Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 756345, 9 pages
http://dx.doi.org/10.1155/2015/756345

Research Article

Prediction of High-Risk Types of Human Papillomaviruses Using Statistical Model of Protein "Sequence Space"

Cong Wang,[1] Yabing Hai,[1] Xiaoqing Liu,[2] Nanfang Liu,[3] Yuhua Yao,[1] Pingan He,[4] and Qi Dai[1,5]

[1] *College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[2] *College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, China*
[3] *Department of Gynecological Oncology, Zhejiang Cancer Hospital, Hangzhou 310022, China*
[4] *College of Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, China*
[5] *Center for Systems Biology, University of Texas at Dallas, Richardson, TX 75080, USA*

Correspondence should be addressed to Qi Dai; daiailiu04@yahoo.com

Received 19 September 2014; Accepted 31 March 2015

Academic Editor: Edward J. Perkins

Discrimination of high-risk types of human papillomaviruses plays an important role in the diagnosis and remedy of cervical cancer. Recently, several computational methods have been proposed based on protein sequence-based and structure-based information, but the information of their related proteins has not been used until now. In this paper, we proposed using protein "sequence space" to explore this information and used it to predict high-risk types of HPVs. The proposed method was tested on 68 samples with known HPV types and 4 samples without HPV types and further compared with the available approaches. The results show that the proposed method achieved the best performance among all the evaluated methods with accuracy 95.59% and *F1*-score 90.91%, which indicates that protein "sequence space" could potentially be used to improve prediction of high-risk types of HPVs.

1. Introduction

Cervical cancer is one of the leading causes of cancer morbidity and mortality in women worldwide [1]. Approximately, 500,000 new cases of cervical cancer were diagnosed each year, with 280,000 deaths [2]. It has become the second most common cancer among women especially for developing countries [3, 4]. Some studies have shown that human papillomavirus (HPV) is strongly associated with cervical cancer, and some types of HPVs can cause abnormal tissue growth in the form of warts (papillomas) and some HPVs are associated with certain cancers and precancerous conditions [5–7].

Human papillomaviruses are icosahedral, nonenveloped particles that contain a small, double-stranded circular DNA of approximately 8000 nucleotide base pairs [8] and belong to the Papillomavirus family (papilloma, polyoma, and simian vacuolating viruses) [9]. The diameter of circular DNA is approximately 55 nm [10–13]. Up to now, there are more than 150 HPV types, and some new types will be identified when they have significant homology differences with the defined HPV types [14–16]. Epidemiologic studies have shown that genital human papillomaviruses have a strong relationship with cervical cancer, independent of other risk factors. According to their relative malignancy, the genital tract HPVs can be grouped into two or three types: low-risk type, intermediate-risk type, and high-risk types [17]. But HPVs are usually divided into two types in clinical association study: high-risk or low-risk types. Low-risk viral types are more closely related with low-grade lesions, and high-risk viral types are associated with high-grade cervical lesions and cancers [17]. High-risk type is composed of 20 HPV types, such as HPV-16, HPV-18, HPV-26, HPV-31, HPV-33, HPV-35, HPV-39, HPV-45, HPV-51–53, HPV-56, HPV-58, HPV-59, HPV-66, HPV-68, HPV-70, HPV-73, HPV-82, and HPV-85 [18]. HPV-16 and HPV-18 are responsible for about 62.6% and 15.7% of cervical cancers [19]. Therefore, discrimination of high-risk types of HPVs becomes one of the most important things for diagnosis and therapy of cervical cancers.

Because of the importance of the HPV types, many epidemiological and experimental methods have been proposed to identify them [5, 20–22]. They are mostly based on the polymerase chain reaction (PCR), a sensitive technique for the detection of very small amounts of HPV nucleic acids in clinical specimens. With rapid increasing of the HPV data in protein and DNA databank, there is a great need to develop some reliable and effective computational methods to predict the high-risk types of HPVs directly from the available data.

Recently, some research works found the correlations between these data and high-risk types of HPVs and proposed some computational methods to predict the high-risk types of HPVs. Eom et al. learned the most informative subsequence segment sets of DNA sequences and used genetic algorithm to classify the risk types of each HPV [23]. Joung et al. classified the risk type of HPVs based on the hidden Markov model and the support vector machines using the protein sequences [24, 25]. Park et al. proposed a classification of the risk type of human papillomavirus by decision tree [26]. Kim and Zhang introduced the string kernel and Gap-spectrum kernel to compute the distances of amino acids pairs and further used them to classify HPV risk types based on E6 protein sequences [7, 9]. Kim et al. proposed an Ensemble support vector machine to classify HPV risk types based on the differential subsequences of protein secondary structures [13]. Esmaeili et al. calculated Chou's pseudo amino acid composition of E6 protein sequences and used ROC to predict HPV risk types [27]. Alemi et al. analyzed the physiochemical properties of all early and late proteins in high- and low-risk HPV types and introduced support vector machines to classify high-risk HPV types based on receiver operating characteristic analysis [28].

These methods have achieved promising results in high-risk types of HPVs prediction, but challenges for information extraction of HPVs still remain. The widely used information of HPVs in high-risk type prediction is sequence-based or structure-based information from the given DNA or protein sequence, and the information from related proteins or family has not been explored until now. With this problem in mind, we presented a novel scheme to predict high-risk types of HPVs using word statistical model of protein sequence space and support vector machine. We first constructed a "sequence space" of the given protein sequence with help of mutation matrices. We then extracted the information of HPV from the protein "sequence space" with the proposed word statistical model. At last, the extracted information was fed into support vector machine to predict high-risk types of HPVs. Through several experiments, we want to address how well the proposed prediction method performed when comparing with the available ones and whether the prediction abilities of the proposed prediction method depends on the choice of the mutation matrices.

2. Materials and Methods

2.1. Datasets. All types of HPV share a common genomic structure which is arranged into the upstream regulatory region (URR) and eight open reading frames (ORFs) encoding the viral early and late genes [11]. URR contains long control region, TATA signal 1 and TATA signal 2. There are polyA signal 1 and polyA signal 2 between early and late genes. Late gene expression produces the structural proteins L1 and L2 [12], which assemble into the viral capsid structure, whereas early gene activity translates into the regulatory proteins E1, E2, E4, E5, E6, and E7. In this paper, we constructed seven datasets of HPV protein sequences: E1, E2, E4, E6, E7, L1, and L2, respectively. Here, we did not use HPV E5 because the lengths of its protein sequences are too small. All the HPV datasets were downloaded from the Human Papillomaviruses Compendium published by Los Alamos National Laboratory (LANL).

There are total 72 types of HPVs in each dataset, but some HPV sequences are missing in LANL. So we downloaded the missing sequences from taxonomy browser in the National Center of Biotechnology Information. For example, HPV 43, 67, 75, 76, 77, and 80 protein sequences are missing in L2 dataset; we obtained these sequences from taxonomy browser. But we could not find the missed sequences of the E4 dataset in the National Center of Biotechnology Information, so the total number of HPV sequences is 71 in the E4 dataset. Among HPV sequences, four sequences (HPV 26, 54, 57, and 70) are selected as the predicting data and others are the training data [13]. Here, HPV risk types are manually determined based on the HPV compendium, in which seventeen HPV types are classified as high-risk types (HPV 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 61, 66, 67, 68, and 72) and the remaining are low-risk types.

2.2. Construction of Protein "Sequence Space". It is well known that there are over 20 amino acids and each one is different from the others. Mutation matrices represent the similarities among amino acids. Let AA_i and AA_j denote two amino acids from the set Ω, and their score was defined as follows:

$$S\left(AA_i, AA_j\right) = \text{Mutation}\left(AA_i, AA_j\right), \tag{1}$$

where $\text{Mutation}(AA_i, AA_j)$ represents the "normalized probability" that the amino acid AA_i mutates into the amino acid AA_j. In evolutionary biology, the score describes the rate at which one amino acid in a protein sequence changes to other amino acids states over time. That is to say the sequence similarity depends on the amino acids' scores represented in above definition. Usually, two amino acids AA_i and AA_j are considered similar if their score is more than zeros. It is worth noting that the similarity between AA_i and AA_j is symmetric, but it is not a transitive relation. For example, AA_i is similar to AA_j and AA_j is similar to AA_k, but AA_i is not similar to AA_k.

Taking amino acids' scores into mind, we classified 20 amino acids into several overlapping classes. Here, star sets were introduced, in which the properties are known between vertices and center. Given an amino acid AA_i, its star set was defined as follows:

$$\text{Star}S\left(AA_i\right) = \bigcup_{\alpha \in \Omega} \text{sig}\left(S\left(AA_i, AA_\alpha\right)\right) \cdot \alpha, \tag{2}$$

Table 1: Star sets of 20 amino acids based on PAM250 substitution matrix.

Matrix	Star set				
PAM250	{AGPST}	{C}	{DEGHNQ}	EDHNQ	{FILY}
	{GADS}	{HDENQR}	{IFLMV}	{KNQR}	{LFIMV}
	{MILV}	{NDEHKQS}	{PAS}	{QDEHKNR}	{RHKQW}
	{SAGNPT}	{TAS}	{VILM}	{WR}	{YF}

where sig is a function that returns the sign of a number, indicating whether the number is positive or zero. If number is greater than zero, 1, otherwise, to zero. For example, 20 amino acids can be partitioned into several star sets based on PAM250 mutation matrix, which was presented in Table 1.

We wanted to go further with the star sets and found some related protein sequences that have a high similarity among them. Suppose $S_1 = s_1^1 s_2^1 \cdots s_n^1$ and $S_2 = s_1^2 s_2^2 \cdots s_n^2$ are two given protein sequences; they are related if they satisfy the following condition:

$$\forall s_k^1 \in \mathrm{Star}S\left(s_k^2\right), \quad \forall s_k^2 \in \mathrm{Star}S\left(s_k^1\right), \quad 1 \le k \le n. \qquad (3)$$

From the above definition, it is easy to note that if two protein sequences have more similar sequences, they should be more closely related. With help of definition of related sequences, we constructed the "sequence space" of given sequence $S = s_1 s_2 \cdots s_n$, denoted by SP_S, as follows.

Step 1. Given a null-set, denoted by ϕ, add its star sets to ϕ and obtain protein "sequence space" SP_S.

Step 2. A prefix s_1 was added to ϕ and obtained its star set $\mathrm{Star}S(s_1)$. We checked whether the star set of the prefix s_1 is empty or not. If its star set is a nonempty set, we added a symbol "–" after $\mathrm{Star}S(s_1)$ and updated the protein "sequence space" SP_S.

Step 3. We repeated Step 2 until the end of the given sequence $S = s_1 s_2 \cdots s_n$ and obtained its protein "sequence space" SP_S as follows:

$$\mathrm{SP}_S = \bigcup_{k=1}^{n} \left[\bigcup_{\alpha \in \Omega} \mathrm{sig}\left(S\left(AA_k, AA_\alpha\right)\right) \cdot \alpha \right] - . \qquad (4)$$

In the construction of the protein sequence space, all the protein sequences were closely related to the given protein sequence. That is to say all the information on the related proteins or family could be explored through the construction of the protein "sequence space."

2.3. Word Statistical Model in Protein "Sequence Space". Word statistical model is one of the most widely used methods for sequence analysis [29–32]. In this model, each sequence is first mapped into an m-dimensional vector according to its word frequencies, and sequence similarity can be measured by distance measures, such as Euclidean distance [33], Mahalanobis distance [34], Kullback-Leibler discrepancy [35], and Cosine distance [36]. When the words occurring in biological sequence are estimative probabilities rather than the frequencies, they are more readily optimized by more complex models, such as Markov model [37–39], mixed model [40], and Bernoulli model [41]. These complex models could be considered to be the modification of traditional word-based models.

A biological sequence can be described as a succession of symbols, and a word is a series of k consecutive letters in the sequence. For a sequence $S = s_1 s_2 \cdots s_n$, the count of its word $W_k = w_1 w_2 \cdots w_k$, denoted by $c(W_k)$, is the number of occurrence of the word W_k in the sequence S. Here, we constructed a word statistical model in protein "sequence space." First of all, a position function of an occurrence of the word W_k was defined as follows:

$$\aleph_i\left(s_i, w\right) = \begin{cases} 1, & \text{if } s_i = w, \\ 0, & \text{otherwise.} \end{cases} \qquad (5)$$

The count of the word W_k in the protein "sequence space" can be defined from the random indicators of occurrence as follows

$$\begin{aligned} \Phi\left(W_k\right) = \sum_{i=1}^{n-k+1} \sum_{\alpha_1 \in \mathrm{Star}S(s_i)} \sum_{\alpha_2 \in \mathrm{Star}S(s_{i+1})} \\ \cdots \sum_{\alpha_k \in \mathrm{Star}S(s_{i+k})} \aleph_i\left(\alpha_1, w_1\right) \times \aleph_{i+1}\left(\alpha_2, w_2\right) \times \cdots \\ \times \aleph_{i+k}\left(\alpha_k, w_k\right). \end{aligned} \qquad (6)$$

In order to eliminate the effects of space size, we normalized the word contents with the size of the space and got word frequencies of protein "sequence space," denoted as $F_k^{\mathrm{SP}_S}$. Consider

$$\begin{aligned} F_k^{\mathrm{SP}_S} &= \left(f^{\mathrm{SP}_S}\left(W_{k,1}\right), f^{\mathrm{SP}_S}\left(W_{k,2}\right), \ldots, f^{\mathrm{SP}_S}\left(W_{k,Y}\right)\right) \\ &= \left(\frac{\Phi\left(W_{k,1}\right)}{\prod_{i=1}^{n-k+1} \prod_{j=1}^{k} \left|\mathrm{Star}S\left(s_{i+j}\right)\right|}, \right. \\ &\qquad \frac{\Phi\left(W_{k,2}\right)}{\prod_{i=1}^{n-k+1} \prod_{j=1}^{k} \left|\mathrm{Star}S\left(s_{i+j}\right)\right|}, \ldots, \\ &\qquad \left. \frac{\Phi\left(W_{k,Y}\right)}{\prod_{i=1}^{n-k+1} \prod_{j=1}^{k} \left|\mathrm{Star}S\left(s_{i+j}\right)\right|} \right), \end{aligned} \qquad (7)$$

where $|\mathrm{Star}S|$ is the size of the star set and Y is the total number of the words that appear in the protein "sequence space" SP_S.

2.4. Prediction Algorithm. There are two types of HPV protein sequences: high-risk type and low-risk type. Let $Y = [y_1, y_2, \ldots, y_n]^T$ denote the type labels of n samples, where $y_i = k$ indicates the ith sample being risk type k, where $k = 1, 2$ denotes two different risk types ($y_i = 1$ indicates the ith sample being high-risk type, and $y_i = 2$ indicates the ith sample being low-risk type). Let x_{ij} be the jth word frequency in protein "sequence space" for the ith sample, where $j = 1, 2, \ldots, m$; $X = (x_{ij})_{n,m}$ denotes all the statistical information of "sequence space" for all samples,

$$X = \begin{array}{c|cccc|} & \text{index1} & \text{index2} & \cdots & \text{index}m \\ x_1 & x_{11} & x_{12} & \cdots & x_{1m} \\ x_2 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n1} & x_{n2} & \cdots & x_{nm} \end{array}, \quad (8)$$

where $x_1, x_2, \ldots, x_n$ are n samples and $x_i = [x_{i1}, x_{i2}, \ldots, x_{in}]$ and $x_i \in R^m$. With help of the support vector machine (SVM), the prediction problem of HPV types was formulated as follows:

$$\begin{aligned} \min_{w,b,\xi} \quad & J(w, b, \xi) = \frac{1}{2}\left(w^T w\right) + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & y_i \left[w^T \varphi(x_i) + b\right] \geq 1 - \xi_i, \quad i = 1, 2, \ldots, n, \\ & \xi_i \geq 0, \quad i = 1, 2, \ldots, n, \end{aligned} \quad (9)$$

where w is defined as a Linear combination of the set of nonlinear data transformations

$$w = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i), \quad (10)$$

b is a bias term, C is a regularization metaparameter, and ξ_i denotes the training error for the ith sample. This optimization problem derived in a dual space can be written as

$$\begin{aligned} \max_{\alpha} \quad & J(\alpha) = \max_{\alpha} \sum_{i=1}^{n} \alpha_i \\ & \quad - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \varphi(x_i)^T \varphi(x_j) \\ \text{subject to} \quad & \sum_{i-1}^{n} \alpha_i y_i = 0, \quad i = 1, 2, \ldots, n, \\ & 0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, n. \end{aligned} \quad (11)$$

In this paper, we used the Gaussian radius basis function kernel to calculate the $\varphi(x_i)^T \varphi(x_j)$ instead of calculating either $\varphi(x_i)$ or $\varphi(x_j)$ explicitly. Then the optimal separating problem was modeled as

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b. \quad (12)$$

And the classifier takes the form

$$y(x) = \operatorname{sign}[f(x)]. \quad (13)$$

After training the model, a test sample $x \in R^m$ will be assigned to a risk type according to the following decision function:

$$y(x) = \begin{cases} 1, & \text{if } f(x) > 0, \\ 2, & \text{if } f(x) \leq 0. \end{cases} \quad (14)$$

When $y(x)$ is 1, it means that the test sample x is the high-risk type of HPV; otherwise, x should be low-risk type. Here, we selected the parameters for the sake of getting the highest overall prediction as possible. A simple grid search strategy based on 10-fold cross-validation for each dataset was performed to get the optimal values of α_i and b for prediction algorithm.

3. Results and Discussion

3.1. Evaluation Measures. Subsampling test, independent dataset test, and jackknife test are three widely used cross-validation methods to evaluate prediction's capability. The jackknife test always yields a unique outcome, which facilitates examining the quality of various predictors. Hence, we chose jackknife test to evaluate the performance of the proposed method and introduced the accuracy for each class, overall accuracy, and $F1$-score as standard performance measures, which were defined as follows:

$$\begin{aligned} \text{specificity}\,(\text{accuracy of high} - \text{risk type}) &= \frac{a}{a + c}, \\ \text{sensitivity}\,(\text{accuracy of low} - \text{risk type}) &= \frac{d}{b + d}, \\ \text{accuracy of totality} &= \frac{a + d}{a + b + c + d} \cdot 100\%, \\ F1\text{-score} &= \frac{2 \cdot a/(a + b) \cdot a/(a + c)}{(a/(a + b)) + (a/(a + c))} \cdot 100\%, \\ &= \frac{2a^2}{2a^2 + ac + ab} \cdot 100\%, \end{aligned} \quad (15)$$

where a is the number of true positives, c is the number of false positives, d is the number of true negatives, and b is the number of false negatives. From their definition, it is interesting to note that $F1$-score will be higher if a is bigger. That is to say $F1$-score will be better to reflect the efficiency of HPV risk type prediction capacity.

3.2. Comparison of Early and Late Proteins' Performances in HPV Type Prediction. The HPV genome encodes a number of early (E1, E2, E4, E5, and E6) and late (L1 and L2) proteins [3, 5]. Several methods classified the high-risk and low-risk HPVs using the information from protein sequences, secondary structure, and pseudo amino acid composition [23–28]. But most of them used E6, E7, or L1 proteins. In this

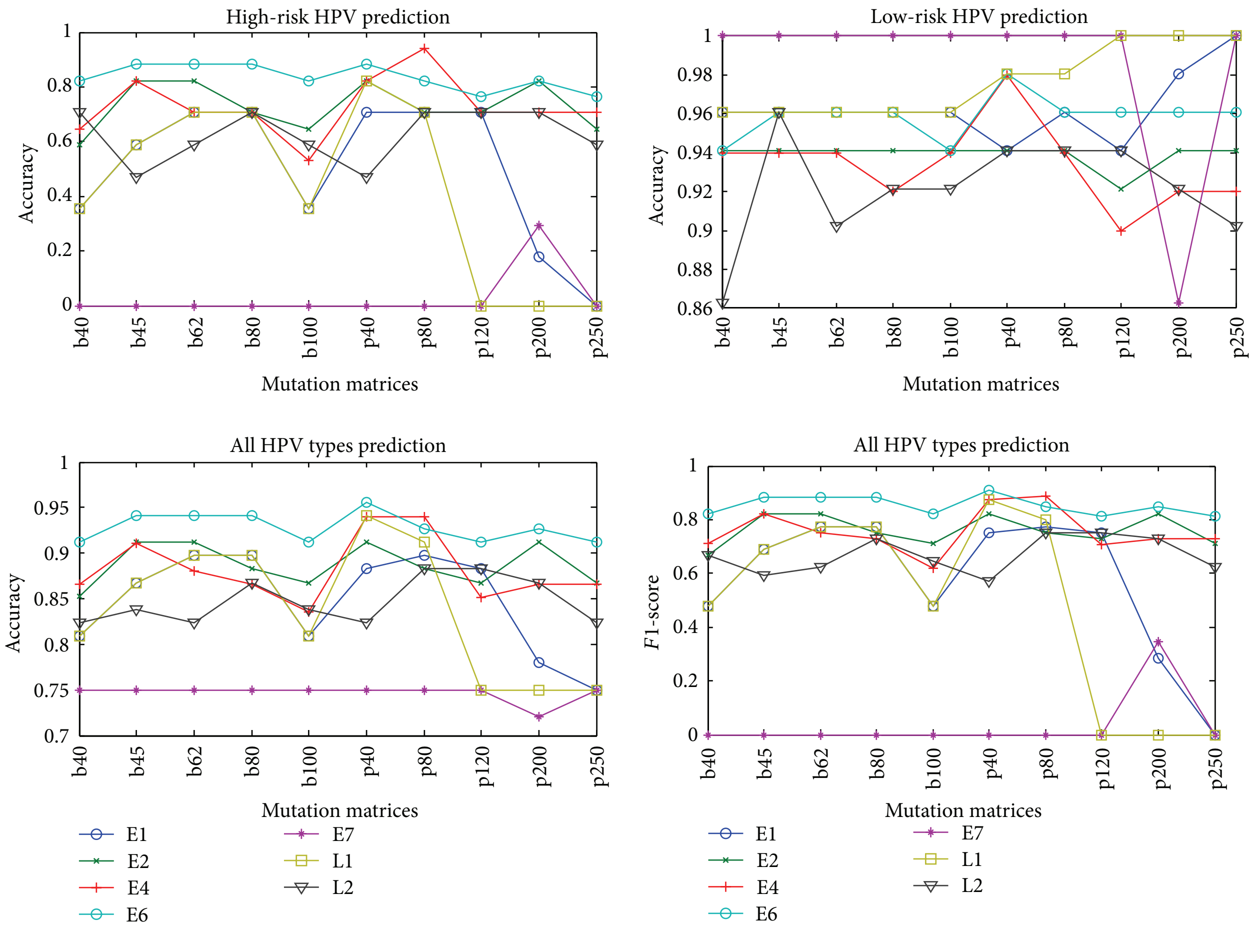

Figure 1: Comparison of prediction accuracy of each class, overall accuracy, and $F1$-score of all the early and late proteins. The mutation matrices in X-coordinate are BLOSUM 40, BLOSUM 45, BLOSUM 62, BLOSUM 80, BLOSUM 100, PAM 40, PAM 80, PAM 120, PAM 200, and PAM 250.

study, we constructed seven protein datasets of E1, E2, E4, E6, E7, L1, and L2 and compared their performance in HPV type prediction. The proteins of E5 were not included because their lengths are too small. The accuracy of each class, overall accuracy, and $F1$-score of all the early and late proteins were summarized in Figure 1.

From Figure 1, it is easy to observe that the accuracies of low-risk type are higher than that of high-risk type. For the low-risk type prediction experiment, E7 performs better than other HPV proteins expect for mutation matrix p200. But as for high-risk type prediction and all-type prediction experiments E6 achieves the best performance among all the HPV proteins according to the accuracies and $F1$-scores. Some experiment studies have shown that E5, E6, and E7 proteins of high-risk HPV play an important role in disease progression and cancer [14]. E5 protein enhances half-life and activity of epidermal growth factor receptor (EGFR). E6 and E7 proteins inactivate p53 and Rb functions [42]. The results also highlight that the sequences of E6 protein are more suitable for HPV high-risk type prediction and E7 protein is more reliable for HPV low-risk type prediction in the proposed model.

3.3. Comparison of Mutation Matrices in HPV Type Prediction.

The proposed word statistical model was constructed based on protein "sequence space" that relies heavily on the mutation matrix. In order to evaluate the influence of different mutation matrices, we adopted ten mutation matrices including PAM 40, PAM 80, PAM 120, PAM 200, PAM 250, BLOSUM 40, BLOSUM 45, BLOSUM 62, BLOSUM 80, and BLOSUM 100. The accuracy of each class, overall accuracy, and $F1$-score of the proposed prediction method based on ten mutation matrices were represented in Figure 1.

Figure 1 largely confirms that the proposed prediction method possesses different performances based on the different mutation matrices. The changes of high-risk type and all-type prediction experiments are similar, but there is a bit of difference in the low-risk type prediction experiments. As for the BLOSUM mutation matrices, BLOSUM 45 and BLOSUM 62 perform better in the prediction of high-risk type of HPVs. For PAM mutation matrices, PAM 40 and PAM 80 achieve the better performance in the high-risk type prediction experiments. Judging from prediction accuracy, it is easier to recognize that PAM 40 achieves the best performance based on E6 protein among PAM and BLOSUM

Table 2: Comparison of the real risk types (REAL) and the prediction results using the proposed approach.

Types	Real	Predicted	Types	Real	Predicted	Types	Real	Predicted	Types	Real	Predicted
HPV 39	High	High	HPV 7	Low	Low	HPV 34	Low	Low	HPV 50	Low	Low
HPV 72	**High**	**Low**	**HPV 30**	**Low**	**High**	HPV 44	Low	Low	HPV 5	Low	Low
HPV 33	High	High	HPV 73	Low	Low	HPV 43	Low	Low	HPV 20	Low	Low
HPV 51	High	High	HPV 6	Low	Low	HPV 32	Low	Low	HPV 23	Low	Low
HPV 16	High	High	HPV 27	Low	Low	HPV 24	Low	Low	HPV 19	Low	Low
HPV 56	High	High	HPV 13	Low	Low	HPV 8	Low	Low	HPV 47	Low	Low
HPV 18	High	High	HPV 55	Low	Low	HPV 48	Low	Low	HPV 22	Low	Low
HPV 59	High	High	HPV 2	Low	Low	HPV 12	Low	Low	HPV 25	Low	Low
HPV 52	High	High	HPV 10	Low	Low	HPV 49	Low	Low	HPV 9	Low	Low
HPV 35	High	High	HPV 42	Low	Low	HPV 15	Low	Low	HPV 36	Low	Low
HPV 68	High	High	HPV 28	Low	Low	HPV 21	Low	Low	HPV 41	Low	Low
HPV 58	High	High	HPV 40	Low	Low	HPV 4	Low	Low	HPV 63	Low	Low
HPV 31	High	High	HPV 3	Low	Low	HPV 65	Low	Low	HPV 1	Low	Low
HPV 66	**High**	**Low**	HPV 11	Low	Low	HPV 37	Low	Low	HPV 80	Low	Low
HPV 45	High	High	HPV 29	Low	Low	HPV 38	Low	Low	HPV 77	Low	Low
HPV 61	High	High	HPV 74	Low	Low	HPV 60	Low	Low	HPV 76	Low	Low
HPV 67	High	High	HPV 53	Low	Low	HPV 17	Low	Low	HPV 75	Low	Low

matrices except for PAM 80 with E4 protein. These results maybe give us some suggestion on how to choose a suitable mutation matrix for the prediction of high-risk type of HPVs based on the different protein sequences.

3.4. HPV Classification. In this study, we extracted the information using the word statistical model of E6 protein "sequence space" that was constructed based on PAM 40 mutation matrix. Leave-one-out cross-validation was applied to determine the prediction performance for all experimental results. HPV types were grouped into two classes, high-risk and low-risk. Table 2 shows the comparison of the manually tagged answer and the results from the proposed prediction approach.

Table 2 shows that the proposed prediction method achieves better performance, in which the prediction results of 65 HPV types are consistent with their real risk types. HPV 66 and HPV 72 are high-risk types, but they are predicted as low-risk type, and HPV 30 is low-risk type, but predicted as high-risk type using the proposed prediction method. In order to highlight the prediction differences, we further compared our results with Kim's results [13]. As for Kim's prediction, HPV 72 was predicted as possible high-risk type, but it was predicted as "low-risk type" in the proposed method; HPV 56 was predicted as possible high-risk type, while we predicted it as high-risk type; HPV 53 and HPV 73 were predicted as possible high-risk types, but they are low-risk types in our results. Phylogenetic analysis showed that HPV 30 was grouped closely with the established carcinogenic type HPV 56, which indicates that HPV 30 is more likely high-risk type. From the comparison, it is easy to note that the results obtained with the proposed method are more consistent with the real risk types.

To further evaluate the performance of the proposed prediction method, we computed the overall accuracy and $F1$-score and compared them with the published results in Figure 2. The methods evaluated here are as follows: SVM using Mismatch kernels (Mismatch) that have been reported in Joung et al. [24], SVM with Linear kernel method (Linear) [13], SVM classifier with the Gap-spectrum kernel (Gap) [7], BLAST predictions with a slight modification of the k-nearest neighbor method [13], Ensemble SVM (Ensemble) based on protein secondary structures [13], and two text-based prediction methods AdaCost [26] and nave Bayes [26].

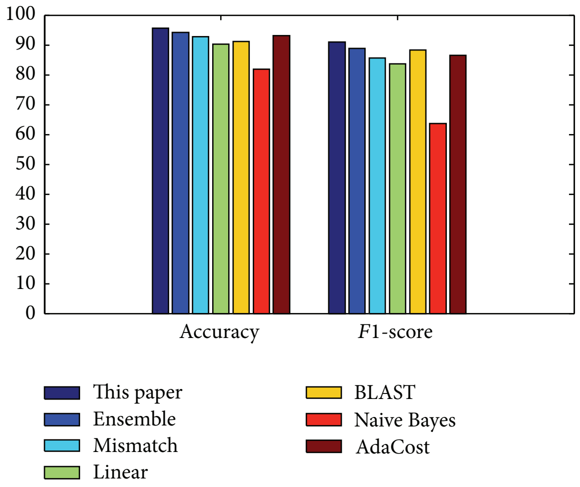

Figure 2: Comparison of overall accuracy and $F1$-score of all the evaluated prediction methods for HPV high-risk viral types.

The proposed approach achieved 95.59% accuracy and 90.91% $F1$-score, while the Ensemble SVMs obtained 94.12% accuracy and 88.89% $F1$-score, and SVM with Mismatch kernel achieved 92.70% accuracy and 85.70% $F1$-score, and SVM using Linear kernel with 90.28% accuracy and 83.72% $F1$-score and BLAST with 91.18% accuracy and 88.24% $F1$-score.

Table 3: Prediction results of the HPVs with unknown types using the proposed prediction methods and from the available methods.

Types	Prediction methods						
	Mismatch [24]	Linear [13]	Gap [7]	Genetic [23]	PseAAC [27]	Ensemble [13]	This paper
HPV 26	Low	Low	High	Low	High	High	High
HPV 54	Low	Low	Low	Low	Low	Low	Low
HPV 57	Low	Low	Low	Low	Low	Low	Low
HPV 70	High	High	High	Low	Low	High	High

As for text-based prediction method, AdaCost [26] achieved better performance with 93.05% accuracy and 84.490 % $F1$-score, and Naive Bayes [26] lags behind with 81.94% accuracy and 63.64% $F1$-score. According to the prediction accuracy and $F1$-score, the proposed prediction method achieved the best performance among all the evaluated prediction methods; the next best prediction approach is Ensemble SVMs, and the others lag behind. It is worth mentioning that the proposed approach is based on protein sequences as well as Mismatch, Linear, and Gap, while Ensemble uses the information from the predicted protein secondary structures. We also noted that text-based prediction does not provide superior results compared with other prediction methods. Although text-based prediction methods have an advantage in having explicit key words in the document, but they rely on the evidence obtained from the literature. When there is no available document for HPVs with unknown risk type, it is impossible to predict them. This comparison also indicates that the proposed word statistical model based on protein "sequence space" is more effective to classify risk types of human papillomaviruses.

3.5. Prediction for Unknown HPV Types. The most important task of this paper is to predict high-risk types of new HPV. Here, we downloaded the E6 protein sequences of HPVs with unknown types from the LANL database and used them to further evaluate the performance of the proposed approach. Table 3 shows the prediction results of all the HPVs with unknown types.

From Table 3, HPV 26 and HPV 70 were predicted as high-risk types and HPV 54 and HPV 57 as low-risk types using the proposed method. In order to compare with the existing methods, we also represented the prediction results of available approaches in Table 3. From the HPV classification, we knew that the proposed prediction method achieved the best performance, and it is followed by Ensemble SVMs. From Table 3, it is easy to note that the proposed method and Ensemble SVMs achieve the same results. As for the HPV 54 and HPV 57, all the methods predicted them as low-risk types. For HPV 26, the proposed method, PseAAC [27], Ensemble [13], and Gap [7] predicted it as high-risk type, while Mismatch [24], Linear [13], and Genetic [23] predicted it as low-risk type. According to the reliabilities of the prediction approaches, HPV 26 should be high-risk type. As for HPV 70, all the prediction methods predicted it as high-risk type except for Genetic [23] and PseAAC [27]. These results show that the proposed method can provide a simple but efficient guideline for the investigation of potentially high-risk HPVs.

4. Conclusion

Genital human papillomaviruses have a strong relationship with cervical cancer, especially high-risk viral types of HPVs. Therefore, discrimination of HPV risk type plays an important role in the diagnosis and remedy of cervical cancer. This paper proposed a computational scheme to predict high-risk types of HPVs with word statistical model of protein "sequence space." With help of mutation matrices, we first constructed a sequence space of the given protein sequences. Instead of only using sequence-based or structure-based information of protein sequences, we extracted the information of HPV from the protein "sequence space" with word statistical model to predict high-risk types of HPVs. The proposed method was tested on 68 samples with known HPV types and 4 samples with unknown HPV types. The results show that the proposed method achieved better performance in comparison to the previous methods.

The main goal of our research is to investigate a new prediction method based on protein "sequence space." The first contribution can be seen from comparison of early and late proteins' performances in HPV type prediction; we found that the "sequence space" of E6 protein is more suitable for HPV high-risk type prediction, while that of E7 protein is more reliable protein for HPV low-risk type prediction. The second contribution can be indicated from comparison of mutation matrices in HPV type prediction; we noticed that PAM 40 achieves the best performance with the sequences of E6 protein among PAM and BLOSUM matrices except for PAM 80 with E4 protein. The third contribution can be deduced from HPV classification and prediction for unknown HPV types; we found that the proposed prediction method achieved the best performance among all the evaluated prediction methods, with 95.59% accuracy and 90.91% $F1$-score, which can be contributed to the introduction of the protein "sequence space." Thus, this understanding can be used to guide development of more powerful method for prediction of high-risk types of HPVs.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors thank the referees for many valuable comments that have improved this paper. This work is supported by National Natural Science Foundation of China

(61370015, 61170316, and 61272312), research grants from Zhejiang Provincial Natural Science Foundation of China (LY14F020046), Medicine and Health Foundation of Zhejiang Province (2011-2011RCA012), and 521 Talent Cultivation Plan of Zhejiang Sci-Tech University.

References

[1] E.-K. Yim and J.-S. Park, "Role of proteomics in translational research in cervical cancer," *Expert Review of Proteomics*, vol. 3, no. 1, pp. 21–36, 2006.

[2] O. Peralta-Zaragoza, V. H. Bermúdez-Morales, C. Pérez-Plasencia, J. Salazar-León, C. Gómez-Cerón, and V. Madrid-Marina, "Targeted treatments for cervical cancer: a review," *OncoTargets and Therapy*, vol. 5, pp. 315–328, 2012.

[3] A. Jemal, F. Bray, M. M. Center, J. Ferlay, E. Ward, and D. Forman, "Global cancer statistics," *CA: A Cancer Journal for Clinicians*, vol. 61, no. 2, pp. 69–90, 2011.

[4] D. Formana, C. de Martel, C. J. Lacey et al., "Global burden of human papillomavirus and related diseases," *Vaccine*, vol. 30, supplement 5, pp. F12–F23, 2012.

[5] F. X. Bosch, M. M. Manos, N. Munoz et al., "Prevalence of human papillomavirus in cervical cancer: a worldwide perspective," *Journal of the National Cancer Institute*, vol. 87, no. 11, pp. 796–802, 1995.

[6] M. H. Schiffman, H. M. Bauer, R. N. Hoover et al., "Epidemiologic evidence showing that human papillomavirus infection causes most cervical intraepithelial neoplasia," *Journal of the National Cancer Institute*, vol. 85, no. 12, pp. 958–964, 1993.

[7] S. Kim and J.-H. Eom, "Prediction of the human papillomavirus risk types using gap-spectrum kernels," in *Advances in Neural Networks—ISNN 2006*, vol. 3973 of *Lecture Notes in Computer Science*, pp. 710–715, Springer, Berlin, Germany, 2006.

[8] C. L. Pang and F. Thierry, "Human papillomavirus proteins as prospective therapeutic targets," *Microbial Pathogenesis*, vol. 58, pp. 55–65, 2013.

[9] S. Kim and B.-T. Zhang, "Human papillomavirus risk type classification from protein sequences using support vector machines," in *Applications of Evolutionary Computing*, vol. 3907 of *Lecture Notes in Computer Science*, pp. 57–66, Springer, Berlin, Germany, 2006.

[10] J. Haedicke and T. Iftner, "Human papillomaviruses and cancer," *Radiotherapy & Oncology*, vol. 108, no. 3, pp. 397–402, 2013.

[11] J. Peng, L. Gao, J. Guo et al., "Type-specific detection of 30 oncogenic human papillomaviruses by genotyping both E6 and L1 genes," *Journal of Clinical Microbiology*, vol. 51, no. 2, pp. 402–408, 2013.

[12] M. S. Longworth and L. A. Laimins, "Pathogenesis of human papillomaviruses in differentiating epithelia," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 362–372, 2004.

[13] S. Kim, J. Kim, and B.-T. Zhang, "Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures," *Computers in Biology and Medicine*, vol. 39, no. 2, pp. 187–193, 2009.

[14] E.-M. de Villiers, C. Fauquet, T. R. Broker, H.-U. Bernard, and H. Zur Hausen, "Classification of papillomaviruses," *Virology*, vol. 324, no. 1, pp. 17–27, 2004.

[15] K. Münger, A. Baldwin, K. M. Edwards et al., "Mechanisms of human papillomavirus-induced oncogenesis," *Journal of Virology*, vol. 78, no. 21, pp. 11451–11460, 2004.

[16] M. L. Eide and H. Debaque, "HPV detection methods and genotyping techniques in screening for cervical cancer," *Annales de Pathologie*, vol. 32, no. 6, pp. e15–e23, 2012.

[17] M. F. Janicek and H. E. Averette, "Cervical cancer: prevention, diagnosis, and therapeutics," *CA: A Cancer Journal for Clinicians*, vol. 51, no. 2, pp. 92–114, 2001.

[18] M. D. Kaspersen, P. B. Larsen, H. J. Ingerslev et al., "Identification of multiple HPV types on Spermatozoa from human sperm donors," *PLoS ONE*, vol. 6, no. 3, Article ID e18095, 2011.

[19] P. Guan, R. Howell-Jones, N. Li et al., "Human papillomavirus types in 115,789 HPV-positive women: a meta-analysis from cervical infection to cancer," *International Journal of Cancer*, vol. 131, no. 10, pp. 2349–2359, 2012.

[20] H. Furumoto and M. Irahara, "Human papilloma virus (HPV) and cervical cancer," *Journal of Medical Investigation*, vol. 49, no. 3-4, pp. 124–133, 2002.

[21] R. D. Burk, G. Y. F. Ho, L. Beardsley, M. Lempa, M. Peters, and R. Bierman, "Sexual behavior and partner characteristics are the predominant risk factors for genital human papillomavirus infection in young women," *The Journal of Infectious Diseases*, vol. 174, no. 4, pp. 679–689, 1996.

[22] N. Muñoz, F. X. Bosch, S. De Sanjosé et al., "Epidemiologic classification of human papillomavirus types associated with cervical cancer," *The New England Journal of Medicine*, vol. 348, no. 6, pp. 518–527, 2003.

[23] J. H. Eom, S. B. Park, and B. T. Zhang, "Genetic mining of DNA sequence structures for effective classification of the risk types of human papillomavirus (HPV)," in *Neural Information Processing: Proceedings of the 11th International Conference, ICONIP 2004, Calcutta, India, November 22–25, 2004*, vol. 3316 of *Lecture Notes in Computer Science*, pp. 1334–1343, Springer, Berlin, Germany, 2004.

[24] J.-G. Joung, O. Sok June, and B.-T. Zhang, "Prediction of the risk types of human papillomaviruses by Support Vector Machines," in *Proceedings of the 8th Pacific Rim International Conference on Artificial Intelligence (PRICAI '04)*, pp. 723–731, August 2004.

[25] J.-G. Joung, O. S. June, and B.-T. Zhang, "Protein sequence-based risk classification for human papillomaviruses," *Computers in Biology and Medicine*, vol. 36, no. 6, pp. 656–667, 2006.

[26] S.-B. Park, S. Hwang, and B.-T. Zhang, "Mining the risk types of human papillomavirus (HPV) by AdaCost," in *Database and Expert Systems Applications: 14th International Conference, DEXA 2003, Prague, Czech Republic, September 1–5, 2003. Proceedings*, vol. 2736 of *Lecture Notes in Computer Science*, pp. 403–412, Springer, Berlin, Germany, 2003.

[27] M. Esmaeili, H. Mohabatkar, and S. Mohsenzadeh, "Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses," *Journal of Theoretical Biology*, vol. 263, no. 2, pp. 203–209, 2010.

[28] M. Alemi, H. Mohabatkar, and M. Behbahani, "In silico comparison of low- and high-risk human papillomavirus proteins," *Applied Biochemistry and Biotechnology*, vol. 172, no. 1, pp. 188–195, 2014.

[29] B. Liao, W. Chen, X. Sun, and W. Zhu, "A binary coding method of RNA secondary structure and its application," *Journal of Computational Chemistry*, vol. 30, no. 14, pp. 2205–2212, 2009.

[30] B. Liao and T.-M. Wang, "Analysis of similarity/dissimilarity of DNA sequences based on nonoverlapping triplets of nucleotide bases," *Journal of Chemical Information and Computer Sciences*, vol. 44, no. 5, pp. 1666–1670, 2004.

[31] Y. Zhang, J. K. Hao, C. G. Zhou, and K. Chang, "Normalized Lempel-Ziv complexity and its application in bio-sequence

analysis," *Journal of Mathematical Chemistry*, vol. 46, no. 4, pp. 1203–1212, 2009.

[32] Y. Zhang and W. Chen, "Invariants of DNA sequences based on 2DD-curves," *Journal of Theoretical Biology*, vol. 242, no. 2, pp. 382–388, 2006.

[33] B. E. Blaisdell, "A measure of the similarity of sets of sequences not requiring sequence alignment," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 14, pp. 5155–5159, 1986.

[34] T.-J. Wu, J. P. Burke, and D. B. Davison, "A measure of DNA sequence dissimilarity based on Mahalanobis distance between frequencies of words," *Biometrics*, vol. 53, no. 4, pp. 1431–1439, 1997.

[35] T.-J. Wu, Y.-C. Hsieh, and L.-A. Li, "Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition," *Biometrics*, vol. 57, no. 2, pp. 441–448, 2001.

[36] G. W. Stuart, K. Moffett, and S. Baker, "Integrated gene and species phylogenies from unaligned whole genome protein sequences," *Bioinformatics*, vol. 18, no. 1, pp. 100–108, 2002.

[37] T. D. Pham and J. Zuegg, "A probabilistic measure for alignment-free sequence comparison," *Bioinformatics*, vol. 20, no. 18, pp. 3455–3461, 2004.

[38] X. Wu, X.-F. Wan, G. Wu, D. Xu, and G. Lin, "Phylogenetic analysis using complete signature information of whole genomes and clustered Neighbour-Joining method," *International Journal of Bioinformatics Research and Applications*, vol. 2, no. 3, pp. 219–248, 2006.

[39] A. Apostolico and O. Denas, "Fast algorithms for computing sequence distances by exhaustive substring composition," *Algorithms for Molecular Biology*, vol. 3, no. 1, article 13, 2008.

[40] Q. Dai, Y. C. Yang, and T. M. Wang, "Markov model plus k-word distributions: a synergy that produces novel statistical measures for sequence comparison," *Bioinformatics*, vol. 24, no. 20, pp. 2296–2302, 2008.

[41] G. Lu, S. Zhang, and X. Fang, "An improved string composition method for sequence comparison," *BMC Bioinformatics*, vol. 9, no. 6, article S15, 2008.

[42] J. R. Gage, C. Meyers, and F. O. Wettstein, "The E7 proteins of the nononcogenic human papillomavirus type 6b (HPV-6b) and of the oncogenic HPV-16 differ in retinoblastoma protein binding and other properties," *Journal of Virology*, vol. 64, no. 2, pp. 723–730, 1990.