# Prediction of the Risk Types of Human Papillomaviruses by Support Vector Machines

Je-Gun Joung[1,2], Sok June O[2,4], and Byoung-Tak Zhang[1,2,3]

[1] Biointelligence Laboratory, Graduate Program in Bioinformatics
[2] Center for Bioinformation Technology (CBIT)
[3] School of Computer Science and Engineering
Seoul National University, Seoul 151-742, Korea
[4] Department of Pharmacology, College of Medicine
Inje University, Busan 614-735, Korea
Phone: +82-2-880-5890, Fax: +82-2-883-9120
{jgjoung,juno,btzhang}@bi.snu.ac.kr

**Abstract.** Infection by high-risk human papillomaviruses (HPVs) is associated with the development of cervical cancers. Classification of risk types is important to understand the mechanisms in infection and to develop novel instruments for medical examination such as DNA microarrays. In this paper, we classify the risk type of HPVs by using the protein sequences. Our approach is based on the hidden Markov model and the Support Vector Machines. The former searches informative subsequence positions and the latter computes efficiently to classify protein sequences. In the experiments, the proposed classifier was compared with previous methods in accuracy and F-cost, also the prediction result of four unknown types is presented.

## 1 Introduction

Human papillomaviruses (HPVs) are small DNA viruses that infect epithelial tissues and relate to the diverse malignant tumors. Especially high-risk types could induce more than 95% of cervical cancer in woman. HPVs have a double-stranded DNA genome of approximately 8,000 bps that codes for 10 viral proteins, eight early gene products and two late gene products. More than 85 different HPV types have been described, with new types characterized because of significant differences in sequence homology compared with other defined HPV types [1]. Recently more than 120 have been partly reported [2]. The HPV types are often classified as low-risk or high-risk [3]. Low-risk viral types are associated with low-grade lesions such as condylomata and not cancers. On the other hand, high-risk viral types are associated with high-grade cervical lesions and cancers [4].

The most urgent and important thing for diagnosis and therapy is to discriminate which HPV genotypes are highly risky. Currently, the HPV risk types are classified manually by some experts. Furthermore, there is no method to test immediately if the new HPVs are detected from patients.

In this paper, we propose a novel method to classify HPV risk types, using protein sequence information. Our approach is based on the hidden Markov

C. Zhang, H.W. Guesgen, W.K. Yeap (Eds.): PRICAI 2004, LNAI 3157, pp. 723–731, 2004.

models (HMMs) and the support vector machines (SVMs). The former is suitable to search informative subsequence positions and the latter provides efficient computation to classify protein sequences. HMM is one of the most successful methods for biological sequence analysis. Especially, it has been quite successful in detecting conserved patterns in multiple sequences [5][6]. Whereas HMM is a generative model, the kernel-based classifier is a discriminant model. Ultimately, the proposed method uses the generative model to get easily distinguishable sequence source and the discriminant model to maximize classification ability.

The proposed SVM includes the string kernel that deals with protein sequences. The string kernel is an inner product in the feature space consisting of all subsequences of length $k$ and maps to feature space from sequences. The string kernel-based approach is efficient to analyze the biological sequence data, because it can extract important features from biological sequences. Recently, several string kernel approaches have been studied in bioinformatics and these have been mostly applied to analyze the protein sequences. For example, the string kernel has been applied to the peptide cleavage site recognition and remote homology detection, outperforming other conventional algorithms [7][8][9][10]. In this paper, SVMs learn a linear decision boundary between the two classes (high-risk and low-risk viral types).

Our work addresses how to classify the viral protein through the kernel-based machine learning approach. It can provide a guide to determine the risk type, when someone finds a novel virus. The paper is organized as follows. In Section 2, the data set is summarized and data pre-processing using HMM is described. Then the kernel method for HPV sequence analysis is presented in Section 3. In Section 4, the experimental results are provided by the proposed method applied to HPV sequence data sets. Concluding remarks and directions on further research are given in Section 5.

## 2 Data Set

### 2.1 Data Resource

The data set was extracted from the HPV sequence database at Los Alamos National Laboratory (LANL). High-risk HPV types can be distinguished from other HPV types based on the structure and function of the E6 and E7 gene products. For this reason, we got sequences corresponding to the 72 types from E6. E6 is an early gene product and plays an important role in cellular transformation. E6 products from oncogenic types of HPV can bind to and inactivate the cellular tumor suppressor gene products. This process plays an important role in the development of cervical cancer. Fifteen HPV types were labeled as high-risk types [11]. The rest were labeled as low-risk types.

### 2.2 Data Pre-processing Using HMM

The training and test data sets consist of subsequences that are estimated as more informative segments in the whole E6 sequence. The procedure for data preprocessing is as follows. First, all known risk type sequences are aligned by the multiple alignment tool such as ClustalW [13]. Second, they are divided into positive and negative sequences, then an HMM is constructed from positive segments. Each segment is the subsequence that is a window with size $w$ and is aligned over the same position by using the multiple alignment tool. Third, the log-likelihoods of positive and negative segments are calculated from the HMM model. Fourth, score is calculated by the difference between positive and negative log-likelihoods. The second and third steps are performed as the window shifts. Finally, the data set for classification is extracted from position that has the high score.

The biological sequence analysis has developed a reasonably successful solution using HMMs. HMMs are one of statistical sequence comparison techniques. They calculate the probability that a sequence was generated by a given model. In our approach, scoring is done by evaluating the probability that presents difference sequences by comparing the positive and negative segments.

## 3 Classifying by Support Vector Machines

### 3.1 Support Vector Machines

After the data-preprocessing, the string kernel-based SVM is trained on the HPV sequence data set and tested on the unknown sequences. Support vector machines were developed by Vapnik for classification of data based on a trained model [12]. Recently they have found several applications in biological data analysis. Given a kernel and a set of labelled training vectors (positive and negative input examples), SVMs learn a linear decision boundary in the feature space defined by the kernel in order to discriminate between the two classes. Any new unlabelled example is then predicted to be positive or negative depending on the position of its image in the feature space relatively to the linear boundary.

SVMs learn non-linear discriminant functions in an input space. This is achieved by learning linear discriminant function in a high-dimensional feature space. A feature mapping $\phi$ from the input space to the feature space maps the training data $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$ into $\Phi(S) = \{\Phi(\mathbf{x}_i), y_i\}_{i=1}^n = \{\mathbf{z}_i, y_i\}_{i=1}^n$. In the feature space, SVMs learn $f(z) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ so that the hyperplane separates the positive examples from negative ones. Here if $f(\mathbf{z}) > 0$ ($f(\mathbf{z}) < 0$) then the example is classified as positive (negative). The decision boundary is the hyperplane $\langle \mathbf{w}, \mathbf{z} \rangle = 0$ and the margin of the hyperplane is $\frac{1}{\|\mathbf{w}\|}$. Among normalized hyperplanes, SVMs find the maximal margin hyperplane that has the maximal margin.

According to the optimization theory, SVM optimization problem is solved by the following dual problem:

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \mathbf{z}_i, \mathbf{z}_j \rangle, \tag{1}$$

$$\text{subject to} \quad \alpha \geq 0 (1 \leq i \leq n), \sum_{i=1}^{n} \alpha_i y_i = 0, \tag{2}$$

where parameters $\alpha_i$ are called *Lagrange multipliers.* The parameters $(\mathbf{w}, b)$ are determined by the optimal $\alpha_i$. For a solution $\alpha_1^*, \ldots, \alpha_n^*$, the maximal margin hyperplane $f^*(\mathbf{z}) = 0$ can be expressed in the dual representation in terms of these parameters:

$$f^*(\mathbf{z}) = \sum_{i=1}^{n} \alpha_i^* y_i \langle \mathbf{z}_i, \mathbf{z} \rangle + b. \tag{3}$$

The dual representation allows for using kernel techniques. In the dual representation, the feature mapping $\phi$ appears in the form of inner products $\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle$.

### 3.2 Kernel Function

In this paper, the function of SVM is the mismatch-spectrum kernel. The mismatch-spectrum kernel is a new string kernel that was used to detect remote homology detection [8]. It is very simple and efficient to compute. In order to capture significant information from the sequence data, mismatch-spectrum kernels use the spectrum. The mismatch-spectrum kernel is the extended version of the spectrum kernel by adding the biologically important idea of mismatches. The $k$-spectrum kernel is based on a feature map from the space of all finite sequences to the vector space. Here the space of all finite sequences consists of an alphabet $\mathcal{A}$ of size $|\mathcal{A}| = l$ and the vector space is the $l^k$-dimensional vectors indexed by the set of $k$-length subsequences ($k$-mers) from $\mathcal{A}$.

For a simple feature map, the coordinate indexed by $\alpha$ of $k$-mer is the number of times $\alpha$ occurs in $x$. The $k$-spectrum feature map $\Phi_{(k)}(x)$ can be defined as:

$$\Phi_{(k)}(x) = (\phi_\alpha(x))_{\alpha \in \mathcal{A}^k}. \tag{4}$$

where $\phi_\alpha(x)$ is the number of occurrences of $\alpha$ in $x$ and $\mathcal{A}^k$ is the alphabet of the amino acids constituting $k$-mers. Thus the $k$-spectrum kernel function $K(x, y)$ for two sequences $x$ and $y$ is obtained by taking the inner product in feature space:

$$K_{(k)}(x, y) = \langle \Phi_{(k)}(x), \Phi_{(k)}(y) \rangle. \tag{5}$$

The use of the kernel function makes it possible to map the data implicitly into a high-dimensional feature space and to find the maximal margin hyperplane in the feature space. More biologically realistic kernel is the model allowing mismatch in $k$-mer subsequences. A fixed $k$-mer subsequence of amino acids is defined as $\alpha = a_1 a_2 ... a_k$, with each $a_i$ a character in $\mathcal{A}$. The $(k, m)$-neighborhood generated by $\alpha$ is the set of all $k$-length sequences $\beta$ from $\mathcal{A}$ that differ from $\alpha$ by at most $m$ mismatches. This set is denoted by $N_{(k,m)}(\alpha)$. The feature map $\Phi_{(k,m)}$ is defined as follows:

$$\Phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))_{\beta \in \mathcal{A}^k}, \tag{6}$$

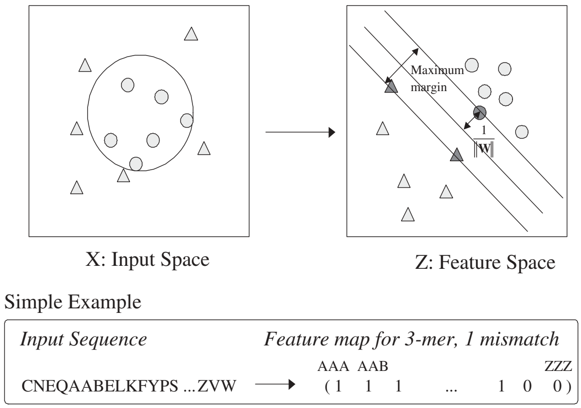

**Fig. 1.** The maximal margin classifier (or SVM) learns a linear discriminant function in high-dimensional feature so that the hyperplane optimally separates with maximum margin. For the mismatch-spectrum kernel, the feature map $\phi_\beta(\alpha)$ is given from input space into a high-dimensional feature space (vector space). The feature map is indexed by all possible $k$-mers.

where $\phi_\beta(\alpha) = 1$ if $\beta$ belongs to $N_{(k,m)}(\alpha)$, $\phi_\beta(\alpha) = 0$ otherwise. Thus, a $k$-mer contributes weight to all the coordinates in its mismatch neighborhood.

The feature map on an input sequence $x$ is defined as the sum of the feature vectors assigned to the $k$-mers in $x$:

$$\Phi_{(k,m)}(x) = \sum_{k-\text{mers } \alpha \text{ in } x} \Phi_{(k,m)}(\alpha) \tag{7}$$

Note that the $\beta$-coordinate for $\Phi_{(k,m)}(x)$ is just a count of all instances of the $k$-mer $\beta$ occurring with up to $m$ mismatches in $x$. The $(k,m)$-mismatch kernel $K_{(m,k)}$ is the inner product in feature space of feature vectors:

$$K_{(k,m)}(x,y) = \langle \Phi_{(k,m)}(x), \Phi_{(k,m)}(y) \rangle. \tag{8}$$

Mismatch kernels are used in combination with the SVM. Fig. 1 shows the classification task of discriminating the positive sequence class from the negative class. SVMs employing the mismatch-spectrum kernel perform the learning in a high-dimensional feature space.

For $(k,m)$ mismatch kernel $K_{(k,m)}$, if Eq. (5) is applied, then the learned SVM classifier is represented as:

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i \langle \Phi_{(k,m)}(x_i), \Phi_{(k,m)}(x) \rangle + b. \tag{9}$$

Here $x_i$ are the training sequences, $y_i$ are labels, and $\alpha_i$ are weights. It can be implemented by pre-computing and storing per $k$-mer scores so that the prediction can be calculated in linear time by look-up of $k$-mer scores [8].

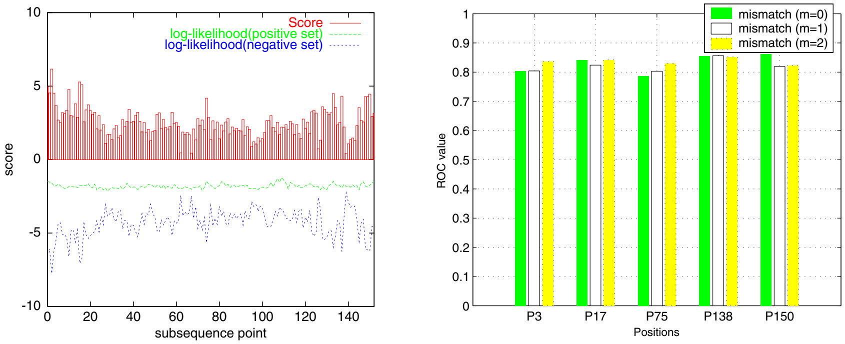

**Fig. 2.** High scores of subsequences through HMM learning in E6 (left). Point of High scores are 3, 17, 75, 138, and 150. These points are possible to play an important role in the tumor-related suppression or activation. The ROC scores of subsequences that present high scores by HMM (right). The point 138 indicates the highly conserved sequence position so that the highly conserved regions are associated with the classification performance.

# 4 Experiments

## 4.1 Searching Informative Subsequences

Left in Fig. 2 shows the scores that are computed through the HMM model to find more informative subsequence positions. Each score is the difference between the log-likelihood of the positive subsequences and one of negative subsequences. The positive data and negative data were selected from the believable types (the number of the positive data set: 15, the number of the negative data set: 11) The window size $w$ is 12 and the number of shifted segments is 153.

In this figure, high scores by HMM are points 3, 17, 75, 138, and 150 that are positions of the starting residues of subsequences. These points are probably motifs that play an important role. The point 138 is zinc-binding region of E6 [16]. In E6, the zinc-binding region is necessary for *trans*-activation and transformation, and is involved in protein-protein interactions. E6 binds to p53 that is a cellular tumor suppressor protein [14]. E6 from high-risk HPV binds p53 with higher affinity than that from low-risk HPV, and mediates the degradation of p53 through the ubiquitin-dependent system.

## 4.2 Prediction Performance of Subsequences

Fig. 2 (right) shows the ROC (receiver-operating characteristic) values of subsequences that present high scores in Fig. 2 (left). An ROC represents the joint values of the true positive ratio (sensitivity) and false positive ratio (1−specificity). Each bar is an average value after 100 runs. The size of $k$-mers is 4. At this test, the point 138 has high accuracy for tree mismatches ($m$=0, 1, 2). Each ROC value is 86 ($m$=0), 86 ($m$=1) and 85 ($m$=2), respectively. The point 138 indicates the highly conserved sequence region as described in the above section. The result suggests that searching the highly conserved region improves the accuracy of the classifier.

**Table 1.** The performance comparison of sequence based classification (SVMs) and text based classification (AdaCost, AdaBost, navie Bayes).

| | Sequence based classification | Text based classification | | |
|---|---|---|---|---|
| Method | SVMs | AdaCost | AdaBoost | naive Bayes |
| Accuracy | 93.15 | 93.05 | 90.55 | 81.94 |
| F-score | 85.71 | 86.49 | 80.08 | 63.64 |

Our approach was compared with textmining approaches in the classification performance. Table 1 shows the comparison of four learning methods. Three methods (AdaCost, AdaBoost, nave Bayes) had been reported in previous study that presented methods to classify the risk type from text data [15]. All results were values predicted by one-leave-out cross-validation. Here the F-score is usually used for Information Retrieval (IR) performance measures. The F-score is computed for given precision $p$ and recall $r$ as F-score $= (2pr)/(p + r)$. SVMs shows 93.15% of accuracy and 85.71% of F-score.

This result supports that sequence based classification can show higher classification performance than text-based classification or similar performance. When the documents of some types are available, text based approach can perform. However new types are detected from patients, text based approach is useless. Our approach can be used generally without additional information such as comments or descriptions.

### 4.3 Classification of Risk Types

Table 2 shows the comparison between the manually tagged answers and the string kernel based predictions. The manually tagged answers are based on the human papillomavirus compendium (1997 version) and Muñoz's [11] paper. Seventeen HPV types were classified as high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 61, 66, 67, 68, and 72). If the type belongs to the skin-related or cutaneous HPV group reported at the human papillomavirus compendium, it is classified as a low-risk type. There was a good agreement between our epidemiologic classification and the classification based on phylogenetic grouping. In this table, symbol ? denotes the risk type that can not be determined, and there are three unknown types. The prediction is the result of one-leave-out cross-validation.

The most interesting is that the classifier predicted high-risks for HPV70. According to the previous study on HPV [17], the document contains that HPV70 also detected in genital intraepithelial neoplasia from one patient. This is very important result because the classifier in this paper provides the probability whether the unknown HPV types are high-risk or not.

The prediction of types 30, 32, 53, 66 and 68 has different answers for the manually tagged answer. HPV30 and HPV32 were associated specifically with a laryngeal carcinoma and Heck's disease, respectively. They were not classified as high-risk, but are probably associated with a high risk of carcinogenesis. In contrast to the two types, the prediction for HPV53, HPV66, and HPV68 is

**Table 2.** Comparison between the manually tagged answer (Man.) and the string kernel based prediction (Class.).

| Type | Man. | Class. | Type | Man. | Class. | Type | Man. | Class. | Type | Man. | Class. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HPV1 | Low | Low | HPV20 | Low | Low | HPV38 | Low | Low | HPV57 | ? | Low |
| HPV2 | Low | Low | HPV21 | Low | Low | HPV39 | High | High | HPV58 | High | High |
| HPV3 | Low | Low | HPV22 | Low | Low | HPV40 | Low | Low | HPV59 | High | High |
| HPV4 | Low | Low | HPV23 | Low | Low | HPV41 | Low | Low | HPV60 | Low | Low |
| HPV5 | Low | Low | HPV24 | Low | Low | HPV42 | Low | Low | HPV61 | High | High |
| HPV6 | Low | Low | HPV25 | Low | Low | HPV43 | Low | Low | HPV63 | Low | Low |
| HPV7 | Low | Low | HPV26 | ? | Low | HPV44 | Low | Low | HPV65 | Low | Low |
| HPV8 | Low | Low | HPV27 | Low | Low | HPV45 | High | High | HPV66 | High | Low |
| HPV9 | Low | Low | HPV28 | Low | Low | HPV47 | Low | Low | HPV67 | High | High |
| HPV10 | Low | Low | HPV29 | Low | Low | HPV48 | Low | Low | HPV68 | High | Low |
| HPV11 | Low | Low | HPV30 | Low | High | HPV49 | Low | Low | HPV70 | ? | High |
| HPV12 | Low | Low | HPV31 | High | High | HPV50 | Low | Low | HPV72 | High | High |
| HPV13 | Low | Low | HPV32 | Low | High | HPV51 | High | High | HPV73 | Low | Low |
| HPV15 | Low | Low | HPV33 | High | High | HPV52 | High | High | HPV74 | Low | Low |
| HPV16 | High | High | HPV34 | Low | Low | HPV53 | Low | High | HPV75 | Low | Low |
| HPV17 | Low | Low | HPV35 | High | High | HPV54 | ? | Low | HPV56 | Low | Low |
| HPV18 | High | High | HPV36 | Low | Low | HPV55 | Low | Low | HPV77 | Low | Low |
| HPV19 | Low | Low | HPV37 | Low | Low | HPV56 | High | High | HPV80 | Low | Low |

sure to make a mistake. HPV type 53 was detected in genital specimens of the 16 of the patients [18]. However, it is probably not associated with a high risk of carcinogenesis. To be exact in prediction, there is a need for the data set to contain sequences for E7 or L1 gene.

## 5 Conclusion

We proposed the use of a kernel based method to classify HPV risk types. The proposed kernel-based classifier includes the mismatch string kernel. The string kernels function as a mapping to feature space from sequences. These kernels compute sequence similarity based on shared occurrences of $k$-mer. The string kernel-based classifier is very powerful to analyze the biological sequence data, because it can extract important features from input sequences. When the classifier learns informative subsequences, the accuracy is better. The informative subsequences could indicate the highly conserved regions.

We predicted the risk type for all types via one-leave-out cross-validation. The most interesting question is ‘what is the risk type of HPV70’. This paper provides the probability whether the unknown HPV types are high-risk or not. Our approach can provide a priori knowledge for probe selection in designing genotyping DNA-microarrays. In other words, it can catch specificity to classify high-risk and low-risk viral infection. For more accurate prediction, the input sequence data could be combined with information of the protein structure. Using the secondary structure is possible to have on effect HPV classification.

## Acknowledgments

This research was supported in part by NRL, the Korean Ministry of Science and Technology (Korean Systems Biology Research Grant, M10309000002-03B5000-00110).

## References

1. H. Pfister, J. Krubke, W. Dietrich, T. Iftner, P. G. Fuchs, Classification of the papilomavirues-mapping the genome, *Ciba Found. Symp.*, Vol. 120, pp. 3–22, 1986.
2. H. zur Hausen, Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis, *Journal of National Cancer Inst.*, Vol. 92, pp. 690–698, 2000.
3. IARC Monographs on the Evaluation of the Carcinogenic Risks to Humans, Lyon, France: IARC Scientific Publications, 1995.
4. M. F. Janicek, H. E. Averette, Cervical cancer: prevention, diagnosis, and therapeutics, *A Cancer Journal for Clinicians*, Vol. 51, pp. 92–114, 2001.
5. P. Baldi, Y. Chauvin, et. al., Hidden Markov models of biological primary sequence information, *PNAS*, Vol. 91, pp. 1059–1063, 1994.
6. S. Eddy, Multiple alignment using hidden Markov models, *ISMB 95*, pp. 114–120, 1995.
7. C. Leslie, E. Eskin, W. Noble, The spectrum kernel: a string kernel for SVM protein classification, *Proceedings of the Pacific Symposium on Biocomputing 2002*, pp. 564-575, 2002.
8. C. Leslie, E. Eskin, J. Weston, W. Noble, Mismatch String Kernels for Discriminative Protein Classification, *Bioinformatics*, Vol. 20, pp. 467-476, 2004.
9. J.-P. Vert, Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings, *Proceedings of the Pacific Symposium on Biocomputing 2002*, pp. 649–660, 2002.
10. T. Jaakkola, M. Diekhans, D. Haussler, A discriminative framework for detecting remote protein homologies, *Journal of Computational Biology*, 2000.
11. N. Muñoz, F. X. Bosch, et. al., Epidemiologic classification of human papillomavirus types associated with cervical cancer, *N. Engl. J. Med.*, Vol. 348, pp. 518–527, 2003.
12. V. N. Vapnik, Statistical Learning Theory, Springer, 1998.
13. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res.*, Vol. 22, pp. 4673–4680, 1994.
14. T. Ristriani, M. Masson, et al., HPV oncoprotein E6 is a structure-dependent DNA-binding protein that recognizes four-way junctions, *J. Mol. Biol.*, 10 (296), pp. 1189–1203, 2000.
15. S.-B. Park, S. Hwang, and B.-T. Zhang, Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost, *Lecture Notes in Computer Science*, Vol. 2736, pp. 403-412, 2003.
16. C. G. Ullman, P. I. Haris, et al., Predicted -helix/ -sheet secondary structure for the zinc-binding motifs for human papillomavirus E7 and E6 proteins by consensus prediction averaging and spectroscopic studies of E7, *Biochem. J.*, Vol. 319, pp. 229–239, 1996.
17. M. Longuet, S. Beaudenon, G. Orth, Two novel genital human papillomavirus (HPV) types, HPV68 and HPV70, related to the potentially oncogenic HPV39, *J. Clin. Microbiol.*, 34 (3), pp. 738–744, 1996.
18. T. Meyer, R. Arndt, et al., Distribution of HPV 53, HPV 73 and CP8304 in genital epithelial lesions with different grades of dysplasia, *Int. J. Gynecol. Cancer.*, 11 (3), pp. 198–204, 2001.