Analysis of Human Papillomavirus Using Datamining - Apriori, Decision Tree, and Support Vector Machine (SVM) and its Application Field

Younghoon Cho[1], Seungwon Burm[1], Nayoung Choi[1] and Taeseon Yoon[2]

[1]*Department of Natural Science, Hankuk Academy of Foreign Studes, 50, Oedae-ro 54beon-gil, Mohyeon-myeon, Cheoin-gu, Yongin-si, Gyeonggi-do, Republic of Korea*
[2]*Department of Computer Science and Engineering, Korea University, Seoul, Repuulic of Korea*

Abstract. Human Papillomavirus(HPV) has various types (compared to other viruses) and plays a key role in evoking diverse diseases, especially cervical cancer. In this study, we aim to distinguish the features of HPV of different degree of fatality by analyzing their DNA sequences. We used Decision Tree Algorithm, Apriori Algorithm, and Support Vector Machine in our experiment. By analyzing their DNA sequences, we discovered some relationships between certain types of HPV, especially on the most fatal types, 16 and 18. Moreover, we concluded that it would be possible for scientists to develop more potent HPV cures by applying these relationships and features that HPV virus exhibit.

1 Introduction

Human Papillomavirus, or HPV, is a DNA virus which is involved in the papillomavirus family. It can induce a number of infections in keratinocytes of the skin or mucous membranes. HPV is known worldwide to induce various genital cancers, such as cervical cancer, vaginal cancer, and anal cancer. Most viruses pass through cell membranes and progress self-proliferation to survive. During this process, HPV produces several kinds of proteins that make host cells infinitely progress the cell-division. E6 protein inhibits p53 which induce an apoptosis. And E7 protein inhibits Rb which restricts cell cycle. Therefore, host cells for HPV keep cell division permanently and eventually turn into cancer cells.

According to numerous studies concerned, HPV has been proved its importance in revealing the mechanism of cancer development. It is the necessary cause of human cancer, especially cervical cancer. In other words, without the infection of HPV and the presence of HPV DNA, several types of cancer will not develop. This fact implies that if we can curb the HPV infection and gene expression, it is possible to prevent and cure cervical cancer and other types of genital cancer. Since these cancers are highly fatal to human, scientists are trying to reveal the overall process of human papillomavirus's expression in human body in order to develop the medication for cervical cancer. Recently, HPV vaccines were invented, and are being offered for girls aged under 15. However, to minimize the adverse effects and maximize the efficiency of HPV vaccines, it is crucial to reveal the molecular structure of Human Papillomavirus to create a best-suited medication for HPV only. The molecular structure of HPV is determined by the sequence of amino acids. And the sequence of amino acids is encoded in mRNA using 4 types of bases: A, G, C, and T. In this paper, we analyzed RNA sequences of human papillomaviruses classified as 'highest risk' and 'probably high risk' in order to find out similarities and differences of those two categories.

2 Relative Research

2.1 Human Papillomavirus (HPV)

Human papillomavirus(HPV) causes many different types of diseases and it is classified by the risk levels into different types. By analyzing various cases, HPV types were classified as 15 high-risk types(16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, 82), 3 probable high-risk types(26, 53, 66), and 12 low-risk types(6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, CP6108)[1]. It has been found that DNA of human papillomavirus (HPV) types 16 and 18 is closely associated with human genital cancer. It supports the concept that HPV type 16 and 18 are key factors in the aetiology of genital cancer. Furthermore, searching about the vaccine [2] HPV 16/18 AS04-adjuvanted vaccine was immunogenic, generally well tolerated, and effective against HPV-16 or HPV-18 infections and the research analyzed efficacy in the final event-driven analysis of the women who were vaccinated at months 0, 1, and 6. The HPV 16/18 AS04-adjuvanted vaccine showed high efficacy against CIN2+ associated with HPV-16/18 and non-vaccine oncogenic HPV types [3]. In the field of cervical cancer, genital HPV has the

role as the central etiologic factor in cervical cancer worldwide [5]. Also the presence of HPV in virtually all cervical cancers implies the highest worldwide attributable fraction so far reported for a specific cause of any major human cancer [6]. Furthermore, ingration of HPV-16 DNA, which occurs in cervices, can result in the increased expression of the viral E6 and E7 oncogenes through altered mRNA stability and occur cervical cancers. Also the demonstration that more than 20 different genital HPV types are associated with cervical cancer has important implications for cervical cancer-prevention strategies that include the development of vaccines targeted to genital HPVs [4].

2.2 Datamining Algorithms

In this study, we used three different datamining algorithms: Decision Tree Algorithm, Apriori Algorithm, and Support Vector Machine Algorithm.

2.1.1 Support Vector Machine (SVM)

Support Vector Machines are supervised learning models with associated learning algorithms which analyze data and recognize patterns, utilized for classification and regression analysis.

2.1.2 Apriori Algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It operates by recognizing the frequent individual items in the database and stretching them out to bigger and bigger item sets as long as those item sets show up sufficiently often in the database.

2.1.3 Decision Tree Algorithm

A decision tree is a decision support tool which uses a tree-like graph or model of decisions and their possible results. It is one way to display an algorithm. They are generally used in operations research, especially in decision analysis in order to help identify a strategy which is most likely to reach a goal.

3 Method

In this study, we used 11 base sequences of different types of HPV (HPV 16, HPV 18, HPV 26, HPV 31, HPV 33, HPV 35, HPV 53, HPV 66, HPV 68a, HPV 68b, HPV 82). As we mentioned in the relative research, three of them (HPV 26, HPV 53, HPV 66) are probable high-risk group, and the others are high-risk group. If we discover several similarities and differences in base sequences between these two groups, it means that we could find amino acids that play a dominant role in evoking cervical cancer. So, we extracted full RNA sequences of these viruses from NCBI (National Center for Biotechnology Information). And we applied several algorithms to figure out similarities and differences between these viruses.

In this paper, we used 3 types of datamining algorithms: Decision Tree Algorithm, Apriori Algorithm, and Support Vector Machine (SVM). Decision Tree Algorithm has its strength to clarify distinct differences between amino acid sequences, while Apriori Algorithm has its strength to declare similarities. And SVM can provide more accurate results since it can use higher dimensions than dimensions other algorithms use.

We conducted three experiments for each algorithm. The base sequence of each virus is too vast to analyze at once. So, we have to divide it into several parts. So, when we analyze HPV base sequences, we conducted three experiments: 9-windows, 13-windows, and 17-windows.

Especially, for SVM algorithm, we carried out the experiments with 10 - fold cross validation, and we applied four types of functions: normal, polynomial, polynoima2, and RBF.

4 Results

4.1 Decision Tree Algorithm Results

4.1.1 9 window results

Table 1 shows the results of 9-windows Decision Tree algorithm. It is remarkable that every virus possesses its unique rule, and that every extracted rule has the probability of at least 0.75. This value is high enough to affirm that every HPV virus has its own distinguishable trait. Also, we can see frequent repeatance of Threonine, Leucine, and Valine in Table 1. It may indicate that these three amino acids play a key role in HPV virus. Moreover, we can find that the amino acids extracted from position 2 and position 9 represent the distinguishable rule of each virus. So, we can find that posiion 2 and position 9 is an important factor that makes HPV viruses different.

Table 1. 9 window rule (POS=Position)

HPV	Rule Content
HPV 16	pos2=D pos3=T pos6=L
HPV 18	pos6=I pos7=L pos9=F
HPV 26	pos3=K pos4=T pos8=E
HPV 31	pos1=L pos4=S pos9=T
HPV 33	pos2=F pos4=V pos9=T
HPV 35	pos1=L pos2=P pos9=S
HPV 53	pos2=V pos7=Y pos8=L
HPV 66	pos1=I pos2=Y pos3=R
HPV 68a	pos1=I pos2=H pos9=K
HPV 68b	pos2=P pos4=T pos9=V
HPV 82	pos1=T pos5=G pos6=F

4.1.2 13 window results

Table 2 shows the results of 13-windows Decision Tree algorithm. All extracted rules have the probability of 0.75, which is quite high as we mentioned before. Table 2 is similar to Table 1 in that all rules are composed of 3 amino acids. What is different is first, in Table 2, position 7 are more frequent than position 2 or 9, and second, relatively various types of amino acids appeared. For instance, HPV 26 has Glutamine, which doesn't appear in Table 1. Also, HPV 31, HPV 68a, and HPV 82 have their unique figure of Lysine.

Table 2. 13 window rule (POS=Position)

HPV	Rule Content
HPV 16	pos2=F pos7=V pos10=Y
HPV 18	pos7=P pos9=Y pos12=I
HPV 26	pos3=E pos7=Q pos13=T
HPV 31	pos4=S pos7=K pos9=L
HPV 33	pos5=V pos7=P pos11=S
HPV 35	pos1=S pos7=I pos11=C
HPV 53	pos2=L pos7=N pos9=A
HPV 66	pos1=L pos7=Y pos8=Y
HPV 68a	pos3=P pos7=E pos13=K
HPV 68b	pos2=T pos5=A pos7=P
HPV 82	pos7=L pos8=K pos13=C

4.1.3 17 window results

Table 3 shows the results of 17-windows Decision Tree algorithm. It is noticeable that many rules of various HPV are constituted of 2 amino acids. Also, in 17-windows results, position 6 plays a key role in extracting rules. Furthermore, HPV 16 and HPV 33 have Cysteine, and this finding is unique to these two viruses.
So, it may implicate the key difference between HPV viruses.

Table 3. 17 window rule (POS=Position)

HPV	Rule Content
HPV 16	pos6=C pos7=L pos16=T
HPV 18	pos6=L pos13=V pos14=L
HPV 26	pos2=P pos9=E
HPV 31	pos5=Q pos6=S
HPV 33	pos6=C pos7=Q
HPV 35	pos5=F pos6=K pos15=T
HPV 53	pos6=G pos17=A
HPV 66	pos4=L pos5=V pos17=L
HPV 68a	pos6=E pos10=D
HPV 68b	pos1=S pos6=V pos12=I
HPV 82	pos6=A pos17=A

4.2 Apriori Algorithm

4.2.1 9 window results

Table 4 shows the results of 9-windows Apriori Algorithm. According to Table 4, it is clear that Leucine plays a dominant rule in all HPV virus types. All HPV viruses have a considerable rate of Leucine, and this fact indicates that we can utilize Leucine Synthesis Inhibitor to treat cervical cancer. Also, HPV 33 and HPV 68b have Isoleucine as a key component. Since Isoleucine is an isoform of Leucine, we may find the relationship between these two amio acids. Moreover, we can find Valine in HPV 16, HPV 18, HAV 35, and HPV 68a and Threonine in HPV 26, HPV 31, HPV 68b, and HPV 82. It is predictable that two amino acids, Valine and Threonine, are exclusive. In other words, when one HPV virus has valine, it doesn't have thereonine. This rule may be able to discover the difference between lethal and non-lethal Human Papillomaviruses. Also, this table implies the differences between high-risk group and probable high-risk group. High-risk group has both Leucine and Valine, while probable high-risk group doesn’t have. This fact implies that Leucine and Valine plays a key role in evoking high-risk cervical cancer.

Table 4. 9 window rule (Amino Acid(Frequency))

HPV	Rule Content
HPV 16	L(44) V(30)
HPV 18	L(34) V(30)
HPV 26	L(32) T(31)
HPV 31	L(30) S(29) T(29)
HPV 33	I(29) L(31)
HPV 35	L(30) V(30)
HPV 53	L(35)
HPV 66	L(43)
HPV 68a	L(40) V(33)
HPV 68b	I(30) T(29)
HPV 82	L(36) T(32)

4.2.2 13 window results

Table 5 shows the results of 13-windows Apriori Algorithm. According to Table 5, it is also clear that Leucine plays a dominant rule in all HPV virus types.

Also, Isoleucine, Valine and Threonine: these three amino acids play a significant role in extracting rules. Especially, HPV 68b has 6 amino acids, and those 6 amino acids have almost equal importance. Moreover, in 13-windows results, unlike 9-windows results, Threonine and Valine co-exist in many HPV viruses.

Table 5. 13 window rule (Amin Acid(Frequency))

HPV	Rule Content
HPV 16	L(44) V(30)
HPV 18	L(34) V(30)
HPV 26	L(32) T(31)
HPV 31	L(30) S(29) T(29)
HPV 33	I(29) L(31)
HPV 35	L(30) V(30)
HPV 53	L(35)
HPV 66	L(43)
HPV 68a	L(40) V(33)
HPV 68b	I(30) T(29)
HPV 82	L(36) T(32)

4.2.3 17 window results

Table 6 shows the results of 17-windows Apriori Algorithm. According to Table 6, it is also clear that Leucine plays a dominant rule in all HPV virus types. Also, unlike Table 4 and Table 5, extracted rules include more amino acids. Also, Threonine and Valine, and Serine take part in extracting distinguishable rules.

Table 6. 17 window rule (Amino Acid(Frequency))

HPV	Rule Content
HPV 16	I(22) L(35) V(22)
HPV 18	I(21) L(26) T(21) V(22)
HPV 26	L(25) S(22) T(20)
HPV 31	L(20) S(20) T(21)
HPV 33	I(22) L(22) V(22)
HPV 35	I(20) L(28) T(21) V(22)
HPV 53	I(21) L(26)
HPV 66	L(30) V(24)
HPV 68a	L(29) T(20) V(21)
HPV 68b	I(25) L(21) R(20) S(20) T(21) V(23)
HPV 82	L(27) V(21)

4.3 Support Vector Machine Results

Considering the fact that results of Apriori and Decision tree don't have a lot in common, we conducted the third experiment using SVM. However, because of the excessive data, infinite loop was created and we were only able to use the normal types. We used 10 fold cross validation, and the phrase "Average loss on test set 90.0000 Zero/one-error on test set 90.00% (35 correct, 315 incorrect, 350 total)" was repeated 10 times. The correct ratio was low due to the large amount of data.

After experiencing the difficulty of experiment using such big data, we decided to compare a few specific types considered to have big differences. From the results made by Apriori and Decision tree, we found that among the high-risk types, HPV 68b has the least similarity with HPV 16, 18 which is most frequently in cases of cancer. As a result, we repeated the experiment HPV 18 and HPV 68b in depth.

4.3.1 9 window results

During experiment, we made data types of <HPV 18 : +1, HPV 68B : -1>. In 9-windows, +1 has 292 data units and -1 has 291 data units. For training, we used 240 data units, while for test, we used 60 data units.

In Table 7, it represents 9-windows results of SVM algorithms. Average of Accuracy rate is highest in RBF, and lowest in Polynomial2 and Normal. Since the accuracy rate is quite low (about half), it is clear that HPV 18 and HPV 68b is not clearly divided into two parts. However, this result is still meaningful since the result of RBF has 76.416%, which means that HPV 18 and HPV 68b have different properties in amino acid sequences.

Table 7. 9 window rule (Accuracy on Test(%))

	1st	2nd	3rd	4th	5th	average
Normal	57.5	51.6	51.6	49.1	46.6	51.834
	50.8	56.6	50.0	50.8	53.3	
Poly-1	54.1	50.8	60.0	51.6	45.8	52.25
	52.5	48.3	59.1	49.1	50.8	
Poly-2	58.3	55.8	50.0	45.8	49.1	51.749
	43.3	50.8	59.1	55.0	50.0	
RBF	79.1	80.0	73.3	73.3	78.3	76.416
	70.8	79.1	75.0	75.8	79.1	

4.3.2 13 window results

In 13-windows, +1 has 202 data units and -1 has 201 data units. For training, we used 160 data units for one time, while for test, we used 40 data units.
In Table 8, it represents 13-windows results of SVM algorithms. The accuracy rate of the experiment is highest in RBF, and lowest in polynomial2. The overall accuracy rate of 13-windows is generally higher than that

of 9-windows. Also, this result indicates that HPV 18 and HPV 68b have some different traits, but it is not clearly divided.

Table 8. 13 window rule (Accuracy on Test(%))

	1st	2nd	3rd	4th	5th	average
Normal	58.7	56.2	66.2	51.2	63.7	56.7508
	65.0	50.0	50.0	55.0	51.2	
Polynomial	55.0	57.5	63.7	57.5	57.5	57.75
	63.7	62.5	57.5	52.5	50.0	
Poly-2	55.0	57.5	57.5	55.0	55.0	54.875
	53.7	52.5	48.7	55.0	58.7	
RBF	75.0	72.5	77.5	76.2	80.0	78.00
	76.2	75.0	82.5	77.5	87.5	

4.3.3 17 window results

In 17-windows, +1 has 155 data units and -1 has 154 units. For training, we used 120 data units for one time, while for test, we used 30 data units.
In Table 9, it represents the results of 17-windows SVM algorithms. Unlike 9 and 13 windows, the average accuracy of experiment is highest in polynomial2, and lowest in normal. Since polynomial2 function can use higher dimensions than normal function, it seems that they can classify different data sets into two groups. Also, polynomial2 and RBF function represents relatively high accuracy rate (76.833 and 76.499). So, we can conclude that HPV 18 and HPV 68b have several stark differences in their base sequences.

Table 9. 17 window rule (Accuracy on Test(%))

	1st	2nd	3rd	4th	5th	average
Normal	63.33	58.33	60.00	60.00	56.67	58.16
	58.33	53.33	66.67	45.00	60.00	
Polynomial	65.00	76.67	66.67	58.33	70.00	61.5
	60.00	65.00	56.67	38.33	58.33	
Polynomial2	90.00	78.33	73.33	76.67	75.00	76.833
	80.00	73.33	73.33	76.67	71.67	
RBF	78.33	81.67	73.33	75.00	83.33	76.499
	81.67	75.00	70.00	73.33	73.33	

5 Conclusion

In this study, we have found out that leucine, isoleucine, threonine, valine are the most dominant amino acids in Human Papillomavirus. Leucine acts on building muscles and regulating blood sugar. One of the major functions of isoleucine is proteinogenesis in the body. Threonine helps to maintain the proper protein balance in the body. Valine has a stimulant effect, so it is needed for muscle metabolism, tissue repair, and the maintenance of a proper nitrogen balance in the body. Most of the factors that induce cancer other than cervical cancer are mostly composed of leucine. As a result, inhibitors of leucine's synthesis are widely used as anti-cancer medicine, but it also distracts the synthesis of leucine that is essential to human body. Unlike other viruses, HPV is created with diverse amino acids. Malaria for instance, is mostly made of Leucine, which is a main amino acid which creates human muscle tissue. However HPV consists of Leucine, Valine, Threonine, Isoleucine, Cysteine, and many others in equal manner. This implicates that HPV can operate in various types of tissues composed of different amino acids. Furthermore, since the sequence and the types of amino acid is very different in every types of HPV, it will produce diverse kinds of proteins, which will cause different symptoms in human body. Since the extracting rules and components are distinct between all types of HPV, in further research, finding out the functions of proteins produced from different sequence of amino acids that attacks human body is necessary.
Through this experiment, we figured out that there are differences and similarities between viruses that we analyzed. From the initial part of this study, we have classified Human Papillomaviruses into three groups: 15 high-risk types, 3 probable high-risk types, and 12 low-risk types. Especially, HPV16 and HPV18, which are the key factors in inducing cervical cancers, show high percent of accordance in its amino acid composition, especially considering the percentage of Leucine and Valine. Also, HPV 18 and other Human Papillomavirus show some remarkable differences. This result indicates that we can make medications that have less adverse effect when we consider these structural differences between them. Regarding this fact, we concluded that we can treat cervical cancers by inhibiting the synthesis of certain kinds of amino acids which are prevalent in HPV16 and HPV18.

References

1. N. Muñoz, F. X. Bosch, S. Sanjosé, R. Herrero, X. Castellsagué, K. V. Shah, J.F. Snijders, Chris J.L.M. Meijer, N Engl J Med **348 :6** 518-27 (2003)
2. E. Schwarz, U. K. Freese, L. Gissmann, W. Mayer, B. Roggenbuck, A Stremlau, H. Hausen. Nature **314(6006)**, 111-4 (1985)
3. J. Paavonen, P. Naud, J. Salmeron, C.M. Wheeler, S.N. Chow, D. Apter, H. Kitchener, X. Castellague, J.C. Teixeira, S.R, Skinner, J. Hendrick, U. Jaisamrarn, G. Limson, S. Garland, A. Szarewski, B.Romanowski, F.Y. Aoki, T.F. Schwarz, W.A. Poppe, F.X. Bosch, D. Jenkis, K.Hardt, T. Zahaf, D. Descamps, F. Struyf, M. Kehtinen, G. Dubin, Lancet, **374(9686),** 301-14 (2009)
4. S. Jeon, P.F. Lambert, Proc Natl Acad Sci U.S.A, **92(5),** 1654-8 (1995)
5. J.M. Walboomers, M.V. Jacobs, M.M. Manos, F.X. Bosch, J.A. Kummer, K.V. Shah, P.J. Snijders, J. Peto, C.J. Meijer, N. Munoz, J Pathol, **189(1),** 12-9 (1999)
6. F.X. Bosch, M.M. Manos, N. Munoz, M. Sherman, A.M. Jansen, J. Peto, M.H. Schiffman, V. Moreno, R. Kurman, K.V. Shah, J Natl Cancer Inst., **87(11),** 796-802 (1995)