Classifying and Identifying HPV Based on Maximal Common Subsequences

Sheng-Lung Peng* and Jyun-Jhao Chen

Department of Information Engineering and Computer Science

National Dong Hwa University

Hualien 974, Taiwan, ROC

*slpeng@mail.ndhu.edu.tw

Received 1 August 2007; Accepted 1 September 2007

Abstract. The topic of genetic classification and identification is an important issue. Different approaches are brought forth including signature. The Human Papilloma Viruses (HPV) are the main factor that cause the cervical cancers and they are reported as the one of the largest ten murderers that bring to women's cancers. People keep trying to use various ways to detect and prevent this horrible disease. In this paper we show that a set of maximal common subsequences can be used to classify and identify HPV.

Keywords: HPV, DNA sequence, Suffix tree, Identification, Classification

1 Introduction

In the world, over 511,000 women are stricken with cervical cancer each year and approximately one half die from it. Nowadays, it is one of the most deadly cancers for women in the developing countries. Human Papilloma Viruses (HPV) are the viruses that cause genital warts, called condyloma.

HPV is found inside cancer cells in the cervix. However, scientists have not yet found out how it causes abnormal cell diversification and then the cervical cancer. There are more than 100 kinds of Human Papilloma Viruses. There are fourteen or more types are closely related to the cervical cancer. It is believed that these high-risk HPVs are related to the cervical cancer, including T16 (Type 16), T18, T31, T33, T35, T52, T58, and so on. Nevertheless, over 90 percent of the condylomas cases are related to T6 and T11. If women are infected with these viruses, they have more chances to suffer cervical cancer.

Although the use of cytomorphological screening of cervical smears (the Papanicolaou test) has reduced the incidence of cervical cancer significantly, the test still has some limitations with respect to sensitivity and specificity. False negative rate for cervical preamlignant lesions and cervical cancer between 15% and 50% and false positive rate of about 30% have been reported. The enforcement of the HPV identification method and the use of colposcope would improve the sensitivity in detecting the cervical cancer.

In recent years, many organisms of whole genomes have been sequenced. As a result of bio-technology progressing, the quantity of genome sequencing is improved. As mentioned to the subsequent huge amount of genomic data, the aspect of bioinformatics now faces the post-genomic era. Many important topics are working on, including sequence's signature (characterization [1]), protein's 2-Dimension or 3-Dimension structure (prediction) and so on. In this study, we focus on the genomic research and try to find out something interesting behind that.

In 1995, Karlin and Burge [2] made a basic observation that each genome has a characteristic "signature" defined by the ratios among the observed dinucleotide frequencies and the frequencies expected. In 1999, Deschavanne et al. [3] explored DNA structures of genomes by means of a new tool derived from the *Chaos Game Representation* (CGR), which allows the depiction of frequencies of oligonucleotides in the form of images. By using CGR, each meaningful graph (or chart) could identify a species. According to that information, we can look for its characteristic and distinguish which it is. Parts of research are applying signal processing operations on the DNA sequences [4,5]. The signature on genome can be taken as basis of classification of viruses. And it also provides a basis to identify novel virulence genes for biologists.

In this paper, we propose a new concept for viruses classification according to the common ordered subsequences and use them to analyze the category in HPV. Then, we propose another concept for identifying viruses based on also the common subsequences. We implement this idea on HPV and obtain a surprising result, it is encouraged that HPV can be identified by the distribution of common subsequences.

* Correspondence author

2 Materials and Tools

2.1 Human Papilloma Viruses

Human Papilloma Viruses (HPV for short) are found inside the cancer cells in the cervix but scientists have not yet to find out how the viruses cause abnormal cell diversification and the cervical cancer [6,7]. By means of bio-molecular technology, HPV in both the high-risk and the low-risk can be identified. Some HPV still cannot be identified by test-kit. In general, they can be classified into three classes as shown in the following table.

Table 1. The classes of HPV.

class	property	# of sequences
1	high-risk	14
2	cannot be identified by test-kit	11
3	low-risk	49

Tests are available to detect high-risk HPV. Hybrid Capture 2 is currently the most important testing product for high-risk HPV approved by the US Food and Drug Administration in 2003.

Researchers believed that a second experimental cervical cancer vaccine will appear to broadly protect against infections and risky precancerous conditions for more than two years. Further, the disease could be progressively eradicated in the global campaign much like smallpox and polio.

Patients given the vaccine sustained a high level of immune response against the viruses that spread cervical cancer. Besides, it would prevent infection for many years. Whether revaccination ultimately would be needed must be determined by an additional long trial.

Cervical cancer is caused by human papilloma viruses, which spread through sex. There are dozens of HPV strains, but two of them, T16 and T18, account for more than 70 percent of cervical cancers.

The GlaxoSmithKline vaccine is designed to prevent infection from both major strains. In a study researchers recruited 1,113 women at 32 clinics beginning in 2000. The participants, aged from 15-25, had no signs of infection. Their cases were followed for 27 months. About half of the women received the vaccine, while the rest received a placebo. None of the vaccinated women developed infections or cervical precancerous lesions. The vaccine also protected 93 percent against abnormal Pap tests.

Our experimental data contain 74 HPV downloaded from GenBank (http://www.ncbi.nlm.nih.gov/Genbank/). There are 14 types in Class 1, 11 types in Class 2, and 49 types in Class 3. The taxonomic positions were determined for each species using the NCBI taxonomy database.

2.2 Suffix Trees

DNA (deoxyribonucleic acid) is the most important molecule in living cells and it contains all of the information that a cell needs to make an existence and to propagate itself. DNA contains four different nitrogenous bases, namely, adenine (A), thymine (T), cytosine (C), and guanine (G). Thus, we can treat DNA sequences as strings whose alphabet contains only A, T, C, and G.

Given three sequences S_1 = ATCCGTGAACCGCAT, S_2 = TTGACCGCTTGGAT, and S_3 = TCGCAAATCCCTGAAG, we say that AT, TGA, CGC are the *common subsequences* of these sequences. A common subsequence is *maximal* if there is no other common subsequence containing it as a subsequence.

A suffix of sequence S of length n is the empty string when $n = 0$ or a subsequence of S that begins at position i where $1 \leqq i \leqq n$ and ends at position n. For example, the suffixes of string TCG are TCG, CG, and G.

Weiner [8] first proposed a linear-time algorithm for building suffix trees. Following the paper, some papers proposed techniques for minimizing the space requirement of the algorithm [9,10].

A suffix tree is a tree-like data structure that is defined as follows:

1. Suffix tree T is a rooted tree with m leaves numbered from 1 to m.
2. Each internal node, excluding the root, of T has at least 2 children.
3. Each edge of T is labeled with a nonempty subsequence of S.
4. No two edges out of a node can have edge-labels starting with the same character.
5. For any leaf i, the concatenation of the edge-labels on the path from the root to leaf i exactly spells out a suffix of S that starts at position i and ends at position m.

A *generalized suffix tree* is committed to the same rules but it contains more than one sequence. That is, it stores k sequences and all of their suffixes. By using the generalized suffix tree, we can easily obtain common subsequences for the input sequences [11].

The idea is primarily motivated by elementary biological considerations. When comparing two or more DNA sequences, regions that are well preserved are often of particular biological interest. They might encode impor-

tant functional domains of the corresponding protein or point to important regulatory elements. Moreover, they may represent suitable sites for PCR-primers that work across all of the considered sequences.

When given a number of sequences, we wish to know the maximal subsequence that is shared by all of them. Maximal common subsequences are observed among sequences. Repeats within a single sequence are also interesting in molecular biology.

3 Classification method

The reason of choosing HPV as our target is that their average length in sequences is between 7000 and 8000 bp and most of them are slightly different. Furthermore, their common subsequences are also fit in length.

Although the length of a SARS sequence is also not too long, its similarity is rather high. Furthermore, the common subsequences are too long to probe in practical aspect. On the contrast, the length of a bacterium is also too long. Nevertheless, the sequence of an HIV is too short, making it difficult to probe with the short common subsequence. Concerning the length of the common subsequences in HPV is just fit and not diversified as much.

Cervical cancer is ranked now the topmost among women's cancers. Besides, the disease condylomas is also caused by HPV. Nowadays, there are vaccines for cervical cancer. They may prevent cervical cancer and condylomas derived from certain types of HPV. However, we still do not know exactly the mechanism therein.

In this paper, we study the common subsequences in HPV. Furthermore, we use common ordered subsequences (respectively, appearing sequence of common subsequences) for classifying (respectively, identifying) HPV. A set of subsequences of S is called *ordered sequences* if they appear sequentially in S.

HPV can be categorized into three classes. However, there is no good approach to make a distinction. Our idea is as follows. For each class, we find a set of ordered sequences and it only appears in the current class of HPV, but does not appear in other classes of HPV. By using the ordered sequences, we may confirm that the claimed class of HPV is identified. For convenience, we use $C_1 \rightarrow C_2 \rightarrow C_3$ to denote the ordered sequences $\{C_1, C_2, C_3\}$ such that subsequence C_1 appears first, then C_2, and finally C_3.

3.1 Classifying Class 1 HPV

Class 1 HPV has been testified that almost 100% of them are related to many kinds of cervical cancers. We test 14 types of HPV which are all in Class 1.

The experiment results obtain 1532 sets of common ordered subsequences. By choosing C_1 = TAAAAGGTGA, C_2 = TATTTTTT, and C_3 = TATGTGT, we obtain the distribution of these common subsequences in Fig. 1. In Fig. 1, the number i denotes the subsequence C_i.

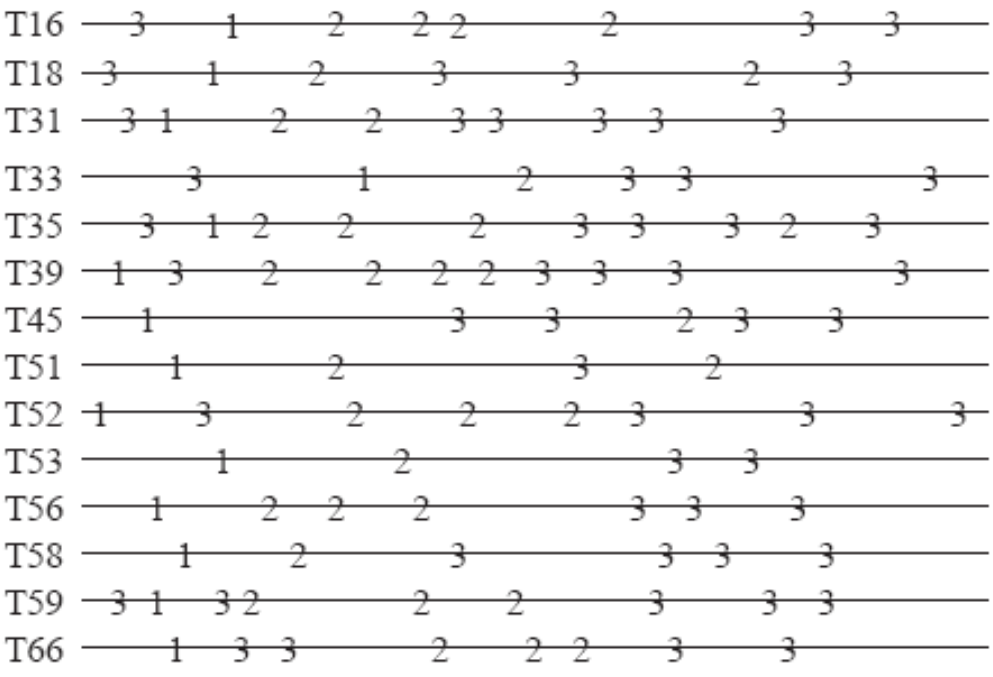

Fig. 1. The distribution of the chose common subsequences in Class 1 HPV.

By our experiment, we observe that Class 1 HPV have the $C_1 \rightarrow C_2 \rightarrow C_3$ common ordered subsequences. By testing the $C_1 \rightarrow C_2 \rightarrow C_3$ ordered subsequences on Classes 2 and 3, we find that they do not have this property. Thus $C_1 \rightarrow C_2 \rightarrow C_3$ can classify HPV of Class 1.

3.2 Classifying Class 2 HPV

We test 11 HPV of Class 2. The result of the experiment obtains 97 sets of common ordered subsequences. By choosing C_1 = TTTAGAT, C_2 = TATTTATT, and C_3 = TTTCTA, we obtain the distribution of these common subsequences in Fig. 2.

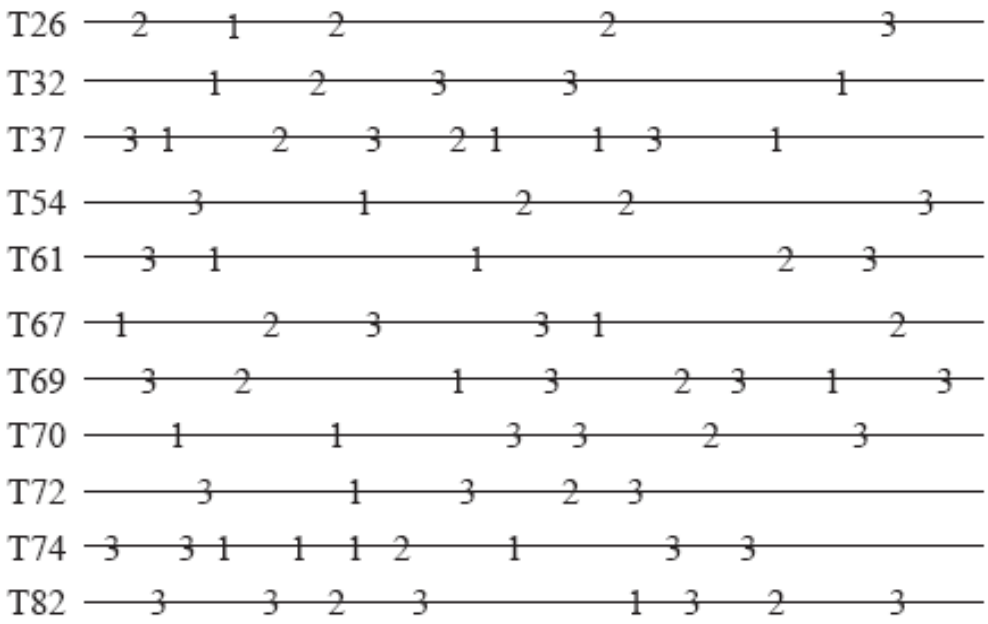

Fig. 2. The distribution of the chose common subsequences in Class 2 HPV.

According to Figure 2, we can find that Class 2 HPV have the $C_1 \rightarrow C_2 \rightarrow C_3$ common ordered subsequences. By testing the $C_1 \rightarrow C_2 \rightarrow C_3$ ordered subsequences on Classes 1 and 3, we find that they do not have this property. Thus $C_1 \rightarrow C_2 \rightarrow C_3$ can classify HPV of Class 2.

3.3 Classifying Class 3 HPV

We test 49 HPV of Class 3. The result of the experiment obtains 10267 sets of common ordered subsequences. By choosing C_1 = AAAAAG and C_2 = AAACAT, we obtain the distribution of these common subsequences in Fig. 3.

By a careful checking, we find that Class 3 HPV have the $C_1 \rightarrow C_2$ common ordered subsequences. By testing the $C_1 \rightarrow C_2$ ordered subsequences on Classes 1 and 2, we find that they do not have this property. Thus $C_1 \rightarrow C_2$ can classify HPV of Class 3.

4 Identification method

Given a set R of subsequences of sequence S, the disjoint occurrences of the subsequences in S is called the *appearing sequence* of S with respect to R. Note that if there is an overlap between two given subsequences in R, we only record the one with smaller starting position in its appearing sequence. For example, the appearing sequence of the sequence depicted in Fig. 4 is 13213. In the case of $R=\{r\}$, the corresponding appearing sequence is equivalent to the repeating sequence of r for S. Repeating sequences are studied in [12]. The concept of appearing sequence is an extension of [12]. With respect to R, if every appearing sequence in a set of viruses is different from others, then they can be treated as a signature.

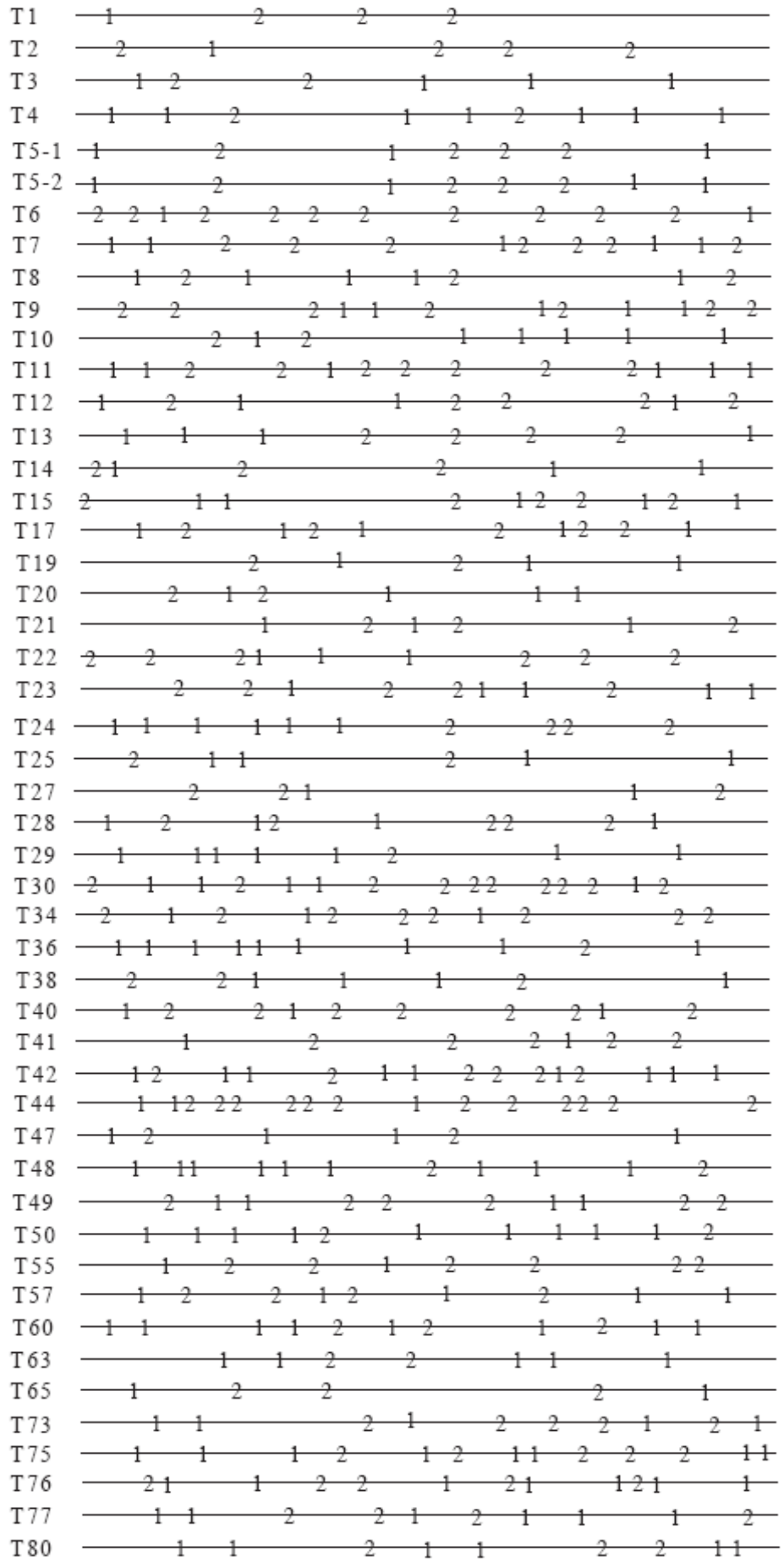

Fig. 3. The distribution of the chose common subsequences in Class 3 HPV.

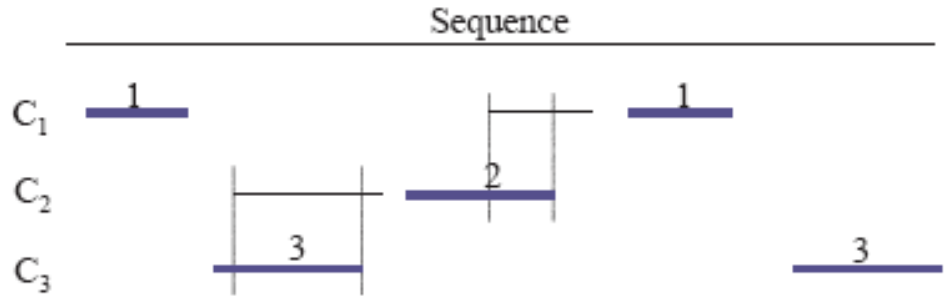

Fig. 4. Appearing sequence is 13213.

4.1 Identifying Class 1 HPV

According to the result of Section 3.1, we have 1532 sets of common ordered subsequences that can be used to classify Class 1 HPV. However, by exploring these sets, we find that there are 861 sets which are available to identify Class 1 HPV. By choosing C_1 = TAAAAGGTGA, C_2 = TATTTTTT, and C_3 = TATGTGT, we can distinguish Class 1 HPV as shown in Table 2.

Table 2. Appearing sequences for Class 1 HPV.

T16	31222233	T51	1232
T18	3123323	T52	13222333
T31	312233333	T53	1233
T33	312333	T56	1222333
T35	312223323	T58	123333
T39	1322223333	T59	313222333
T45	133233	T66	13322233

4.2 Identifying Class 2 HPV

According to the result of Section 3.2, we have 97 sets of common ordered subsequences that can be used to classify Class 2 HPV. However, by exploring these sets, we find that there are 78 sets which are available to identify Class 2 HPV. By choosing C1 = TATATAG, C2 = TATTTATT, and C3 = TTTCTA, we can distinguish Class 2 HPV as shown in Table 3.

Table 3. Appearing sequences for Class 2 HPV.

T26	1212233	T69	133121131323
T32	11233	T70	1331231
T37	132323	T72	311323
T54	132213	T74	13312133
T61	1323	T82	3213233
T67	1123312		

4.3 Identifying Class 3 HPV

According to the result of Section 3.3, we have 10267 sets of common ordered subsequences that can be used to classify Class 3 HPV. However, by exploring these sets, we find that there are only 741 sets which are available to identify Class 3 HPV. By choosing C1 = AAAAAG and C2 = AAACAT, we can distinguish Class 3 HPV as shown in Table 4.

Table 4. Appearing sequences for Class 3 HPV.

T1	1222	T19	21211	T44	1122222122222
T2	21222	T20	212111	T47	121121
T3	122111	T21	121212	T48	11111121112
T4	112112111	T22	222111222	T49	2112221122
T5-1	12111212	T23	22122111211	T50	11112111112
T5-2	1212221	T24	1111112222	T55	122122222
T6	221222222221	T25	211211	T57	122121211
T7	112221222112	T27	22112	T60	11112121211
T8	12111212	T28	1212112221	T63	1122111
T9	222112121122	T29	11111211	T65	12221
T10	21211111	T30	211211222222212	T73	1121222121
T11	1122122222111	T34	21212221222	T75	1112121122211
T12	121122212	T36	111111121	T76	211221211211
T13	11122221	T38	2211121	T77	1122121112
T14	212211	T40	1221222212	T80	112112211
T15	2112122121	T41	1222122		
T17	1212121221	T42	121121122212111		

5 Conclusion

In this paper, we propose methods to identify and classify HPV according to their maximal common subsequences. For classification, we use different ordered common subsequences to classify the three classes of HPV. For identification, we use appearing sequences of given maximal common subsequences to identify each type of HPV. The concept of appearing sequences is a generalization of repeating sequences [12].

In conclusion, our study shows that the common subsequences in DNA sequences can be used as a tool for classifying and identifying viruses. The occurrence of distinctive sequences would continue to be evolved in the process of classification and identification.

Acknowledgement

This research was supported by National Science Council under Grant NSC 92-2213-E-259-001.

References

[1] R. Sandberg, G. Winberg, C. I. Bränden, A. Kaske, I. Ernberg, and J. Cöster, "Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier," *Genome Research,* Vol.11, pp.1404-1409, 2001.

[2] R. Karlin and C. Burge, "Dinucleotide relative abundance extremes: a genomic signature," *Trends in Genetics,* Vol.11, pp.283-290, 1995.

[3] P. J. Deschavanne, A. Giron, J. Vilain, G. Fagot, and B. Fertil, "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences," *Molecular Biology and Evolution,* Vol.16, pp.1391-1399, 1999.

[4] R. Sandberg, C. I. Bränden, I. Ernberg, and J. Cöster, "Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content," *Gene,* Vol.311, pp.35-42, 2003.

[5] L. Kaderali and A. Schliep, "Selecting signature oligonucleotides to identify organisms using DNA arrays," *Bioinformatics,* Vol.18, pp.1340-1349, 2000.

[6] J. Cuzick, "Human papillomavirus testing for primary cervical cancer screening," *The Journal of American Medical Association*, Vol.283, pp.108-109, 2000.

[7] M. M. Manos, W. K. Kinney, L. B. Hurley, M. E. Sherman, J. Shieh-Ngai, R. J. Kurman, et al., "Identifying women with cervical neoplasia: using human papillomavirus DNA testing for equivocal Papanicolaou results," *The Journal of American Medical Association*, Vol.281, pp.1605-1610, 1999.

[8] P. Weiner, "Linear pattern matching algorithms," *Proceedings of* 14th *IEEE Annual Symp. on Switching and Automata Theory*, pp. 1-11, 1973.

[9] E. M. McCreight, "A space-economical suffix tree construction algorithm," *Journal of Algorithms*, Vol.23, pp.262-272, 1976.

[10] E. Ukkonen, "On-line construction of suffix trees," *Algorithmica*, Vol.14, pp.249-260, 1995.

[11] D. Gusfield, *Algorithms on strings, trees, and sequences*, Cambridge University Press, New York 1999.

[12] Y. C. Chang and C. H. Chang, "Common repeat sequences in bacterial genomes," *Journal of Medical and Biological Engineering*, Vol.23, pp.65-72, 1998.