In-Depth Understanding of Norovirus and Rotavirus by Employing Apriori Algorithm in Case of DNA Sequence

Cheolho Heo, Daeseop Kim, and Taeseon Yoon

***Abstract*—An Infectious disease of Norovirus and Rotavirus have occurred annually and cause influence into all range of ages. In case of pediatrics, these virus spread quickly because of outbreak in gastroenteritis and institution, which make economic loss socially. They show lots of similar communicable diseases. To prevent us from these viruses, knowing and recognizing them intrinsically is the most significant thing that we have to discern. This why we do this research through employing apriori algorithm and Support Vector Machine (SVM). We can compare and contrast DNA sequence of both norovirus and rotavirus and find some resemblance. Moreover, doing SVM experiments, we get more information about these viruses. Our research has an importance on seeking for some probability of accurate and non-side effect treatment.**

***Index Terms*—Apriori algorithm, support vector machine, norovirus, rotavirus.**

I. Introduction

In 2012, in japan, more than 10000 people were infected and passed away in this year because of these viruses. Because of this affair, government of Japan had to make an effort to restore massive damage and to build some prevention measures. We can confirm the state of seriousness from this website: http://www.iph.pref.osaka.jp/infection/Noro/noro.html. However, after 2 years from that situation, those viruses become spread again in Korea, nearest country of Japan.

Norovirus is a main cause of epidemic viral gastroenteritis to all age around the world [1]. Seriously, this virus can survive at high temperature environment unlike other virus. Furthermore, only through touching objects like food and water that infectee made contact, other people can be infected easily [2]. After 24~48 hours of incubation period, suddenly, vomiting, diarrhea and nausea arise and persist. Moreover, overall body symptoms like headache, fever, chills and myalgia can be accompanied.

Rotavirus, whose major infection route is excrement and mouth, is also a main cause of epidemic viral gastroenteritis [3]. This virus cause vomiting, fever, watery diarrhea (not with blood) and dehydration. Usual target of Rotavirus is infants and children but sometimes, to adults and the aged, mass attack and aids diarrhea arise. If Rotavirus get into human body, more than 30% infectees get upper 39 Celsius degree fever [4]. In infant's case, they can be killed by serious human dehydration. Unfortunately, this virus cannot be detected through blood and white blood cell of stool tests.

In common with other virus, Norovirus and Rotavirus also can't be cured by antibiotics and there are no antiviral agents. Unluckily, these viruses have so many kinds that infect one experienced person again and again [5]. This is why inventing and developing vaccine of Norovirus and Rotavirus are backbreaking and strenuous. Usually, these can be cured by itself but momentary symptoms are very painful for us and also, there are a number of people having critical symptoms because of their genetic properties [6].

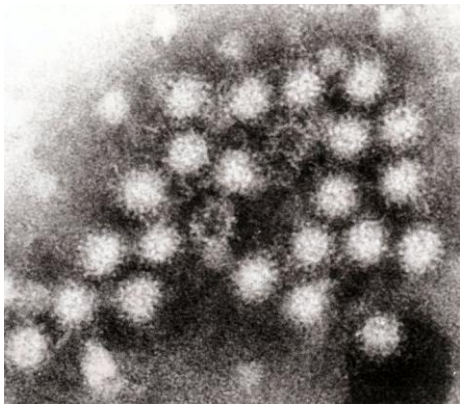

Fig. 1. Transmission electron micrograph of Norovirus.

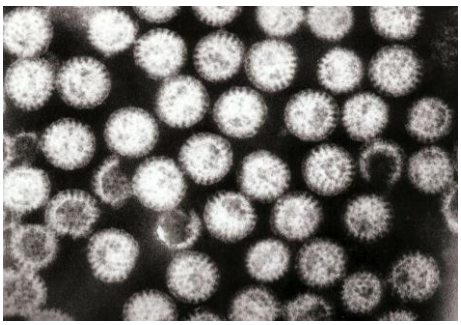

Fig. 2. Transmission electron micrograph of Rotavirus.

II. Materials

We will use FASTA format of viruses, norovirus and rotavirus. FASTA format is one of the way of expressing DNA sequence or Protein sequence. It can show those sequence as a word string. This word string should be no longer than 120 characters and usually do not exceed 80 characters. Each character has its own meaning, such as "G" means "Guanine", "T" means "Thymine." To check the overall similarity and resemblance of the norovirus and the rotavirus, their DNA sequences will be compared with this crucial format. Also, we got this information of virus from NCBI (National Center for Biotechnology Information) which have precise DNA database. We got genome

Manuscript received March 26, 2014; revised May 30, 2014.

Cheolho Heo, Daeseop Kim, and Taeseon Yoon are with Hankuk Academy of Foreign Studies, Korea (e-mail: dydakchry@naver.com).

sequences of norovirus from http://www.ncbi.nlm.nih.gov/nuccore/551460473?report=fasta and genome sequence of rotavirus from http://www.ncbi.nlm.nih.gov/nuccore/470378?report=fasta

III. Experiments

Our approach to norovirus and rotavirus using SVM follows next steps below. First, extract DNA sequence of norovirus, and rotavirus by apriori algorithm. Next, make algorithm to learn standard judgment, hyper plane, by using Support Vector Machine. Finally confirm similarity of optional DNA information of mitochondria based on the classifier, the learned standard judgment.

Normally DNA string bit information of virus from first step has inseparability problem. SVM solves this problem by applying Soft Margin SVM [7], [8]. Soft Margin SVM lowers the generalization errors of the classifier. It defines new high infinite dimensional space, and remaps the complicate data in there [9], [10]. In the high infinite dimensional space, SVM is able to separate data easily, and find out patterns that cannot be recognized in two dimensional domain spaces. To support this computational logic reasonable, kernel function is suggested. Normal function, polynomial function, sig function and RBF function are representative kernel functions which are being used frequently in SVM [11]. SVM also provides kkt-condition which automatically deletes error to carry out perfect classification.

A. Apriori Algorithm

Apriori algorithm is the outstanding algorithm to find various associations from the data [12]. Apriori algorithm finds correlation based on the frequency of each data [13]. The basic rule of Apriori algorithm is that all subsets of the frequency set have high frequency. The process of the algorithm is easy. First, it counts the frequency set of data. Next it finds the minimum support value. Then it generates the candidate set. Finally it repeats the counting, and setting candidate sets. Based on Using breadth-first search and a Hash tree structure, apriori algorithm is able to count candidate item sets efficiently, and find frequencies accurately. Using breadth-first search and a Hash tree structure follows next procedure. First, it generates candidate item sets of length k from item sets of length k-1. Then some candidates contain an infrequent sub patterns. These are deleted. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. Finally, it scans the transaction database so that we can determine frequent item sets among the candidates. Also Apriori algorithm uses a "bottom up" approach. In the approach, frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm operates until no further successful extensions are found.

B. Method

We analyze DNA sequence of these two virus employing apriori algorithm. Using the result from this process (window5, window7, window9), we classify amino acid and express into graph. We count the frequency number of amino acid of both virus and arrange them. After that, we set two graph on same graph surface. To compare them more accurately, we adjust data to proportion data because total amount numbers of result of each data are different.

Furthermore, we progress SVM experiment. In 5window, we conduct test data (9 amino acid of norovirus and rotavirus) and training data (90 amino acid of norovirus and rotavirus). Similarly, in 7window, we conduct test data (6 amino acid of norovirus and rotavirus) and training data (90 amino acid of norovirus and rotavirus) and in 9window, we conduct test data (5 amino acid of norovirus and rotavirus) and training data (90 amino acid of norovirus and rotavirus).

IV. Results and Conclusions

In graph of results, x axis represent kinds of amino acid of each virus, such as ‘A’, ‘G’. At first comparison, y axis represents numbers of frequency. However, at second comparison, we apply proportion of the amino acid. This is because total amounts of amino acid of both viruses are different fairly. There are three graphs of result, 5window, 7window and 9window.

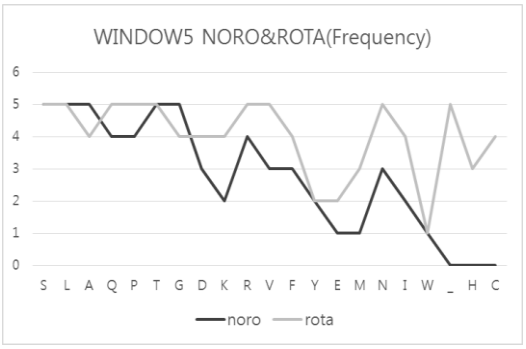

Fig. 3. Frequency graph of 5window.

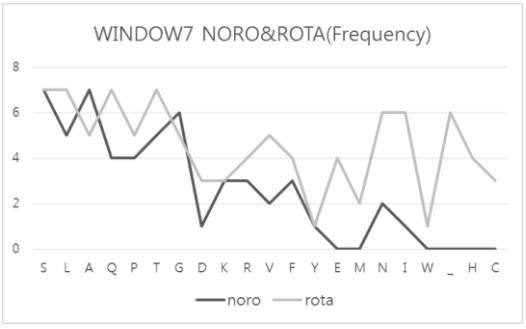

Fig. 4. Frequency graph of 7window.

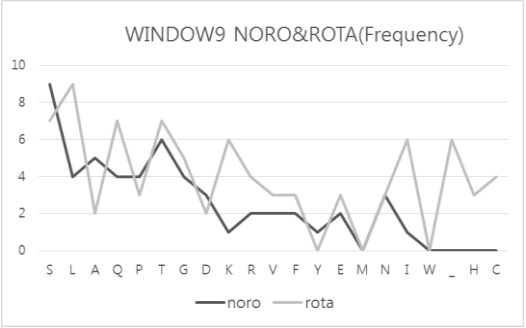

Fig. 5. Frequency graph of 9window.

At these three Graph representing frequency of amino acid, we can find only a little similarity. Two aspects of graph line

look alike; however, there are some amino acid which have very different result in every graph. As mentioned before, the difference of amounts of total amino acid is cause. So, next, we adjust data from frequency to proportion.

At this result meaning proportions, we can find lots of likeness in two virus, norovirus and rotavirus. Not only aspects of graph line, especially, but also some amino acid like Q, D, V and others have similar data results. In window9, this result is one of the best outcomes that have more resemblance. However, at amino '_', 'H' and 'C', they have a little gap in outcome because, in our database of norovirus, there is no amino acid mentioned upper line. Except, these parts, overall forms are very analogous.

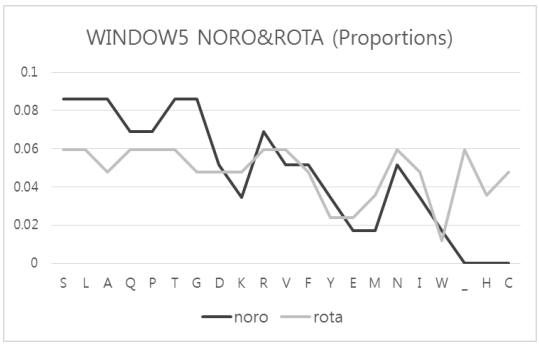

Fig. 6. Proportion graph of 5window.

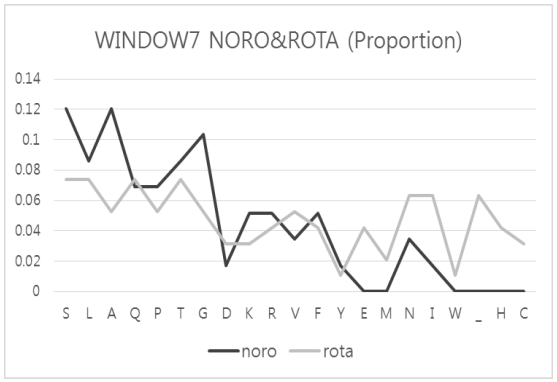

Fig. 7. Proportion graph of 7window.

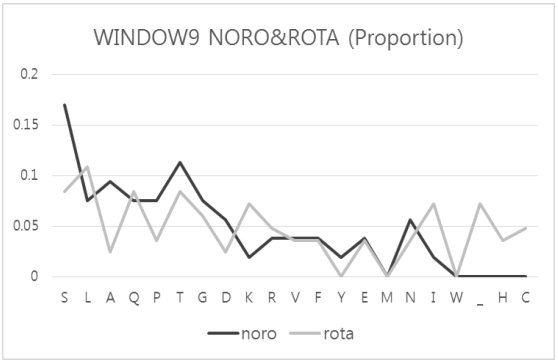

Fig. 8. Proportion graph of 9window.

It manifested that ribosomal composition of rota virus and noro virus is high from apriori algorithm. To discuss this similarity in depth, we progressed the study with support vector machine. Normal kernel, one-dimensional polynomial function, two-dimensional polynomial function, radial basis function, and signoid function are used in support vector machine.

The result with normal kernel shows that rota virus and noro virus have high conformity degree in each of window 5, window 7, and window 9. Support Vector Machine could not classify the data exactly. In other word, linear analysis demonstrates same results with apriori algorithm. The results of one-dimensional polynomial function support vector machine was alike this except the number of support vector it had used during the classification. However the result of two-dimensional polynomial function support vector machine was different. The certainty of classifying had extraordinarily increased in all conditions of windows. Support vector machine could separate data very successfully. The signification of this availability of separation is existence of different characteristic of ribosomal pattern between rota virus, and noro virus. The accuracy was 66.67% in window 5 which is about 10% higher than normal kernel support vector machine, and one-dimensional polynomial support vector machine. The accuracy in window 7 was 91.67%. Comparing that accuracy of normal kernel support vector machine, and one-dimensional polynomial support vector machine is 75.00%, it is about 16% higher, and defines the pattern difference between viruses. Although accuracy of normal kernel support vector machine, and one-dimensional polynomial support vector machine is 50.00%, the accuracy in window 9 is 90.00%. This numerical value attracts the attention because this also exceeds the value of radial basis function support vector machine which normally presents high percent of separate ability. It can be interpreted that 5 window is too short to handle pattern analysis so that the accuracy of two-dimensional polynomial support vector machine in window 5 is relatively lower than window 7 and window 9. As the length of ribosomal data gets longer to window 7 and window 9, amount of data and proteinic patterns increases. Therefore the classifying accuracy becomes rapidly appreciated. To sum up, rota virus and noro virus is composed by similar ratio of ribosomal identity in appearance. However they are undoubtedly different blocks of protein.

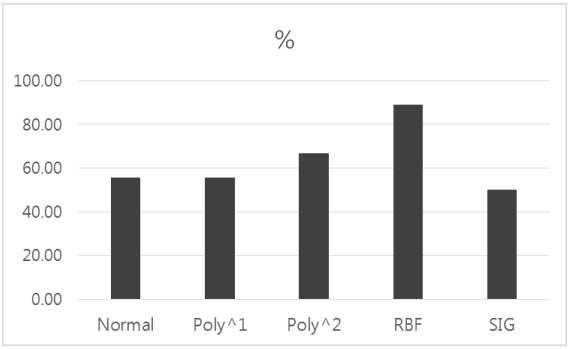

Fig. 9. Percentage graph of 5window.

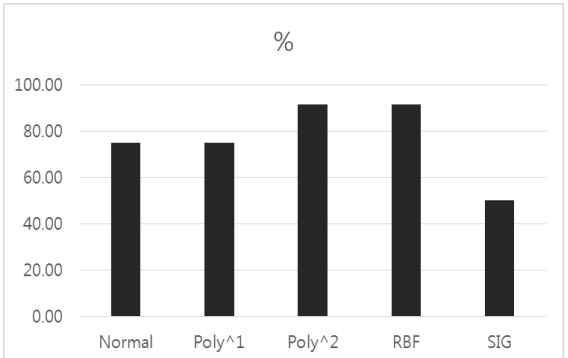

Fig. 10. Percentage graph of 7window.

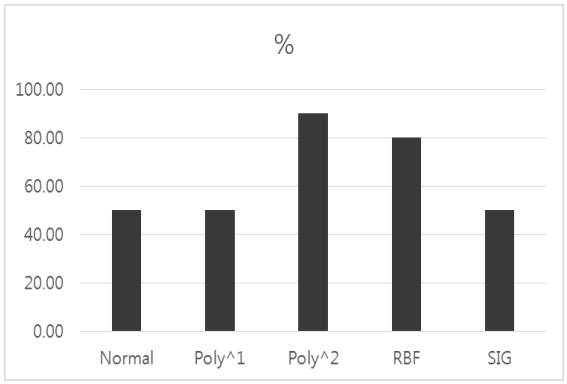

Fig. 11. Percentage graph of 9window.

REFERENCES

[1] A. King, E. Lefkowitz, M. J Adams, and E. B. Carstens, *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, Amsterdam: Elsevier, 2011, pp. 981–982.

[2] R. Goodgame, "Norovirus gastroenteritis," *Curr Gastroenterol Rep*, vol. 8, no. 5, pp. 401–08, 2006.

[3] P. H. Dennehy, "Transmission of rotavirus and other enteric pathogens in the home," *Pediatr. Infect. Dis. J.*, vol. 19, no. 10 (Suppl), pp. S103–S105, 2000.

[4] D. I. Bernstein, "Rotavirus overview," *The Pediatric Infectious Disease Journal*, vol. 28, no. 3 (Suppl), pp. S50–S53, 2009.

[5] J. Y. Tang, L. Y. Chuang, E. Hsi, Y. D. Lin, C. H. Yang, and H. W. Chang. (July 2013). Identifying the association rules between clinicopathologic factors and higher survival performance in operation-centric oral cancer patients using the apriori algorithm. *BioMed Research International*. [Online]. Available: http://www.hindawi.com/journals/bmri/2013/359634/

[6] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proc. the Fifth Annual Workshop on Computational Learning Theory - COLT '92*, 1992, p. 144.

[7] G. Y. Cui, C. Fang, and K. Han, "Prediction of protein-protein interactions between viruses and human by an SVM model," *BMC Bioinformatics*, vol. 13, suppl. 7, p. s5, 2012.

[8] Bayardo and J. Roberto, "Efficiently mining long patterns from databases," *ACM Sigmod Record*, vol. 27, no. 2, ACM, 1998.

[9] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273, 1995.

[10] L. Griffith, E. van den Heuvel, I. Fortier *et al*., "Harmonization of Cognitive Measures in Individual Participant Data and Aggregate Data Meta-Analysis," Agency for Healthcare Research and Quality (US), 2013

[11] Austin, *Madame Curie Bioscience Database*, Landes Bioscience, 2000

[12] W. H. Press and S. A. Teukolsky, "Support Vector Machines," *Numerical Recipes: The Art of Scientific Computing*, 3rd ed., New York: Cambridge University Press, 2007

[13] M. L. Wang, K. Wang, A. X. Yan, and C. Y. Yu, "Classification of HCV NS5B polymerase inhibitors using support vector machine," *Int J Mol Sci.*, vol. 13, no. 4, pp. 4033-4047, 2012.

Cheolho Heo is currently a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Daeseop Kim is currently a student in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in chemistry and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.

Taeseon Yoon was born in 1972. He is currently a teacher in science major of Hankuk Academy of Foreign Studies, Korea. He is mostly interested in engineering and biology. He has been studying pattern analysis and computer programming and its application to chemistry and biology.