Analysis of Hepatitis C Virus using Data mining algorithm -Apriori, Decision tree

Ha YeongJeong and Tae Seon Yoon

Hankuk Academy of Foreign Studies, Yongin, Gyeonggi, Republic of Korea

ABSTRACT

Hepatitis C, which presents with symptoms such as acute fatigue and jaundice, is highly likely to become chronic, and is the main cause of liver cancer, attracting much public attention. Moreover, the number of infected people is increasing worldwide nowadays. However, we found that there are 6 different genotype in hcv. In vaccine and medicine developing for viruses, analysis of them is most important. Therefore, we decided to compare 6 genotype using Apriori algorithm and Decision tree algorithm. We tried to find out some difference between genotype 1 and others by analyzing the genotype 1, since genotype 1 is most common, and tried to find out the correlation between the genotype 1b and 2a with the highest number of infections in Korea. With these algorithm, we were able to find several rules and differences between them.

KEYWORDS

Hepatitis C virus, Apriori algorithm, Decision tree algorithm, virus, Bioinformatics

1.INTRODUCTION

Hepatitis C is an infectious disease caused by HCV. Unlike Hepatitis B, 75% of it becomes chronic disease and 20% of it becomes liver cancer and cirrhosis. Nowadays ,Hepatitis C, a major cause of liver cancer and chronic fatigue, unlike HAV and HBV, HCV is not focused on one place but spreads all over the world., and there is much interest in its treatment and prevention. [1] Now, at least 1% of HCV carriers. However, there are six genotypes in HCV that are classified into six genotypes and they are differ by 34% of the nucleotide sites over the complete genome, but all genotypes, RNA's 5 'and 3' end sites, UTR (Untranslated region), which plays important role in translation and viral RNA replication, is well preserved. Therefore, in our study, we usedrnacode of hcv in the experiment to compare hcv genotype.HCV genotype 1,hcv genotype 1b and 2a are the most common in Korea. In our study, the characteristics of the genotype a of each genotype through 1 to 6 and genotype 1b were classified using the data base of the genotype of hcv using the data mining algorithm (;Apriori algorithm, Decision tree). Furthermore, we conducted several experiments to discover differences and distinct characteristics of genotype 1b, 2a of the most patients in Korea have.

2.RELATIVE RESEARCH

2.1 Hepatitis C Virus

Hepatitis C Virus (HCV) belongs to the genus of Hepacivirus and is a member of Flaviviridae Family. HCV is a virus causing Hepatitis C. There are six genotypes in hcv. And 70% of hcv patients become infected by genotype 1 20% of patients become infected by genotype 2, and approximately 1% by other genotypes each. [2] In early infections, there are no symptoms or only mild symptoms, but when the inflammation becomes severe, symptoms such as chronic fatigue, abdominal pain, and yellow tinged skin appear. HCV, having a positive sense single stranded RNA genome. RNA genome consists of a single open reading frame that is 9600 nucleotide base

long. At both ends, it has a UTR that plays an important role in transcription and translation. UTR is a transcriptional initiation site that has an internal ribosome entry site (IRES) that initiates the translation of the virus polyprotein and initiates the synthesis of positive-strand viral RNA. The viral RNA between the 5 'and the 3' UTR synthesizes a polyprotein composed of 3,000 amino acids. They are divided into structural proteins (; C, E1, E2, p7) and non-structural proteins (NS2, NS3, NS4A, NS4B, NS5A, NS5B). The route of infection includes medical practices using non-sterile devices, dental treatment, and folk remedies. In many countries this is most common transmission of Hepatitis C. In Australia, however, Hepatitis C is infected when they use steroid-containing drug delivery devices with others. Other routes of infection include the use of needles that have not been sterilized or tattooed with non-sterile devices, or mother with hepatitis C can infect hepatitis C in the fetus during childbirth. In addition, blood products such as razors or toothbrushes can also be one of the infection route. [3],[4],[5] There is a growing interest in the prevention, diagnosis and treatment of hepatitis C, and the development of therapeutic agents is being accelerated nowadays. The only way to diagnose hepatitis c is blood inspection.There is no vaccination for hepatitis C yet, but several treatments has been made. Nowadays, we use sofosbuvir or simeprevir for chronic infection treatment. The cure rate is about 85%. Before these medications, we used a combination of peginterferon and ribavirin but its cure rate was 50%, which is lower to former and had much side effects. [6]

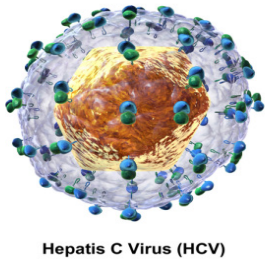

Figure 1 Hepatitis C virus

2.2 Data mining algorithm

In our study, we applied 2 data mining algorithm to each experiment (; Apriori Algorithm, Decision Tree Algorithm)

2.2.1 Apriori algorithm

Apriori algorithm is a data mining algorithm that derives association rules by finding frequent itemsets from a transaction dataset. Since algorithm is quiet easy to implement and the result is meaningful, this algorithm is one of the popular algorithm.[7] The association rule derived from Apriori algorithm is notable since we can define a minimum frequency as the minimum support and handle only the meaningful association rules on it.[8]

2.2.2 Decision tree algorithm

Decision tree algorithm is one way to display a algorithm which is a typical analysis in data mining. It uses decision tree as forecasting model which associates observed value and the desired value about certain topic. It is generally used to classify given things but not used in prediction. [9], [10]

3.MATERIALS AND METHODS

HCV is classified into 6 genotypes. Each genotype is divided into several subtypes which is represented by lowercase alphabet. (Genotype 1a, 1b, Genotype 2a, 2b, 2c ,2d , Genotype 3a, 3b, 3c, 3d, 3e, 3f, Genotype 4a, 4b, 4c, 4d, 4e, 4f, 4g, 4h, 4i, 4j, Genotype 5a, Genotype 6a) Genotype 1 is most common type worldwide but in Korea, Genotype 1b and 2a is most common. Our study focus on comparing genotype of hcv and analyzing difference various RNA sequence. Since there are much subtypes in genotype, we choose subtype a in each genotype as representative samples and genotype 1b since genotype 1 is most widely infected (in Korea and worldwide). So if we find evident difference between genotype 1 and others, we can say that we found amino acid which acts important role in hcv. Therefore, we extracted full RNA sequences of this virus from NCBI, and we applied 2 different data mining algorithm, apriori algorithm and decision tree to compare these genotype. We selected Apriori algorithm to discover similarity between amino acid sequences and Decision tree algorithm to figure out distinct differences between them. Since complete RNA code is so vast to inspect at once, we divided RNA code into several parts with criterion. We carried out three experiments with each algorithm ; 7-windows, 9-windows and 13-windows. For decision tree, we conducted 3 experiments with 10 fold cross validation.

4.RESULTS

4.1 Apriori Algorithm

4.1-(1) 7 window

Genotype	Position and frequency
1a	1. position5=P 63 2. position 3=S 61 3. position 1=P 58 4. position 2=S 55 5.position 7=S 53 6.position 7=P 51
1b	1. position6=S 75 2. position2=S 56 3. position4=P 56 4. position7=S 56 5. position3=S 53 6. position7=P 53 7. position3=R 52 8. position4=S 52 9. position4=R 51
2a	1. position5=G 67 2. position4=G 66 3. position7=G 62 4. position3=L 58 5. position4=R 58 6. position1=G 57 7. position6=R 56 8. position2=R 55 9. position5=L 55 10. position2=G 54 11. position3=R 51 12. position6=G 51
3a	1. position1=L 58 2. position4=L 58 3. position1=R 54 4. position2=L 52 5. position3=L 52 6. position4=R 51 7. position6=L 51 8. position7=S 51
4a	1. position4=L 52 2. position2=A 47
5a	1. position1=G 54 2. position1=L 53 3. position6=L 51
6a	1. position1=G 58 2. position5=R 57 3. position4=R 56 4. position5=G 51

Table 1 Rule extraction under 7 window using Apriori algorithm.

In this table, similarity and differences between genotypes are well represented. It is noticeable that all genotype show their unique characteristics. Genotype 1 was mainly consists of Serine(S) and Phenylalanine(P). They both have Serine in position 2, position 3 and position 7, Phenylalanine in position 7 in high proportion. According to this result, we can assume that Serine and Phenylalanine and play key roles in its genotype. However, genotype 1b also has Arginine (R) in mass proportion unlike genotype 1a had no rules include Arginine. Genotype 2a has Glycine (G) and Arginine in high proportion. We can find that 2a and 1b both have Arginine in considerable proportion, and both are the genotype which are dominant in Korean Hepatitis C patients. In Genotype 3a, 4a, and 5a, we can find Leucine (L) takes up a large proportion, and Leucine is an amino acid which can be found in most of viruses. Except Genotype 1, 3, 4, Glycine takes up a considerable proportion in virus. Genotype 1 is the most frequent and dominant genotype, which shows very distinctive feature from other genotype. And according to the result, we can assume that Serine and Phenylalanine may play the key role which make genotype 1 dominant around the world.

4.1-(2) 9 window

Genotype	Position and frequency
1a	1. position5=P 53 2. position8=S 49 3. position3=S 48 4. position4=P 48 5. position6=P 47 6. position7=S 47 7. position6=S 46
1b	1. position9=S 54 2. position7=P 48 3. position8=P 48
2a	1. position9=G 52 2. position8=R 51 4. position3=G 48 5. position6=G 48 6. position7=G 46 7. position1=P 45
3a	1. position1=L 48 2. position7=L 45
4a	1. position6=L 48
5a	1. position3=A 40 2. position3=L 40
6a	1. position3=R 47 2. position5=G 47

Table 2 Rule extraction under 9 window using Apriori algorithm

Most results has a tendency alike to results of experiment under window 7. Serine and Phenylalanine take up a considerable proportion in genotype 1. Leucine is a big part of amino

acid in genotype 3, 4, and 5. And Glycine takes up a lot of proportion in genotype 2 and 6. It is noticeable that Alanine (A) has found in only genotype 5a. And, as in genotype 1, phenylalanine occupies a large part in 2a. Therefore, since Genotype 1 and 2 are the most common genotypes of hepatitis C patients, this results are meaningful.

4.1-(3) 13 window

Genotype	Position and frequency
1a	1. position12=S 38 2. position1=P 33 3. position2=P 33 4. position7=P 33 5. position10=R 33
1b	1. position13=S 41 2. position2=P 35 3. position8=S 35
2a	1. position13=G 37 2. position6=G 36 3. position8=G 35 4. position11=G 35
3a	1. position4=P 37 2. position5=R 35 3. position5=L 35 4. position12=R 33
4a	1. position4=L 30 2. position5=V 29 3. position4=A 28 4. position2=A 27
5a	1. position4=A 32 2. position7=L 32 3. position4=G 29 4. position11=V 29
6a	1. position5=R 31 2. position8=L 31 3. position4=G 30 4. position6=G 30 5. position10=R 30 6. position12=G 30

Table 3 Rule extraction under 13 window using Apriori algorithm

In this table, similarity between genotype 4a and 5a is noticeable. Unlike other genotype, Valine (V) and Alanine (A) occupy large proportion in genotype 4a and 5a. And genotype 3a shows a tendency different from the previous experiment. Phenylalanine and Arginine occupies considerable proportion in this experiment unlike Leucine was the major amino acid in the

previous experiment. And each genotype has very distinctive features of amino acid that can be clearly classified into each genotype.

4. 2 Decision Tree algorithm

4.2.-(1) 7 window

Genotype	Rule
1a	position5 =P position6 =A position7 =G
2a	position3 =G position5 =A position7 =C position2 =H position5 =A position7 =G position5 =H position6 =D
3a	position1 =Q position4 =R position5 =E position2 =G position5 =S position7 =L position2 =R position3 =L position5 =S position3 =S position5 =H position6 =G
4a	position1 =L position2 =G position5 =L position1 =V position5 =A position7 = A position2 =A position4 =T position5 = G
5a	position2 =Q position5 =L position7 =F position2 =L position4 =C position5 =L position1 =R position2 =C position5 =S
6a	position4 =Q position5 =G position7 =K position5 =H position6 =C
1b	position4 =G position5 =I position6 =P position2 =Q position5 =L position7 =R position3 =S position4 =S position5 =R position3 =P position5 =A position7 =R position2 =N position5 =M

Table 4 Rule extraction under 7 window using Decision tree algorithm

In first experiment in decision tree, we found some rules with frequency over 0.75. In this table, several rules have found with experiments under 7 windows. Generally, position5 appeared in all genotype. According to this result, we can assume the fact that position 5 is the main factor differ genotype because it showed in every rule that separate genotype.

4.2.-(2) 9 window

Genotype	Rule
1a	position3 =S position6 =L position7 =S
	position4 =R position6 =C position7 =S
2a	position2 =V position5 =R position6 =Q
	position2 =G position5 =A position6 =D
	position1 =A position5 =C position6 =G
3a	position2 =H position5 =L position6 =G
	position3 =P position6 =L position8 =H
	position3 =S position4 =P position6 =C
4a	position6 =A position8 =L position9 =Y
	position1 =V position6 =H position7 =R
	position5 =D position6 =Q
5a	position2 =R position5 =G position6 =I
	position6 =A position9 =N
	position3 =L position5 =N position6 =I
	position6 =K position9 =A
6a	position3 =L position6 =R position7 =G
	position6 =E position8 =A position9 =H
1b	position3 =R position6 =L position8 =P
	position2 =G position5 =P position6 =S
	position2 =R position4 =Q position6 =S

Table 5 Rule extraction under 9 window using Decision tree algorithm

In this experiment, we found several rules with frequency over 0.75. And in this table position 6 was a main factor that differs each genotype since all genotype had rules with position 6. Also in this table, we can see that the amino acid types at position 6 are different for each genotype, so we might assume that each genotype has very distinct and unique characteristics.

4.2.-(3) 13 window

Genotype	Rule
1a	position4 =T position9 =T position12 =P position10 =A position11 =S position12 =L
2a	position4 =R position9 =A position12 =P position2 =D position12 =R
3a	position1 =H position12 =P position2 =A position4 =W position12 =R position2 =S position4 =S position12=P
4a	position7 =Y position12 =L position10 =A position12 =K position9 =C position12 =S position10 =K position12 =K
5a	position7 =N position12 =S position7 =S position9 =S position12 =S position4 =K position12 =A
6a	position4 =L position8 =I position12 =T position10 =C position12 =C position10 =I position11 =H position12 =L
1b	position4 =A position8 =L position12 =P position5 =R position6 =L position12 =G position12 =Q position13 =D

Table 6 Rule extraction under 13 window using Decision tree algorithm

We could also extract some rules with frequency over 0.75. In this experiment, position 12 was a critical factor that distinguishes genotype.All rules in 13 window had position 12 as constituent. Each genotype had its unique genotype, however, even genotype 1a and 1b were also distinguishable, since the component amino acids were similar, but the positions were different and the position 12 = P were only common part between them.

5. ANALYSIS

Using Apriori algorithm, we found that genotype were easy to distinguish since they all had distinct characteristics. However, in genotype 1, Serine and Phenylalanine took up a large proportion of amino acid but not common in other genotype. Since genotype 1 is the most widely spread around the world, Serine and Phenylalanine might be an important amino acid in virus to work, to survive and to spread widely. However, this was not proven experimentally in this study, therefore, further study of how Serine and Phenylalanine works in them is needed.

Applying Decision tree algorithm, first, we found the rules with frequency over 0.75. And we could extract numerous rules in those results. In each experiment, it showed that with certain position, each genotype was able to be distinguished easily. Since genotype of hcv is so diverse and there are about 50 subtypes, vaccine is hard to develop and until now vaccine for hcv have not been developed. And our data might help people to find specific rules to find differences and similarities between varioushcv, however, further study for detailed differences is necessary.In addition, in our study, certain correlation between genotype 1b and 2a was not found. Therefore, different form of study should be conducted to understand the certain relationship between hcv 1b

and 2a, which are highly prevalent in Korea.

6. CONCLUSION

Hepatitis C, which has a huge impact on health and daily life, has already attracted a great deal of attention, and research on therapeutic agents is being accelerated. However, as I mentioned before, due to the nature of the virus, there are many mutations and various genotype and subtype have distinct characteristics. Therefore, the vaccine has not yet been developed and the side effects of the therapeutic agent cannot be ignored. Although prevention is of course the most important, there is a need for improved medicine with a good understanding of hepatitis c virus. Additionally, in this study, we conducted experiments using full rna code of hcv, and derived several results. Further study is necessary to compare genotype and study deeper into hcv using the rna code of specific proteins that directly makes virus viral, such as NS2, NS3, NS4A, NS4B, NS5A, and NS5B.

7. REFERENCE

[1] HWANG SUN BONG "HEPATITIS C VIRUS AS AN EMERGING INFECTIOUS VIRUS."

[2] ROSEN, HR (2011-06-23). "CLINICAL PRACTICE.CHRONIC HEPATITIS C INFECTION".THE NEW ENGLAND JOURNAL OF MEDICINE. 364 (25): 2429–38

[3] POST, JEFFREY J., ET AL. "ACUTE HEPATITIS C VIRUS INFECTION IN AN AUSTRALIAN PRISON INMATE: TATTOOING AS A POSSIBLE TRANSMISSION ROUTE." MEDICAL JOURNAL OF AUSTRALIA 174.4 (2001): 183-184.

[4] PUIG-BASAGOITI, FRANCESC, ET AL. "PREVALENCE AND ROUTE OF TRANSMISSION OF INFECTION WITH A NOVEL DNA VIRUS (TTV), HEPATITIS C VIRUS, AND HEPATITIS G VIRUS IN PATIENTS INFECTED WITH HIV." JOURNAL OF ACQUIRED IMMUNE DEFICIENCY SYNDROMES (1999) 23.1 (2000): 89-94.

[5] CHEN, TRONG-ZONG, ET AL. "INJECTION WITH NONDISPOSABLE NEEDLES AS AN IMPORTANT ROUTE FOR TRANSMISSION OF ACUTE COMMUNITY-ACQUIRED HEPATITIS C VIRUS INFECTION IN TAIWAN." JOURNAL OF MEDICAL VIROLOGY 46.3 (1995): 247-251.

[6] "HEPATITIS C FAQS FOR HEALTH PROFESSIONALS". CDC. JANUARY 8, 2016. RETRIEVED 4 FEBRUARY 2016.

[7] AGRAWAL, RAKESH, AND RAMAKRISHNAN SRIKANT. "FAST ALGORITHMS FOR MINING ASSOCIATION RULES." PROC. 20TH INT. CONF. VERY LARGE DATA BASES, VLDB. VOL. 1215. 1994.

[8] WU, XINDONG, ET AL. "TOP 10 ALGORITHMS IN DATA MINING." KNOWLEDGE AND INFORMATION SYSTEMS 14.1 (2008): 1-37.

[9] WU, XINDONG, ET AL. "TOP 10 ALGORITHMS IN DATA MINING." KNOWLEDGE AND INFORMATION SYSTEMS 14.1 (2008): 1-37.

[10]HTTPS://EN.WIKIPEDIA.ORG/WIKI/DECISION_TREE_LEARNING

Authors

HayeongJeongwas born in Seoul, Korea, in 1999. She is a student in science major of HankukAcademy of Foreign Studies, Korea. She is mainly interested in virology and bio-science and applying bio-informatics to analyze and discover numerous facts about virus.

Taeseon Yoon was born in Seoul, Korea, in 1972. He was a Ph.D. candidate with the 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 a adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a computer science and statistics teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-do, Korea, 2013.