

Mining Fuzzy Amino Acid Associations in Peptide Sequences of Herpes Simplex Virus

Rishu Gupta*

Abstract

Herpes is a usually mild recurrent skin condition in which most infections are unrecognized and undiagnosed. The mechanism of disease is still not well understood. The analysis of peptide sequences of herpes can reveal information which may be useful for understanding the mechanism of disease. In this paper an attempt has been made to develop a model for mining fuzzy amino acid associations in peptide sequences of herpes virus. The uncertainty arising due to variation in length of sequences and this is handled by employing fuzzy sets. Total 9160 sequences were taken from National Centre for Biotechnology Information. After that around 4004 non-redundant peptide sequences of herpes virus filtered to form the dataset. This dataset is transformed to fuzzy transaction dataset and their fuzzy support and confidence have been computed. The patterns generated from this model can be useful in understanding the structure, function and interaction of the protein in the disease.

Keywords: HSV, ARM, Frequent Pattern, Threshold
Abbreviations: ARM-Association Rule Mining, HSV- Herpes Simplex Virus

1. Introduction

1.1. Herpes Simplex Virus

Herpes simplex is a viral disease of skin or mucous membrane caused by both herpes simplex virus type1

(HSV-1) and type2 (HSV-2) both of which belong to a wider group called herpesviridae of the herpes virus family that infects humans. An acute infectious disease characterized by superficial vesicles containing clear fluid in the skin and mucous membranes particularly of the buccal area and sometimes on the conjunctiva, cornea and genitalia, but may occur anywhere on the body. HSV-1 which produces most cold sores also called oral herpes or herpes labialis. The visible symptoms of which are colloquially called fever blisters, is an infection of the face, mouth, throat, eye, lips or gums[5]. Genital herpes simply known as herpes is another most common form of herpes that causes primarily genital infection. They can be spread when an infected person is producing and shedding the virus [2,3].

Herpes simplex virus infection of the eye results in keratoconjunctivitis, a serious condition that sometimes leads to corneal blindness. After a primary infection, the virus travels to nerve cell ganglion where it persists in a dormant phase. Various factors such as sun exposure, chapping or abrasion of the skin, fever, stress, fatigue, or menstruation can reactivate the virus. In genital infections (asymptomatic) recurrences are common while herpes simplex can occur in seemingly healthy people, patients with cancer, acquired immunodeficiency syndrome (AIDS), and other diseases associated with impaired immune function are especially prone to such infections[5,8].

HSV have three main structural components. First is a central core holds the viral DNA, second is an inner core which is surrounded by an envelope that is made of viral

* M.Tech Scholar, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, India.
E-mail: altruismrishugupta@yahoo.in

glycoproteins and host cell membranes, and a capsid. The tegument is located between the capsid and the envelope (Spear *et al.* 2004) and various proteins that are delivered into the infected cell upon cell fusion (Smiley *et al.* 2004).

Herpes simplex is easily transmitted by direct contact with a lesion or the body fluid of an infected individual. Transmission may also occur through skin to skin contact during periods of asymptomatic shedding. Transmission is likely to occur during symptomatic reactivation of the virus that causes visible and typical skin sores. Asymptomatic reactivation means that the virus causes atypical, subtle or hard to notice symptoms that are not identified as an active herpes infection. HSV-1 is usually acquired orally during childhood, but may also be sexually transmitted. HSV-2 is primarily a sexually transmitted infection but rates of HSV-1 genital infections are increasing. HSV-1 classic lesions tend to form as small fluid filled blisters that can appear as a single blister or in a cluster. Sores may also appear inside the mouth or on the back of the throat and the lymph nodes in the neck may swell. Oral herpes is not always limited to the area. This occur by age 15 around 25% of UK population and by age 30 around 50%.

Worldwide rate of HSV infection are between 65% and 90%. HSV1 is more common than HSV2 with rates of both increasing as people age. An estimated 536 million people worldwide were infected with HSV-2 in 2003, with the highest rates in sub-Saharan Africa and the lowest rates in Western Europe.

In the UK in 2004, there were 19,180 cases of a genital herpes. In the US, 57.7% of the population is infected with HSV-1 and 16.2% are infected with HSV-2. Among that HSV-2 seropositive, only 18.9% were aware that they were infected. During 2005–2008, the prevalence of HSV-2 was 39.2% in blacks and 20.9% in women.

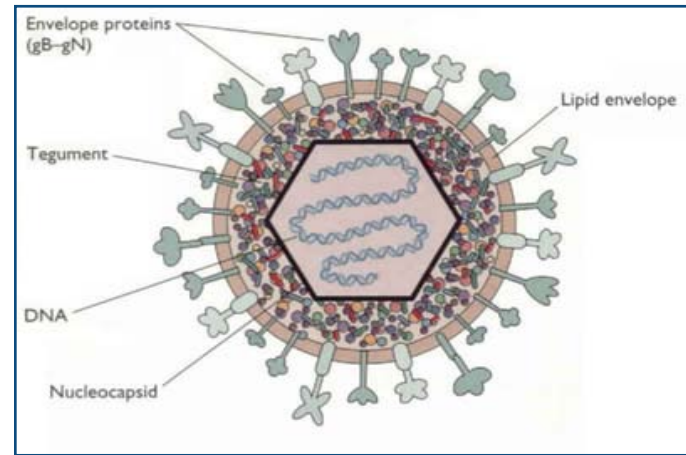
Till now no treatment is available for this virus and also no method is available to eradicate herpes virus from the body; but some antiviral medications are available to reduce the frequency and duration.

Association rule mining was first proposed by Agrawal *et al* in 1993 [11]. The study of amino acid association patterns can give us insights about the mechanism of disease and its treatment.

A number of algorithms are reported in the literature for association rule mining. Some researchers have made an

attempt to explore amino acid and nucleotide association patterns in peptide and nucleotide sequences but no attempt is reported in the literature on the study of amino acid associations in peptide sequences of herpes simplex virus.

Figure 1. Life Cycle of Herpes Simplex Virus [14]



In this paper an attempt has been made to study the amino acid association patterns in peptide sequences of herpes simplex virus [1,4,5,7,9].

2. Preliminaries

Database: The protein database is a collection of sequences from several sources, including transformation. The database of protein can be easily obtained from public websites. In this study, the data of herpes simplex virus were obtained from NCBI; total 9160 sequences have been retrieved from this site. The length of the sequences was varying. Therefore to remove the bias in degree of relationships among amino acids due to variation of length of sequences normalized is performed for each sequence. After normalization we get 4004 non redundant sequences.

Item set: A collection of one or more items.

Transaction: Each transaction is a list of items or each transaction is a subset of item set.

$$T_j = \{ f_j(A_i); i=1 \text{ to } 20 \} \text{ for all } j=1 \text{ to } 20$$

Support: Fraction of transactions that contain both item set X and item set Y

$$\text{Support}(X \rightarrow Y) = \text{Support}(X \cap Y)$$

Confidence: It measures how often item set Y appear in transaction that contain item set X

$$\text{Confidence } (X \rightarrow Y) = \text{Support } (X \cap Y) / \text{Support } (X)$$

Frequent item set: An item set whose support is greater than or equal to threshold value.

Fuzzy item set: A set of one or more fuzzy items.

Fuzzy Transaction: Each transaction is a set or subset of fuzzy item set.

$$T_j = \{ \mu_j(A_i); i=1 \text{ to } 20 \} \text{ for all } j=1 \text{ to } 20$$

Fuzzy Frequent item set: A fuzzy item set whose support is greater than or equal to threshold value.

3. Method

The membership of i^{th} sequence for each amino acid is calculated by

$$\mu_i = f_i(A)/L_i$$

Where, L_i is the length of sequence and $f_i(A)$ is the frequency of amino acid A in first sequence.

The calculated threshold value was based on the assumption that there are 20 amino acids and each will have equally likely chance of appearing in a sequence. The frequent amino acids are the amino acids that were above the support threshold value.

Thus the ordinary threshold value has been calculated by

$$\text{Tr} = [\max(\text{frequency support}) + \min(\text{frequency support})]/2$$

Thus the fuzzy threshold value has been calculated by

$$\begin{aligned} \text{Tr} &= (1/20) * N \\ &= .05 * N \end{aligned}$$

Where Tr is the threshold and N is the number of sequences.

The amino acid association patterns for peptide sequences can be calculated by support and confidence as follows:

Table 1. Frequency Support and Membership

Amino Acid	Ordinary Frequency Support	Fuzzy Result	Fuzzy Result
		Frequency	Membership
Glycine [G]	143215	143215	319.13058
Alanine [A]	216509	216509	436.6497
Valine [V]	124783	124783	265.00296
Leucine [L]	172978	172978	363.1411
Isoleucine [I]	59683	59683	130.92943
Methionine [M]	33969	33969	77.75225
Proline [P]	159379	159379	366.26627
Phenylalanine [F]	62904	62904	130.83112
Tryptophan [W]	20638	20638	47.638584
Asparagine [N]	45603	45603	99.019066
Glutamine [Q]	59528	59528	127.785255
Serine [S]	115992	115992	268.93707
Threonine [T]	116014	116014	265.93707
Tyrosine [Y]	51527	51527	110.83533
Cysteine [C]	30963	30963	70.659584
Aspartic Acid [D]	96416	96416	205.22684
Glutamic Acid [E]	95186	95186	202.78477
Lysine [K]	45296	45296	99.41711
Arginine [R]	144753	144753	315.37317
Histidine [H]	45809	45809.0	99.06875

Table 2. Analysis of Frequent Pattern of Peptide Sequences of HSV

	Ordinary	Fuzzy
Total no of non-redundant peptide sequences	4004	4004
Threshold	118573.5	200.2
No of amino acids above threshold	6	10
Maximal no of frequent pattern	4	5
No of two frequent pattern	12	28
No of three frequent pattern	8	44
No of four frequent pattern	2	23
No of five frequent pattern	-	4

Table 3. Physicochemical Based Analysis of Peptide Sequences of HSV

Parameters	Responsible Amino Acids	Ordinary	Fuzzy
Hydropathicity(Gravy)		-0.27188376	-0.27188376
Aliphatic Index		78.2225	78.2225
Molecular Weight		49877.375	49877.375
Absorbance		0.99282426	0.99282426
Aromaticity	W Y F	7.2253966%	7.2253966%
Protein stability	R D C E K	22.314302%	22.314302%
C-Beta Branched	V I T	16.53019%	16.53019%
Polarity	R N D C E Q H K S T G Y	46.570026%	46.570026%
Salt bridged	R N E K	17.896982%	17.896982%
Helix	M A L E K R Q H	43.00637%	43.00637%
Beta sheet	V I T C W F Y	25.520235%	25.520235%
Coil	N D P S G	31.423506%	31.423506%
Positive charged	K R H	10.359398%	10.359398%
Negative charged	D E	10.190108%	10.190108%

The Support of single amino acid set is calculated by

The Support of 2-amino acid set is calculated by

The Support of n-amino acid set is calculated by

The confidence is calculated by

The confidence for n-amino acid is calculated by

All above expressions are used to compute amino acid association's patterns.

4. Tool

To implement above method, a tool is developed that require Java EE platform. Here Netbeans 7.2 (a popular JAVA Editor) is used for performing this task. This tool takes amino acid sequences as an input (all sequences

should be in FASTA format) and generates frequent patterns as an output.

5. Results and Discussion

The total sequences of HSV are taken from NCBI and after normalization we have 4004 non-redundant sequences remains in data set.

In above table gray color highlights the frequent amino acids with their frequency support and membership. Amino acid G, A, V, L, P, S, T, D, E and R are more frequent in dataset of herpes virus.

According to table 2 maximal frequent patterns varies that is frequent patterns depends on the length of the sequences as well as on number of sequences.

The number of frequent amino acid in fuzzy result is more than the ordinary result and also more maximal number of frequent patterns in fuzzy result. Therefore fuzzy result is more usable and better for understanding of associations.

Table1 shows the frequent amino acids in ordinary and fuzzy result. Amino acid Glycine, Alanine, Valine, Leucine, Proline, and Arginine are frequent in ordinary as well as in fuzzy result but from fuzzy result few more amino acids Threonine, Serine, Aspartic acid and Glutamic acid are also frequent.

The physiochemical properties also vary according to length of sequences. From above study maximal ten amino acids are frequent in fuzzy result. As hydrophobicity or gravity values are negative in both ways that indicates the sequences of HSV are hydrophilic in nature. Amino acid Aspartic acid (D), Glutamic acid (E) and Arginine (R) are responsible for protein stability. In structure formation only four amino acid Alanine (A), Leucine (L), Glutamic acid (E) and Arginine (R) are responsible for helix formation; Valine(V) and Threonine (T) are responsible for Beta sheet. Coils are highly formed in structure of HSV. There is no difference between physicochemical properties result i.e. in both ways result is same.

6. Conclusion

The fuzzy model for mining frequent amino acids in peptide sequences of Herpes virus was developed with the help of Apriori Algorithm. The approach is appropriate in view of uncertainty arising due to variation in the number of sequences as well as in length of the sequences. The mining association patterns have been obtained successfully to generate association's rules for predicting the structure and properties of sequences. The obtained fuzzy result indicates that amino acid Glycine (G), Alanine (A), Valine (V), Leucine (L), Proline (P), Serine (S), Threonine (T), Aspartic Acid [D], Glutamic Acid (E) and Arginine (R) are more frequent in dataset of HSV. Amino acid Glycine (G) is more frequent which is essential for HSV; and Lysine (L) reduce the number of recurring outbreaks of cold sores. From ordinary result only six amino acids are frequent and from fuzzy ten amino acids are frequent. Number of maximal frequent pattern is also higher in fuzzy result. Therefore, it can be conclude that fuzzy result is more efficient as compared to ordinary result.

The found result indicates that the sequences of Herpes Simplex Virus are hydrophilic in nature as hydrophobicity is negative. Therefore the discovered association relationship and patterns can be useful for development of mechanisms for diagnosis and treatments of the disease.

7. Acknowledgement

This research paper is made possible through the help and support from everyone, including: parents, teachers, family, friends, and in essence, all sentient beings.

First and foremost, I would like to thank Dr. Kamal Raj Pardasani for their most support and encouragement. They kindly read my paper and offered invaluable detailed advices on organization, and the theme of the paper.

References

1. Kumari, T. & Pardasani, K. R. (2012). Mining fuzzy associations among amino acids of class A GPCR. *Online Journal of Bioinformatics*, 13(2), 202-213.
2. Genital Herpes. (2012). Herpes - genital, Herpes simplex - genital, Herpesvirus 2, HSV-2. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001860/>
3. Herpes Oral. (2012). Cold sore, Fever blister, Oral herpes simplex, Herpes labialis, Herpes simplex. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmedhealth/PMH0001631/>
4. Chiu, H. W. & Hung, F. (2008). Association Rule Mining from Yeast Protein Interaction to Assist Protein-Protein Interaction.
5. Wang, M., Shang, X. & Li, Z. H. (2008). Sequential Pattern Mining for Protein Function Prediction.
6. Lopez, F. J., Blanco, A., Garcia, F., Cano, C. & Marin, A. (2008). *Fuzzy association rules for biological data analysis*.
7. Gupta, N. (2006). Mining quantitative association rules in protein sequences, *Data Mining, LNAI*, 3755, 273-281.
8. Gaby, A. R. (2006). Natural Remedies for Herpes Simplex.
9. Besemann, C., Denton, A. & Yekkirala, A. (2004). Differential Association Rule Mining for the Study of Protein Interaction Networks.
10. Stancik, L. M. & Slonczewski, J. (2001). Antiviral Therapy Resistance of Herpes Simplex Virus1 and 2.

11. Agrawal, R. & Psaila, G. (1995). *Paper presented at Active Data Mining*. Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining, Montreal.
12. Gibbs, J. S., Chiou, H. C., Bastowt, K. F., Chengt, Y. C. & Coen, D. M. (1988). Identification of amino acids in herpes simplex virus DNA polymerase involved in substrate and drug recognition.
13. Database retrieved from <http://www.ncbi.nlm.nih.gov/protein?term=herpes%20simplex%20virus>
14. <http://www.bio.davidson.edu/people/sosarafova/assets/bio307/jehodge/page01.html>