N6)
H6	N1	A/duck/Germany/1868/68(H6N1)
H7	N7	A/fowl plague virus/Dutch/27(H7N7)
H7	N1	A/chick/Brescia/1902(H7N1)
H7	N9	A/chick/China/2013(H7N9)
H7	N3	A/turkey/England/639H7N3)
H7	N1	A/fowl plague virus/Rostock/34(H7N1)
H8	N4	A/turkey/Ontario/6118/68(H8N4)
H9	N2	A/turkey/Wisconsin/1/66(H9N2)
H9	N6	A/duck/Hong Kong/147/77(H9N6)
H9	N6	A/duck/Hong Kong/147/77(H9N6)
H9	N8	A/manishsurpur/Malawi/149/77(H9N8)
H9	N7	A/turkey/Scotland/70(H9N7)
H10	N8	A/quail/Italy/1117/65(H10N8)
H11	N6	A/duck/England/56(H11N6)
H11	N9	A/duck/Memphis/546/74(H11N9)
H12	N5	A/duck/Alberta/60/76/(H12N5)
H13	N6	A/gull/Maryland/704/77(H13N6)
H14	N4	A/duck/Gurjev/263/83(H14N4)
H15	N9	A/shearwater/Australia/2576/83(H15N9)

B. Decision Tree

The method which would be mainly used in this research is decision tree learning. Decision tree learning is a method commonly used in data mining that uses a decision tree as a

Manuscript received April 9, 2014; revised July 1, 2014.

Seung Jae Lim, Cheolho Heo, Yunsik Hwang, and Taeseon Yoon are with the Hankuk Academy of Foreign Studies, South Korea (e-mail: tonylim0930@gmail.com).

predictive model [8]. Traditionally, decision tree is drawn in manual forms in real life with leaves with branches [9], yet in this research it will be realized by the template of informatics [10].

In this case of data mining, assume that each feature of a class (Each feature is identified as 'Class') should be able to be separated by distinct, finite criteria [11], [12]. Decision tree is consisted of internal node, arcs, and leaf. In this diagram, internal node is labeled with an input feature. The arcs are labeled with each of the possible values of the feature. Each leaf of the tree is labeled with a class or a probability distribution [13], [14].

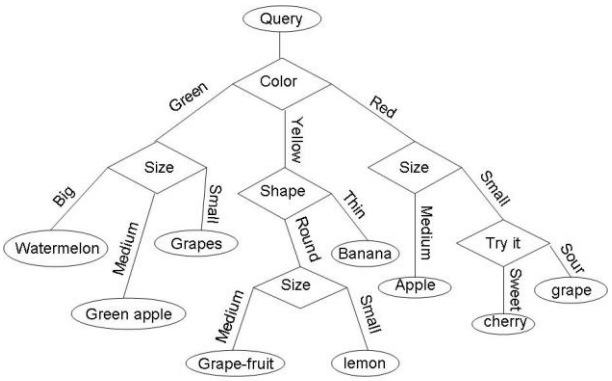

Fig. 1. Manual form of decision tree [15].

In this method of data mining, a tree can be "learned" by recursively repeated process of splitting the source set into subsets based on a number of criteria. Recursion partitioning is completed when the outcome of subset has identical value of the target variable, or when the process of splitting doesn't give additional value. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and it is by far the most common strategy for learning decision trees from data [16]. Fig. 1 and Fig. 2 are simple form of decision tree.

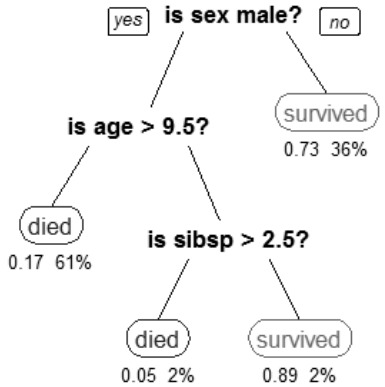

Fig. 2. Example of decision tree: Death toll of Titanic [17].

In data mining, decision trees can be described as the combination of mathematical and computational form, after the process of categorization and generalization. Generally, the diagram gradually comes in records of the form [18]:

$$(X, Y)=(x_1, x_2, x_3, \ldots, x_k, Y) \quad (1)$$

The dependent variable, Y, is the target variable that we are trying to understand, classify or generalize. The vector x is composed of the input variables, x_1, x_2, x_3 etc., that are used for testing.

II. Experiment Object

A. H1N1

H1N1 Virus is the most common case of influence that found in human body [19]. This virus, in early 18th century, occurred a very fatal influenza called Spanish flu. Influenza virus have about 34,400 amino-acid. When Spanish flu occurred, it was found that H1N1 Virus have 25~30 amino-acid mutants [20]. Fig. 3 is structure of H1N1.

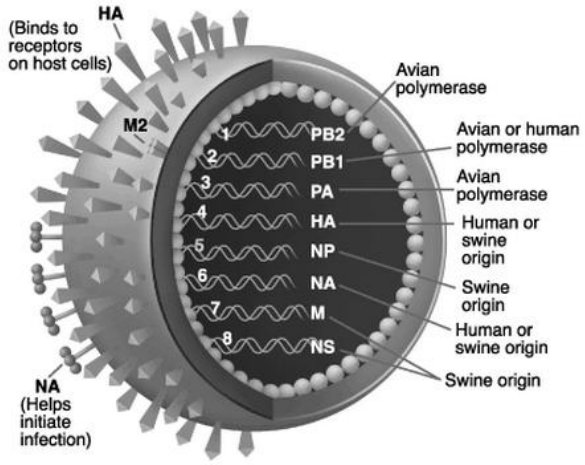

Fig. 3. Structure of H1N1 [21].

B. H5N1

H5N1 Virus is one kind of H1N1 mutants and known for its dangerousness containing high-pathogenicity avian influenza. In the past, people believed that this virus never infect human [22]. However, in 1997, there was 6 victims because of Hong Kong avian influenza. This virus have some specialized features: they don't be shifted from human to human and only through direct contact with birds. This is very similar with normal influenza so it is hard to distinct them. In case of South-East Asia, this shows more than 80% fatality [23], [24].

C. H5N8

H5N8 virus is also commonly known as the "bird flu" in that infects mostly avian species although some have been found in mammals as well. These viruses range in the level of severity. This virus is one of the many subtypes. One of the main reasons for concern when it comes to these viruses is that they undergo constant change. This makes vaccine manufacturing almost impossible. In 1983, there was an outbreak that included diarrhea, nervousness and depression. As a result, 8,000 turkeys, 28,020 chickens and 270,000 ducks were slaughtered. For the most part symptoms of the H5N8 virus are respiratory. There are the common "flu-like" symptoms of fever, chills, headache, coughing and weakness. There are reports of there being conjunctivitis associated with the virus as well [25].

D. Neuraminidase

In this research, several kinds of virus (H1N1, H5N8, H5N1)'s gene that express the property of Neuraminidase

will be mainly analyzed. Neuraminidase, refer to Fig. 4, which acts as a crucial component for the virus's survival, is a hydrolase enzyme that cleaves the surface of its host. This is particularly used for the dissociation of glycoside bond.

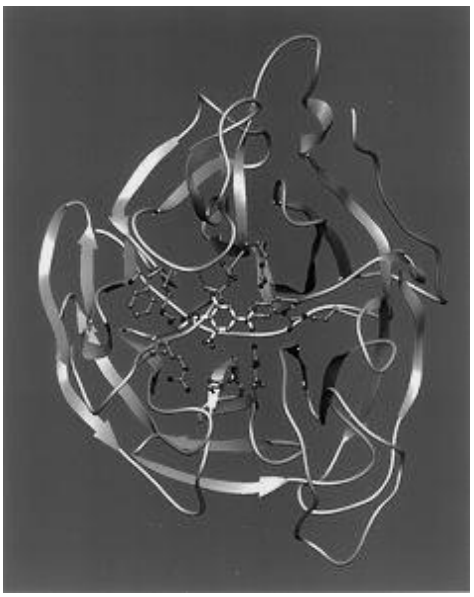

Fig. 4. Neuraminidase diagram [26].

As widely known, after a virus efficiently complete propagation process in the host, it comes out of the host, and repeatedly permeates into another. In this process, the surface of the host's cell should be decomposed to enable the virus to come out. Neuraminidase facilitates the hydrolysis of the surface of the cell by embarking on the decomposition of glycoside bond. Consequently, the virus's antigenicity often depends on this enzyme's existence. Neuraminidase activities also include assistance in the mobility of virus particles through the respiratory tract mucus and in the elution of virion progeny from the infected cell [27], [28]. As the importance of this enzyme stood out, the existence as well as the intensity of Neuraminidase became one of the standards of virus's classification.

III. Experiments

A. Decision Tree

By employing decision tree(C 5.0) algorithm, 10 fold cross validation experiment held with 3 classes(H1N1, H5N1, H5N8).

B. Rule Extraction

With the result of decision tree, we've selected high frequency data set which overpass frequency rate 0.75. We've selected amino acidic rules among each window and each class.

IV. Conclusion

A. Results

According to Table II, it's noticeable that all of the influenza A shown their rules extracted with amino acid at position 7. We could assume that the amino acid of position 7 makes each subtype differentiate from each other. Considering the fact that neuraminidase takes part in virus exodus from host cell, we could also assume that their species specificity partially came from amino acidic features of position 7.

TABLE II: Rule Extraction under 7 Window

Type	Rule	Frequency
H1N1	pos6 = P pos7 = N	0.8
	pos4 = D pos7 = E	0.75
	pos4 = F pos7 = E	0.75
	pos1 = N pos7 = G	0.75
	pos3 = C pos7 = F	0.75
	pos3 = Q pos7 = C	0.75
	pos4 = S pos7 = A	0.75
H5N1	pos6 = S pos7 = N	0.75
	pos1 = S pos7 = G	0.75
	pos2 = R pos7 = E	0.75
	pos6 = S pos7 = G	0.75
	pos1 = F pos7 = P	0.75
H5N8	pos4 = L pos7 = D	0.75
	pos5 = L	0.75
	pos3 = S pos7 = C	0.75
	pos6 = Q pos7 = G	0.75

TABLE III: Rule Extraction under 9 Window

Type	Rule	Frequency
H1N1	pos6 = D	0.833
	pos7 = S	0.75
	pos6 = G	0.75
	pos7 = A	0.8
	pos7 = D	0.75
H5N1	pos7 = F	0.75
	pos6 = T	0.75
	pos9 = S	0.75
	pos7 = G	0.75
	pos3 = D	0.8
	pos7 = S	0.8
H5N8	pos9 = T	0.75
	pos7 = D	0.75

It's noticeable that rule extraction under 9 window also shown their rules extracted with amino acid at position 7. To be specific, the rule which is pos7=D extracted in all type of the virus and the rule which is pos7=S extracted in H1N1 and H5N1. These results support that among the subtype of the influenza A (H1N1, H5N1, H5N8) have similarities with amino acids. Also they support the formal hypothesis we've made which is amino acid of position 7 makes difference among virus.

According to the result of rule extraction under 13 window, there are no specific experimental features. The result suggest that among them exist amino acidic differences which differs them from each other. Also, these differences evoke different pathological features. Considering all of the results, H1N1, H5N1 and H5N8 have similarities in some of amino acidic features while there are noticeable differences.

TABLE IV: Rule Extraction under 13 Window

Type	Rule	Frequency
H1N1	pos5 = N	0.8
	pos5 = F	0.75
H5N1	pos1 = D	0.8
	pos3 = C	0.75
	pos1 = W	0.75
	pos5 = P	0.75
	pos1 = S pos5 = C	0.75
H5N8	pos3 = E	0.8
	pos3 = H	0.75
	pos1 = I pos11 = G	0.8

B. Expectations

Recently, the spread of H5N8 in the South Korea evoked extensive damages. Being conscious of its biological features can help to invent effective vaccines or preventing diseases. At this perspective, the results of this paper suggest H5N8's independent features which differentiate it from other subtypes of influenza A and also similarities as well. Plus, trial of employing computer intelligence based algorithm to comparing various influenza A and recognizing their patterns is rare. Thus, this research can contribute in noticing influenza A virus more accurately.

REFERENCES

[1] Avian influenza strains are those well adapted to birds, European Centre for Disease Prevention and Controls.

[2] S. L. Knobler, A. Mack, A. Mahmoud, and S. M. Lemon, *The Threat of Pandemic Influenza: Are We Ready? Workshop Summary*, Institute of Medicine of the National Academies Press, 2005.

[3] The Writing Committee of the World Health Organization (WHO) Consultation on Human Influenza A/H5, "Avian influenza A (H5N1) infection in humans," *New England Journal of Medicine*, vol. 353, pp. 1374-1385, September 29, 2005.

[4] Confirmed Human Cases of Avian Influenza A(H5N1). [Online]. Available: http://web.archive.org/web/20040218103733/www.who.int/csr/disease/avian_influenza/country/en/

[5] Y. Hiromoto *et al.*, "Evolutionary characterization of the six internal genes of H5N1 human influenza a virus," *The Journal of General Virology 81 (Pt 5)*, pp. 1293–1303, 2000.

[6] *First Human Avian Influenza A (H5N1) Virus Infection Reported in Americas*, CDC, January 8, 2014.

[7] C. N. Kawaoka, "22," in *Topley and Wilson's Microbiology and Microbial Infections*, B. Mahy and L. Collier, Eds. 1998, p. 415.

[8] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Pub Co Inc., 2008.

[9] Decision Tree. Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Decision_tree(website)

[10] J. R. Quinlan, "Induction of decision trees," *Journal Machine Learning*, vol. 1, pp. 81-106, 1986.

[11] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[12] L. Rokach and O. Maimon, *Data Mining with Decision Trees: Theory and Applications*, World Scientific Pub Co Inc., 2008.

[13] J. J. Rodriguez, L. I. Kuncheva, and C. J. Alonso, "Rotation forest: A new classifier ensemble method," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1619-1630, 2006.

[14] G. V. Kass, "An exploratory technique for investigating large quantities of categorical data," *Applied Statistics*, vol. 29, no. 2, pp. 119–127, 1980.

[15] Decision Tree. [Online]. Available: http://www.wikipedia.org/

[16] R. C. Barros, M. P. Basgalupp, A. C. P. L. F. de Carvalho, A. A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews*, vol. 42, no. 3, pp. 291-312, 2012.

[17] S. H. Cha and C. C. Tappert, "A genetic algorithm for constructing compact binary decision trees," *Journal of Pattern Recognition Research*, pp. 1-13, 2009.

[18] Decision Tree Learning. Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Decision_tree_learning

[19] Influenza Summary Update 20, 2004–2005 Season, FluView: A Weekly Influenza Surveilance Report, Centers for Disease Control and Prevention.

[20] P. Palese, "Influenza: Old and new threats," *Nat. Med.*, vol. 10, December 2004.

[21] Structure of H1N1. [Online]. Available: http://ksj.mit.edu/tracker/2009/04/swine-flu-pigflu-h1n1-who-now-prefers-wh

[22] K. S. Li *et al.*, "Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia," *Nature*, vol. 430, pp. 209–213, 2004.

[23] K. Ungchusak, P. Auewarakul, S. F. Dowell *et al.*, "Probable person-to-person transmission of avian influenza A (H5N1)," *N. Engl. J. Med.*, vol. 352, no. 4, pp. 333–340, January 2005.

[24] K. S. Li, Y. Guan, J. Wang *et al*., "Genesis of a highly pathogenic and potentially pandemic H5N1 influenza virus in eastern Asia," *Nature*, vol. 430, pp. 209–213, 2004.

[25] Avian influenza A (H5N1)-update 31: Situation (poultry) in Asia: need for a long-term response, comparison with previous outbreaks, *Epidemic and Pandemic Alert and Response*, World Health Organization, 2004.

[26] Neuraminidasa. [Online]. Available: http://es.wikipedia.org/wiki/Neuraminidasa

[27] P. Palese, K. Tobita, M. Ueda, and R. W. Compans, "Characterization of temperature sensitive influenza virus mutants defective in neuraminidase," *Virology*, vol. 61, no. 2, pp. 397–410, October 1974.

[28] C. Liu, M. C. Eichelberger, R. W. Compans, and G. M. Air, "Influenza type A virus neuraminidase does not play a role in viral entry, replication, assembly, or budding," *Journal of Virology*, vol. 69, no. 2, pp. 1099–106, February 1995.

Seung Jae Lim was born in 1996. He is currenty a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Cheolho Heo was born in 1997. He is currently a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Yunsik Hwang was born in 1996. He is currently a student of Hankuk Academy of Foreign Studies, Korea. He is mainly interested in both organic chemistry and molecular biology. he has been studying pattern analysis and computer programming and its application to advanced biology.

Taeseon Yoon was born in Seoul, Korea, in 1972. He was a Ph.D. candidate with the degree in computer education from the Korea University, Seoul, Korea, in 2003. From 1998 to 2003, he was with 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 a 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.