Analysis of Two Different Sequences of Old Arena Viruses by Decision Tree, Apriori Algorithm, and Shannon Entropy

Yuree Chung*, Yujin Moon, Taeseon Yoon
Department of International, Hankuk Academy of Foreign Studies, Yongin, South Korea.

* Corresponding author. Tel.: 01095249324; email: yuripunks525@gmail.com
Manuscript submitted January 5, 2015; accepted September 8, 2015.
doi: 10.17706/ijcce.2016.5.4.269-275

Abstract: LASV (Lassa virus) and LCMV (Lymphocytic Choriomeningitis virus), which show the different mortality in spite of same symptoms and origin, are introduced into the human population by rodents which shed the virus in urine and droppings. Direct contact with these materials, touching virus-infected objects, eating contaminated food, or exposure to open cuts or sores develop infection. In this paper, we analyzed 4 different proteins of LASV and LCMV: glycoprotein, Z protein, nucleoprotein, and L protein. Also, we investigated the similarities between them based on the frequency of amino acids by decision tree. Furthermore, we look for one amino acid's frequency of relating rates to another amino acid to find the difference between two viruses by Apriori algorithm.

Keywords: Apriori algorithm, decision tree, LASV (Lassa virus) and LCMV (Lymphocytic Choriomeningitis virus), Shannon entropy.

1. Introduction

LASV (Lassa virus) and LCMV (Lymphocytic Choriomeningitis virus) are included in the category of Arenavirus, which is the same category with the Ebolavirus that has shocked the world for its high fatality and relatively short survival period after the occurrence of disease. Desired results of Ebolavirus such as developing the vaccine 'ZMapp' by many countries' repeated research have been achieved after its sudden appearance. Even though LASV and LCMV show similar symptom as Ebolavirus because of the similar rodent which carries the disease, the research has not been achieved in the extent that we desired. To give help for the researches which try to prevent chaotic stage from outbreak of disease, we made this paper to help the research of LASV and LCMV by comparing the sequence of glycoprotein, Z protein, nucleoprotein, and L protein.

2. Materials

There are two types of Arenavirus which are 'Old World Arenavirus' and 'New World Arenavirus'. LCMV and LASV are both included in Old World arena virus. Many of Old World Arenavirus take rodents as the natural host, particularly an 'Old World rodents' that live in African continent. Sequences of two viruses have some similar early stage of symptoms such as meningitis , encephalitis or meningoencephalitis (inflammation of both the brain and meninges) [1].

2.1. Arena Virus

Arena virus is composed of nucleocapsid with two-segmented RNA segments. Each two-segmented RNA are denoted Small (S, include nucleoprotein, glycoprotein) and Large (L, Z protein, L protein) [2]. It is divided into two groups: old world (found in the eastern hemisphere in places such as Europe, Asia, and Africa) and new world (Argentina, Bolivia, Venezuela, Brazil, and the United States). Arenavirus infects rodent and occasionally human, and at least eight Arenaviruses are known to cause human disease [3].

2.2. Lassa Virus(LASV)

Lassa virus, whose host isMastomys natalensis, causes a fatal hemorrhagic fever that had been first occurred in 1969 in the town of Lassa, in Borno State, Nigeria [4]. The West Africa is mainly a raging place. In particular, LASV has been raged in Guinea, Liberia, Sierra Leone, as well as Nigeria. Every year, about 5,000 people died from 300,000-500,000 infected people [5], which means 15 to 20 percent of mortality rate. Though there is an antiviral drug called Ribavirin, it is fatal to the pregnant women. Furthermore, the early symptoms of LASV are misdiagnosed as symptoms of other African hemorrhagic fever like Malaria, Yellow fever, etc.

2.3. Lymphocytic Choriomeningitis Virus (LCMV)

Lymphocytic Choriomeningitis virus is categorized in arena virus family. It is known that LCMV causes fever, malaise, lack of appetite, muscle aches, headache, nausea, and vomiting. LCMV is naturally spread by the common house mouse, Mus musculus [6]. The mortality rate of LCMV is less than 1%. LCMV is a prototype of more severe hemorrhagic fever viruses.

3. Methods

3.1. Decision Tree

Decision tree is a form of multiple variable analysis which provides unique capabilities to supplement, complement, and substitute for traditional statistical forms of analysis (such as multiple linear regression), a variety of data mining tools and techniques (such as neural networks). It is based on algorithms that identify various ways of splitting a data set into branch-like segments. These segments make an inverted decision tree that originates with a root node at the top of the tree. The goal of analysisis reflected in this root node as a simple, one-dimensional display in the decision tree interface [7]. In our experiment, we applied decision tree method to analyze the difference between LASV and LCMV. The extracted rules mean there are differences between these two viruses.

3.2. Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases [8]. We used this algorithm to analyze the similarity of frequent kinds of amino acidof virus. We conducted our experiment on three different windows: 5-window, 7-window, and 9-window. The number of window represents the number of sequences of virus which we have disposed before applying algorithm. For example, 5-window means that we have split the whole sequence of specific virus into five sized sequence and disposed these sequences in row. We can surmise different combinations of amino acids that we can obtain from applying Apriori algorithm to specific sequence. In our experiment, we compared the sequences of glycoprotein, nucleoprotein, L protein, and Z protein.

3.3. Shannon Entropy

Shannon entropy means an absolute limit on the best possible lossless compression of any communication, with restraint: treating messages to be encoded as a sequence of independent and identically-distributed random variables [9]. Shannon's source coding theorem proves that, in the limit, the

entropy divided by the logarithm of the number of symbols in the target alphabet is the average length of the shortest possible representation to encode the messages in a given alphabet [10]. We used Shannon Entropy to analyze the differences of two different viruses through a more precise calculation.

4. Helpful Hints

4.1. Figures and Tables

Table 1. Rule Extraction of Glycoprotein, Z Protein, Nucleoprotein, L Protein of LASV and LCMV under 5 Window

Window5	*LASV*		*LCMV*	
	Rule	**frequency**	**rule**	**frequency**
glycoprotein	pos1 = P	0.800	pos3 = S	0.824
	pos3 = E	0.800		
	pos3 = N	0.778		
z protein	pos2 = G pos2 = K	0.800	pos2 = S	0.857
	pos5 = P	0.857	pos2 = I	0.750
	pos3 = A	0.833	pos5 = S	0.833
			pos3 = P	0.750
nucleoprotein	pos3 = Y	0.857	pos3 = C	0.857
	pos2 = P	0.818	pos3 = E	0.750
	pos3 = M	0.857	pos2 = N	0.833
			pos3 = P	0.750
l protein	pos2 = A	0.786	pos3 = K	0.833
	pos2 = N pos4 = L	0.833	pos4 = G	0.857
	pos4 = W	0.800	pos2 = W	
	pos1 = I pos5 = K	0.833		

Table 1 shows that position 3 is the important factors which differentiate each protein.It shows that serine is the most dominantly extracted amino acid. Also we assume that position 2 is the important factors which differentiate two viruses at Z protein and L protein.

Table 2. Rule Extraction of Glycoprotein, Z Protein, Nucleoprotein, L Protein of LASV and LCMV under 7 Window

Window 7	*LASV*		*LCMV*	
	Rule	**frequency**	**rule**	**frequency**
glycoprotein	pos7 = K	0.800	pos6 = P	0.750
			pos1 = K	0.750
			pos1 = P	0.750
z protein	pos3 = A pos3 = T	0.750	pos3 = L	0.833
	pos3 = N	0.833		
nucleoprotein	pos3 = A	0.833	pos1 = L pos3 = K	0.833
l protein	pos4 = A	0.762	pos1 = L pos6 = V	0.833
	pos4 = D pos7 = R	0.833	pos3 = L pos4 = K	0.875
	pos4 = L pos5 = N	0.833	pos4 = L pos5 = L	0.818
	pos1 = P pos5 = K	0.833		

Table 2 shows that position 1 is the important factors which differentiate each protein, and leucine is the most dominantly extracted amino acid.It shows that leucine is the most dominantly extracted amino acid in L protein of LCMV. L protein of LASV, position 4 is the important factors differentiate two viruses.

According to Table 3, it is noticeable that there were no rules of glycoprotein and nucleoprotein between LASV and LCMV. This result is due to lack of unique amino acidic features among proteins.

Table 3. Rule Extraction of Glycoprotein, Z Protein, Nucleoprotein, L Protein of LASV and LCMV under 9 Window

Window 9	*LASV*		*LCMV*	
	Rule	**frequency**	**rule**	**frequency**
Glycoprotein	Not extracted		Not extracted	
z protein	pos9 = P	0.833	pos9 = N	0.750
	pos1 = T	0.800	pos1 = Q	0.800
	pos7 = A	0.750	pos9 = S	0.750
			pos7 = I	0.800
nucleoprotein	Not extracted			
l protein	pos6 = Q	0.800	pos9 = C	0.800
	pos7 = H	0.833	pos6 = P	0.900
	pos5 = W	0.800	pos1 = N pos7 = L	0.857
	pos9 = W	0.800		
	pos4 = N pos6 = E	0.833		
	pos4 = T pos8 = L	0.833		
	pos5 = Y	0.857		
	pos1 = E pos3 = F	0.857		
	pos1 = L pos8 = V	0.857		

4.2. Apriori Algorithm

Place figure captions below the figures; place table titles above the tables. If your figure has two parts, include the labels "(a)" and "(b)" as part of the artwork. Please verify that the figures and tables you mention in the text actually exist.

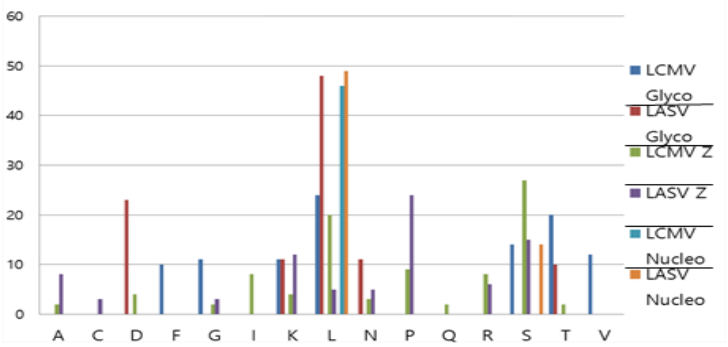

Fig. 1. 5-window LCMV and LASV glycoprotein, Z protein, and Nucleoprotein sequences.

Fig. 1 is an analysis of each glycoprotein, Z protein, and nucleoprotein of LCMV and LASV under 5-window. Amino acid L and S are more frequently appear in LCMV Z protein sequence, but amino acid P is more frequently found in LASV Z protein. The result shows that amino acid L has been extracted the most frequently in both LCMV and LASV glycoprotein sequences and nucleoprotein sequences.

Fig. 2 is the result from 7-window experiment. We compared each LCMV and LASV glycoprotein, Z protein, and nucleoprotein. From this experiment, we can observe that amino acid S and T are more frequently extracted from LCMV glycoprotein sequence than LASV glycoprotein sequence. Also, amino acid S is more frequent in LCMV Z protein sequence where amino acid P is more frequent in LASV Z protein sequence. However, we could not find the distinct difference between LCMV and LASV nucleoprotein sequences.

Fig. 3 is the graph which shows the result of the experiment under 9-window. The method was same as the experiments mentioned above (5-window and 7-window). Amino acid P was more frequently appear in LASV glycoprotein sequence and Z protein sequence than those of LCMV. Amino acid S is, however, found more frequently in LCMV Z protein sequence. In the case of nucleoprotein, amino acid G is more frequent in LCMV and amino acid T was more frequent in LASV sequence. Overall, in nucleoprotein sequence, Leucine is

the most dominant and frequent amino acid. Leucine is most frequently found in glycoprotein sequence under window 5 and 7, but this pattern seems to decrease under window 9.

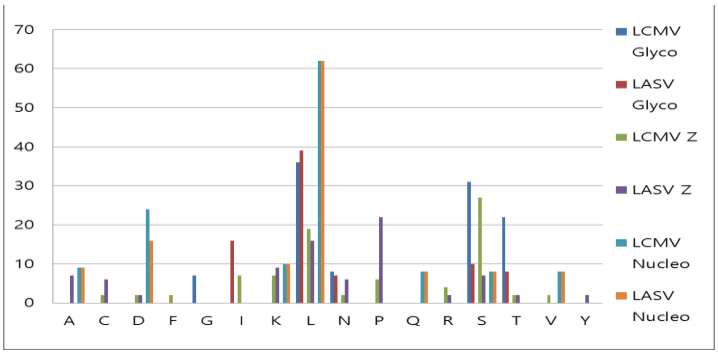

Fig. 2. 7-window LCMV and LASV glycoprotein, Z protein, and nucleoprotein sequences.

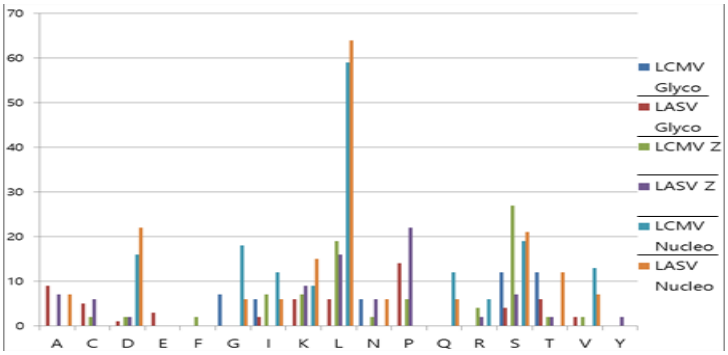

Fig. 3. 9-window LCMV and LASV glycoprotein, Z protein, and nucleoprotein sequences.

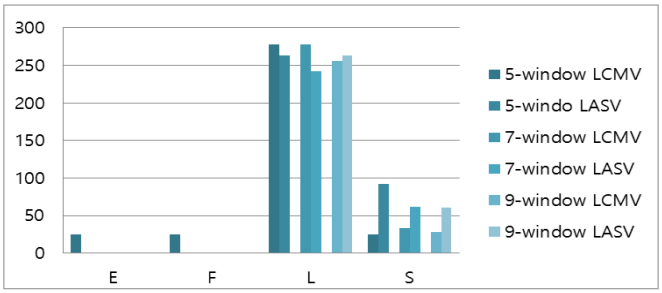

Fig. 4. 5, 7, and 9-window L protein sequences.

Fig. 4 is the result of L protein sequences of LCMV and LASV under each 5,7, and 9- window experiments. Since the total amount of amino acid sequence is far more than that of other proteins, we made a separate graph of L protein from other three categories.Compared to other categories of sequence, L protein has comparably more number of sequence which is concentrated on amino acid L and S under all three experiments. Overall experiments show that there is no distinct difference between L protein sequence of two viruses.

4.3. Shannon Entropy

The uncertainty probability from the random variables can be calculated by using Shannon entropy theorem [9]:

$$H(X)=\sum_{i=1}^{n} p(x_i)I(x_i)=\sum_{i=1}^{n} p(x_i)\log_b \frac{1}{p(x_i)}=-\sum_{i=1}^{n} p(x_i)\log_b\ p(x_i)$$

The following charts are the results:

Table 4. Shannon Entropy Results of LASV

	Shannon Entropy	Normalized Shannon Entropy
Glycoprotein	6.0452450350808	0.97559905494257
Z protein	4.4181794671772	0.96149384809796
Nucleoprotein	6.1816113682288	0.97442116578607
L protein	7.5469648107622	0.97957048563563

Table 5. Shannon Entropy Results of LCMV

	Shannon Entropy	Normalized Shannon Entropy
Glycoprotein	6.051931950051	0.97445204569513
Z protein	4.4580571792945	0.97895865114782
Nucleoprotein	Non	Non
L protein	7.5474167215253	0.98008880733689

Table 4 and Table 5 show the results of each Shannon entropy calculation of LASV and LCMV sequence respectively. To sum up the results from two charts, the mathematical results of each protein of Shannon entropy and normalized Shannon entropy between LASV and LCMV do have many differences.

5. Conclusion

Even though the symptoms are similar, the mortality rates between LASV and LCMV are different (LASV: 15~20%, LCMV: less than 1%). According to our results, glycoprotein, Z protein, nucleoprotein, and L protein sequence show similarities in amino acid sequences because LASV and LCMV are categorized in arenavirus. In the approach of decision tree method, glycoprotein and Z protein's amino acid sequence are similar because their rules are not extracted in window 9, the highest window version which shows more detailed information about protein sequences of LCMV and LASV. This result is not surprising because glycoprotein is located on the surface of cell, and it is closely related to the receptor of host cell; therefore, similarity of the host of two viruses may affect the result which shows much similarity in glycoprotein sequence. As window versionupgraded, LASV and LCMV show lack of unique features. we hope these rules would help the further research to illuminate the factor which differentiates the mortality of LCMV and LASV.

In the approach of Apriori algorithm, we can find that there are many different kinds of amino acid that has been extracted. We conducted Apriori algorithm experiment under three different windows just as decision tree method. Since the purpose of this paper is to find out the difference of LCMV and LASV through computer algorithm, we focused on the noticeable difference between two viruses when analyzing the Apriori algorithm results. However, we could not integrate all the results from both decision-tree algorithm and Apriori algorithm and also we could not provide the actual mechanism of how different amino acid can affect the function of virus. Therefore, our next research purpose will be searching for the algorithm which can integrate the whole results from different algorithm such as decision-tree algorithm, and Apriori algorithm so that we can predict the mechanisms of amino acid of specific sequence more precisely.

References

[1] Centers for Disease Control and Prevention. (2014). *Lymphocytic Choriomeningitis (LCM)*.

[2] Briese, T., Paweska, J. T., McMullan, L. K., *et al*. (May 2009). Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated Arenavirus from southern Africa.

[3] Botten, J., Whitton, J. L., Barrowman, P., Sidney, J., Whitmire, J. K., Alexander, J., Kotturi, M. F., Sette, A., & Buchmeier, M. J. (2010). A multivalent vaccination strategy for the prevention of old world Arena virus infection in humans. *Journal of Virology, 84*.

[4] Frame, J. D., Baldwin, J. M., Gocke, D., J., & Troup, J. M. (July 1970). Lassa fever, a new virus disease of man from West Africa: Clinical description and pathological findings. *Am. J. Trop. Med. Hyg*.

[5] Delgado, S., Erickson, B. R., Agudo, R., *et al*. (April 2008). Chapare virus, a newly discovered Arena virus isolated from a fatal hemorrhagic fever case in Bolivia.

[6] Zhou, X., Ramachaundran, S., Mann, M., & Popkin, D. L. (2012). Role of lymphocytic Choriomeningitis virus (LCMV) in understanding viral immunology: Past, present and future. *Viruses*.

[7] Decision trees — What are they? From http://support.sas.com/publishing/pubcat/chaps/57587.pdf

[8] Rak, & Ramakrishnan, S. (September 1994). Fast algorithms for mining association rules in large databases. *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 487-499). Santiago, Chile.

[9] Shannon, C. E. (July–October 1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*.

[10] Goise, F., & Olla, S. (2008). Entropy methods for the Boltzmann equation. Centre Émile Borel, Institut H. Poincaré, Paris.

Yuree Chung was born in Seoul, South Korea in 1997. She is attending at Hankuk Academy of Foreign Studies now. She is a president of HAFS Bioinformatics Club 'GATTACA' in 2014 and translated a research titled *Essential Bioinformatics* into Korean this year.

Also, she founded Korean Youth Bioinformatics Association with her colleagues in 2015. She got the HAFS Scholarship Award and Excellence Award in computer science and many other subjects. In 2013, she conducted an experiment about 'bacteriostatic action of effective microorganisms.'

Yujin Moon was born in Taeback, SouthKorea, in 1997. She is attending at Hankuk Academy of Foreign Studies now. She was selected to participate in Kyungam Bio Youth Camp which was held by Korean Society for Molecular and Cellular Biology. She got HAFS Scholarship Award and Excellence Award in Biology and many other subjects. She interned at the Korea Blood Cancer Association to find out the cause of blood cancer and research the treatment with the members in KBCA in 2014.

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 is a computer science and statistics teacher.