Deeper Understanding of Norovirus and Astrovirus by Analyzing Genome and Coding Sequences Using Apriori Algorithm

Eun-Young Kim, Min-Ji Kim, Eun-Ki Hong, Ye-Rim Cho, and Tae-Seon Yoon

***Abstract*—Norovirus infection is an epidemic stomach disorder caused by norovirus. Humans can be infected by norovirus regardless of age, gender and location. Infection is now occurring sporadically throughout the world. Norovirus maintains its infectiousness even after being heated in 60 degrees Celsius for 30 minutes. Symptoms of norovirus infection include diarrhea, stomach, vomiting, and nausea. Although astrovirus is less known than norovirus, it also causes the very similar symptoms as norovirus. Norovirus causes dehydration, pyrexy, drowsiness, diarrhea, and etc. Furthermore, both viruses do not have specific antiviral agents.**

For the research, our goal is to compare and contrast norovirus and astrovirus for deeper understanding; we seek for any possibilities of connections in the treatment of the two viruses by using Apriori algorithm. Finding any relationship between the treatments of two viruses will devote to the overall enhancement of the health of humankind.

***Index Terms*—Norovirus, astrovirus, fasta, apriori algorithm.**

I. Introduction

Norovirus, also known as NVLs (Norwalk-like viruses), is named after Norwalk, Ohio, where it was first detected from a cholera morbus patient by an electron microscope. Norovirus remains its infectiousness even after being heated for 30 minutes under 60 ^{0}C. Also, it has strong resistibility that it does not inactivated by the chlorine concentration of common tap water. The virus can be easily transmitted through drinking contaminated water and contacting with objects with stained surfaces. Only about 10 virions can cause gastroenteritis in human. 90 percent of Gastroenteritis is caused by norovirus. Norovirus is highly infectious. Its contagiousness is the strongest when its symptoms are manifested and it remains its infectious for 3 to 14 more days after recovery. The symptoms appear about 24~48 hours after infection and the common symptoms are nausea, vomiting, diarrhea, and stomach. By taking a biopsy of human Jejunum, casehardened villus, vacuolization of cytoplasm and infiltrating monocyte can be observed. This intestinal damage prevents water and nutrient absorption and causes diarrhea [1], [2].

Astrovirus is a type of RNA virus which was first discovered in 1975 using an electron microscope following a diarrhea outbreak [3]. It is named after the Greek word "Astron" meaning "star" because of its five- or six-pointed star shape. Similar to the norovirus, astrovirus can cause diarrhea, nausea, fever and vomit in both young humans and animals. However, it is not well known as much as norovirus. Acute gastroenteritis caused by astrovirus infection usually occurs in infants and most of the viruses spread are serotype 1 [4]. There are two main types of astroviruses: mamastrovirus and avastrovirus. The former affect mammals and the latter affect birds.

Beyond similar symptoms, norovirus and astrovirus also share similar structure [5]. They are both positive strand RNA viruses, whose viron RNA is same with their mRNA. So, they function as mRNA and can be directly translated. This means that as soon as the host is infected, those viruses start the protein synthesis immediately. Also, norovirus and astrovirus are not cured by antibiotics and antiviral agent for them is yet to be found. Since there are many mutants of those two viruses, reinfection can occur after people once being infected. This makes it difficult to develop vaccines.

We set our purpose of comparing genomic and coding sequences of norovirus and astrovirus after observing preceding researches analyzing similar viruses. For example, "HIV, HBV, HCV and STIs: similarities and differences" explains similarities and differences among the three major blood-borne viruses: human immunodeficiency virus (HIV), hepatitis B virus (HBV) and hepatitis C virus (HCV). "*Structural similarities between influenza virus matrix protein M1 and human immunodeficiency virus matrix and capsid proteins: an evolutionary link between negative- -stranded RNA viruses and retroviruses*" compares the structure of the HIV matrix protein with the membrane-binding N domain of M1 of influenza virus.

As both astrovirus and norovirus have actually no precise precautions or vaccines yet, deeper approach toward them should be progressed. Like previous studies [6], [7], in order to understand more about norovirus and astrovirus, our aim is to analyze and compare their amino acid sequences and base sequences by using apriori algorithm.

II. Materials

A. FASTA Format

We used the FASTA format, the text-based format that represents either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The simplicity of FASTA format makes it convenient to manipulate and parse sequences. A sequence in

Manuscript received December 31, 2013; revised March 14, 2014. This work was supported in part by Hankuk Academic of Foreign Studies.

The authors are with Hankuk Academic of Foreign Studies, Yongin, Korea (e-mail: 96eun@naver.com, mjkim961001@naver.com, sylviaeunki@ naver.com, dpfla4148@ naver.com, tsyoon@hafs.hs.kr).

FASTA format has a series of lines, each of which should be no longer than 120 characters and usually do not exceed 80 characters. Each character has its own meaning, such as "A" means "Adenine", "C" means "Cytosine," In order to check the overall similarity of the norovirus and the astrovirus, their genomic sequences and coding sequences will be compared with this format.

B. Virus Sequences

Coding sequence is a portion of a gene's DNA or RNA, which is composed of exons and it codes for protein. It is bounded by the five prime untranslated region and the three prime untranslated region. As norovirus and astrovirus is both RNA virus, its coding sequence all a part of RNA sequence. Genome sequence is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time. As both viruses are retrovirus, we can use DNA sequence.

We used 6 sequences, 1637F, WH1859, CS-E1, 532642969, 146199230, and 4HS191. Among these sequences, 1637F and 532642969 represent genome sequence of astrovirus, WH1859 and 146199230 represent coding sequence of astrovirus, CS-E1 represents coding sequence of norovirus, and 4HS191 represents genome sequence of norovirus.

We got genome sequences of norovirus from http://www.ncbi.nlm.nih.gov/nuccore/429346005?report=fasta and genome sequence of astrovirus from http://www.ncbi.nlm.nih.gov/nuccore/532642969?report=fasta.

III. Experiments

A. Apriori Algorithm

We used apriori algorithm in order to compare Norovirus and Astrovirus. Apriori algorithm is the first developed algorithm developed to find association between the data. Also it is one of the most frequently used algorithms. Based on the frequency of each data, apriori algorithm analyzes the data mathematically to find correlation between them.

Apriori algorithm has four steps to find association rule. First step is finding the frequency set. Second is identifying the minimum support. Third is generating the candidate set. Final step is repeating second step and third step. The result of algorithm finds strong item set whose reliability is 100%.

Apriori uses a "bottom up" approach to breadth-first search direction. The computation starts from frequent 1-itemsets and continues until all maximal length frequent itemsets are found. In the breadth-first search, frequent subsets extend one item at a time (a step known as candidate generation) and groups of candidates are tested against the data. The algorithm operates until no further successful extensions are found.

Using breadth-first search and a Hash tree structure, the apriori algorithm, we could count candidate item sets efficiently. The process of it is as follows. First, it generates candidate item sets of length k from item sets of length k-1. Then it prunes the candidates that contain an infrequent sub pattern. According to the downward closure lemma (if itemset I is not frequent, then any candidate that contains I is guaranteed to be not frequent) 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.

In using apriori we used 5, 7, and 9 windows to find out periodicity of comparing sequences. 5window is data divided at every five amino acid, 7window is divided at every 7amino acid, and 9window at every 9acid. Therefore, it shows more accurate similarity by considering three different kinds of periodicity.

For example, when we analyze the genome sequence of norovirus by this algorithm, the result comes out as:

1) amino2=G 53 acc:(0.99453)
2) amino4=L 53 acc:(0.99453)
3) amino3=S 50 acc:(0.99447)

B. Methods

After analyzing the 6 sequences by the apriori algorithm, we arranged the outcomes according to the types of Amino acids. Then, we chose two sequences to compare: genome sequences and coding sequences of norovirus and astrovirus. After that, we picked out the line from the amino arrangement that has the most similar sorts of amino acids.

By using the outcomes of the algorithm we compared the number of each amino acids repeated in the amino line. Then, we made a graph to compare the similarity of the two viruses easily.

C. Results

Our assumption was that the more similar a frequency of the amino acid is, the more likely that acids will be applied same method to take care.

For all the graphs, x-axis represents the kinds of amino acid. For the y-axis, first graph represents frequency while second and third graph shows the proportion of the amino acid. The reason was the difference between total amounts of the amino acid that we analyzed. This difference made us unable to compare properly, so instead we used the proportion approach.

1) CS-E1 & WH1859

First, after comparing the coding sequence of norovirus (CS-E1) and astrovirus (WH1859), it shows great similarities on all kinds of windows (see Fig. 1-Fig. 3).

When comparing 5window apriori, we used amino 4. Both of them have the great percentage of amino acid 'L'. As we analyze the trend line of both viruses, involution trend line of 4HS191 is

$$y = 19.895x^{-0.599}$$

and that of 532643969 is

$$y = 22.936x^{-0.726}.$$

In the case of 7window apriori, we also used amino 4 and both have amino acid 'L' as the most frequently appearing acid. Involution trend line of 4HS191 is

$$y = 14.555x^{-0.547}$$

and that of 532643969 is

$$y = 12.511x^{-0.518}.$$

In 9window, we compared amino3 and 'L' is also repeated the most in the arrangement. Involution trend line of 4HS191 is

$$y = 0.0742x^2 - 1.5678x + 10.636$$

and that of 532643969 is

$$y = 0.1096x^2 - 2.0031x + 11.367.$$

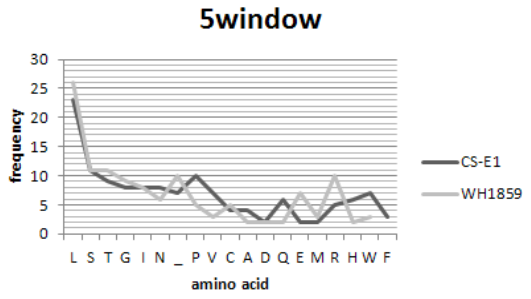

Fig. 1. CS-E1 and WH1859: 5window apriori.

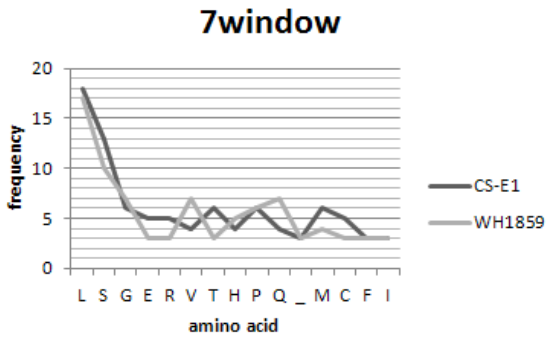

Fig. 2. CS-E1 and WH1859: 7window apriori.

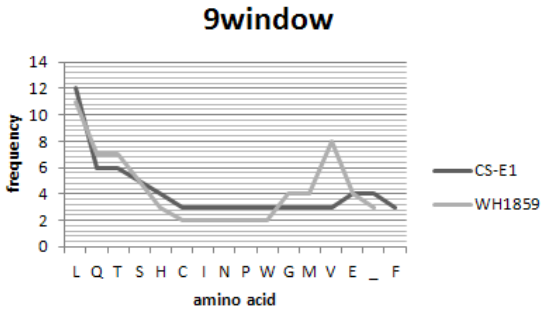

Fig. 3. CS-E1 and WH1859: 9window apriori.

2) 4HS191 & 1637F

Next, we compared genome sequence of norovirus (4HS191) and astrovirus (1637F). This shows some discordance between the two sequences.

When comparing 5window apriori (Fig. 4-Fig. 6), we used amino 3. Both of them has the greatest percentage of amino acid 'S'. Cubic equation trend line of 4HS191 is

$$y = -0.0006x^3 + 0.0238x^2 - 0.3382x + 2.3535$$

and that of 532643969 is

$$y = -0.0018x^3 + 0.0677x^2 - 0.7616x + 3.2058$$

In the case of 7window apriori, we used amino 2. In norovirus, acid 'L' appeared most frequently but in astrovirus, acid 'S' appeared most frequently. Cubic equation trend line of 4HS191 is

$$y = -0.0003x^3 + 0.0082x^2 - 0.1344x + 1.7129$$

and that of 532643969 is

$$y = -0.0002x^3 + 0.0197x^2 - 0.4119x + 2.7324$$

In 9window, we also compared amino2 and 'L' is also the most repeatedly used acid. Cubic equation trend line of 4HS191 is

$$y = -0.001x^3 + 0.0303x^2 - 0.3494x + 2.2604$$

and that of 532643969 is

$$y = -0.0005x^3 + 0.0256x^2 - 0.4269x + 2.6837$$

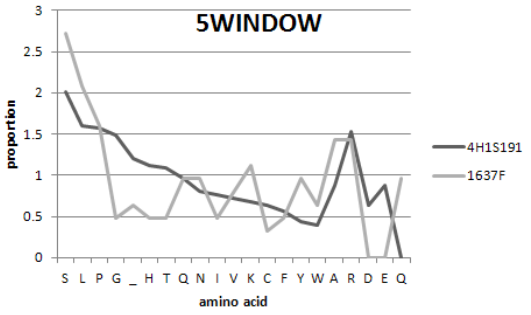

Fig. 4. 4HS191 and 1637F: 5window apriori.

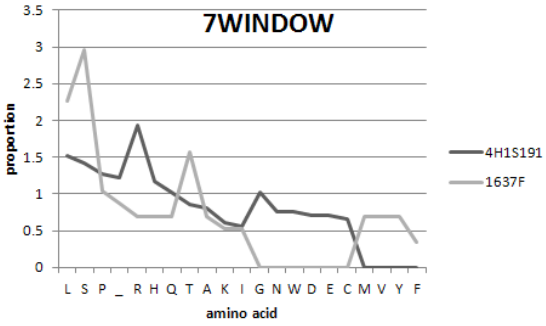

Fig. 5. 4HS191 and 1637F: 7window apriori.

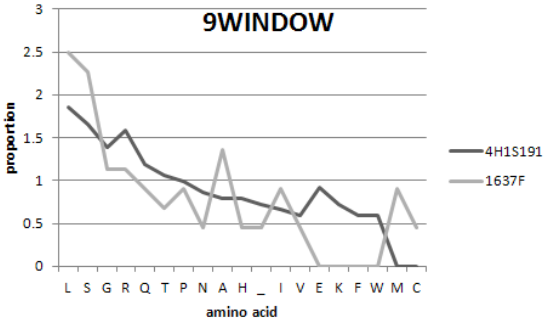

Fig. 6. 4HS191 and 1637F : 9window apriori.

3) 4HS191 & 532642969

Third, we compared genome sequence of norovirus (4HS191) and another genome sequence of astro virus (532642969). In this case, the two graphs show some similarities in their approximate shapes.

When comparing 5window apriori (Fig. 7- Fig. 9), we used amino 4. Both of them have the great percentage of amino

acid 'L'. Linear trend line of 4HS191 is

$$y = -0.0826x + 1.867$$

and that of 532643969 is

$$y = -0.0893x + 1.9381$$

These two results showed very similar slope.

In the case of 7window apriori, we used amino 2. In norovirus, acid '*R*' was, in the case of astrovirus, acid 'L' was the most frequently repeated. Linear trend line of 4HS191 is

$$y = -0.0759x + 1.6439$$

and that of 532643969 is

$$y = -0.0799x + 1.8317.$$

In 9window, we compared amino5 and norovirus has acid 'R', and astrovirus has acid 'L' as the amino acid that appears the most. Linear trend line of 4HS191 is

$$y = -0.0911x + 1.6119$$

and that of 532643969 is

$$y = -0.0869x + 1.5616.$$

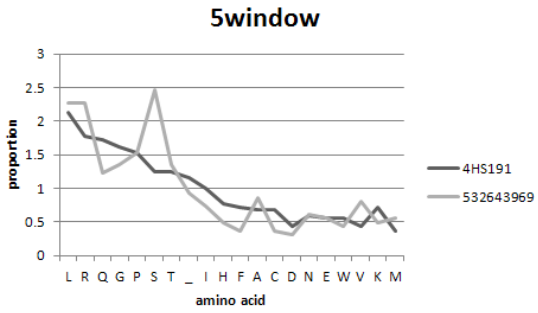

Fig. 7. 4HS191 and 532642969: 5window apriori.

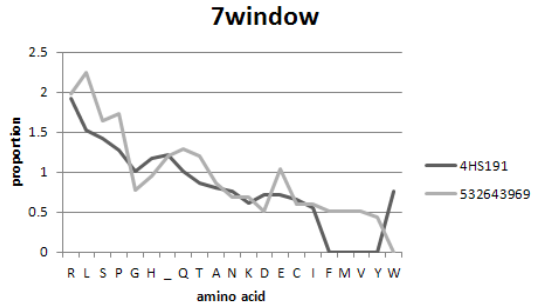

Fig. 8. 4HS191 and 532642969: 7window apriori.

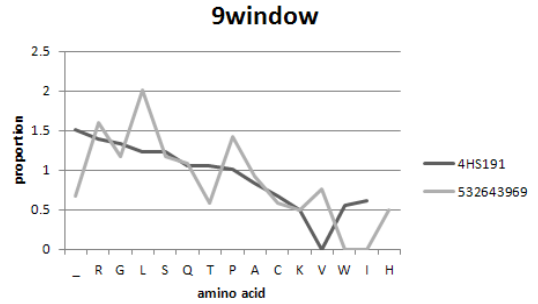

Fig. 9. 4HS191 and 532642969: 9window apriori.

Through analysis of the data, we found out that comparing of coding sequence shows more similarities than that of genome sequence. Coding sequence is part of the RNA base sequences which has genetic information that can be transcribed to make protein, while non-coding sequence is the sequences which doesn't have any information. As genome sequence includes both the coding and the non-coding sequences of the RNA, we assume that the differences in the comparison of the two genome sequences are partly influenced by non-coding sequences. So, we focused on the similarities between coding sequences that involve real information.

IV. Conclusion

In conclusion, according to high resemblance between the coding sequences of norovirus and astrovirus, further studies about genetically coincident base sequences or transcriptional regulatory factors can be progressed. Since the two viruses belong to the same group, according to the international classification and trigger similar symptoms, we can suppose that they might share common genetic factors that arouse the illness. So, our results can be used to assure that there might be the common illness-responsible genetic part in both viruses. Based on our results, we can make further study to make antiviral agents by attacking the shared part of the sequences. As many people suffer illness and even die from the varieties of the two viruses every year, this possibility of antiviral agents can be an ardent desire to the medical world.

Through this research using Apriori algorithm, we expected to find similarities in the coding sequence and genome sequence of the two viruses, human astrovirus and norovirus. However, a number of errors occured in our experiment results for several reasons.

First, the coding sequences and genome sequences we used in this experiment are just a single case of countless possible circumstances. Therefore more trials are required to generalize this experiment's results. In addition, the sequences used in the experiment were not various enough.

Also, since we selected the compared sequences based on our own judgments, the examined set may have not been the most similar sequence set. Besides, we only used the specific part of the sequences, not the whole.

The conformity of norovirus and astrovirus' coding sequences opens the possibility of complementary cooperation in the research of the two different viruses. For instance, the research data of norovirus can be used as a reference further astrovirus studies. Also, if treatment for astrovirus is developed, we might be able to apply it also on norovirus with slight reform.

References

[1] S. R. Finkbeiner, C. D. Kirkwood, and D. Wang. "Complete genome sequence of a highly divergent astrovirus isolated from a child with acute diarrhea," *Virol J*, vol. 5, no. 1, 2008, p. 117.

[2] P. T. Gia *et al*., "Human astrovirus, norovirus (GI, GII), and sapovirus infections in Pakistani children with diarrhea," *Journal of medical virology*, vol. 73, no. 2, 2004, pp. 256-261.

[3] H.-N. Yan *et al*., "Detection of norovirus (GI, GII), Sapovirus and astrovirus in fecal samples using reverse transcription single-round multiplex PCR," *Journal of virological methods*, vol. 114, no. 1, 2003, pp. 37-44.

[4] M.-T. Martha *et al*., "Molecular analysis of a serotype 8 human astrovirus genome," *Journal of General Virology*, vol. 81, no. 12, 2000, pp. 2891-2897.

[5] L. Catriona *et al.*, "Real-time reverse transcription PCR detection of norovirus, sapovirus and astrovirus as causative agents of acute viral gastroenteritis," *Journal of virological methods*, vol. 146, no. 1, 2007, pp. 36-44.
[6] G. Nick *et al.*, "Likelihood-based tests of topologies in phylogenetics," *Systematic Biology*, vol. 49, no. 4, 2000, pp. 652-670.
[7] H. Audray *et al.*, "Structural similarities between influenza virus matrix protein M1 and human immunodeficiency virus matrix and capsid proteins: an evolutionary link between negative-stranded RNA viruses and retroviruses," *Journal of general virology*, vol. 80, no. 4, 1999, pp. 863-869.

Eun-Young Kim was born in Korea. Kim is now studying at Hankuk Academy of Foreign Studies (Yongin, Korea) as third grade student in natural science program.

Min-Ji Kim was born in Korea. Kim is now studying at Hankuk Academy of Foreign Studies (Yongin, Korea) as third grade student in natural science program.

Eun-Ki Hong was born in Korea. Hong is now studying at Hankuk Academy of Foreign Studies (Yongin, Korea) as third grade student in natural science program.

Ye-Rim Cho was born in Korea. Cho is now studying at Hankuk Academy of Foreign Studies (Yongin, Korea) as third grade student in natural science program.