ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 160-165
doi: 10.18178/wcse.2017.06.028

Data Mining of Influenza A: H3N8, H7N3, and H7N7

Seunghee Han, Jungeun Huh and Taeseon Yoon[+]

[1] Hankuk Academy of Foreign Studies

Abstract. Each type of Influenza A has specific target animals. Among them, H3N8 only infects equine and H7N3 infects only human, but H7N7 infects both species. However, all of them are very fatal to the environment. Thus, by analyzing the amino acid sequence of H3N8, H7N3, and H7N7 using Apriori, Decision tree, K-means, and SVM algorithm, we found differences and similarities among the viruses that contribute to their target animals.

Keywords: Influenza A, apriori, decision tree, K-means, SVM

1. Introduction

There are many kinds of Influenza A, each targeting a specific type of animals. Equine influenza, a type of influenza A, is known to infect horses but not humans. For instance, H3N8 is an equine influenza virus endemic in birds, horses, and dogs but not in humans. However, some strains of viruses in birds and pigs are thought to potentially affect both humans and horses. H7N7 is one of those viruses and it can take away lives of humans and horses. On the other hand, H7N3, a virus with the same hemagglutinin number as H7N7, has the ability to infect birds and humans but does not infect horses. In this paper, the variation among the viruses that leads the difference in target species is verified by analysing the DNA sequences of three viruses, H3N8, H7N3, and H7N7 using the Apriori algorithm, decision tree, K-means, and SVM.

2. Materials

H3N8 virus, a subtype of Influenza A that infects mostly horses, is classified as equine influenza (El). Transmission between human beings and equines has not been reported. H7N3 virus, a subtype of Influenza A virus, infects humans and birds. H7N7 is known to infect humans, birds, and horses. A lot of people were infected by H7N7. One person died due to this virus and many people showed trivial symptoms.

3. Methods

3.1. Window

Window is a regularly divided region of peptide sequence. Figure 1 below is one example of 13-window groups. Original peptide sequence is divided into the windows of fixed size 13. The selection of optimal window size is essential in data analysis since it shows the rules that are the characteristics of the gene. Adequately sized windows ensure high reliability of patterns by eliminating the variability that can be caused by unique window. Too narrow or wide window can cause the miss of genomic features of interest.

3.2. FASTA Format

FASTA format is used frequently in the bioinformatics field. It represents- nucleotide sequences or peptide sequences in a text based form. In other words, it expresses nucleotides and amino acids in a single letter codes. The FASTA for scripting languages like Python, Ruby and Peri.

[+] Corresponding author. Tel.: + (82)10-2325-1642; fax: +(82)31-332-0042.
E-mail address: (seunghee991105@gmail.com,).

3.3. Apriori Algorithm

Apriori is a type of algorithm used to identify the frequency of individual items in the database and the associated relationships among items. It is a "bottom up" approach which extends the frequent subsets one at a time. Apriori is designed to operate on databases containing transactions. By using these measures, the similarity among influenza A could be revealed.

3.4. Decision Tree

Decision tree is a tool to display the decisions and their possible consequences with the chance included. Decision nodes are represented by squares, chance nodes are represented by circles, and the end nodes are represented by triangles. Each internal node shows a test on an attribute, while each branch shows the outcome of the test and each leaf node represents a class label. It is mostly utilized in operations research, and especially useful in identifying a strategy.

3.5. K-means Algorithm

K-means algorithm is a vector quantization method of partitioning the observations into clusters. When n observations are given, K-means algorithm partitions the observations into k($\leqq$n) clusters that maximize the cohesion level of objects within the cluster. In other words, purpose of this algorithm is to find set S that minimizes the square sum of the distances between μ_i and objects of each set S_i, when μ_i is set as a mean of set S_i. The algorithm starts by setting the initial μ_i. After setting the mean, two steps are repeated: assignment step and update step. In assignment step, we calculate the Euclidean distances of the observations to the means of each cluster and assign the observations to the cluster of closest distance. Next, in update step, we reset the center of mass as new mean (μ_i) of the cluster. The procedure is repeated until the cluster does not change.

3.6. SVM

SVM, which is short for support vector machine, is a useful algorithm for classification, prediction and regression problems. It is a supervised learning model and has the ability to locate training examples into a new category. In other words, it is a non-probabilistic binary linear classifier. It could be shown as a group of points in a space and this shows how examples are divided. The data is plotted into n-dimensional space. The number n represents the number of features. Thus, each value of feature is the value of a specific coordinate. Then, the two classes can be differentiated by finding the hyper-plane.

4. Result

4.1. Apriori Algorithm

The genome of H3N8, H7N3, and H7N7 are first analyzed using Apriori algorithm in 5, 7, and 9 windows. Minimum support is set as 0.1 for each window in order to regard the associations that appear in more than 10% of the whole instances as best rules. For example, 78 instances are found in 5 window thus only the rules with over than 8 instances are regarded as best rules. Rules are the tendency of specific amino acids to appear in specific positions, notated as Pos#=N. # is the position number and N is the amino acid. The minimum metric confidence level is 0.9 and 18 cycles are performed in each experiment for reliable analysis.

4.1.1 Apriori Algorithm in 5window

Apriori algorithm in 5window is shown in Table 2. In H3N8, 78 instances are found so only the rules with more than 8 instances are regarded as best rules. 70 instances are found in H7N3 that only the rules with more than 7 instances are regarded as best rules. In H7N7, there are 36 instances that rules with more than 4 instances are selected as best rules. The selected rules are categorized by position number and are organized in the order of largest instances from the top.

Table 1: 5window apriori experiment result (amino acid/instance)

Species	pos1		pos2		pos3		pos4		pos5	
H3N8					G S	9 9	V G S	10 9 9	G	11

H7N3	G L	7 7	E G	7 7	G S	11 8	I T S	8 8 7	G	9
H7N7	A G I	5 5 5	E Q D	5 5 4	G I	5 4	E N T	6 5 5	A E R	4 4 4

Table 2 shows 5window apriori experiment result of three viruses H3N8, H7N3, and H7N7. Many rules involve Glycine in H3N8 and H7N3 with large instances in many positions suggesting that Glycine is essential in the entire genome of H3N8 and H7N3. On the other hand, in H7N7, rules with Glutamic acid and Glycine have relatively large instances but there is no remarkable amino acid that is dominant in the genome.

4.1.2 Apriori Algorithm in 7window

Apriori algorithm in 7window is shown in Table 3. In H3N8, 56 instances are found so only the rules with more than 6 instances are regarded as best rules. 50 instances are found in H7N3 that only the rules with more than 5 instances are regarded as best rules. In H7N7, there are 26 instances that rules with more than 3 instances are selected as best rules. The selected rules are categorized by position number and are organized in the order of largest instances from the top.

Table 2: 7window apriori experiment result (amino acid/instance)

Species	**pos1**		**pos2**		**pos3**		**pos4**		**pos5**		**pos6**		**pos7**	
H3N8	G S	7 6	G	10			V	7	T S	7 6	G	7	S G	7 6
H7N3	G T	8 6	I L	6 5	L A R	6 5 5	T	6	G V A	9 6 5	G S T	8 5 5	S D	7 6
H7N7	G E	6 4	R E T	4 3 3	I V	5 3	D E N	4 4 4	A N R	3 3 3	E Q I	5 4 3	D I Q	3 3 3

Species	Association Rule	Instance	Confidence	Lift
H7N7	pos3=I & pos4=D	3	1	1

Table 3 shows 7window apriori experiment result of three viruses H3N8, H7N3, and H7N7. Again, in H3N8 and H7N3, many rules involve Glycine with large instances in many positions suggesting that G is essential in the entire genome of H3N8 and H7N3. On the other hand, in H7N7, rules with Glutamic acid and Glycine have relatively large instances but there is no remarkable amino acid that is dominant in the genome. Additionally, concurrence of Isoleucine in position 3 and Aspartic acid in position appeared 3 times but the lift value of 1 indicates the irrelevance of the events.

4.1.3 Apriori Algorithm in 9window

Apriori algorithm in 9window is shown in Table 4. In H3N8, 44 instances are found so only the rules with more than 4 instances are regarded as best rules. 39 instances are found in H7N3 that only the rules with more than 4 instances are regarded as best rules. In H7N7, there are 20 instances that rules with more than 2 instances are selected as best rules. The selected rules are categorized by position number and are organized in the order of largest instances from the top.

Table 3: 9window apriori experiment result (amino acid/instance)

Species	**pos1**		**pos2**		**pos3**		**pos4**		**pos5**		**pos6**		**pos7**		**pos8**		**pos9**	
H3N8	G S N Q	7 6 4 4	G R E	7 6 4	A C G T	5 4 4 4	G V	5 4	D V I N S	6 6 4 4 4	V D T	6 4 4	V G K S	9 4 4 4	S I V	6 5 4	T D G S	5 4 4 4
H7N3	R D K L	8 5 5 4	I S T G	5 5 5 4	G L	8 4	G F	7 4	G N S T	4 4 4 4	G P I	6 5 4	D N	4 4	A	6	T L	8 5
H7N7	D N V G L R	5 3 3 2 2 2	E L A I W V	4 3 2 2 2 2	E D F I N	3 2 2 2 2	G V	3 3	G I N	3 3 3	E Q	4 3	S	3	T	3	I	5

Table 4 shows 9window apriori experiment result of three viruses H3N8, H7N3, and H7N7. Again, in H3N8 and H7N3, relatively many instances of Glycine are involved in many positions suggesting that Glycine is essential in the entire genome of H3N8 and H7N3. On the other hand, in H7N7, rules with Glutamic acid and Glycine have relatively large instances but there is no remarkable amino acid that is dominant in the genome.

4.2. Decision Tree

We defined H3N8 as class 1, H7N3 as class 2, and H7N7 as class 3. Class 3 was too small to be compared with the others so the data was doubled and then, experimented.

Table 4: Decision tree experiment result

Species	5 Window Rule	7 Window Rule	9 Window Rule	
H3N8	pos3 = A pos4 = G & pos5 = K pos3 = H pos1 = C	pos4 = W	pos2 = F pos2 = R pos1 = N & pos2 = G pos1 = S	pos4 = P pos9 = D pos1 = C
H7N3		pos2 = I	pos4 = F pos2 = M pos2 = T	pos1 = R & pos2 = G pos1 = R & pos4 = G
H7N7	pos4 = E pos2 = Q & pos4 = A pos3 = I & pos4 = D	pos4 = M pos1 = Y	pos2 = E & pos6 = E pos7 = H pos2 = W	pos1 = V pos1 = L & pos8 = Q pos1 = R & pos2 = E

Table 4 indicates the result of 5, 7, and 9window decision tree algorithm. 5window shows that H3N8 and H7N7 have unique characteristics that can distinguish the two. These characteristics were the rules that had the probability of at least 0.800. This value is high enough to conclude that H3N8 and H7N7 possess their distinguishable trait. However, this result concludes that H7N3 does not have any remarkable unique characteristics. The amino acids extracted from position 3 are distinguishable rule in H3N8 while position 4 represents a distinguishable rule in H7N7. Thus, position 3 and position 4 are important factors in each H3N8 and H7N7. 7window shows that all H3N8, H7N3, and H7N7 have unique characteristics that can distinguish them. This value is high enough to conclude that H3N8, H7N3, and H7N7 possess their distinguishable trait. The percentages of each rule on Table 7 are high enough to consider the rules as essential rules of the genome. 9window shows that H3N8, H7N3, and H7N7 all have their unique characteristics that can distinguish them. This value is high enough to conclude that three genomes all possess their distinguishable trait. The percentages of each rule on Table 7 are high enough to consider the rules as essential rules of the genome. Also, Arginine and Glycine are frequently repeated in H7N3, while Glutamic acid is frequently repeated in H7N7. This indicates that Arginine especially plays a significant role in H7N3. Moreover, the amino acids extracted from position 1 represent a distinguishable rule of H3N8. In H7N3, position 2 is the distinguishable rule. In H7N7 both position 1 and position 2 are distinguishable rule.

4.3. K-means

We also used K-means algorithm to verify our results from the previous experiments. We conducted the K-means algorithm by dividing the sequence structurally into two clusters.

Table 5: Two clusters in 5window K-means

Attribute	Full Data (220.0)	0 (134.0)	1 (86.0)
pos1	G	G	I
pos2	E	E	D
pos3	G	G	G
pos4	T	T	G
pos5	G	E	R

Table 6: Two clusters in 7window K-means

Attribute	Full Data (158.0)	0 (102.0)	1 (56.0)	Attribute	Full Data (158.0)	0 (102.0)	1 (56.0)
pos1	G	G	G	**pos5**	A	S	A
pos2	G	G	R	**pos6**	G	G	E
pos3	I	Y	G	**pos7**	D	S	D
pos4	D	D	E	**cleavage**	1	2	3

Table 7: Two clusters in 9window K-means

Attribute	Full Data (123.0)	0 (94.0)	1 (29.0)	Attribute	Full Data (123.0)	0 (94.0)	1 (29.0)
pos1	D	R	D	**pos6**	G	G	F
pos2	E	E	I	**pos7**	S	V	S

pos3	G	G	E	**pos8**	A	T	N
pos4	G	G	S	**pos9**	T	T	F
pos5	N	N	G	**cleavage**	1	1	3

Table 5 above is the result of K-means in 5window. There are total 220 instances. Total three iterations are held for precise result. This shows that the H3N8, H7N3, and H7N7 can be divided into two clusters, which are mentioned as 0 and 1 in the table. The full data indicates that Glycine is the characteristic amino acid among the three others since it appears three times. Similarly both separated clusters have shown that Glycine is the prominent characteristic amino acid. However, in the first cluster Glutamic acid is also shown as a characteristic amino acid. As shown on the graph, both of the amino acids are analyzed twice. In the second cluster Glycine is also the characteristic amino acid. 69% of entire genome is grouped into the first cluster, whereas 31% of the genome is grouped into the second clusters. This indicates that about seven-tenths and three-tenths of the genome have completely different characteristics. Table 6 above is the result of the K-means algorithm in 7window. 158 instances were experimented and three iterations are held for better result. Table 6 is organized in a similar matter with table 8. The full data shows that Glycine is the characteristic amino acid. Glycine is also the characteristic amino acid in the first cluster. However, Glutamic acid and Glycine is the characteristic amino acid in the second cluster. 65% of entire genome is grouped into the first cluster, while the rest are grouped into the second clusters. Table 7 above is a graph explaining about the K-means algorithm in 9window. 123 instances were experimented this time and three iterations were shown. Table 7 is also categorized in the same way with table 8 and table 10. The full data indicates that Glycine is the main characteristic amino acid. Glycine is shown three times. In the first cluster Glycine is the main characteristic amino acid among the others since it is mentioned three times. However in the second cluster, Glycine is only mentioned once. Instead, it's main characteristic amino acid is Phenylalanine. 76% of the entire genome is grouped into the first cluster while the rest of them are clustered into the second cluster.

4.4. SVM

From Apriori algorithm, we figured out that H3N8, H7N3, and H7N7 are very similar, possessing Glycine as their main amino acid. Therefore, for more accurate results, we did the third experiment utilizing SVM algorithm. We conducted the algorithm in four types: normal, polynomial, RBF, and sigmoid. The experiment is done in 5window, 7window, and 9window. The experiment method was 10 fold cross validation.

Table 8: SVM experiment result (Average)

	Normal	**Poly**	**RBF**	**Sig**
Window 5	62.60	61.00	22.80	71.80
Window 7	60.33	60.34	26.00	61.33
Window 9	63.25	64.50	27.50	71.75

During the experiment, we made data types of <H3N8, H7N3, and H7N7 >. In 5window, we used 180 training sets and 50 test sets. Table 8 shows the results of SVM algorithm. Normal, polynomial, and sigmoid have high average percentage over 60%. However, the average percentage of RBF is remarkably low, implying a meaningful result that the data set is dividable. Thus, this graph proves that H3N8, H7N3, and H7N7 have different properties. In 7window, we used 120 training sets and 30 test sets. Normal, polynomial, and sigmoid have high average percentage over 60%. However, the average percentage of RBF is remarkably low, implying a meaningful result that the data set is dividable. Thus, this graph proves that H3N8, H7N3, and H7N7 have different properties. In 9window, we used 90 training sets and 20 test sets. Normal, polynomial, and sigmoid have high average percentage over 60%. However, the average percentage of RBF is remarkably low, implying a meaningful result that the data set is dividable. Thus, this graph proves that H3N8, H7N3, and H7N7 have different properties.

5. Conclusion and discussion

Our research is composed of 4 data mining algorithms: Apriori, Decision Tree, K-means, SVM. The result of apriori and K-means has shown that because all three viruses have Glycine as main amino acid the three viruses are very similar. However, the result of SVM has proved that although they are very similar they are capable of separation. Moreover, decision tree algorithm revealed that H7N3 has Arginine as its unique main amino acid. Although our experiment has failed to show the connection between H3N8 and H7N7 nor between H7N3 and

H7N7, we were able to see the common genome characteristics among the three viruses through the apriori algorithm and K-means algorithm. Moreover, we were able to conclude from the SVM results that the three viruses each possess unique characteristics that enable the classification. Further studies could be focused to find the direct genetic connections between the viruses to figure out which aspect of virus gene contributes to the specific types of species that can be infected from that virus.

6. References

[1] "Avian Influenza A Virus (H7N7) Associated with Human Conjunctivitis and a Fatal Case of Acute Respiratory Distress Syndrome." *Proceedings of the National Academy of Science*. N.p., n.d.

[2] Gusnanto, A., C. C. Taylor, I. Nafisah, H. M. Wood, P. Rabbitts, and S. Berri. "Estimating Optimal Window Size for Analysis of Low-coverage Next-generation Sequence Data." *Bioinformatics* 30.13 (2014): 1823-829.

[3] Lee, Min Young, and Taeseon Yoon. "Analysis of Methicillin-resistant Staphylococcus Aureus Using Apriori, DBSCAN, and K-means Algorithms." *Proceedings of the 3rd International Conference on Biomedical and Bioinformatics Engineering - ICBBE '16* (2016): n. pag.

[4] Myers, Christine, and W. David Wilson. "Clinical Techniques in Equine Practice." *Equine Influenza Virus* 5.3 (2006): 187-96.

[5] Yamanaka, Takashi, Manabu Nemoto, Hiroshi Bannai, Koji Tsujimura, Takashi Kondo, Tomio Matsumura, Sarah Gildea, and Ann Cullinane. "Assessment of Antigenic Difference of Equine Influenza Virus Strains by Challenge Study in Horses." *Influenza and Other Respiratory Viruses* 10.6 (2016): 536-39.