

Analysis of HSV-1 and HSV-2 that Cause Herpes Simplex with Apriori Algorithm, Decision Tree, and Support Vector Machine

Juhyun Lee*, Yeojin Jung*, Seungho Shin*, and Taeseon Yoon**

*International course, HAFS (Hankuk Academy of Foreign Studies)

**Faculty

juhyun62015@gmail.com, alwaysujj@gmail.com, leoshin99@naver.com, tsyoon@hafs.hs.kr

Abstract— The Herpes Simplex is a common fluid-intermittent disease that is common among not only adults but also children. Several symptoms include blisters and sores forming on infection sites that are mainly located in oral and genital body parts. The two main types of viral infection, commonly referred to as HSV-1 and HSV-2, are both part of the same Herpes Simplex Virus group (HSV), but they exhibit different characteristics in the main target of infection and the location of the infection sites. Normal cases of HSV-1 cause oral lesions and has less dangers of cross-generation infection when the female is pregnant, whereas HSV-2 cause genital lesions and usually is co-infected with human immunodeficiency virus type 1(HIV-1). Currently, no specific permanent vaccination is available for this disease and since it is chronic and can be transmitted as long as symptoms are present, the development of vaccination is significant at this stage. In order to provide further possibilities for vaccination development, we analyzed the genomes of HSV-1 and HSV-2 by using Apriori algorithm, Decision tree algorithm, and Support Vector Machine. Thus, we extracted results that showed the similarity and differences of the two genomes, mainly focusing on the prominent distribution of amino acids in each of the 3 categories we have chosen – dividing the genomes up in to 13, 17, and 19 windows. Through this, we have showed further potential in developing effective vaccination more specific to each type of Herpes Simplex.

Keywords— Herpes Simplex, HSV-1(herpes simplex virus type 1), HSV-2(herpes simplex virus type 2), Apriori algorithm, Decision Tree, SVM(Support Vector Machine)

I. INTRODUCTION

The Herpes Simplex, commonly known as the Herpes, can be classified into two types mainly by the differences in infection and symptoms caused by HSV-1 and HSV-2. Herpes simplex virus type 1 (HSV-1) causes a highly contagious infection, which is endemic throughout the world. Approximately 3.7 billion people under age 50 (67%) have HSV-1 infection globally. Majority of HSV-1 infections are oral herpes transmitted by oral-to-oral contact, but it can also cause genital herpes if transmitted by oral-genital contact. HSV-1 can possibly lead to a more severe complication such as encephalitis. Herpes simplex virus type 2 (HSV-2) have globally infected patients, with an estimated infection of 417

million people aged 15-49 (11%) worldwide. It is mainly transmitted during sex, through genital contact with someone infected with the virus, but also rarely transmitted from mother to infant during delivery. One of the possible complications regarding HSV-2 is that it can increase the chance of HIV infections. Symptoms of herpes caused by both HSV-1 and HSV-2 include painful blisters or open sores called ulcers, at the site of infection. The herpes infections are mostly asymptomatic, and people with infection cannot often notice that they are infected. Both HSV-1 and HSV-2 infections are life-long, and are most contagious when symptoms are present; however, they can still be transmitted without any particular symptom [1].

HSV is a double-stranded DNA virus with an enveloped, icosahedral capsid, and this paper investigates the similarities and the differences of the genomic sequences between the two distinct types of HSVs using decision tree, apriori algorithm, and SVM. Also, in the process of analyzing each genome sequence, we can further bring up potentials of developing vaccines related to respective types of viruses.

II. RELATED RESEARCH

As Herpes simplex virus types 1 and 2 (HSV-1 and HSV-2) cause prevalent, chronic infections to individuals, a lot of researches have been done on the subject. Due to the serious outcomes of the disease, recent studies are focused heavily on developing vaccines. Several protein subunit vaccines based on HSV-2 envelope glycoproteins are currently being processed through advanced-phase clinical trials. It is necessary to analyze the effect of vaccination on HSV-2 infection in that this infection emerges during asymptomatic shedding. These studies aim to identify whether these antigens enhance the level of cellular immunity [2].

Most viral studies done on the herpes simplex is on HSV-2. As HSV-1 and HSV-2 share similarities in pathogenetic methods and host interactions, studies done on developing antigens for HSV-2 provide partial protection from HSV-1 as well, even though the reverse process does not apply as

efficiently. This possibility has led to some skewed research done on vaccination itself.

In the case of actual vaccine development, two parallel studies showed that recombinant gD2 subunit vaccine with an alum/monophospholipid adjuvant (AS04) also induced neutralizing antigens and CD4+ immune responses. Across multiple trials, some showing negative results overall, gave us deeper understanding on the optimal level of population size, relevant design, and possible results for future observations to take place. This includes acquisition of HSV-2 infection to measure the vaccine's efficacy in reducing viral shedding frequency and in inducing sterilizing immunity [3].

Not only this, one study on quantitative analysis of HSV-1 and HSV-2 stated that the genomes of both viruses shared high degrees of similarities, but showed explicit, pathological differences in both human and animal hosts as well as cell cultures. Through primer extension analysis, mostly low relative abundance of HSV-2 RNA and general simultaneity in timing of active expression with HSV-1 was shown. According to the process of generating recombinant viruses with critical exchanges in promoters, specific differences in promoter elements of the two. In this study, the authors indicated that more extensive recombinants will be needed to analyze the potential different stabilities in each mRNA strand. Also, the differences in mRNA accumulation were shown through results, but the reasons for this observation were not clarified [4].

Another study clarified the differences between continents on both HSV-1 and HSV-2. During the process of investigating HSV infections in different countries, results have shown that the glycoprotein sequence diversity within HSV-2 is relatively lower than that of the HSV-1 strains. Also, less selective pressure, or the degree of evolutionary freedom, is applied on HSV-2 glycoproteins when compared to HSV-1. In the same study, differences in viral symptoms are shown. HSV-1 infection is mainly related to oral lesion and is frequently diagnosed in children, whereas HSV-2 is primarily associated with genital lesions. This relation to external wounds also influences the methodology in infection for the two types of virus. Finally, the study points out that HSV-2 usually causes co-infection of human immunodeficiency virus type 1 (HIV-1) when compared with that of HSV-1, the percentage of co-infection estimating around 50-90%. Thus, HSV-2 lesion—mainly genital lesion—causes HIV transmission and acquisition [5].

III. METHODS

A. Apriori Algorithm

The Apriori Algorithm is an algorithm used mainly for mining most-frequent data sets over transactional boundaries. This method is, simply put, a “bottom up” approach. First, finds frequent individual items from a data set, then proceeds to extend the items to larger and larger data sets. The data sets

are extended one at a time, and relevant data sets are selected by choosing those items of larger size that do appear sufficiently often in the database.

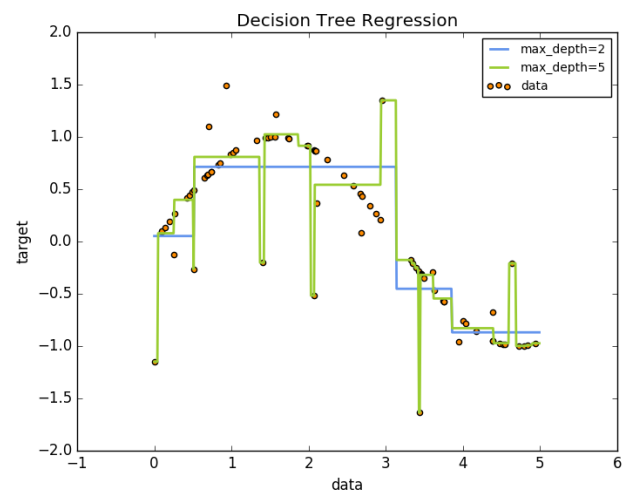
Apriori uses breadth-first search and a Hash-tree structure to efficiently determine relevant candidate items. In other words, it generates itemsets of length k from sets of length $k-1$. Then, according to the downward closure lemma, the candidate set now contains all relevant itemsets of length k . This process repeats, and the algorithm terminates when no larger extensions are applicable [6].

B. Decision Tree

A decision tree is decision support model that depicts possible outcomes, and is most often used to determine the most efficient and most likely strategy to reach a certain goal. A set of nodes represent different tests that decide certain attributes, each branch of the tree represents predicted outcomes, each leaf node represents the decision taken after computing all attributes, and the overall structure composed of these elements resembles a flowchart [7].

The decision tree can be linearized into decision rules. The outcome is the leaf node and the conditions of the path form a conjunction of the if clause, thus forming a statement like: if condition 1 & condition 2 & condition 3 outcome

It can also be expressed using a flowchart, where it is made easy for people to understand.



In the analysis process, the decision tree is used for classification and regression. It aims to create a model that predicts a target variable's value through simple decision rules. For example, according to the if-then-else decision rules, decision trees learn data like the graph below. When the maximum depth is deeper, the model is considered more accurate [8].

C. SVM (Support Vector Machine)

The Support Vector Machine is a method of classifying data in machine learning with associated learning algorithms that assign new examples to one group or the other between two separate groups. When we separate data types in mainly 2 groups, it is not necessarily linearly separable. Thus, by

defining a point using a kernel function, $k(x,y)$, the decision function of the SVM is fully specified by a subset of training samples, or the support vectors. The kernel function, taking different forms for different data types, represents a dot product of input at points fixed into higher dimensional feature space in transformation. In this case, it can be an adjusted parameter in certain functions. A certain data point is viewed as p -dimensional, and the purpose of the SVM is to decide whether we can classify the data with a p -1-dimensional hyperplane, which is called a linear classifier. A reasonable choice using the SVM would be a data set that sets the two groups furthest apart, or in other words, the one that represents the largest margin. When such a hyperplane exists, it is called the maximum-margin hyperplane, and the linear classifier is the maximum margin classifier [9].

Generally, SVM can be classified into 4 main categories according to the different error functions. Classification SVM type 1 minimizes the error of function $12wTw + Ci=1Ni$ with constraints $y_i(wT(x_i)+b)-1 \leq 0$, $i=1, \dots, N$ and type 2 minimizes function $n 12wTw-v + Ci=1Ni$ with constraint $y_i(wT(x_i)+b)-1 \leq 0$, $i=1, \dots, N$ and 0. In a regression SVM, one can estimate dependence of variable y on independent variables x . For the basic functions, it is $y=f(x) + \text{noise}$ and therefore the task is to find a functional form for f that predicts new cases in SVM.

IV. RESULTS

We extracted and analyzed the genome sequences of HSV-1 and HSV-2 from the National Center for Biotechnology Information (NCBI) [10], [11]. By using apriori algorithm, decision tree, and SVM, we focused on the amino acid composition in each virus to find out the similarities and differences between two viruses.

There were mainly 4 amino acids present in both HSV-1 and HSV-2: A, R, G, and P. A is an abbreviation for alanine, whose main functions are the primary amino acid in sugar metabolism and a boost in immune system in producing antibodies. R is short for the amino acid arginine that mainly functions as an essential factor for normal immune system activity and production of growth hormones. As it speeds up the rate of viral growth, it exerts influence in both HSV-1 and HSV-2. For P, or proline, it mainly functions as a critical component of cartilage directly relating to joints, tendons, and ligaments. Finally, glycine, G, mainly functions as part of the structure of hemoglobin, and part of cytochromes, enzymes involved in energy production.

A. Apriori Algorithm

We use the Apriori Algorithm to find the common features of two viruses: HSV-1 and HSV-2. First, we compared the frequencies of amino acids present in each of the viruses. Most similar were the frequencies of glycine in both HSV-1 and HSV-2. However, the amount of alanine differed by 1001 amino acids with less alanine present in HSV-1, and the

frequency of proline differed by 771 amino acids, with more proline in the genome of HSV-2. Now, let us compare the amount of amino acid in each arbitrary position.

We have extracted data for each window of the division in 13, 17, and 19 windows. Then, we compiled data into positions of both viruses, HSV-1 and HSV-2, and set the standard of the two frequencies being similar as with the difference in frequency to be less than 10. The results are as follows.

TABLE 1. RULE EXTRACTION UNDER 13, 17, 19 WINDOWS APRIORI

Window 13			Window 17		Window 19	
	HSV-1	HSV-2	HSV-1	HSV-2	HSV-1	HSV-2
Amino acid G						
Amino 1	504	501				
Amino 2	481	482				
Amino 3	506	500	399	396		
Amino 4	519	510	378	388		
Amino 5	491	485	383	382		
Amino 6			365	374		
Amino 8	503	496				
Amino10	485	476				
Amino12	508	502				
Amino19					354	351
Amino acid R						
Amino 1			432	429		
Amino 5			439	440	402	401
Amino 6	571	568				
Amino 7	565	567			411	414
Amino 8			423	430	363	355
Amino 9	574	570	459	460		
Amino acid P						
Amino 2	443	453			351	358
Amino 3					334	331
Amino 5			351	355		
Amino 7	472	465				
Amino10	498	489			331	324
Amino acid A						
Amino 2	443	453				

Through this, we can see that in divisions to 13 and 17 windows showed more similar data frequencies in glycine, division to 19 windows showed slightly more variation (frequency increases more when boundary is extended to 20). Also, Amino acid R showed relatively similar frequency in windows all windows when arbitrarily divided into 13, 17, and 19 windows. Thus, we can deduce the more consistent trends, which is that both HSV-1 and HSV-2 possess relatively similar levels of arginine and glycine.

It is noticeable that the same amino acid that can potentially inhibit the Herpes Simplex when bioengineered correctly is present in large quantities in both Herpes Simplex 1 and 2. A conclusion can be drawn that potential cures that are introduced

such as L-Arginine are effective as it could substitute the arginine present in both types of the virus. [5]

Also, glycine present, according to a previous research done on the effects of conserved glycine in transmembrane transfer processes, as glycine has prominent effects on the TM domain conformation of specific proteins, the presence of glycine in the genomes of HSV-1 and HSV-2 would have relationships with this particular function. However, as it did not directly affect the degree of infection, we can conclude that this does not have a major impact on the overall infection process, but rather subtly alters the viral infection during its process.

B. Decision Tree

We conducted 10-fold cross validation experiment using two viruses, HSV-1 and HSV-2, to observe certain patterns. HSV-1 is class 1 and HSV-2 is class 2. Similar to apriori algorithm, we used 13, 17, and 19 windows for efficient rule extraction.

Among extraction results, we divided into two classes, gathered the rules of patterns, and then selected data with high frequencies. We set the frequency equal to or above 0.85 since above 0.85 is almost regarded as fixed rule patterns to influence the experiment results.

TABLE 2. RULE EXTRACTION UNDER WINDOW13

	Amino acid	Frequency	Tally
<i>Class 1</i>			
<i>Rule 1</i>	Pos2=V Pos3=Y	0.857	4
<i>Rule 2</i>	Pos3=H Pos5=K	0.857	3
<i>Rule 3</i>	Pos4=P Pos5=Y	0.857	3
<i>Class 2</i>			
<i>Rule 1</i>	Pos5=E Pos13=Q	0.866	5

Contrasting to apriori algorithm, decision tree gives more detailed information about the amino acid in the inner segments of viruses. Through specified rules of patterns, we identified HSV-1 and HSV-2 are different in some aspects. Analysis of window 13 involves fewer rules with high tallies. When classifying class 1 virus, or the HSV-1, the positions of high-frequency amino acids were located in the upper windows of the 13. Also, in class 1, tyrosine amino acid, Y, located in both positions 3 and 5 of class1, while no rule of class 2 exhibited such characteristics. Finally, position 5 is present in three of the rules: class 1 rule 2, class 1 rule 3, and class 2 rule 1. We have concluded that position 5 is important decision factor that differentiates HSV-1 and HSV-2.

TABLE 3. RULE EXTRACTION UNDER WINDOW17

	Amino acid	Frequency	Tally
<i>Class 1</i>			
<i>Rule 1</i>	Pos4=I Pos12=R	0.8762	7
<i>Rule 2</i>	Pos2=M Pos12=A	0.857	3
<i>Class 2</i>			
<i>Rule 1</i>	Pos9=T Pos10=R Pos12=S	0.869	3

The rule extraction under window 17 involves position 12, which differentiates the two types of viruses. HSV-1 includes amino acid A, whereas HSV-2 notably includes amino acid S. Hence, position 12 is crucial in distinguishing HSV-1 and HSV-2. In window 17, similarities were not found.

TABLE 4. RULE EXTRACTION UNDER WINDOW19

	Amino acid	Frequency	Tally
<i>Class 1</i>			
<i>Rule 1</i>	Pos1=Q Pos6=S	0.890	6
<i>Rule 2</i>	Pos5=S Pos13=H	0.907	4
<i>Rule 3</i>	Pos1=R Pos10=R Pos13=L	0.863	3
<i>Rule 4</i>	Pos1=G Pos15=R Pos17=R	0.857	3
<i>Class 2</i>			
<i>Rule 1</i>	Pos1=A Pos5=M	0.886	5
<i>Rule 2</i>	Pos1=R Pos9=Q	0.893	5
<i>Rule 3</i>	Pos1=F Pos15=W	0.875	3

According to this table, it's notable that more rules of patterns were found than windows 13 and 17. Analysis of window 19 using decision tree revealed 4 major rules for HSV-1 and 3 rules for HSV-2. Using Rule 3 in class 1 and Rule 2 in

class 2, arginine, R, is observed in both position 1 in class 1 and class 2. In addition, amino acid R is observed in various positions including 1, 10, 15, and 17 in class 1. While position 1 in class 1 contains amino acid Q and G, position 1 in class 2 contains amino acid A and F. Not only this, class 1's rule of Pos5=S has frequency rates of 0.907. This is markedly high frequency, which means this pattern is almost fixed except rare cases. In contrary, class 2 has rule of Pos5=M. Thus, these differences evoke different pathological features.

Apriori algorithm and decision tree proves that HSV-1 and HSV-2 are similar member of Herpesviridae in overall point of view though when observed thoroughly, they reveal difference of composition of amino acid in each position.

C. SVM

In the process of conducting SVM experiment, we used 13, 17, and 19 windows. Through mapping the data with four ways that are normal, poly, sigmoid, and R.B.F, we concentrate on the accuracy to prove their linear and nonlinear classification.

TABLE 5. CLASSIFICATION OF ACCURACY RATE UNDER SVM

Window	SVM Function			
	NORMAL	POLY	SIGMOID	R.B.F
Window 13	52.62	51.25	50	78.62
Window 17	52.92	49.83	50	78.75
Window 19	53	49.9	50	73.1

To find out the scientific accuracy to classify two viruses, we analyze the normal, poly, sigmoid, and R.B.F function. The table shows that NORMAL function divided into approximately 52.85%, POLYNOMIAL function into 50.3%, SIGMOID function into 50% in all three cases of 9 window, 13 window, and 17 window, and R.B.F into 76.83%. Thus, non-linear classification, R.B.F, proves HSV-1 and HSV-2 most accurately differentiated.

V. CONCLUSIONS

In conclusion, we can conclude that the specific functions relating to the infectious degrees and symptoms of the virus has a lot to do with, first, the presence of two amino acids, arginine and glycine, which we could infer through the analysis of the Apriori algorithm, and the position 5 of window 13, position 13 of window 17, and position 1 of window 19. During the analysis of the Apriori algorithm, we found the overall trends in similar numbers of amino acids roughly equivalent to the trends in each of the 3 window divisions (13, 17, and 19). For the Decision tree algorithm, as the genomes we are comparing is relatively large with approximately 150,000 amino acids, to get sufficient number of rules from the collection of data, we set the frequency rate to 0.85. The number of rules found when frequency rates are 0.80 or 0.75, both of which are sufficient when deciding the accuracy of the data. Also, with regards to

the SVM classification, the data provided is not linear nor polynomial-differentiated data, but is most accurately divided by the R.B.F. method, or non-linear method by an accuracy rate of 78.62, 78.85, and 73.1 in each of 13, 17, and 19 windows.

Thus, we can point to the specific loci which decide the different characteristics of the two viruses through the sum of data. Also, we can further elaborate on the efficiency of previously existing treatment and medicine of the Herpes Simplex. Previous uses of L-arginine and other relevant substances relates to the composition of both genomes of the HSV- 1 and HSV- 2.

Further studies point to the possibility and specific functions of each window division and the impact each presence of amino acid has on each window division. As the loci of genetic sequence that includes each codon of the amino acid is crucial in deciding the function of the viral substance, this research will further specify and elaborate on the sequence analysis we have performed. Even though distinct treatment can be given accordingly when treating the disease, no significant vaccine has been found for the Herpes Simplex, through the analysis of prevalent amino acids as well as specific characteristics for both, we can specify the reasons why certain antigens are ineffective in different types of the Simplex virus.

REFERENCES

- [1] "Herpes Simplex Virus." World Health Organization. World Health Organization, n.d. Web. 13 Nov. 2016.
- [2] Koelle, David M., and Lawrence Corey. "Recent Progress in Herpes Simplex Virus Immunobiology and Vaccine Research." *Clinical Microbiology Reviews*. American Society for Microbiology, Jan. 2003. Web. 13 Nov. 2016.
- [3] Johnston, Christine, David M. Koelle, and Anna Wald. "HSV-2: In Pursuit of a Vaccine." *The Journal of Clinical Investigation*. American Society for Clinical Investigation, 01 Dec. 2011. Web. 13 Nov. 2016.
- [4] "Quantitative Comparison of the HSV-1 and HSV-2 Transcriptomes Using DNA Microarray Analysis." *Quantitative Comparison of the HSV-1 and HSV-2 Transcriptomes Using DNA Microarray Analysis*. N.p., n.d. Web. 13 Nov. 2016.
- [5] Lamers, Susanna L., Ruchi M. Newman, Oliver Laeyendecker, Aaron A. R. Tobian, Robert C. Colgrove, Stuart C. Ray, David M. Koelle, Jeffrey Cohen, David M. Knipe, and Thomas C. Quinn. "Global Diversity within and between Human Herpesvirus 1 and 2 Glycoproteins." *Journal of Virology*. American Society for Microbiology, 15 Aug. 2015. Web. 13 Nov. 2016.
- [6] "10 Apriori." Apriori. N.p., n.d. Web. 13 Nov. 2016.
- [7] "11 Decision Tree." Decision Tree. N.p., n.d. Web. 13 Nov. 2016.
- [8] "1.10. Decision Trees¶." 1.10. Decision Trees — Scikit-learn 0.18.1 Documentation. N.p., n.d. Web. 13 Nov. 2016.
- [9] "Voltage Problem Area Classification Using Support Vector Machine SVM." *International Conference Data Mining, Civil and Mechanical Engineering (ICDMCME'2014)*, Feb 4-5, 2014 Bali (Indonesia) (2014): n. pag. Web.
- [10] "Human Herpesvirus 1 Complete Genome - Nucleotide - NCBI." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 13 Nov. 2016.
- [11] "Herpes Simplex Virus Type 2 (strain HG52), Complete Genome - Nucleotide - NCBI." National Center for Biotechnology Information. U.S. National Library of Medicine, n.d. Web. 13 Nov. 2016.



Juhyun Lee was born in Republic of Korea. She is currently attending HAFS. She is passionate in learning chemistry, biology, and pharmacology. This is her first trial to write a paper. Although she faced some difficulties, she didn't give up and kept working. She wants to develop a new medicine in the future, especially a medicine that can cure cancer.



accomplish her goal.

Yeojin Jung is currently a junior at HAFS. She is interested in a variety of fields including bioinformatics, neuroscience, marketing, and economics, and hopes to merge these fields together in to a new area of study in the future. Through her unique perspective in viewing things around her and original thoughts in linking ideas of different fields, she is currently trying hard to accumulate expertise in a variety of fields as to



Seungho Shin was born in Seoul, Korea, in 1972. Since 2015, he has been studying in the International Department, Hankuk Academy of Foreign Studies. He is especially interested in the field of bioinformatics, and he wishes to study artificial intelligence and robotics in the future.



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.

Taeseon Yoon was born in Seoul, Korea, in 1972. He received his 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