International Proceedings of Chemical, Biological and Environmental Engineering, Vol. 97 (2016)
DOI: 10.7763/IPCBEE. 2016. V97. 12

Comparison of amino acid sequence of FMDV between East Asia Countries: China, South Korea, North Korea, Mongolia, Japan by using Decision Tree and Apriori Algorithm

Sung jin Kim [1], Seung Hye Song [1] and Taeseon Yoon[1 +]

[1] Hankuk Academy of Foreign Studies, South Korea

Abstract. Every year, considerable amount of FMDV infected bovids, suids, ovids, caprids and other cloven-hoofed animals are buried as a temporary solution of the disease. However, due to this burial, secondary problem such as pollution of drinking water arises and efforts to solve this secondary problem are influencing our whole society. In addition, not only in Korea, FMDV is worldwide problem of each country due to the impact of the virus. So in order to find better ways to cope with the virus or to develop a cure, we decided to genetically compare FMDV in 5 East Asia countries: China, South Korea, North Korea, Mongolia, and Japan. We used apriori algorithm to find similarities among the five subjects and used decision tree to find out their distinctive characteristics. In the results of apriori algorithm, we could categorize the subjects into two groups: one consists of Mongolia and North Korea, and the other consists of China, Japan, and Korea. Each group members shared significantly similar features in their groups. For example, In 9window and 17window results, Mongolia and North Korea identically had R as their main amino acids, while China, Japan, and Korea had L and T as their main amino acid in 9window and 17window results. In the results of decision tree, we could not find any accordance of most frequent amino acid among the fice subjects, nor accordance of their positions. We assumed that although all the subjects were from East Asia, each of the five subjects had quite distinctive feature from country to country. According to the results, we were able to expect that similar traits found between three countries, might have been the result of the geological nearness and we were able to predict that these factors made interaction between 3 country's FMDV possible, which made their traits to be similar. By this, we suggest that research teams from the three countries cowork in their study of FMDV and develop a basic vaccine which reflects the common features of FMDV from each nation. And after that, they can specialize those basic vaccine based on the distinctive traits they could find in each FMDV. By this process, all three nations will be able to develop higher quality of vaccine and make more adequate cure for FMD of each country.

Keywords: FMDV, East Asia, China, South Korea, North Korea, Mongolia, Japan, Decision Tree, Apriori Algorithm.

1. Introduction

The foot-and-mouth disease virus (FMDV) is the pathogen that causes foot-and-mouth disease which accompanies symptoms such as vesicles (blisters) in the mouth and feet of bovids, suids, ovids, caprids and other cloven-hoofed animals[1], [2]. Since they are highly infectious and a major plague of animal farming, every year due to FMDV many stockbreeding farmhouse get mental and financial harm. By research we were able to find out that foot-and-mouth disease virus occurs in seven major serotypes which are O, A, C, SAT-1, SAT-2, SAT-3, and Asia-1. And since FMDV is RNA virus, like other RNA viruses it exhibits high mutation rates during replication[3], [4]. We were able to find out that there are many kinds of mutation even though it's FMDV outbroke in the same region.

Since O serotype is most common, by comparing the breakout during 2010-2011 we decided to compare FMDV O serotype amino acid sequence of East Asia countries: China, South Korea, North Korea, Mongolia,

+ Corresponding author. Tel.: + 821056967108; fax: +82313240700.
E-mail address: tsyoon@hafs.kr.

Japan. The genomic data of FMDV O stereotype were provided from NCBI online, which is National Center for Biotechnology Information. By this research we are trying to find out similarities and differences of each country's FMDV amino acid sequence and will try to find out the aspect of transition and predict future mutation of such virus by using a proper algorithm which is decision tree and apriori algorithm

2. Materials and Methods

2.1 FMDV

The foot-and-mouth disease virus (FMDV) is a single-strand RNA picornavirus which belongs to the *Aphthovirus* genus. Its size is mainly 25-30nm, and has icosahedral capsid made of protein but not envelope.[4], [5] The virus causes foot-and-mouth disease which is highly infectious among cloven-hoofed animals including domestic and wild bovids and suids. An individual which caught on the disease gets vesicles in the mouth and feet. The disease is highly infectious and can be spread in following courses: through aerosols, through contact with contaminated object, and by domestic and wild predators[6].

FMDV comes in seven major serotypes: O, A, C, SAT-1, SAT-2, SAT-3, and Asia-1. These serotypes appears to be regional, and the O serotype is most common.

2.2 Apriori and Decision Tree

Apriori[7] is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.

Decision tree is a decision support tool that generates a tree-like graph of decisions, in order to apply it into a certain algorithm[8]. In this study, we used this method to extract the specific patterns of amino acid sequences in each virus. The procedure done is called the 'Data mining.' Data mining refers to extracting patterns from raw or warehoused data, in other words, efficiently extracting previously unknown but potentially useful patterns from data. The extracted patterns can be used to classify the data, understand the characteristics of each data sets, and predict outcomes for future situations. In this process, decision tree is used as a predictive model which decides the value of given data sets.

In both methods, we examined 9,13, and 17window results to analyze properly.

3. Results and Discussion

3.1 Apriori Algorithm

Class1 refers to FMDV O serotype of Korea, class2 refers to FMDV O serotype of North Korea, Class3 refers to FMDV O serotype of Mongolia, Class4 refers to FMDV O serotype of China, and Class5 refers to FMDV O serotype of Japan. In the experiment we used 10-fold cross validation.

Table 1, 2, and 3 shows each results of 9window, 13window, and 17window. In the tables, all amino acids in the sequence of amino acid are shown, from high to low frequency. To see the frequency of amino acids clearly, Fig.1, 2, and 3, which refers to 9window, 13window, and 17window results, are attached below each table.

Table 1: Rule extraction of apriory algorithm(9window)

Nation	**Rule**
China	L 32
Japan	A 35 /L 35
Korea	L 35
Mongolia	R 47 /P 37 /L 30
North Korea	R 46 /P 33 /H 32 /L 32 /G 30 /Q 30

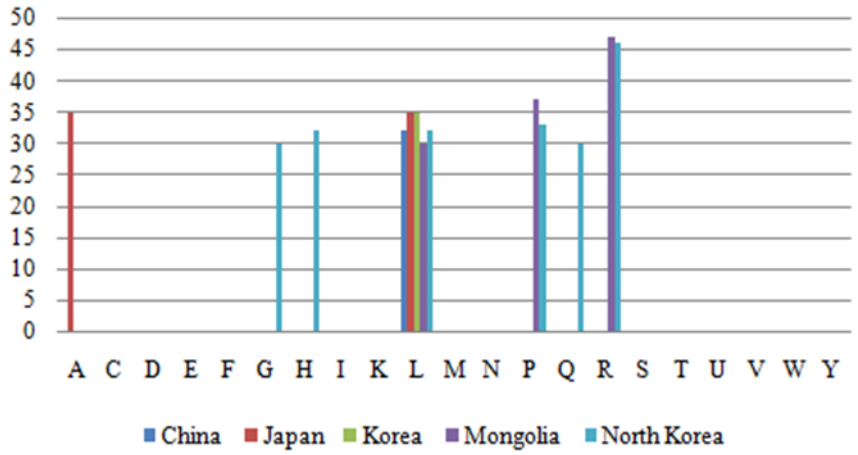

Fig. 1: Main Amino Acid Tendency in 9window result
L was common feature in all the five subjects. P and R were shared exceptionally in Mongolia and North Korea. North Korea solely has G, H and Q.

Table 2: Rule extraction of apriory algorithm(13window)

Nation	**Rule**
China	A 24 /L 24 /V 24
Japan	L 25 /V 25 /A 24
Korea	L 25 /V 25 /A 24
Mongolia	R 35 /H 27 /G 22
North Korea	P 33 /R 29 /H 25 /G 22 /L 21

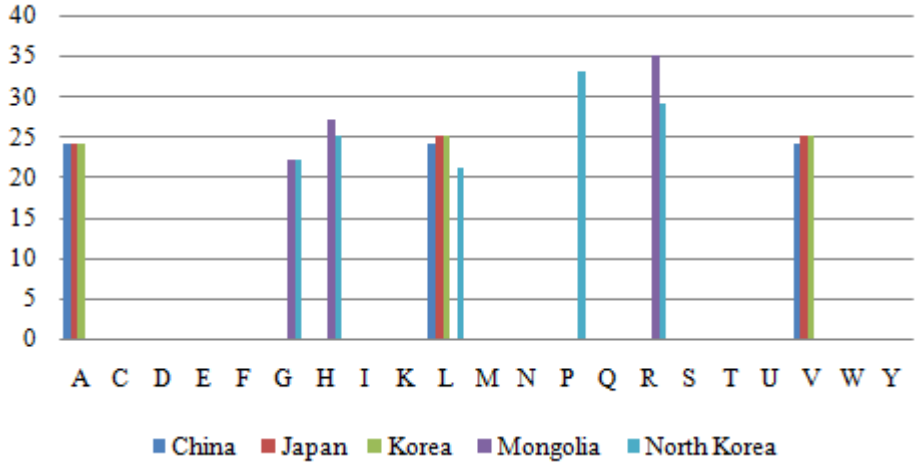

Fig. 2: Main Amino Acid Tendency in 13window result
L was shown in all subjects, except for Mongolia. Clear distinction between the two groups is shown: China, Japan, and Korea group and Mongolia and North Korea group. A and V were common features among China, Japan, and Korea, but were not found in Mongolia and North Korea. Mongolia and North Korea had G, H, R as their common features, which were not shown in China, Japan and Korea.

Table 3: Rule extraction of apriory algorithm(17window)

Nation	**Rule**
China	T 21 /A 20 / L 20 /G 19 /V 18 / R 16
Japan	T 21 /A 20 /G 20 /V 19 /L 18
Korea	T 21 /A 20 /G 20 /V 20 /L 18
Mongolia	R 28 /H 20 /L 20
North Korea	R 25 /G 22 /H 22 /L 22 /P 16

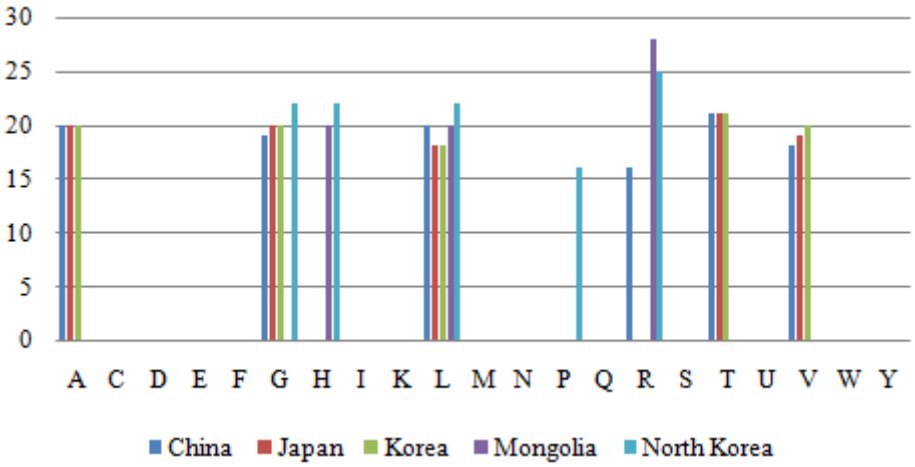

Fig. 3: Main Amino Acid Tendency in 17window result

L was common feature among the five subjects. China, Japan, and Korea shared A, T, and V, which were not found in Mongolia and North Korea. Mongolia and North Korea shared H and R, which were not found in China, Japan, and Korea.

Using apriori algorithm, we could find out the traits of each subject FMDV according to the country of outbreak and compare one another with their extracted amino acids.

In the results, there was common characteristic across the subjects. Amino acid L was found in all countries FMDV in all three windows experimented, except for Mongolia in the 13window results (Table 2).

However, by arranging extracted amino acid and their frequency value, we were able to assort FMDV into two groups by their similarities; one consists of Mongolia and North Korea, and the other consists of China, Japan, and Korea.

Similarities in each group are revealed in main amino acids of 9window and 17window results. In Table 1, which is 9window results, North Korea and Mongolia commonly had R as their most frequent amino acid while China, Japan, and Korea had L as their most frequent one. In Table 3, which is 17window results, R was also found to be the main amino acid of the North Korea and Mongolia, while other three, China, Japan, and Korea had T as their main amino acid.

The similarities were also found not only in their most frequent amino acids, but also in their generally extracted amino acids. North America and Mongolia commonly had P in 9window(Table 1), H and G in 13window(Table 2), and H in 17window(Table 3). However, these common features are not found in other three subjects, China, Japan, and Korea.

China, Japan, and Korea were also found to share similar characteristics. In 13window results, they all had A, L and V at almost the same frequency level between 24 -25(Table 2). In 17window results, A, G, L and V were commonly included and demonstrated similar frequencies(Table 3).

3.1. Decision tree

Table 4 refers to 9window, Table 5 refers to 13window, and Table 6 refers to 17window.

Table 4: Rule extraction of decision tree(9window)

Class	Rule	Frequency
Class1	Pos2=D pos3=L pos7=A	.75
Class2	Pos3=R pos6=P	.75
Class3	Pos2=G pos3=P pos8=H	.75
Class4	Pos1=T pos6=K	.75
Class5	Pos3=A pos7=K	.75

Table 5: Rule extraction of decision tree(13window)

Class	Rule	Frequency
Class1	N/A	N/A
Class2	pos10 = Q pos13 = P	.8
Class3	Pos7=R	.8
Class4	Pos8=S pos13=Y	.8
Class5	Pos5=F pos12=T	.75

Table 6: Rule extraction of decision tree(17window)

Class	Rule	Frequency
Class1	pos11 = V pos15 = I	.75
Class2	Pos3=H pos8 = P	.8
Class3	Pos8=V pos16=S	.8
Class4	Pos8=G pos16=F	.833
Class5	Pos4=I pos8=Y	.75

By using decision tree we tried to find out distinctive traits that can represent each subject. In all 9window, 13window, and 17window results, we could hardly find same position and same amino acid among the subject FMDVs. This allowed us to conclude that although they all broke out in similar region, East Asia, and also the same period of time, they all had respective features that make them effectively distinguished. All of the subjects are considered to possess strong individual characteristics.

4. Conclusions

We were able to conclude that some FMDVs share some similar traits that let us categorize the five subjects into two groups, but that each FMDV has their own distinct characteristic which allow us to distinguish one from another. According to the results of apriori algorithm, we were able to divide 5 countries into 2 groups according to their extracted amino acid : one consisted of Mongolia, North Korea and the other consisted of China, Japan and Korea. The former group shared mainly amino acid R in 9window and 17window results while the latter group shared amino acid L in 9window and T in 17window. Furthermore, subject FMDVs also shared similar amino acid extractions in each group, which were not identified in the other groups. In decision tree results, however, we weren't able to find similar traits among the subjects and we interpreted it as all of them had their own strong distinctive traits.

According to the infection process of FMDV mentioned above, we were able to assume that similar traits found between the FMDV of three countries, China, Japan and Korea, might have been the result of the geological nearness. We assume that interactions between the three countries' FMDV were possible, making their traits similar. By this, we suggest that research teams from China, Japan, and Korea can cowork in their study of FMDV and develop a basic vaccine which reflects the common features of FMDV in three nations. And after that, they can specialize those basic vaccine based on the distinctive traits they could find in each FMDV. By this process, all three nations will be able to develop higher quality of vaccine and make more adequate cure for FMD of each country.

5. References

[1] Martinez-Salas E, Saiz M, Sobrino F (2008). "Foot-and-Mouth Disease Virus". *Animal Viruses: Molecular Biology*. Caister Academic Press. pp. 1–38. isbn=978-1-904455-22-6.

[2] Arzt, J.; Juleff, N.; Zhang, Z.; Rodriguez, L. L. (2011). "The Pathogenesis of Foot-and-Mouth Disease I: Viral Pathways in Cattle". *Transboundary and Emerging Diseases* 58 (4): 291. doi:10.1111/j.1865-1682.2011.01204.x.

[3] Carrillo, C., et al. "Genetic and phenotypic variation of foot-and-mouth disease virus during serial passages in a natural host." Journal of virology 81. **20** (2007): 11341-11351.

[4] Carrillo C, Tulman ER, Delhon G; et al. (May 2005). "Comparative Genomics of Foot-and-Mouth Disease Virus". *J. Virol.* 79 (10): 6487–504. doi:10.1128/JVI.79.10.6487-6504.2005. PMC 1091679. PMID 15858032.

[5] Arzt, J.; Baxt, B.; Grubman, M. J.; Jackson, T.; Juleff, N.; Rhyan, J.; Rieder, E.; Waters, R.; Rodriguez, L. L. (2011). "The Pathogenesis of Foot-and-Mouth Disease II: Viral Pathways in Swine, Small Ruminants, and Wildlife; Myotropism, Chronic Syndromes, and Molecular Virus-Host Interactions". *Transboundary and Emerging Diseases* **58** (4): 305. doi:10.1111/j.1865-1682.2011.01236.x.

[6] Canadian Food Inspection Agency - Animal Products - Foot-and-Mouth Disease Hazard Specific Plan Archived June 5, 2008, at the Wayback Machine.

[7] Rakesh Agrawal and Ramakrishnan Srikant Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487-499, Santiago, Chile, September 1994..

[8] Quinlan, J. R. "Induction of Decision Trees." *Machine Learning* 1.**1** (1986): 81-106..