An Analysis of the Genomes of Dengue Virus Using Decision Tree and Apriori Algorithm

Hyunseong Kim*, Juyoung Yoo, Taeseon Yoon
Department of International, Hankuk Academy of Foreign Studies, Yongin, Gyeonggi-do, South Korea.

* Corresponding author. Tel.: 82-10-4228-9033; email: kiml11@daum.net
Manuscript submitted March 28, 2015; accepted September 23, 2015.
doi: 10.17706/ijcce.2016.5.4.294-301

Abstract: Dengue fever, caused by the dengue virus, has been a widespread epidemic during the 21st century. A mosquito-borne RNA virus, the dengue virus has four serotypes; all are able to cause the disease. Vaccination for the virus is arduous, since the vaccine must be able to immunize all four serotypes. In order to investigate genomic similarities and differences, we analyzed the genomes of the four serotypes: Dengue virus 1, Dengue virus 2, Dengue virus 3, and Dengue virus 4. We investigated the positions on each genome that had significant differences by using the decision tree. We also tried to find the similarities of the four viruses with the apriori algorithm. Through our experiment, we were able to investigate both the genomic similarities and the differences of each serotype, and were able to reach an interesting conclusion that the viruses, though they possess certain similarities, have an unusually large number of differences amongst themselves.

Key words: Apriori algorithm, decision tree, dengue virus, Leucine.

1. Introduction

Dengue fever, an epidemic which led to forty-thousand deaths in Spain during the 18th century, has once again devastated many regions in the world. The disease threatened the Indian capital of Delhi in 2006, with 448 cases reported, and also severely hit the Philippines in 2008, with a total of 15,061 cases reported. Dengue fever is caused by the dengue virus, an RNA virus of the genus *Flavivirus* that can be transmitted through the bites of infective female *Aedes*mosquitoes. Vaccination is difficult, due to the fact that the vaccine must immunize all four serotypes; the immunization of only one serotype can lead to severe Dengue fever when infected by another serotype as a result of Antibody-Dependent Enhancement. An investigation on the genomic similarities amongst the four serotypes may aid the development of a vaccine that can protect against all four serotypes. This paper investigates the similarities and the differences of the genomic sequences.

2. Materials and Procedure

Materials that were used in the experiment include the genomic sequences of Dengue Virus 1 (DENV 1), Dengue Virus 2 (DENV 2), Dengue Virus 3 (DENV 3), and Dengue Virus 4 (DENV 4), which were all taken from the genetic database of the National Center for Biotechnology Information (NCBI). Tools that were used for the procedure include decision tree and apriori algorithm.

2.1. Dengue Virus

Dengue virus is a small RNA virus that belongs to the genus Flavivirus. Flaviviruses are transmitted through the bite of an infected arthropod. In the case of dengue virus, the virus is transmitted through mosquitoes, and is more efficient in infecting urban mosquitoes such as *Ae. Aegypti* and *Ae. Albopictus* [1]. Symptoms appear 3-14 days after infection. A high fever is followed by severe headache, pain behind the eyes, and pain in the muscle joints [2]. Symptoms for the fever can be distinguished from other diseases by checking clinical features such as variety of cutaneous signs, pulse rate and the presence of pharyngeal injection [3]. Prevention for the disease can only be done by not being bit by mosquitoes, and treatment can only be done by typical flu treatments, which include rest and pain relievers.The genomes for dengue virus encode only ten proteins. Three of these proteins function as the coat and deliver the RNA to the host, while the other seven serve to replicate the virus once the host is infected [4]. The virus targets cellular receptors in human monocytes. The host cell provides surface receptors, shows endocytic activity, and triggers signals for penetration [1]. The E glycoprotein is responsible for virion attachment to the receptor and fusion of the virus envelope with the target cell membrane.

2.2. Comparison of the Four Dengue Virus Serotypes

Four serotypes of dengue virus exist. The four subtypes have come into existence by evolving independently by adapting to peridomestic mosquito vectors and human reservoir hosts [1]. The subtypes of dengue virus have 60-80% homology amongst themselves. The underlying difference between the subtypes is in the surface proteins. First infections by one serotype cause minor symptoms and immunity develops for the serotype. However, secondary infections caused by another serotype result in harsh symptoms. The four serotypes differ in severity of symptoms. One study conducted by Halsey et. al demonstrates the differences in the symptoms among the four serotypes. DENV 2 and DENV 3 have a higher prevalence of malaise compared to the other serotypes. Those infected with DENV 2 experienced more abnormal pain. Those infected with DENV 3 had a higher prevalence of musculo-skeletal and gastrointestinal symptoms. These individuals faced nausea, abdominal pain, vomiting, and diarrhea. Those infected with DENV 1 had an increased prevalence of rhinorrhea. People infected with DENV 4 had a higher prevalence of cutaneous and respiratory symptoms, and an increased prevalence of pharyngeal congestion [5].

2.3. Decision Tree

A decision tree is a structure that poses a series of questions about the features associated with data items. Each question is contained within a node, and each internal node points to its child node which contains possible answers to the question. The internal node is usually named with the name of an input feature. Each leaf of the tree is labelled with a class or a probability distribution. A decision tree is thus grown by continuously adding questions [6]-[8]. A tree is termed "learned" by dividing the source set into subsets which usually branches down recursively. This sort of learning is termed as recursive partitioning. This process of a top-down induction of decision trees is a part of "greedy algorithm", and is one of the most common strategies for learning decision trees [9]. Data from decision trees come in the form: $(X, Y) = (x_1, x_2, x_3,..., x_k, y)$ when the decision tree is expressed as a combination of mathematical and computational techniques. Here, the dependent variable Y needs to be understood, classified, or generalized. The independent variable X is composed of input variables [9].

2.4. Apriori Algorithm

The Apriori algorithm is an influential algorithm that is frequently used for mining frequent itemsets that is contained in boolean association rules. This algorithm is for frequent item set mining and association rule

learning over transactional databases. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database which has applications in domains such as market basket analysis. As mentioned above Apriori is designed to operate on databases containing transactions which include collections of items bought by customers. On the other hand, other algorithms are constructed to find association rules for data that has no transactions. This algorithm uses a "bottom up" approach where frequent subsets are prolonged one at a time, which is also generally named the candidate generation. Each group of candidates is tested repeatedly until no more extensions are found. Utilizing the breadth-first search and a Hash tree structure, this algorithm counts candidate items efficiently by proceeding a few steps. First, it generates candidate item sets of length k from item sets of length k-1, which make candidates contain an infrequent sub patterns. As a result, according to the downward closure lemma, the candidate set will contain all frequent k-length item sets. In the final step, it scans the transaction database to determine frequent item sets among the candidates [10].

2.5. Procedure

We extracted the genome sequences of DENV 1, DENV 2, DENV 3, DENV 4 from the National Center for Biotechnology Information (NCBI). For the decision tree, we used the sequences for a 10-fold cross validation experiment held with 4 classes (DENV 1, DENV 2, DENV 3, DENV 4). We extracted data with very high frequency rates; only data with frequency rates higher than 0.75 were chosen. However, due to the excessiveness of the data, we analyzed data with frequencies higher than 0.83. For the Apriori Algorithm, we extracted the data, and then graphed the data to compare the trends between the four viruses.

3. Results

3.1. Decision Tree

Table 1. Rule Extraction under 9 Window (1)

Class (virus)	DENV 1	DENV 2	DENV 3	DENV 4
DENV 1	65	162	74	76
DENV 2	83	150	73	71
DENV 3	76	153	62	86
DENV 4	94	136	93	54

Several rules with frequencies above 0.75 were found, according to Table 1. Most notable was the number of rules found with comparisons between DENV 2 and the other DENV viruses. This may indicate that DENV 2 may be more different than the other dengue viruses.

Table 2. Rule Extraction under 9 Window (2)

virus	Rule	Frequency	Rule	Frequency
DENV 1	pos1 = L pos4 = I	0.83	pos4 = Q pos7 = W	0.83
	pos1 = G pos2 = T	0.875		
DENV 2	pos2 = I pos4 = T	0.83		
DENV 3	pos1 = E pos2 = L	0.83	pos2 = K pos3 = A	0.83
	pos2 = A pos9 = K	0.80		
DENV 4	pos4 = R pos7 = V	0.83	pos3 = Y pos9 = G	0.83
	pos2 = R pos4 = G	0.83	pos2 = V pos3 = K	0.83
	pos2 = S pos7 = E	0.83	pos3 = E pos4 = M	0.83
	pos2 = M pos7 = S	0.83		

Using Table 2, it can be assumed that position 2 is an important factor that differentiates the viruses since position 2 is observed most frequently. Positions 1 and 4 can also be inferred to be important factors since they are also observed frequently among the viruses.

Table 3. Rule Extraction under 13 Window (1)

Class(virus)	DENV 1	DENV 2	DENV 3	DENV 4
DENV 1	84	45	83	49
DENV 2	105	29	82	45
DENV 3	94	37	67	63
DENV 4	96	38	90	37

Table 4. Rule Extraction under 13 Window (2)

virus	Rule	Frequency	Rule	Frequency
DENV 1	Not extracted		Not extracted	
DENV 2	pos2=G pos9=F	0.83	pos7=L pos12=P	0.83
DENV 3	pos8=G pos12=A	0.86	pos2=A pos8=S	0.86
	pos4=R pos7=T	0.83	pos8=A pos9=I	0.83
DENV 4	pos2=T pos8=S	0.83	pos2=G pos6=G	0.86
	pos3=V pos9=D	0.83	pos9=R pos12=S	0.83

Several rules were also found for the 13 window. However, in comparison to the 9 window, fewer rules were found. Furthermore, unlike Table 1, Table 3 showed more rules when DENV 1 was compared with the other viruses. Surprisingly, DENV 2 showed the least amount of rules when compared with the other viruses. In Table 4, it can be seen that the extremely high standards we set for the frequency caused no rules to be extracted for DENV 1. However, there were several rules for DENV 1 that exceeded a frequency of 0.8, but for consistency, we maintained a standard of 0.83. It can be seen in Table 4 that position 2 was once again the most frequent rule, like Table 2. Yet, it can also be seen that position 9 had rules for all 3 viruses.

Table 5. Rule Extraction under 17 Window (1)

Class (virus)	DENV 1	DENV 2	DENV 3	DENV 4
DENV 1	40	54	53	53
DENV 2	54	54	47	45
DENV 3	47	45	52	56
DENV 4	50	43	59	48

Table 6. Rule Extraction under 17 Window (2)

virus	Rule	Frequency	Rule	Frequency
DENV 1	pos3=A pos5=N	0.83	pos5=N pos15=R	0.83
DENV 2	pos4=I pos5=K	0.86	pos4=I pos8=P	0.83
DENV 3	pos5=C pos11=L	0.83	pos10=T pos13=V	0.83
DENV 4	pos4=L pos16=N	0.83	pos5=Q pos12=T	0.83

Unlike the 9 window and 13 window results, there was no particular virus that showed more rules than the other viruses. It can be seen in Table 5 that the number of rules for each virus is similar.

The 17 window also showed different results for the specific rules, as can be seen in Table 6. While Tables 2 and 4 had position 2 as the important position, Table 6 has position 5 as the important position.

3.2. Apriori Algorithm

The following are the results for the dengue virus strands obtained from the apriori algorithm. The apriori algorithm, unlike the decision tree, finds similarities between the information given. In this case, it would find the similarities between the viral strands.

Table 7. Apriori 9 Window Rules

Type	DENV 1	DENV 2	DENV 3	DENV 4
Rules	amino1=L 42	amino3=L 44	amino6=L 45	amino9=L 44
	amino3=L39	amino1=L40	amino2=L43	amino8=G40
	amino8=L 39	amino9=G39	amino9=G41	amino1=L39
	amino9=G39	amino4=G38	amino8=L 39	amino2=L39=
	amino1=A 38	amino5=L38	amino4=T38	amino3=G38
	amino2=T38		amino7=T 38	amino5=T38
			amino8=T38	amino7=G38

Table 8. Apriori 13 Window Rules

Type	DENV 1	DENV 2	DENV 3	DENV 4
Rules	amino1=L 32	amino2=L 30	amino3=L 34	amino9=G 34
	amino7=V 32	amino6=L 30	amino1=L 31	amino12=L 33
	amino12=G 29	amino13=L 30	amino2=A 29	amino3=L 32
	amino9=T 28	amino4=L 28	amino8=G 28	amino10=L 31
	amino6=L 27	amino10=G 28	amino10=G	28amino13=G 28
	amino13=L 27	amino12=G 28	amino11=L 28	amino11=G 27
	amino3=G 26	amino7=L 27	amino12=G 28	amino2=L 26
	amino4=T 26	amino11=G 27	amino7=E 27	amino3=V 26
	amino8=L 26		amino7=T 27	amino5=L 26
	amino10=L 26		amino8=L 27	amino10=T 26
	amino11=T 26		amino12=L 27	
	amino13=G 26		amino6=L 26	
			amino8=T 26	
			amino9=G 26	
			amino13=G 26	

Table 9. Apriori 17 Window Rules

Type	DENV 1	DENV 2	DENV 3	DENV 4
Rules	amino13=L 29	amino5=T 25	amino10=L 25	amino10=L 27
	amino6=L 26	amino16=G 25	amino17=L 24	amino15=G 25
	amino8=G 25	amino9=A 23	amino7=L 23	amino1=L 24
	amino2=L 24	amino11=L 23	amino8=G 23	amino7=G 24
	amino12=L 24	amino12=L 23	amino9=L 23	amino7=L 24
	amino15=G 23	amino15=G 23	amino3=L 22	amino12=L 24
	amino17=T 23	amino2=L 22	amino4=L 22	amino1=T 23
	amino3=L 22	amino8=G 22	amino6=G 22	amino4=L 23
	amino16=A 22	amino8=L 22	amino1=L 21	amino4=T 23
	amino4=A 21	amino13=L 22	amino3=T 21	amino13=L 22
	amino4=T 21	amino15=L 22	amino4=E 21	amino2=V 21
	amino15=L 21	amino3=V 21	amino9=T 21	amino14=G 21
	amino8=I 20	amino5=E 21	amino6=A 20	amino14=L 21
	amino11=K 20	amino10=T 21	amino8=T 20	amino5=L 20
	amino11=L 20	amino15=T 21	amino13=G 20	amino8=G 20
	amino17=G 20	amino5=L 20	amino15=G 20	amino11=G 20
		amino6=L 20		
		amino10=K 20		

We selected the most representative rule of each of the four viruses for the 9 window, 13 window, and 17 window (showing in Tables 7-9). We did this in order to find the similarities between the viruses. The following Fig. 1-Fig. 3 are the results.

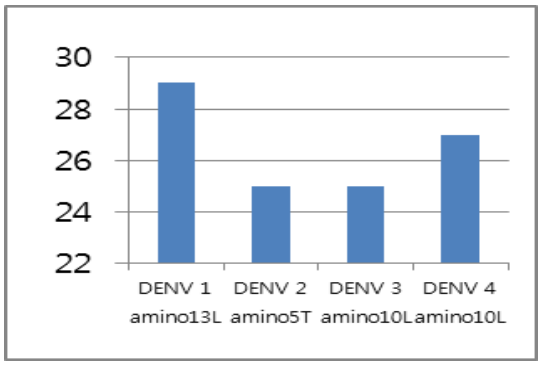

Fig. 1. Apriori 9 window.

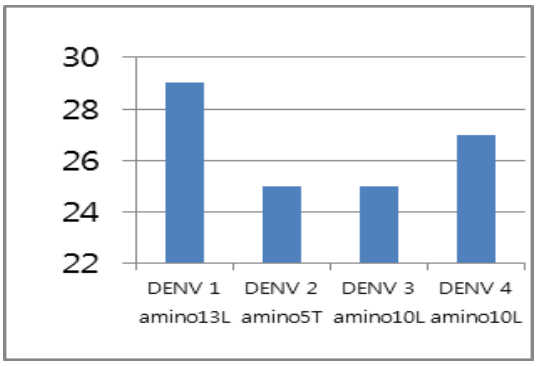

Fig. 2. Apriori 13 window.

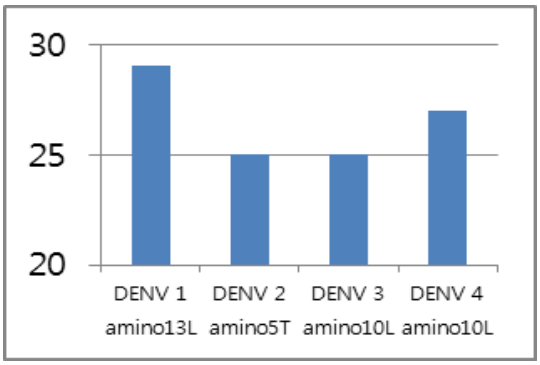

Fig. 3. Apriori 17 window.

We were able to find a total of 136 rules for the apriori algorithm analysis. Compared to the apriori algorithm results from ebola virus, extracted by Go *et al*. [9], which is about 245, dengue virus had a significantly fewer number of rules. This indicates that dengue virus has much less similarities in comparison to other viruses. This may indicate why it is difficult to find a universal vaccine for all four virus strands. Considering that vaccination for one strand leaves an infected organism highly vulnerable to a second infection, the road to finding a universal vaccine may prove to be harder than expected. However, even with the small number of similarities, we were able to find specific amino acids that showed up in each virus. That was leucine. How leucine affects the behavior of Dengue virus was not investigated, but would prove to be a fruitful investigation if conducted. Similar to the data from the decision tree, we did not investigate the specific biological implications of our results. The data from the apriori algorithm provides suggestions towards further investigation on the Dengue virus.

4. Analysis and Discussion

The decision tree data showed several rules. Even when we increased the standard to a frequency of 0.83, we were able to find several rules. Usually, rules that exceed a frequency of 0.75 are acceptable, and we were able to find over one hundred rules for each window. This indicates that the 4 dengue viruses have many differences amongst themselves. This data shows why it is difficult to find a vaccine that cures all four viruses. There are too many differences amongst the viruses that immunity against one strand will not create immunity against another strand. Differences among dengue viruses are greater than those of other

viruses, such as Ebola viruses, as can be seen from the decision tree data provided by Go *et al.* [9]. Our data supports the fact that it is difficult to find a vaccine for dengue virus. With the decision tree data, however, we found that position 2 in dengue viruses presented many rules for each window. Position 2 may be an important factor in distinguishing the 4 dengue viruses. Further experiments on this specific position may allow scientists to find the specific differences between dengue viruses. The data extracted from the apriori algorithm showed that the amino acid leucine was most common among the four viruses. How leucine in dengue virus affects the human body has not been experimentally conducted. Further investigation on the affects of leucine in dengue virus may help lead to vaccine development.

5. Conclusion

The decision tree data showed several rules. Even when we increased the standard to a frequency of 0.83, we were able to find several rules. Usually, rules that exceed a frequency of 0.75 are acceptable, and we were able to find over one hundred rules for each window. This indicates that the 4 Dengue viruses have many differences amongst themselves. This data shows why it is difficult to find a vaccine that cures all four viruses. There are too many differences amongst the viruses that immunity against one strand will not create immunity against another strand. Differences among Dengue viruses are greater than those of other viruses, such as Ebola viruses, as can be seen from the decision tree data provided by Go *et al*. [9]. Our data supports the fact that it is difficult to find a vaccine for Dengue virus. With the decision tree data, however, we found that position 2 in Dengue viruses presented many rules for each window. Position 2 may be an important factor in distinguishing the 4 Dengue viruses. Further experiments on this specific position may allow scientists to find the specific differences between Dengue viruses. Finding a vaccine for Dengue virus, and creating a treatment for Dengue virus has been an arduous task. The data we extracted from the decision tree further substantiated this phenomenon. Yet, we were able to find certain similarities in the viruses, and also found that certain positions were significant factors. We believe the data we extracted will provide help for further investigations and experiments on Dengue virus.

References

[1] Seema, & Jain, S. K. (2005). Molecular mechanism of pathogenesis of dengue virus: Entry and fusion with target cell. *Indian Journal of Clinical Biochemistry*, *20(2)*, 92-103.

[2] *Dengue Fever*. (2014). Northern Territory Government.

[3] Chadwick, D., Arch, B. N., Wilder-Smith, A., & Paton. N. (2005). Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: Application of logistic regression analysis. ChesterRep.

[4] Goodsell, D. S. (2008). *Molecule of the Month: Dengue Virus*. Protein Data Bank.

[5] Halsey, E. S., Morgan, A. M., Gotuzzo, E., Fiestas, V., Suarez, L., Vargas, J., Aguayo, N., Madrid, C., Vimos, C., Kochel, T. J., & Laguna-Torres, V. A. (2012). Correlation of serotype-specific dengue virus infection with clinical manifestations. In S. K. Singh (Ed.), *PLoS Neglected Tropical Diseases*.

[6] *Decision Trees T,* pp. 1-18.

[7] Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? *Nature Biotechnology, 26(9)*, 1011-1013.

[8] Mitchell, T. M. (1997). *Machine Learning*. New York: McGraw-Hill.

[9] Go, E., Lee, S., & Yoon, T. (December 2014). Analysis of Ebolavirus with decision tree and Apriori algorithm. *International Journal of MachineLearning and Computing, 4(6)*.

[10] Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th VLDB Conferencee* (pp. 1-32). Santiago, Chil.

Hyunseong Kim was born in Seoul, South Korea and currently attends the International Department of the Hankuk Academy of Foreign Studies. He has a strong interest in chemistry and biochemistry. Currently, he is undergoing an intensive physics course in order to gain a stronger background on chemistry. He wishes to get a career as a professor or as a researcher for medical firms that research medication and vaccines. This interest led to his publishing of the following paper.

Juyoung Yoo was born in Seoul, South Korea and currently attends the International Department of the Hankuk Academy of Foreign Studies. He has a strong interest in chemistry. He wishes to become a professor for a university.

Taeseon Yoon was born in Seoul, Korea, in 1972. He got a 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.