

Comparison of Flavivirus Using Datamining-Apriori, K-means, and Decision Tree Algorithm

Chan woo Kim*, Se Hwan Ahn*, Taeseon Yoon**

*Natural Science, Hankuk Academy of foreign studies, Yong-in, Republic of Korea

**Computer Science and Engineering, Korea University, Seoul, Republic of Korea

kimcwbest1@gmail.com, brandon07@naver.com, tsyoon@hafs.hs.kr

Abstract—Nowadays, zika virus infection-occurred in brazil-is becoming serious global problem. Temporary treatments-using drug-can alleviate symptoms for a while but cannot be a basic solution for infections. We researched amino acids of Zika virus and other four kinds of flaviviruses-Dengue fever, Yellow fever, Saint Louis, Japanese Encephalitis-with bioinformatic experiment; Apriori, K-means and Decision Tree algorithm. Results show differences and common things among them-Japanese Encephalitis has most representative amino acid.

Keywords - zika virus and flavivirus, genetic treatment, apriori, k-means, decision tree algorithm

I. Introduction

Zika virus is being widely spread in South America and Asia region and it is one of the most popular issues throughout the world[1][2]. Zika virus is mainly transmitted by the Aedes aegypti and Aedes albopictus. Major symptoms of Zika virus are headache, ophthalmitis, arthralgia, nausea, ganglion and in serious cases it can lead to general paralysis [3]. To diagnose, we should check antibody inside the blood or detect RNA of Zika virus.

Zika virus is one of the species in flavivirus genus. There are several viruses which belong to flavivirus genus for instance Yellow fever[4][5], Japanese Encephalitis[6][7], dengue fever[8][9][10], St. Louis Encephalitis[11] etc. Flavivirus usually infected with mosquitoes have similar traits and shares lots in common[12]. The purpose of our research is to show new approach of analyzing Zika virus based on similarities and differences between flavivirus. Therefore we should find out how Zika is similar with other flaviviruses in the aspect of base sequence. In this process, we will contrast each viruses which belongs to flavivirus, using bioinformatic ways-K means, apriori, decision tree algorithms-and find out some similarities and differences among them.

II. EXPERIMENT

A. GENERAL OUTLINE OF EXPERIMENT

We had an experiment of five flavivirus: Dengue fever, Japanese Encephalitis, Saint Louis, Yellow fever, and Zika virus. We used three algorithms to determine similarities and

differences of these viruses : apriori, k-means and decision tree algorithms. We divided each sequences to 5,9,11 windows for the convenience of the study. Surprisingly, results of apriori and k-means could not show the tendency among 5 viruses, but the results of decision tree algorithm showed the tendency between them.

B. Experiment using Apriori

1) Concept of Apriori

Apriori is an algorithm which can search the frequent elements in the database using the set theory[13]. It also enlarges the frequent data set to efficiently analyze the data set. Using Apriori algorithm, we can apprehend the general trend within the database. We tried to find out the tendency among five flaviviruses with apriori algorithms. Main purpose of this algorithm is to figure out protein(amino acid) which has highest frequency in each sequences. It averages the sequences of each windows with statistics ways, and make the outcome; most common amino acid of that virus.

2) Apriori Results

We chose top five amino acid frequency of each viruses and windows. As you can see tables below, each of five virus has different figures.

(A=Dengue fever, B=Japanese Encephalitis, C=Saint Louis, D=Yellow fever, E=Zika virus)

Table 1. Results of 5 windows.

	1 st	2 nd	3 rd	4 th	5 th
A	amino2 =G 14	amino4 =G 14	amino1 =V 11	amino1 =S 10	amino3 =G 10
B	amino2 =E 20	amino1 =G 18	amino2 =R 18	amino2 =V 18	amino5 =V 18
C	amino5 =T 15	amino1 =V 10	amino2 =S 10	amino1 =T 9	amino3 =T 9
D	amino1 =K 9	amino1 =V 9	amino2 =V 9	amino3 =G 9	amino4 =A 9

^a Corresponding author: tsyoon@hafs.hs.kr

E	amino2 =G 14	amino4 =G 14	amino1 =V 11	amino1 =S 10	amino3 =G 10
---	-----------------	-----------------	-----------------	-----------------	-----------------

Table 2. Results of 9 windows.

	1 st	2 nd	3 rd	4 th	5 th
A	amino5 =G 10	amino7 =S 10	amino1 =G 9	amino3 =P 8	amino4 =V 7
B	amino3 =E 13	amino3 =R 13	amino5 =V 13	amino7 =A 12	amino1 =R 11
C	amino2 =A 8	amino3 =T 8	amino4 =T 8	amino5 =L 8	amino1 =G 7
D	amino4 =T 9	amino1 =G 8	amino3 =V 8	amino8 =A 7	amino1 =K 6
E	amino5 =G 10	amino7 =S 10	amino1 =G 9	amino3 =P 8	amino4 =V 7

Table 3. Results of 11 windows.

	1 st	2 nd	3 rd	4 th	5 th
A	amino3 =L 9	amino 7=G 9	amino 1=G 8	amino 5=G 7	amino1 0=D 7
B	amino2 =E 12	amino 1=R 11	amino 6=A 11	amino 7=R 11	amino1 =E 10
C	amino 11=T 7	amino 1=V 6	amino 6=G 6	amino 1=G 5	amino 2=G 5
D	Amino 1=G 6	amino 2=T 6	amino 3=A 6	amino 11=V 6	amino 3=G 5
E	Amino 3=L 9	amino 7=G 9	amino 1=G 8	amino 5=G 7	amino 10=D 7

Numbers of Table 1~3 represent the average figure of amino acids found in each windows.

3) Results analysis

As we mentioned in introduction, purpose of our study is to find out similarity among flaviviruses and work out a way to treat Zika virus. However, we could not find exact common ground. Instead, it is certain that each viruses should be treated with different genetic ways, although all of them belong to same flavivirus.

C. EXPERIMENT USING K MEANS ALGORITHM

1) Concept of K means

K-means clustering is a clustering method which divides N databases into k clusters. It aims to minimize the dissimilarity and cost function in each cluster. While in the process the similarity within the data object increases and similarity

between the data object decreases. Complexity of the algorithm depends on the number of cluster and the sum of squares.

Doing experiment of each 5 cases and 3 windows(5,9,11 windows) takes a long time and too complicated to analyze. So we divided each sequence into pieces, connecting them together and made it to one amino sequence. Then apply it to k-means algorithms. For example, in 5 windows, sequence structure will look like Table.4.

(A=Dengue fever, B=Japanese Encephalitis, C=Saint Louis, D=Yellow fever, E=Zika virus)

Table 4. Structure of sequences for k-means.

A1					A2					A3					A4					A5				
V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V
A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E

V is for virus, and A is for amino acid. For example, in 5 windows algorithm, 1/5 sequences of each virus consists each window.

2) K-means result

Table 5. Results of 5 windows.

Attribute	full data(524.0)	cluster 0(322.0)	cluster 1(202.0)
amino1	V	T	G
amino2	G	G	E
amino3	G	G	E
amino4	G	G	K
amino5	A	A	G
Cleavage	2	5	2

cluster 0: T,Q,G,E,A,5

cluster 1: M,M,G,K,R,2

Table 6. Results of 9 windows.

Attribute	full data(292.0)	cluster 0(202.0)	cluster 1(90.0)
amino1	G	G	V
amino2	A	A	C
amino3	G	G	P
amino4	A	E	A
amino5	G	L	G
amino6	V	V	T
amino7	S	G	S
amino8	K	K	D
amino9	A	A	R
Cleavage	2	2	2

Table 7. Results of 11 windows.

Attribute	full data(239.0)	cluster 0(138.0)	cluster 1(101)
amino1	G	G	V
amino2	E	T	E

amino3	L	L	A
amino4	G	E	G
amino5	G	G	T
amino6	G	G	A
amino7	G	G	R
amino8	T	L	A
amino9	V	V	L
amino10	D	D	I
amino11	V	V	E
Cleavage	2	5	2

3) Result analysis

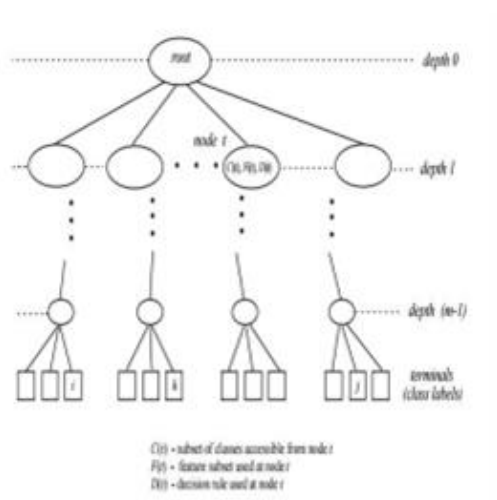
Biggest reason we used k-means algorithm was to find out specific common things-using clusters-among them because we could not find out something special with apriori algorithm. However, no matter cluster is 0 or 1, there was nothing general; all five of them was different. Therefore It is certain that even partial amino acid of them is totally different and they should be treated with different genetic ways, although all of them belong to same flavivirus

D. Experiment using Decision Tree algorithm.

1) Concept of Decision Tree

We used Decision tree algorithm to figure the difference and similarities between five flaviviruses. Decision tree algorithm classifies the data bases by dividing them with certain standards[14]-[15]. Using Decision tree, we could determine the position of amino acid. So, we used this algorithm to analyze the best according position of the amino acid. By analyzing the best according position it is easy to find the inclination within the databases.

Figure 1. example of the decision tree



2) Decision Tree results

(A=Dengue fever, B=Japanese Encephalitis, C=Saint Louis, D=Yellow fever, E=Zika virus)

Table 8. Results of 5 windows

A	B	C	D	E
---	---	---	---	---

7	47	14	9	13
13	126	17	25	2
6	50	14	9	3
6	55	11	5	2
21	45	13	10	1

Table 9. Results of 9 windows

A	B	C	D	E
4	28	7	4	7
10	69	11	11	1
3	34	0	4	2
0	31	7	7	2
11	28	5	7	2

Table 10. Results of 11 windows

A	B	C	D	E
2	23	3	4	9
6	56	10	11	1
2	25	6	3	1
2	30	2	1	1
13	24	3	1	0

Figures in tables mean the number of position that virus have in tree; virus which has largest number takes largest area in tree. In all windows, figures for 'b(Japanese Encephalitis)' was the largest. Therefore, amino acid of other four viruses-except Japanese Encephalitis-resembles to that of Japanese Encephalitis.

4) Result analysis

As we mentioned in introduction, there are several viruses which are similar to Zika and belonged to flavivirus. Through results of decision tree, we could figure that amino acid of Japanese Encephalitis is most representative for flaviviruses. There are some common things between Japanese Encephalitis and others. They usually inhabit south-east Asia and south-pacific area. If human became infected by mosquitoes, in most case, patient suffers from acute high-fever and headache and these can even lead to shock-syndrome. These are similar to Japanese Encephalitis virus-in the aspects of infection route, diagnosis, and symptoms.

III. Conclusion

We had researched about differences among zika virus and others belong to flavivirus to suggest most efficient genetic treatment. We had experiment with three algorithms -apriori, k-means, and decision tree algorithm. In apriori and k-means, we could not find common things among five viruses; amino acids which have highest frequency of each virus were different. Therefore, it is evident that genetic treatment should not be focused to most common amino acid.

Instead, as you can see results of decision tree, Japanese Encephalitis was most representative of five flaviviruses. We could have already predicted that symptoms and treatment of five viruses are similar. Additionally, with decision tree algorithms, we concluded that treatments of japan mosquito

encephalitis could be general solution for flavivirus including Zika virus.

References

- [1] Zanluca, Camila, Vanessa Campos Andrade De Melo, Ana Luiza Pamplona Mosimann, Glauco Igor Viana Dos Santos, Claudia Nunes Duarte Dos Santos, and Kleber Luz. "First Report of Autochthonous Transmission of Zika Virus in Brazil." *Mem. Inst. Oswaldo Cruz* 110.4 (2015): 569-72. Web.
- [2] Hayes, Edward B. "Zika Virus Outside Africa." *Emerging Infectious Diseases* 15.9 (2009): 1347-350. Web.
- [3] Lustig, Yaniv, Ella Mendelson, Nir Paran, Sharon Melamed, and Eli Schwartz. "Detection of Zika Virus RNA in Whole Blood of Imported Zika Virus Disease Cases up to 2 Months after Symptom Onset, Israel, December 2015 to April 2016." *Euro Surveill.* 21.26 (2016): n. pag. Web.
- [4] Rice, C., E. Lenches, Eddy, S. Shin, R. Sheets, and J. Strauss. "Nucleotide Sequence of Yellow Fever Virus: Implications for Flavivirus Gene Expression and Evolution." *Science* 229.4715 (1985): 726-33. Web.
- [5] "Neurologic Disease Associated with 17D-204 Yellow Fever Vaccination: A Report of 15 Cases." *National Center for Biotechnology Information*. U.S. National Library of Medicine, n.d. Web. 10 Nov. 2016.
- [6] Whitley, Richard J., and John W. Gnann. "Viral Encephalitis: Familiar Infections and Emerging Pathogens." *The Lancet* 359.9305 (2002): 507-13. Web.
- [7] Weissenböck, Herbert, Jolanta Kolodziejek, Angelika Url, Helga Lussy, Barbara Rebel-Bauder, and Norbert Nowotny. "Emergence of Usutu Virus, an African Mosquito-Borne Flavivirus of the Japanese Encephalitis Virus Group, Central Europe." *Emerging Infectious Diseases* 8.7 (2002): 652-56. Web.
- [8] "Rapid Detection and Quantification of RNA of Ebola and Marburg Viruses, Lassa Virus, Crimean-Congo Hemorrhagic Fever Virus, Rift Valley Fever Virus..." *National Center for Biotechnology Information*. U.S. National Library of Medicine, n.d. Web. 10 Nov. 2016.
- [9] Endy, T. P. "Epidemiology of Inapparent and Symptomatic Acute Dengue Virus Infection: A Prospective Study of Primary School Children in Kamphaeng Phet, Thailand." *American Journal of Epidemiology* 156.1 (2002): 40-51. Web.
- [10] Kuhn, Richard J., Wei Zhang, Michael G. Rossmann, Sergei V. Pletnev, Jeroen Corver, Edith Lenches, Christopher T. Jones, Suchetana Mukhopadhyay, Paul R. Chipman, Ellen G. Strauss, Timothy S. Baker, and James H. Strauss. "Structure of Dengue Virus: Implications for Flavivirus Organization, Maturation, and Fusion." *Selected Papers of Michael G. Rossmann with Commentaries The Development of Structural Biology* (2014): 317-25. Web.
- [11] Roehrig, J.t., J.h. Mathews, and D.w. Trent. "Identification of Epitopes on the E Glycoprotein of Saint Louis Encephalitis Virus Using Monoclonal Antibodies." *Virology* 128.1 (1983): 118-26. Web.
- [12] Chambers, T. "Flavivirus Genome Organization, Expression, And Replication." *Annual Review of Microbiology* 44.1 (1990): 649-88. Web.
- [13] Ye, Yanbin, and Chia-Chu Chiang. "A Parallel Apriori Algorithm for Frequent Itemsets Mining." *Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06)* (2006): n. pag. Web.
- [14] Safavian, S.r., and D. Landgrebe. "A Survey of Decision Tree Classifier Methodology." *IEEE Transactions on Systems, Man, and Cybernetics* 21.3 (1991): 660-74. Web.
- [15] Olaru, Cristina, and Louis Wehenkel. "A Complete Fuzzy Decision Tree Technique." *Fuzzy Sets and Systems* 138.2 (2003): 221-54. Web.



Chan woo Kim (M'16) was born in Gwang-ju Metropolitan city, Republic of Korea, in 2000 February 26th. He graduated Heung-Deok middle school located in Yong-in, south Korea in 2014. He became a Member of IEEE in 2016. Since 2015, he has been studying in the Department of Natural Science, Hankuk Academy of Foreign Studies, located in Yong-in South Korea. As high school

student, he is interested in bioinformatics fields, including biology and veterinary. Chan woo Kim has no memberships in professional societies and doesn't have any awards and work for IEEE committees and publications.



Se hwan Ahn (M'16) was born in Seoul in 1999 May 27th. He graduated Gyeong won middle school located in Seoul, South Korea in 2014. He became a Member of IEEE in 2016. He is now attending to Hankuk Academy of foreign students, located in Yong-in South Korea. He is interested in researching Biochemistry and he had a previous research of fine dust particle. Se hwan Ahn has no memberships in professional societies and doesn't have any awards and work for IEEE committees and publications.



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