Analysis of Human Herpes Viruses with the Application of Data Mining

Yusin Kim[1], Sung Min Kim[1(✉)], Jiwoo Lee[1], Ann Jeong[2], Jaeuiy Lim[2], and Taeseon Yoon[1]

[1] Natural Science Department, Hankuk Academy of Foreign Studies, Yongin, Korea
{ushin6l2,danielkim98}@naver.com, tsyoon@hafs.kr
[2] International Department, Hankuk Academy of Foreign Studies, Yongin, Korea

Abstract. The human herpes Virus, categorized into 8 types from HHV-1 to HHV-8 is in a DNA virus family, Herpesviridae. They are highly infectious and responsible for many human related diseases. Through this research, we applied data mining, which is the process of finding and analyzing the relationship among the raw data, for the first time to analyze the human herpes virus family. We discovered the patterns regarding the sequence data and analyzed them in associated with the clinical aspects and phylogenetic study results. In addition, we applied decision tree algorithm to support our data. As a result, we were able to stabilize the proof for the bioinformatical categorization. The rules we found were consists of 11 amino acids, Alanine, Leucine, Proline, Valine, Asparagine, Threonine, Phenylalanine, Arginine, Isoleucine, Serine and Lysine. Alanine and Leucine were the two major amino acids from the rules.

Keywords: Human herpes virus · Data mining · Apriori algorithm · Decision tree algorithm

1 Introduction

1.1 The Human Herpes Viruses

The Herpes Virus (also known as Herpesviridae) is a family of DNA viruses and one of the most infectious pathogens in existence. Distinguished by 8 categories, the Human Herpes Virus (HHV) can be analyzed according to its protein amino acid sequence. We believed that in order to acquire the appropriate means of eradicating the symptoms, it is necessary to comprehend the similarities and differences of the protein sequence and relate these comparisons to the characteristics of each virus based on the type. Thus, we analyzed the sequences to find patterns for categorizing the Herpes viruses with the application of data mining, using Apriori algorithm and decision tree algorithm to extract rules.

The Human Herpes Viruses represent a group of DNA viruses that share common biological features that account for oral pathology. The eight categories are named: Herpes Simplex Virus (HSV, HHV-1, 2), Varicella-zoster Virus (VZV, HHV-3),

D.-S. Huang et al. (Eds.): ICIC 2015, Part II, LNCS 9226, pp. 426–435, 2015.
DOI: 10.1007/978-3-319-22186-1_42

Epstein-Barr Virus (EBV, HHV-4), Cytomegalovirus (CMV, HHV-5), and the remaining HHV-6, HHV-7, and HHV-8.

The most common oral HHV infections, HHV-1, 2, as known as Herpes Simplex Virus (alpha herpes viruses), are the causes of Primary Herpetic Gingivostomatitis. This diagnosis is what brings about skin, lip, oral mucosal ulcerations, fever, and lymphadenopathy. It is related to Gingivostomatitis in that it is related to gingival erythma, edema, and erosive lesions. Usually sexually transmitted, HSV causes ulcerations and blistering of the external genitalia and cervix. [4] The practical therapeutic control method for this infection is natural antibodies produced from the patient's body. The recurring episodes of HSV get milder as the presence of antibodies to previous viruses are made. The control process extends over a few weeks; during the first week the illness exacerbates, starting from the second week the condition begins to improve, and on the fourth week, symptoms disappear completely. Another symptom that occurs as a consequent of latent ganglionic HHV-1 infection is the Recurrent Herpes Labialis. This typically affects the lips, bringing cold sores to the infected areas. In this condition, stress, fever, and trauma commonly takes place.

The Varicella-zoster virus, HHV-3 (alpha herpes virus), leads to chicken pox. It is related to oral manifestations of vesicles and ulcerations. The infection usually appears at childhood near the respiratory tract. During the process of secondary multiplication, the vesicles are filled with infectious viruses which may lead to Central Nervous System infection.

The Epstein-Barr virus, HHV-4 (gamma herpes virus), is associated with infectious mononucleosis, lymphomas, nasopharyngeal carcinoma, and oral manifestations. However, in most cases, patients do not realize that they are infected. This virus imposes symptoms that result usually in slight fatigueness or a minor flu. HHV-4 is highly contaminous, spreading through the contact of saliva. HHV-4 has an affinity for the cells of the mouth and throat. Once it enters the B lymphocyte, it controls the DNA. When these cells slough off, the virus is gets dispersed inside the saliva, thus the cause for its communicable aspects.

The Cytomegalovirus, HHV-5 (beta herpes virus) is widespread but usually imposes no harm on adults. Because there are rarely any symptoms or signs of the infection, most people who are infected do not know that they are infected. However, the one detrimental effect that the Cytomegalovirus may have upon humans is during the state of pregnancy. Its infection during pregnancy may cause birth defects such as hearing loss, blindness, epilepsy, mental retardation, or in extreme cases, death.

Human Herpesvirus 6, HHV-6 (beta herpes virus), is a group of two viruses: HHV-6 alpha and HHV-6 beta. These are widespread and infect humans in infantry. Acquiring these viruses may lead to various symptoms such as fever, diarrheas, rashes, and seizures. Like the former herpes viruses depicted above, the HHV-6 is potentially long-lasting and may reactivate after some time.

Human Herpesvirus 7 (beta herpes virus), is similar to the HHV-6. Both viruses affect humans in their childhood; HHV-7 tending to affect children generally 2 years older than the patients of HHV-6. Both viruses show symptoms of fever and rash [3]. It is unknown whether the HHV-7 infection is primarily caused by the HHV-7 virus, or if it is affected by the reactivating HHV-6 virus. Thus, the main inquiry is whether HHV-7 is wholly independent with the HHV-6 virus.

The Human Herpes Virus 8, HHV-8 (gamma herpes virus), also known as the Kaposi's sarcoma-associated herpesvirus, causes Kaposi's sarcoma, a cancer of the blood vessel, and Lymphoma, a cancer of the lymphocyte. It also causes Castleman's disease, which is an infection that enlarges the lymph node. HHV-8 can be transmitted through sexual interaction or oral contact. It can also be transmitted through organ transplantation. Similar to the other Herpes viruses, this virus is widely infected by many people; however, most do not have symptoms.

1.2 The Control Group

We selected control groups for the experiment. Herpes viruses associated with animals from the main three subfamilies of Herpes viruses; alpha, beta, gamma, were selected in consideration of symptoms and significances. These were the Alcelaphine herpesvirus, Bovine herpesvirus 2, and Elephant Endotheliotropic Herpes Virus. These viruses were used to compare the characteristics and to check the uniqueness of the extracted rules from human herpes virus.

Alcelaphine herpesvirus is responsible for malignant Catarrhal fever. Bovine herpesvirus 2 causes bovine mammillitis and pseudo-lumpy skin disease. BoHV-2 is similar in structure to the Human Herpes simplex virus.

Symptoms of pseudo-lumpy skin disease include fever and skin nodules on the face, back, and perineum. Bovine mammillitis is characterized by lesions restricted to the teats and udder. This virus may spread through an arthropod vector, but can also be spread through milkers and milking machines.

The Elephant Endotheliotropic Herpes Virus (EEHV) has affected many elephants, usually infecting young elephants. Early symptoms include bloodshot eyes, head and neck swelling, cyanosis of the tongue and ulcers in the mouth. Later on, EEHV spreads to the heart, liver, and tongue, infecting the microvascular endothelial cells and bringing cardiac failure that may lead to death. The transmission method of EEHV is unknown, and more research is necessary.

Alcelaphine Herpesvirus 1 is a fatal lymphoproliferative disease caused by a group of ruminant gamma herpes viruses. The most common areas of infection are the head and the eyes. Symptoms appear in forms of fever, depression, discharge from the eyes and nose, lesions of the buccal cavity and muzzle, swelling of the lymph nodes, and opacity of the corneas leading to blindness, inappetance and diarrhea. Death usually occurs within ten days. The mortality rate in symptomatic animals is 90 to 100 percent.

1.3 Subfamily of Herpes Virus

There are three known subfamily, alpha, beta, gamma in herpes virus family.

Alphaherpesvirinae (HHV-1, 2, 3) and Bovine herpesvirus 2 (BoHV-2) have many similarities. Both groups of viruses show symptoms of blisters and occur on the face and perineum. However, there is a difference that we can see between HHV and BoHV-2- whether the disease accompanies fever or not. HHV-1, 2, 3 do not cause fever while BoHV does.

Betaherpesvirinae (HHV-5, 6, 7) and Elephantid herpesvirus 1 (EEHV) are similar in that they both target infants or children. (The main target of HHV-5 is the embryo, the main targets of HHV-6 and 7 are infants and children, and the main target of EEHV is a young elephant.) On the other hand, there is a difference that HHV-5, 6, and 7 have similar symptoms of red spots like roseola or "owl eye" while EEHV does not.

Gammaherpesvirinae, (HHV-4, 8) and Alcelaphine herpesvirus 1 (AIHV-1) are common; both groups are associated with symptoms of the lymphoma. Otherwise, the two groups are fairly distinguishable. HHV-4 involves lymphoma and causes diseases such as gastric cancer or colorectal cancer. HHV-8 causes skin cancer which is called kaposi's sarcoma. Lastly, AIHV-1 causes Bovine malignant catarrhal fever which includes fever, blindness, and diarrhea. The features and similarities.

2 Materials and Methods

2.1 Sequence Data

We obtained the protein sequence data from the NCBI (National Center for Biotechnological Information). We chose the Major Capsid Protein as common in all types of the Herpes Virus, which consist about 1000 amino acids.

The structure of the major capsid protein plays an important role in virus categorization. It forms the capsomere structures of the herpes virus, the pentons and hexons, which contain five and six VP5 monomers, respectively. Since major capsid protein is common in all types of the herpes virus, and is known as they take part in crucial functions of virus assembly, we made a conclusion that it can represent the properties of each herpes virus partly. As the data contained the structure information, we could expect that the rules we extracted using apriori algorithm from the sequence can explain the characteristics and similarities of the herpes virus.

2.2 Apriori Algorithm

The Apriori algorithm is a data mining algorithm for finding associated rules. It is well known for simple methods and effective performance. It is related to a set theory and uses breadth-first search and a hash tree structure to count the frequency of the item sets. The basic principle of this algorithm is that the subsets of highly frequent item sets are also highly frequent.

We divided each sequence set into 7, 9 and 11 windows. Using the apriori algorithm, we could extract rules from each MCP sequence data. The rules we found consist of a number of amino acids found in the digit of each divided window. For example, "amino 5L 17" means 17 Leucine were found on the fifth section. We obtained the protein sequence data from the NCBI(National Center for Biotechnological Information). We chose the Major Capsid Protein as common in all types of the Herpes Virus, which consist about 1000 amino acids.

2.3 Decision Tree Algorithm

The Decision Tree Algorithm is one of the most common and effective algorithms for data mining analysis. This algorithm is usually used for data classification. This algorithm is a tree shaped graph (model) of decisions and shows the possible consequences. The application of the algorithm in bioinformatics is mainly for classifying the sequences. The Decision Tree Analysis can extract rules, classify each experimental class and print out the accuracy of the classification. If the classification error rate is high, it means that the classification is difficult, and that the experimental classes are similar. In this research, using the See 5.0 program, we applied the decision tree to support our apriori algorithm analysis and to analyze the provided classification table.

3 Result

3.1 Apriori Algorithm Analysis

Apriori algorithm analysis can extracts rules regarding the index of the divided window. The divided window number is related with the amino acid binding and protein formation. One rule is consists of three information, index, amino acid and the size. For example, let's slice a randomly created sequence …LAAVNGLLEARFPNKII… in 7 windows, …LAAVNGL/LEARFPN/KIIP…, or in 9 windows, … LAAVNGLLE/ARFPNKIIP. There are two Leucine at the first place in the 7window sliced sequence and two Asparagine at the forth place in the 9window sliced sequence. The rules extracted from apriori algorithm are same as the example, amino acids repeated constantly.

The results of the diagrams that display the rules were drawn. The x axis is the amino acids abbreviation (Table 1) and the y axis is the number of rules found. To make a clear description of our results, we combined the rules made of same amino acids.

The similarities of rules are quite clear between HHV1 and 2, the simplex viruses. But HHV3 and bovine2 are not that similar consists of other rules (Figs. 1, 2 and 3).

We analyzed the distribution of the rules using diagram and detailed rules, The rules were abbreviated to index-amino acid. For example, 1-A indicates the alanine repeated at the first place of the sliced window. The list of the extracted rules are omitted in this paper because the amount of data. Approximately 500 rules were extracted.

To begin with, Alanine, Leucine, and Valine each have their own rules that have effects on symptoms, stages of infection, and the viruses themselves. 1-A was found in the 7th window experiment, 1-A, 2-A, 4-A, and 6-A from the 9th window experiment, and 4-A, 5-A, 6-A, and 11-A from the 11th window experiment from HHAv-1, 2 and Bovine2. These patterns of alanine seem to appear widely for Herpes Simplex Virus-HHV-1, 2 and Bovine2; however, this is not true for viruses HHV-6, 7 and Elephantid1, which are the betaherpesvirinae. Leucine usually appears at location 10 and 11 in the 11th window. 5-V, 6-V, 7-V, 8-V and 9-V show obvious stakes, with high frequency in several experiments.

Table 1. Amino acid abbreviation

Amino acid	One-letter abbreviation	Three-letter abbreviation	Amino acid	One-letter abbreviation	Three-letter abbreviation
Alanine	A	Ala	Methionine	M	Met
Cysteine	C	Cys	Asparagine	N	Asn
Aspartic acid	D	Asp	Proline	P	Pro
Glutamic acid	E	Glu	Glutamine	Q	Gln
Phenylalanine	F	Phe	Arginine	R	Arg
Glycine	G	Gly	Serine	S	Ser
Histidine	H	His	Threonine	T	Thr
Isoleucine	I	Ile	Selenocysteine	U	Sec
Lysine	K	Lys	Valine	V	Val
Leucine	L	Leu	Tryptophan	W	Trp
			Tyrosine	Y	Tyr

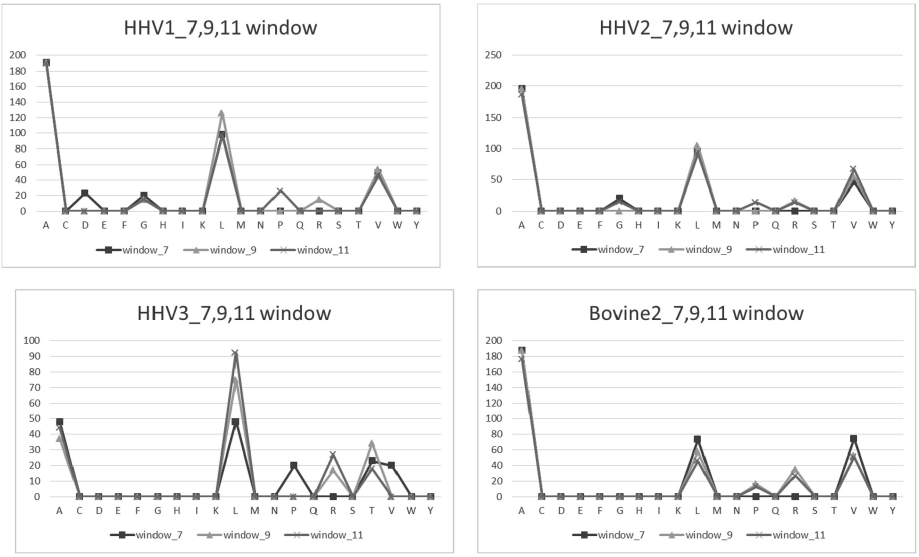

Fig. 1. Apriori algorithm analysis result. Alphaherpesvirinae, HHV 1, 2, 3 and Bovine2

Secondly, HHV-6 and HHV-7 seem to have similarities. The common rules of these two viruses show similar distributions of Asparagine located in the 4th, 7th place and Threonine located in the 6th and 7th place. They also show common rules on 5-L, 7-L, 8-L and 9-N in the 9 window experiment. In the 11 window experiment, HHV-6 and HHV-7 show similar distributions of 1-L, 6-L, 9-L, 10-L and 5-I. Thus, they partly shared their rules with Elphantid1. This relationship between HHV-6 and HHV-7, plus Elephantid1, is not the common features of the betaherpesvirinae and other herpesviruses. They didn't have a large common alanine rules and other minor rules were

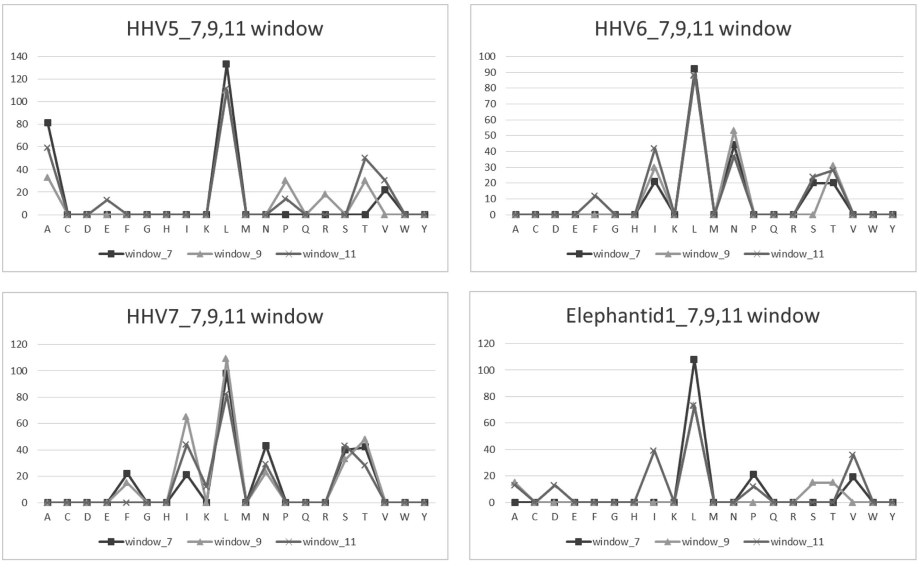

Fig. 2. Apriori algorithm analysis result. Betaherpesvirinae, HHV5, 6, 7 and Elephantid1

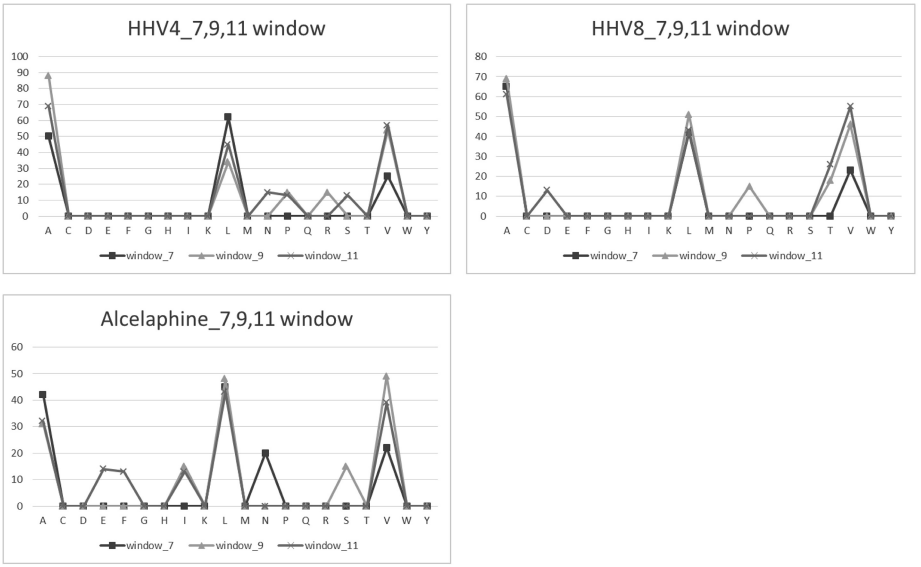

Fig. 3. Apriori algorithm analysis result. Gammapesvirinae, HHV4, 8 and Alcelaphine

different from other herpesviruses. Therefore we concluded that HHV 6, 7 are unique and can be classified differently with other betaherpesvirinae.

Lastly, the comparison of the species and the possibility of classification were found. Based on the distribution of the rules, we concluded that alphaherpesvirus

subfamily is more similar with gammaherpesvirus subfamily and betaherpesvirus subfamily can be separated from them. The difference between human and animal herpes virus is not that clear in alphaherpesvirus subfamily but in beta and gamma-herpesvirus subfamily, the distribution and detailed rules showed a possibility of the classification. For instance, HHV4 and 8 showed different distribution of the rules with the Alcelaphine which is visible in the diagram.

3.2 Decision Tree Algorithm Analysis

Table 2 is the result of the decision tree algorithm experiment and it is automatically drawn by C 5.0 program after the experiment. This table ensembles an 11-by-11 matrix. For example, (1, 4) entry of the matrix, which is 10, indicates that class 4 was classified as class 1 for 10 times. The classification of the decision tree algorithm experiment is based on the obtained rule from the divided sequences. High value of the matrix can be explained as the similarities between the experimental classes. We found high similarities between HHV-6 and HHV-7, HHV-1, 2 and Bovine2.

Table 2. The result of the 7 window decision tree algorithm experiment. Program, C5.0 Release 2.07 GPL Edition, was used. Class 1 to 8 are the human herpes virus 1 to 8 respectively and class 9 to 11 are Alcelaphine(gamma subfamily), Bovine2(alpha subfamily) and Elphantid1(beta subfamily). Values above 30 and peak value from each classes were marked bold.

1	2	3	4	5	6	7	8	9	10	11	Classified as
37	**35**	**23**	10	11	14	7	15	7	**26**	12	class1
29	**35**	**23**	14	16	9	12	15	13	22	8	class2
27	29	19	11	**21**	20	13	12	16	19	13	class3
27	29	13	16	15	20	21	13	9	17	**18**	class4
25	22	25	14	11	26	12	14	15	18	14	class5
9	21	5	13	19	23	**54**	15	8	17	9	class6
12	14	19	12	12	**48**	26	17	12	10	11	class7
26	**31**	19	8	14	14	22	8	25	20	10	class8
25	25	20	15	8	10	23	**23**	13	17	17	class9
37	**43**	19	13	9	16	10	7	10	**26**	8	class10
18	21	14	**18**	7	**31**	20	20	13	14	17	class11

The 10 cross validation was held and the average classification error percentages was 89.3 %, standard error 0.5 % in 7window experiment. 9 and 11 window experiment also had similar results.

It is noticeable that the diagonal entries of the matrix are not the always biggest value. And some results like class 4 and 8 do not completely match with the apriori algorithm experiment. So considering the errors, this matrix cannot supply a solid value. Rather, it can be used to see the tendency of the large datasets.

4 Conclusion

Data mining is the process of finding and analyzing the relationship among the data. Through this project we used the data mining method to analyze the Herpes Virus configuration, and used the found patterns to investigate the similarities of the viruses and their distinguable characteristics. We were not only able to investigate the virus itself, but considered this investigation to be helpful in medical research and symptom analysis.

In the results of analyzing the general data, the rules of Alanine, Valine, and Leucine were the main components. Meticulously analyzing the results, we were able to conclude that these amino acids are distributed in patterns. We were also able to make a prediction that in the case that further investigation of Alanine, Valine, and Leucine is conducted, analysis of the Herpes Viruses and their cure centered around this investigation will develop.

HHV6 and HHV7 showed similar characteristics in several windows, but there were also unique patterns that could only be found in their boundaries. We discovered that, unlike other viruses, HHV6 and HHV7 have liver-related traits, and that this is connected to the fact that HHV6 and HHV7 are deeply related. This also can be supported by the decision tree algorithm experiment (Table 2, (6, 7) and (7, 6)). The patterns that commonly appear for these viruses will assist in identifying the uniqueness of the viruses.

Other than the relationships that showed the major patterns, the minor patterns that were found in the analysis process appeared in various forms. In other words, general patterns could also be seen including the observations that Threonine and Asparagine were located mainly under 4, and that Valine was located in HHV5-9 and its high frequency. We couldn't found the effects and the exact influence of these minor rules. So, the more understanding with the sequence pattern is necessary.

The three top main amino acids, Alanine, Valine and Leucine, were nonpolar amino acids. It is quite remarkable that rules regarding Leucine and Alanine were always extracted from all of the experiment regardless of window. Since amino acids sequence determines function and structure of the protein, we assumed that Alanine and Leucine are the key amino acid forming the major capsid protein. Even though we found a lot of other large and small rules, however, we couldn't utilize them in a solid knowledge. Therefore, a close research regarding the features of repeated amino acid should be held in near future.

The limitation of our research is that the size of the data was not big enough to have sufficient credibility and could only find vague information. Also, other kinds of protein should be analyzed to check the validity of our conclusion. So conducting a deeper research with more dataset is necessary in near future.

References

1. Jones, C.: Bovine herpes virus 1 (BHV-1) and herpes simplex virus type 1 (HSV-1) promote survival of latently infected sensory neurons, in part by inhibiting apoptosis. J. Cell Death **6** (2013)
2. Fatahzadeh, M., Schwartz, R.A.: Human herpes simplex virus infections: epidemiology, pathogenesis, symptomatology, diagnosis, and management. J. Am. Acad. Dermatol. **57**(5), 737–763 (2007)

3. Anzivino, E., et al.: Herpes simplex virus infection in pregnancy and in neonate: status of art of epidemiology, diagnosis, therapy and prevention. Virol. J. **6** (2009)
4. Edelman, D.C.: Human herpesvirus 8 – a novel human pathogen. Virol. J. **2** (2005)
5. Arvin, A., Campadelli-Fiume, G., Mocarski, E., et al. (eds.): Human Herpesviruses: Biology, Therapy, and Immunoprophylaxis. Cambridge University Press, Cambridge (2007)
6. Ykihiro, N.: Herpesvirus genes: molecular basis of viral replication and pathogenicity. Nagoya J. Med. Sci. **59**, 107–119 (1996)
7. Bowman, B.R., et al.: Structure of the herpesvirus major capsid protein. EMBO J. **22**(4), 757–765 (2003). PMC. Web. 1 March 2015
8. Bayardo, Jr., R.J.: Efficiently mining long patterns from databases. In: SIGMOD 1998 Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, pp. 85–93 (2) (1998)
9. RuleQuest Research (2008). http://www.rulequest.com/see5-unix.html