Analysis of Ebolavirus with Decision Tree and Apriori algorithm

Eunby Go, Seungmin Lee, and Taeseon Yoon

***Abstract*—EbolaVirus, which has high facilities up to 90%, is introduced into the human population through close contact with the blood, secretions, organs or other bodily fluids of infected animals. It is mostly occurred in Central and West Africa, near tropical rainforests and their host are bats. In this paper, we analyzed the DNA sequences of 5 Ebolavirus : Bundibugyo Ebolavirus, Reston Ebolavirus, Sudan Ebolavirus, TaiForest Ebolavirus, and Zaire Ebolavirus and investigated the difference between them based on the genus they are involved in. Furthermore, we look for the frequency of the amino acid and find the similarities between the Ebolaviruses.**

***Index Terms*—Ebolavirus, bundibugyo ebolavirus, reston ebolavirus, sudan ebolavirus, taiforest ebolavirus, zaire ebolavirus, decision tree, apriori algorithm.**

I. Introduction

EVD, Ebola virus disease, caused by ebola viruses, broke out recently in the Republic of Guinea, Liberia, West Africa. Number of infectees went beyond 240, while the dead exceeded 100. The first outbreak of Ebola was in 1976, simultaneously in Nzara, Sudan, and in village Yambuku, Democratic Republic of Congo. The village Yambuku was located near the Ebola river, from which the disease was named. Genus Ebolavirus is comprised in the Filoviridae family, order Mononegavirales, classified as 5 distinct species, Bundibugyo ebolavirus (BDBV), Zaire ebola virus(EBOV), Reston ebolavirus (RESTV), Sudan ebolavirus(SUDV) and Tai Forest ebolavirus (TAFV) [1].

II. Materials and Methods

Ebolaviruses are divided into five genera: Bundibugyo Ebolavirus, Reston Ebolavirus, Sudan Ebolavirus, TaiForest Ebolavirus, and Zaire Ebolavirus. Our study focus on common factors and analyzing differences various DNA sequences of infectious Ebolavirus: Bundibugyo Ebolavirus(BDBV), Reston Ebolavirus (RESTV), Sudan Ebolavirus(SUDV), TaiForest Ebolavirus (TAFV), and Zaire Ebolavirus (EBOV) [2].

Manuscript received May 15, 2014; revised July 18, 2014.

Eunby Go is with the Genetics, Conversation and Evolutionary, Biology Program, National Institute of Standards and Technology SEOUL. She is at the HAFS (Hankuk Academy of Foreign Studies) (e-mail: eunbygo@naver.com).

Seungmin Lee is with the Neurobiology, Conversation and Evolutionary, Biology Program, National Institute of Standards and Technology SEOUL. She is enrolled in HAFS (Hankuk Academy of Foreign Studies)(e-mail: tmdalsdl0420@gmail.com).

Taeseon Yoon was with the Korea University, South Korea. He is now with the Hankuk Academy of Foreign Studies, Lecturer, South Korea. (e-mail: tsyoon@hafs.hs.kr).

A. Bundibugyo Ebolavirus (BDBV)

Bundibugyo ebolavirus, confirmed as Ebola in 2007, as new species of Ebola in 2008. It is one of four ebolavirus that causes EVD in humans. EVD due to BDBV infection cannot be differentiated from EVD caused by other Ebolaviruses by clinical observation alone. An epidemiological study determined there were 116 confirmed and probable cases of the new Ebola species, BDBV, and that the outbreak had a mortality rate of 34%(39 deaths) [3].

B. Reston Ebolavirus (RESTV)

Reston Ebolavirus was first described in 1990 as a new "strain" of Ebola virus, a result of mutation from Ebola virus. It is the single member of the species Reston ebolavirus, which is included into the genus Ebolavirus, family Filoviridae, order Mononegavirales. Despite its status as a level-4 organism, Reston virus is non-pathogenic to humans, though hazardous to monkeys; the perception of its lethality was confounded due to the monkey's coinfection with Simian hemorrhagic fever virus (SHFV).

C. Sudan Ebolavirus (SUDV)

Sudan Ebolavirus emerged in 1976; it was first assumed to be identical with the Zaire species. It is believed to have broken out first among factory workers. However the virus was not found in any of the local animals and insects that were tested in response. The carrier is still unknown.

D. TaiForest Ebolavirus (TAFV)

TaiForest Ebolavirus made its first and thus far only known appearance in 1994 during a viral hemorrhagic fever epizootic among western chimpanzees (Pan troglodytes verus) in Ta ï National Park, C ôte d'Ivoire. As more dead western chimpanzees were discovered, many tested positive for infection with an ebolavirus distinct from those already known[4].

E. Zaire Ebolavirus (EBOV)

Electron micrographs of EBOV show them to have the characteristic threadlike structure of a filovirus. EBOV VP30 is around 288 amino acids long. The virions are tubular in general form but variable in overall shape and may appear as the classic shepherd's crook or eyebolt, as a *U* or a *6*, or coiled, circular, or branched; laboratory techniques, such as centrifugation, may be the origin of some of these formations. The RNA genome of EBOV encodes at least 7 structural proteins. The ribonucleoprotein (RNP) complex that mediates transcription and replication of the viral genome comprises 4 of these proteins: nucleoprotein (NP), VP35, VP30, and the RNA-dependent RNA polymerase (L). The 3 other proteins—glycoprotein (GP), VP40, and VP24—are membrane-associated proteins [5].

DOI: 10.7763/IJMLC.2014.V4.470

F. Decision Tree

Decision tree learning is an analysis method commonly used in data mining, in order to create a model that predicts the value of a target variable based on several input variables. This method uses a decision tree as a predictive model. Traditionally, decision tree is drawn in manual forms with leaves and branches to identify a strategy that is most likely to reach a goal. Yet in this research, it will be embodied by the template of informatics [6].

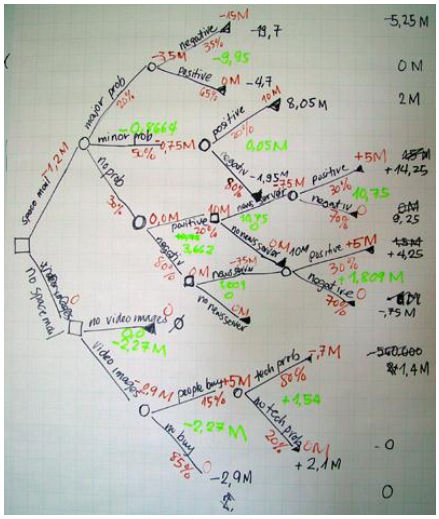

Fig. 1. Manual form of decision tree.

According to Fig. 1, we assume that each feature of a class (each element of the domain of the target feature is identified as 'Class'.) should be able to be separated by distinct, finite criteria [7]. Decision tree consists of internal node, arcs, and leaf. In this diagram, internal node is labeled with an input feature. The arcs from a node 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 [8].

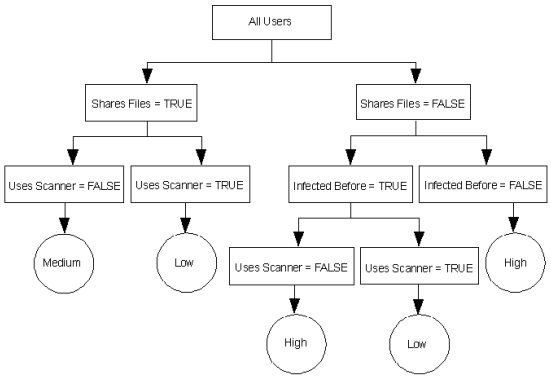

Fig. 2. Example of decision tree: Microsoft users.

A tree can be "learned" by dividing the source set into subsets by an attribute value test, and the process is repeated recursively, which is called recursive partitioning. The process is completed when the subset has the identical value with the target variable, or when recursive partitioning any longer adds value to the predictions. This process of top-down induction of decision trees (TDIDT) is an example of a greedy algorithm, and is definitely the most common strategy for learning decision trees from data [9].

Decision trees can also be expressed as the combination of mathematical and computational techniques in order to assist the description, categorization and generalization of a given set of data.

Data comes in records of the form:

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

The dependent variable, *Y*, is the target variable that needs to be understood, classified or generalized. The variable factor *x* is composed of the input variables, x_1, x_2, x_3 etc., which are used for the process [10].

G. Apriori Algorithm

In this research, analysis bases on Apriori Algorithm, suggesting associations between the five ebolaviruses. Apriori Algorithm is the classic form of algorithm for finding and extending frequent items in a transactional database. This algorithm accentuates the general trends of the database and reveals the association rules [11].

Using "bottom up" approach, frequent subsets are extended one at a time, which is known as candidate generation, and each group of candidates is tested against the data until no further extensions are found. By Breadth-first search and a Hash tree structure, candidate item sets of length k are generated from length *k*-1, and then the candidates with infrequent sub pattern are removed. According to downward closure lemma, the candidate set comprises every frequent k-length item set [12]. Then, transaction database is scanned to reveal the frequent item sets among the candidates. We divided the amino acid sequences of virus into 9, 13, 17 windows and analyzed the frequent sequence pattern respectively [13].

III. Results

A. Decision Tree

TABLE I: Rule Extraction under 9 Window(1)

virus	Rule	Frequency
Bundibugo	Not extracted	
Reston	Not extracted	
Sudan	pos5 = T pos9 = A	0.750
	pos4 = G pos9 = D	0.800
	pos1 = R pos9 = P	0.750
	pos1 = Q pos7 = F	0.750
TaiForest	Not extracted	
Zaire	pos2 = I pos7 = V	0.750
	pos1 = I pos5 = R	0.750

TABLE II: Rule Extraction under 9 Window(2)

virus	Rule	Frequency
Bundibugo	pos7 = L pos8 = K pos9 = S	0.750
Reston	pos4 = L pos7 = L pos8 = Q	0.750
	pos1 = Y pos5 = L pos7 = F	0.750
Sudan	pos3 = T pos5 = R pos9 = S	0.750
TaiForest	Not extracted	
Zaire	Not extracted	

According to Table I, it is noticeable that there were no rules of Bundibugo Ebolavirus, Reston Ebolavirus, and TaiForest Ebolavirus. This result is due to lack of viruses unique amino acidic features among other classes. Also most of the Sudan Ebolavirus shown their rules extracted with amino acid at position 9. We assume that position 9 is the important factors which differentiate each other. Although there are no resemblance between two rules of Zaire

Ebolavirus, position 2 and position 7 come out a lot more than position 1and position 5.

According to Table II it's noticeable that even it is same 9 window, more positions were found than the table I. It shows that their rules extracted with amino acid at position 7. To be specific, the rule which is pos7=L extracted in Bundibugo virus and Reston Virus. These results support that among the subtype of Ebolavirus(Bundibugo virus, Reston Virus) have similarities with amino acids. Also there are similarities in pos9=S between Bundibugo and Sudan Ebolavirus. However, TaiForest Ebolavirus rules were also not found.

TABLE III: RULE EXTRACTION UNDER 13 WINDOW

virus	Rule	Frequency
Bundibugo	Not extracted	
Reston	Not extracted	
Sudan	pos6 = L pos11 = T	0.857
	pos7 = I pos11 = L	0.833
	pos2 = K pos9 = G	0.750
TaiForest	pos5 = A pos9 = V	0.750
Zaire	Not extracted	

According to Table III, 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 difference evoke different pathological features. Considering all of the results, Bundibugyo Ebolavirus, Reston Ebolavirus, Sudan Ebolavirus, TaiForest Ebolavirus, and Zaire Ebolavirus.have similarities in some of amino acidic features while there are noticeable difference exist.

TABLE IV: RULE EXTRACTION UNDER 17 WINDOW

virus	Rule	Frequency
Bundibugo	Not extracted	
Reston	pos6 = H pos12 = F	0.750
Sudan	pos6 = I pos15 = L	0.833
	pos6 = V pos10 = V	0.750
	pos1 = Y pos6 = D	0.750
	pos6 = F pos9 = D	0.750
TaiForest	Not extracted	
Zaire	pos6 = I pos15 = R	0.750

According to Table IV, It's noticeable that most of the Ebolavirus shown their rules extracted with amino acid at position 6. We assume that position 6 is the important factors which differentiate each other. Also Reston Ebolavirus rules were found which was not found in 13 window. However, it seems odd that the rule of TaiForest Ebolavirus was not extracted under 17 window condition while that of TaiForest Ebolavirus was extracted under 13 window condition. Due to increased window condition, TaiForest Ebolavirus shown lack of unique features which can classify itself from others.

B. Apriori Algorithm

By Apriori algorithm, results showed these kinds of patterns.

Zaire ebolavirus

1. Amino9 is L 28
2. Amino2 is L 24
3. Amino6 is L 24
4. Amino4 is L 21
5. Amino13 is L 21
6. Amino1 is L 20
7. Amino7 is L 19
8. Amino12 is L 19
9. Amino3 is S 18
10. Amino10 is L 18
11. Amino1 is S 17
12. Amino1 is T 17
13. Amino4 is R 17
14. Amino9 is S 17

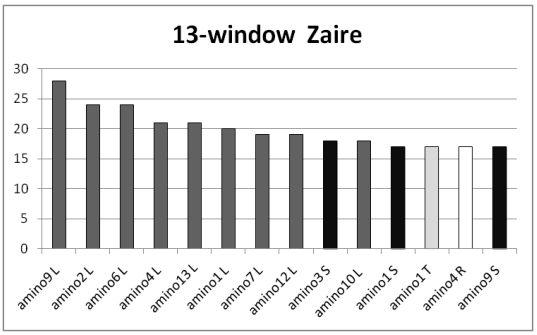

Fig. 3. 13-window Zaire ebolavirus amino sequence.

Fig. 3 is the result drawn by analyzing Zaire ebolavirus, split into 13 windows. Other results are also shown in the similar patterns. Roughly, 49 rules were found in each ebolavirus (9, 13, and 17 windows collectively), so we found total 245 rules in our research. Based on the results above, we arranged the types of amino acids by the frequencies of appearance. We selected the most representative rule of each virus (9, 13, and 17 window respectively) in order to find the similarity in sequence pattern among 5 viruses. The results are in the below.

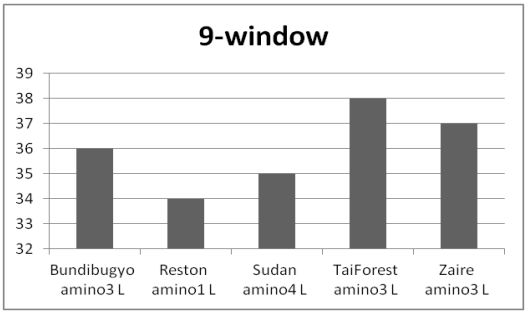

Fig. 4. 9-window Ebolavirus amino sequence.

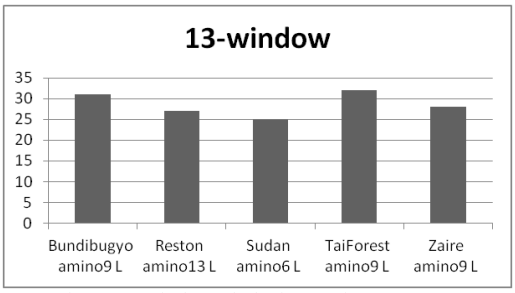

Fig. 5. 13-window Ebolavirus amino sequence.

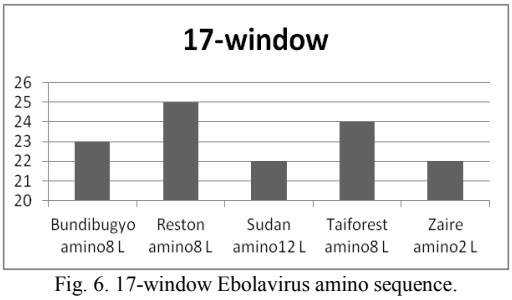

Fig. 6. 17-window Ebolavirus amino sequence.

Fig. 4, Fig. 5, and Fig. 6 show an analysis of 5 ebolavirus sequence in 5, 13, 17 window. 5-window: Comparison between Bundibugyo ebolavirus amino3 Leucine, Reston ebolavirus amino1 Leucine, Sudan ebolavirus amino4 Leucine, Taiforest ebolavirus amino3 Leucine, Zaire ebolavirus amino3 Leucine. 13-window: Bundibugyo ebolavirus amino9 Leucine, Reston ebolavirus amino13 Leucine, Sudan ebolavirus amino6 Leucine, TaiForest amino9 Leucine, Zaire amino9 Leucine. 17-window: Bundibugyo ebolavirus amino8 Leucine, Reston ebolavirus amino8 Leucine, Sudan ebolavirus amino12 Leucine, TaiForest amino8 Leucine, Zaire amino2 Leucine.

IV. Conclusion

In the process of researching and developing vaccines for Ebolavirus which causes serious diseases for human, scientists have been frustrated, for there has not been any effective means to cure the disease because it started to occurred a lot lately and there were less data and references about it [14]. After analysis using Decision Tree, trying 9, 13, and 17 windows of analysis, we thoroughly investigated genetic structure of Bundibugyo Ebolavirus, Reston Ebolavirus, Sudan Ebolavirus, TaiForest Ebolavirus, and Zaire Ebolavirus. [15] As a result of analysis, we found that high windows shows more detailed information about DNA sequences of Ebolaviruses and there are similarities of DNA sequences between the species of it. Since Reston Ebolavirus does not cause vital disease to humans as it does to animals such as pigs and monkeys, we look forward to finding the ways of curing Ebolavirus disease [16].

Also it certainly suggests that Bundibugyo virus, Reston virus, Sudan virus, TaiForest virus, and Zaire virus show similarities in amino acid sequences. When we arranged the amino acid in frequency-descending order, five viruses shared similarities in most of the windows (9-window, 13-window, and 17-window) [17]. As the chart above suggests, the frequency of appearance for Leucine was overwhelming than the others. Also, the result shows that ebolavirus sequence is not much affected by the criterion. As a result, we could assume from this fact that ebolavirus may share similarities among ebolavirus that has same hosts [18].

References

[1] *Ebola virus disease Fact sheet N°103*, World Health Organization, March 2014, pp. 201-256.

[2] M. Eichner, S. F. Dowell, and N. Firese, "Incubation Period of Ebola Hemorrhagic Virus Subtype Zaire OH AND BRETT," *Osong Public Health and Research Perspectives*, June 2011, vol. 2, no. 1, pp. 3-7.

[3] T. Hoenen, A. Groseth, D. Falzarano, and H. Feldmann "Ebola virus: unravelling pathogenesis to combat a deadly disease," *Trends in Molecular Medicine*, 2006 May, vol.12, no. 5, pp. 206-215.

[4] J. W. King and A. A. Khan, "Medscape: Ebola virus, clinical presentation," 2012.

[5] S. P. Fisher-Hoch *et al.*, "Pathophysiology of shock and hemorrhage in a fulminating viral infection (Ebola)," *J. Infect. Dis*, vol. 152, no. 5, pp. 887-894, November 1985.

[6] S. Baize *et al.*, "emergence of zaire ebola virus disease in guinea — preliminary report," *New England Journal of Medicine*, 16 April 2014.

[7] *25 people in Bakaklion, Cameroon killed due to eating of ape.*

[8] X. Pourrut, B. Kumulungui, T. Wittmann, G. Moussavou, A. Délicat, P. Yaba, D. Nkoghe, J. P. Gonzalez, and E. M. Leroy, "The natural history of Ebola virus in Africa," *Microbes and infection*, vol. 7, issue 7-8, pp. 1005-1014, 2005 June.

[9] J. Morvan, V. Deubel, P. Gounon, E. Nakouné, P. Barrière, S. Murri, O. Perpète, B. Selekon, D. Coudrier, A. Gautier-Hion, M. Colyn, and V. Volehkov, "Identification of Ebola virus sequences present as RNA or DNA in organs of terrestrial small mammals of the Central African Republic," *Microbes and Infection*, vol.1, issue 14, December 1999, pp. 1193-1201.

[10] *Fruit bats may carry Ebola virus*, BBC News, December 11, 2005.

[11] R. L. Swanepoel, P. A. Leman, F. J. Burt *et al.*, "Experimental inoculation of plants and animals with Ebola virus," *Emerging Infectious Diseases*, vol. 2, no. 4, pp. 321-325, 1996.

[12] E. M. Leroy, B. Kumulungui, X. Pourrut *et al.*, "Fruit bats as reservoirs of Ebola virus," *Nature*, vol. 438, no. 7068, pp. 531-710.

[13] X. Pourrut, A. Délicat, P. E. Rollin, T. G. Ksiazek, J. P. Gonzalez, and E. M. Leroy, "Spatial and temporal patterns of Zaire ebolavirus antibody prevalence in the possible reservoir bat species," *The Journal of infectious diseases*. vol. 196, issue supplement 2, pp. S176-S183, 2007.

[14] J. Starkey, "90 killed as fruit bats spread Ebola virus across West Africa," *The Times*, April 2014.

[15] D. Taylor, R. Leach, and J. Bruenn, "Filoviruses are ancient and integrated into mammalian genomes," *BMC Evolutionary Biology*, 2010 Jun, vol.10

[16] J. P. Gonzalez, X. Pourrut, and E. Leroy, "Ebolavirus and other filoviruses," *Current topics in microbiology and immunology*, vol. 315, pp. 363-387, 2007.

[17] A. T. Peterson, J. T. Bauer, and J. N. Mills, "ecologic and geographic distribution of filovirus disease," *Emerging Infectious Diseases*, vol. 10, no.1, January 2004.

[18] *Questions and Answers about Ebola Hemorrhagic Fever*, Centers for Disease Control and Prevention, 2009-03-25.

Eunby Go was born in Republic of Korea. She was very curious about all the things, especially about the nature. So she attended HAFS and started to write a paper that she studied and research. But first time to do anything is always hard to everyone. Although it was hard she kept working. She studied biology and chemistry. Genetics was her final choice of her major and she tried to discover the secrets of DNA and other genes. In 2012, she continued working on the project and finished the paper above. She, with 3 fellow workers, is currently working on a paper, concerning H3N2 treatment and its sequences.

Seung Min Lee was born in 1996. She is currently a student in a science major of Hankuk Academy of Foreign Studies, Korea. She is mostly interested in neurobiology and has been recently studying bioinformatics, analyzing the genomic sequences and finding the regularity in order to reveal its features.

Taeseon Yoon was born in Seoul, Korea, in 1972. He received his Ph.D. 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 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.