ANALYSIS OF ZIKA VIRUS AND CACIPACORE VIRUS USING DATAMINING

Chaeun Kim[1] and Taeseon Yoon[2]

[1]Nature Science Course, Hankuk Academy of Foreign Studies, Yongin, South Korea
[2]Hankuk Academy of Foreign Studies, Yongin, South Korea

ABSTRACT

Zika virus is spreaded by mosquito. There is high probability of Microcephaly. In 1947, the virus was first found from Uganda, but it has broken out all around world, especially North and south America. So, apriori algorithm and decision tree were used to compare amino acid sequences of zika virus and cacipacore virus. By this, dissimilarity and similarity about them were found.

1. INTRODUCTION

In Brazil, more than 1.5 million people have become infected in Zika virus since April, 2015. The first outbreak of Zika was in Uganda in 1947 [1, 2]. Now, in recent 2 months, Zika virus has been found in about 39 countries. And, because of zika virus, few people died [3]. Not only Zika virus, but also other viruses (like west nile virus, yellow fever, Cacipacore virus, and dengue virus) are fatal to people. And all of them belong to flavivirus. Cacipacore is a flavivirus that belongs to the large group of Japanese encephalitis viruses (JEV). Unfortunately, vaccines exist for yellow fever and dengue virus [4, 5] but other viruses (west nile virus, Cacipacore virus, and zika virus) are not. An investigation on figuring out similarities and differences between the viruses will lead us to a deeper understanding about the viruses. By using apriori algorithm and decision tree, we can expect to find out the possibility of an effective treatment.

2. MATERIALS AND METHODS

Materials we used in the experiment are Zika Virus, and Cacipacore virus. Their genome sequences were gathered from the National Center for Biotechnology Information (NCBI). And we use apriori algorithm and decision tree to progress the experiment.

2.1 Zika Virus

In 1947, zika virus was first identified in Uganda rhesus monkey. [1, 2] People can find that the virus is shared to homo-sapiens from mosquito. Conducted three to seven days, people who are infected can notice only a minimal symptoms such as rash, acute fever, conjunctivitis and muscle pain appear. But, rougly 80% cases are an inapparent infection [6]. This virus has high probability of Microcephaly. So many countries warn to woman who can become pregnant [3].

2.2 Cacipacore Virus

Cacipacore, ssRNA positive-strand viruses, was first found in Brazil in the municipality of Theobroma, 320 km from the city of Porto Velho. It shows a genetic correlation with other emerging arbo viruses such as: the West Nile virus (WNV) the Rocio virus (ROCV) and the St. Louis encephalitis virus (SLEV). It belongs to the large group of Japanese encephalitis viruses (JEV).

2.3 Apriori algorithm

The Apriori Algorithm is the one of Algorithms used for data mining [14, 16]. In this research, We used apriori algorithm to find frequency of amino acid sequence of Flavivirus. We divided 3 window rules; 13-window, 17- window and 19-window. The number of window represents is the number of sequences of virus which we have disposed before applying algorithm. We can compare Zika virus and Cacipacore virus using frequency of amino acid sequence. In other words, we can find out similarity and dissimilarity about the viruses.

2.4 Decision tree

One of the common data mining methods is decision tree. It uses branching method by drawing leaves and branches to generate model which expects target variable based on several input variables [6, 16]. Decision tree is a graph that every internal node corresponds to input variables and each branch corresponds to possible outcome of the input variables. And leaf node is a value of target variable when each input node has a level of route from root node to leaf node. Decision tree stretches continuously by adding the input variables. It expands until node of subset equals to target variable or the new predictive value cannot be added caused by division. If we use decision tree, it is effective to find out differences between data [13, 17]. So in our experiment, decision tree method is applied in order to compare and contrast the viruses.

2.5 Window

Window is a fixed size of region of molecular sequence in the fields of bioinformatics, and it refers to the specific size of nucleotide numbers [11]. If the size of window is small, the pattern would not be exhibited and most of the windows would exhibit 0 count. If the size of window is wide, the characteristics of patterns and genomes would be seen.

3. RESULT

3.1 Apriori algorithm

Using apirori algorithm, we found out results like these.

Apriori algorithm in 13, 17, and 19 window was used to identify genomes of Zika virus, and Cacipacore virus based on typical amino acids. For analyzing graphically, the graph was used to identify; amino acids are aligned on x-axis in alphabetical order of their FASTA format, and the frequencies of amino acids are aligned on y-axis. The y values are given only to the amino acids of best rules. This research performed 18 cycles of experiment to ensure the accuracy of the results. For each window, the minimum support is set as 0.1(10%). It means that rules that appear in over 10% of whole instances are selected as best rules [12]. Before getting into analysis, the condition is described on the following. There are total 264 instances in the 13 window experiment, 202 instances in 17window, and 181 instances in 19 window.

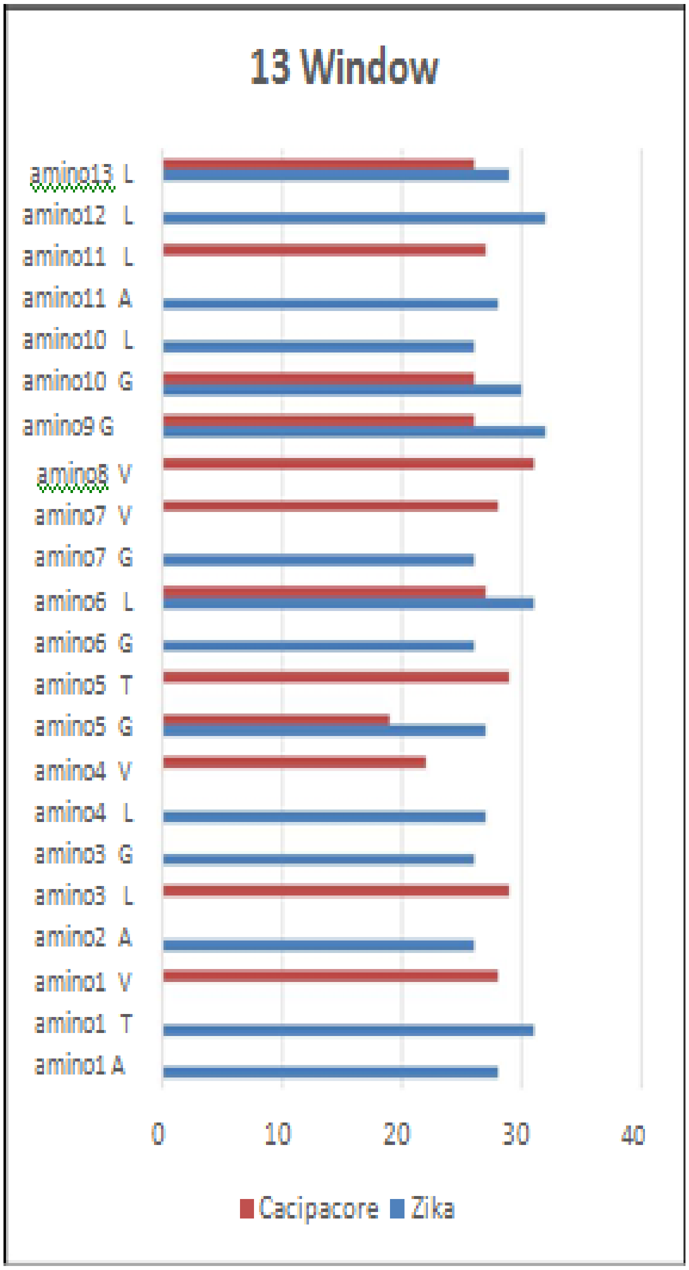

Figure 1. 13-window amino acid sequence

Fig.1 is the result of amino acid sequences of zika and cacipacore virus using apriori algorithm window-13. It shows amino1 to amino 13 so it's attribute is 13(excluded cleavage). Number of best rules found in each cacipacore and zika virus are 10 and 16.

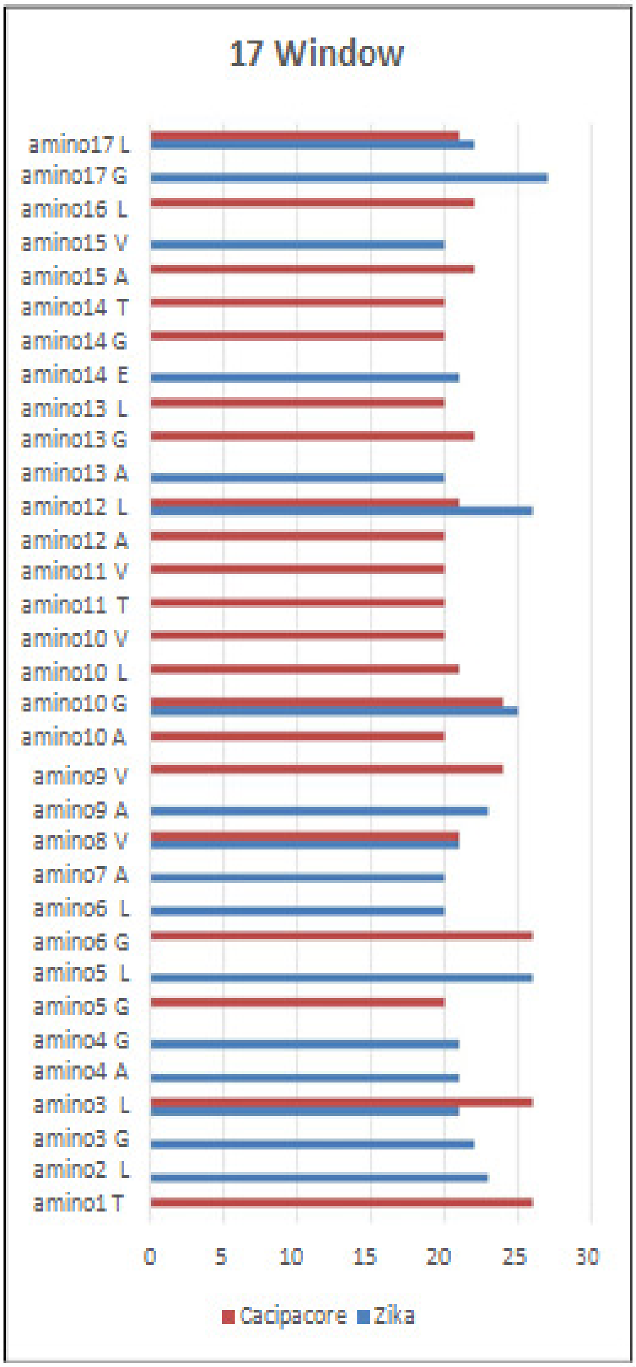

Fig. 2 is the result of amino acid sequences of the viruses using apriori algorithm window-17. It shows amino 1 to amino 17, so it's attribute is 17(excluded cleavage). Number of best rules found in each Cacipacore and Zika virus are 21 and 17. Also viruses have same number of amino8 V.

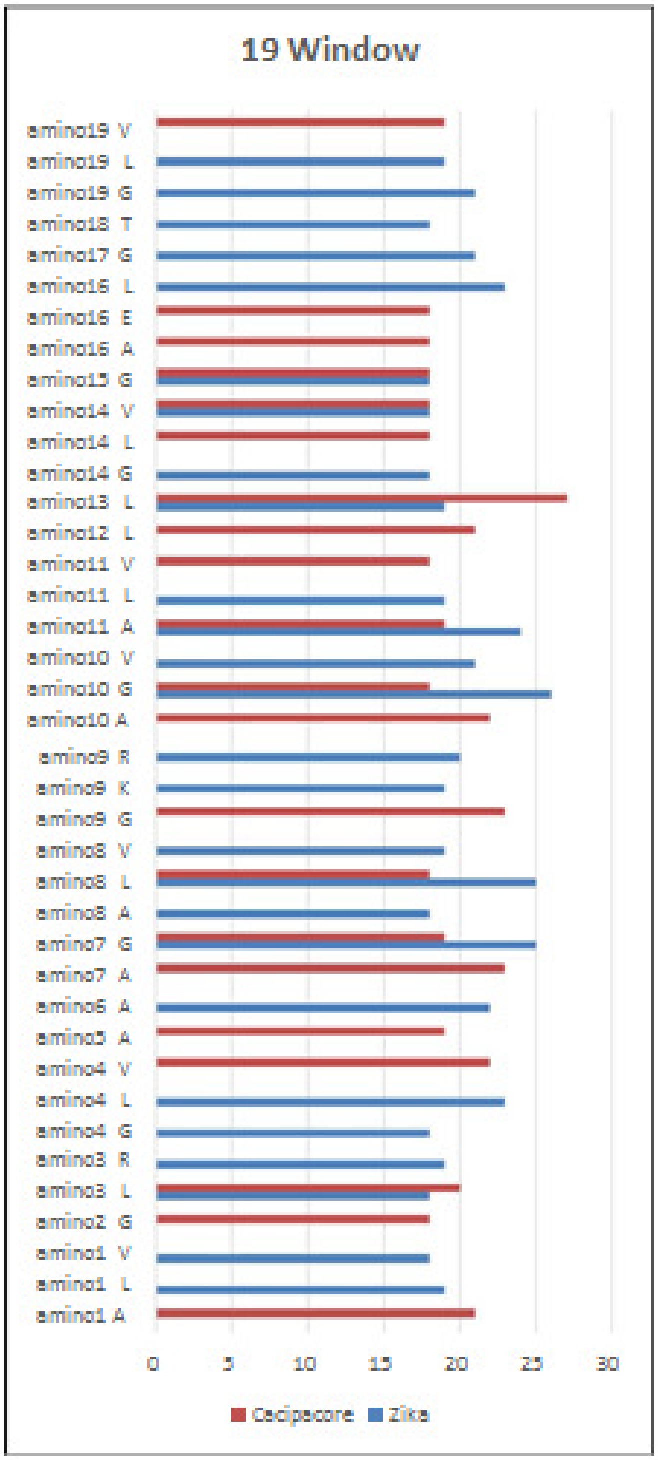

Fig. 3 is the result of amino acid sequences of the viruses using apriori algorithm window-19. It shows amino 1 to amino 19, so it's attribute is 19(excluded cleavage). Number of best rules found in each Cacipacore and Zika virus are 21 and 27. Also viruses have same number of amino15 G and amino14 V.

Table 1. Overall Analysis of Similar Amino Acid Rules

	Similar Amino Acid Rules in Each Viruses
13 Window	amino6 L, amino10 G, amino13 L
17 Window	amino3 L, amino8 V, amino10 G, amino12 L, Amino17 L
19 Window	amino3 L, amino14 V, amino15 G

This is the highest simillar rules between two virues. In 13 window, amino10 G is 30 in Zika and 26 in Cacipacore. In 17 window, amino 8 V is same as 21 and amino 17 L has 1gap between the two viruses. In 19 window amino14 V and amino15 G is same as 18.

3.2 Decision tree

Decision tree allows to figure out the major rules to verify the similarities between two different classes and allows to analyze the amino acids in each viruses [14]. 10-fold cross validation was used for experiment. I set the condition as 13, 17 and 19 windows, and the experimental data is divided into 10 folds, and 10 trials of experiments are held for accuracy. In addition, I set 0.833 as the minimum rate of frequency for getting meaningful data. So, the data which has less than the frequency of 0.833 are abandoned. Frequency refers to the possibility of being classified into certain class.

Table 2. Rule extraction under 13window

virus	**Rule**	**Frequency**
Zika	pos1 = N pos12 = K	0.833
Cacipacore	pos6 = N pos8 = K	0.857
	pos1 = S pos6 = R	0.833

According to Table 2, the result of rule extraction under 13 window, there are no specific experimental features. It shows both Zika and Cacipacore have amino acid N and K. They have position1 but amino acid sequence was different.

Table 3. Rule extraction of 17window

virus	**Rule**	**Frequency**
Zika	pos16 = V pos17 = L	0.833
Cacipacore	pos4 = Lpos17 = R	0.833
	pos6 = Gpos17 = P	0.833

According to Table 3, It's noticeable that two viruses shown their rules extracted with amino acid mostly at position 17. I assume that position 17 is the important factor which differentiate each other. The highest frequency is same as 0.833 and they have amino acid L.

Table 4. Rule extraction under 19 window

virus	Rule	Frequency	
Zika	pos14= Q pos7 = V	0.833	
	pos1 = T pos2 = L	0.833	
	pos3 = T pos14 = G	0.833	0.833
	pos4 = V pos8 = L	0.833	0.833
Cacipacore	pos4 = G pos11 = A	0.875	
	pos2 = S pos12 = G	0.857	
	pos4 = L pos8 = L	0.857	
	pos3 = A pos15 = T	0.857	
	pos4 = V pos8 = G	0.857	
	pos2 = A pos18 = L	0.833	
	pos3 = L pos9 = V	0.833	

According to Table 4, Cacipacore virus has various rules and frequencies while Zika has same frequency as 0.833. Cacipacore has frequency 0.857 the most. Pos8=L was extracted in both of them. In Cacipacore half of the rules include position4 and pos4, 3, 2, 8 was frequent than other positions. Zika also includes position 4, 3, 2, 8, though amino acids are different.

4. DISCUSSION AND CONCLUSION

I measured the frequency of Zika and Cacipacore viruses' amino acid sequences using apriori algorithm and decision tree. Experiments were separated into 3(13-window, 19-window) and were done on each viruses. As a result of an apriori algorithm, all windows showed L(leucine) and G(Glycine) as a common amino acid. Analyzing common amino acids of two viruses, V(valine) which wasn't exist in 13window appeared in 17window and A(Adenine) that wasn't exist before is showed in 19window. By the decision tree it is found that 19window showed more rule extractions than others. So it seems that high window shows more detailed information. In decision tree there was no common amino acid of all windows but 17 and 19 window showed high frequency of G, L, V as apriori algorithm. Wefound Zika and Cacipacore virus are quite similar and have different features. In the process of researching and developing vaccines for Flavivirus which cause serious disease for human, more studies should be progressed. Precedent studies showed many other Flavivirus such as West Nile and Dengue virus are also correlated. Analyzing virus we look forward to finding the means of curing Flavivirus.

REFERENCES

[1] Dick, G.w.s, S.f Kitchen, and A.j Haddow. Transactions of the Royal Society of Tropical Medicine and Hygiene 46.5 209-20 (1952)

[2] Dick, G.w.a. Transactions of the Royal Society of Tropical Medicine and Hygiene 46.5 521-34 (1952)

[3] M. Robert W., J. Homan, M. V. Callahan, J. Glasspool- Malone, L. Damodaran, A. D. B. Schneider, R. Zimler, J.Talton, R. R. Cobb, I. Ruzic,J. Smith-Gagen, D. Janies, and J. Wilson. PLoS Negl Trop Dis PLOS Neglected Tropical Diseases 10.3 (2016)

[4] Mcarthur, Monica A., Marcelo B. Sztein, and Robert Edelman. "Dengue Vaccines: Recent Developments,Ongoing Challenges and Current Candidates." Expert Review of Vaccines 12.8 933-53 (2013)

[5] M. Milton, F. Da Silva Pereira Cruz, M. T. Cordeiro,M. A. Da Motta, K. M. De Melo Cassemiro, R. De Cassia Carvalho Maia, R. C. Bressan Queiroz De Figueiredo,
R. Galler, M. Da Silva Freire, J. T. August, Ernesto T. A.Marques, and Rafael Dhalia. PLoS Negl Trop Dis PLOS Neglected Tropical Diseases 9.4 (2015)

[6] Hayes, Edward B. Emerg. Infect. Dis. Emerging Infectious Diseases 15.9 1347-350 (2009)

[7] B. Cécile, M. Jimenez-Clavero, A. Leblond, B. Durand, N. Nowotny, I. Leparc-Goffart, S. Zientara,E. Jourdain, and S. Lecollinet. International Journal of Environmental Research and Public Health IJERPH 10.11 6039-083 (2013)

[8] Winkelmann, R. Evandro, H. Luo, and T. Wang.F1000Research F1000Res (2016)

[9] M. Ecker, S. L. Allison, T. Meixner, and F. X. Heinz. Journal of General Virology 80.1 179-85 (1999)

[10] P. Vasiliki, S. Geroy, E. Diza, A. Antoniadis, and A.Papa. Vector-Borne and Zoonotic Diseases 7.4 611-16 (2007)

[11] B. Matthias, D.A. Groneberg, D. Klingelhoefer, and A. Gerver. Parasites & Vector Parasit Vectors 6.1 331(2013)

[12] Gardner, L. Christina, and K. D. Ryman. Clinics in Laboratory Medicine 30.1 237-60(2010)

[13] E.B. Go, S.M. Lee, and T.S. Yoon. International Journal of Machine Learning and Computing IJMLC 543-46 (2014)

[14] H.S. Kim, J.Y. Yoo, and T.S. Yoon. International Journal of Computer and Communication Engineering IJCCE 294-301(2015)

[15] R. Rodriguez-Roche, and Ernest A. Gould. BioMed Research International 2013 1-20 (2013)

[16] Y. S. Kim, S. M. Kim, J. W. Lee, J. Ann , J. Lim, and T. S. Yoon. Intelligent Computing Theories and Methodologies Lecture Notes in Computer Science 426-35 (2015)

[17] Jang, S. P., K. H. Park, Y. L. Kim, H. N. Cho, and T.S. Yoon. Journal of Biosciences and Medicines JBM 03.06 49-53(2015)

[18] Y.J. Yang, B.K. Gu, and T.S. Yoon. MATEC Web of Conferences 69, 01005 (2016)