Computers in Biology and Medicine 39 (2009) 187 – 193

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/cbm

Ensembled support vector machines for human papillomavirus risk type prediction from protein secondary structures

Sun Kim[a], Jeongmi Kim[b], Byoung-Tak Zhang[a,*]

[a]School of Computer Science and Engineering, Seoul National University, Seoul 151-744, Republic of Korea
[b]ISU ABXIS CO., LTD, Seoul 120-752, Republic of Korea

ARTICLE INFO

Article history:
Received 30 April 2007
Accepted 8 December 2008

Keywords:
Human papillomavirus risk type prediction
Ensemble methods
Support vector machines
Protein structures

ABSTRACT

Infection by the human papillomavirus (HPV) is regarded as the major risk factor in the development of cervical cancer. Detection of high-risk HPV is important for understanding its oncogenic mechanisms and for developing novel clinical tools for its diagnosis, treatment, and prevention. Several methods are available to predict the risk types for HPV protein sequences. Nevertheless, no tools can achieve a universally good performance for all domains, including HPV and nor do they provide confidence levels for their decisions. Here, we describe ensembled support vector machines (SVMs) to classify HPV risk types, which assign given proteins into high-, possibly high-, or low-risk type based on their confidence level. Our approach uses protein secondary structures to obtain the differential contribution of subsequences for the risk type, and SVM classifiers are combined with a simple but efficient string kernel to handle HPV protein sequences. In the experiments, we compare our approach with previous methods in accuracy and F1-score, and present the predictions for unknown HPV types, which provides promising results.

1. Introduction

Cervical cancer is one of the leading causes of cancer morbidity and mortality in women worldwide [1]. Epidemiologic studies have shown that the association of genital human papillomavirus (HPV) with cervical cancer is strong, independent of other risk factors, and that this is consistent in several countries [2].

The HPV is a relatively small, double-strand DNA tumor virus that belongs to the papovavirus family (papilloma, polyoma, and simian vacuolating viruses). More than 100 human types are specific for epithelial cells, including skin, respiratory mucosa, and the genital tract.

Genital tract HPV types are classified into two or three types by their relative malignant potential as low-, intermediate-, and high-risk types [3]. The common, unifying oncogenic feature of the vast majority of cervical cancers is the presence of HPV, especially high-risk type HPV [4]. Thus, the detection of high-risk HPVs has become one of the most essential strategies for cervical cancer treatment. Since HPV classification is important for medical judgments, there have been many epidemiological and experimental studies to classify the HPV risk types [2,4,5]. These 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.

There have been a few computational studies for HPV risk type prediction. These are all based on machine learning (ML) techniques, but use different approaches. For sequence-based methods, DNA or proteins can be used to discriminate between the risk types. Eom et al. [6] presented a sequence comparison method using HPV DNA. They formulate certain informative subsequences into genetic encoding of evolutionary algorithms. However, they have not separated training and test data clearly. A combined ML technique also has been proposed to predict HPV risk types from protein sequences [7,8]. HPV proteins are first aligned, and the subsequences in high-risk HPVs against low-risk HPVs are selected by hidden Markov models. A support vector machine (SVM) is then used to determine the risk type from the selected subsequences. The main drawback of this study is that the method is biased by one specific sequence pattern. Alternatively, biomedical literature can be used to predict HPV risk types. A boosting method, Adacost based on naïve Bayes classifiers, has been proposed in Park et al. [9], whereas text-mining approaches are limited for the prediction capability because they depend solely on text material to capture the discriminating evidence, and the obvious keywords such as 'high' tend to appear explicitly in HPV-related materials.

The most important aspect of cervical cancer diagnosis is to determine which HPVs are high risk. The HPV risk types are still manually classified by experts, and there is no deterministic method for

* Corresponding author. Tel.: +82 2 880 1847; fax: +82 2 875 2240.
E-mail addresses: skim@bi.snu.ac.kr (S. Kim), jkjeong@isu.co.kr (J. Kim), btzhang@bi.snu.ac.kr (B.-T. Zhang).

0010-4825/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2008.12.005

predicting the risk type of unknown or new HPVs. ML techniques are useful for discovering hidden patterns from given data, and also provide robust results for unknown patterns. Therefore, ML-based approaches can be effective in solving the HPV risk prediction problem. However, in previous ML studies, they only provide binary decisions without any confidence level for the high-risk type.

In this paper, we show that by applying an SVM ensemble, HPV risk type prediction can be improved decisively compared to other methods. We present an ML approach utilizing protein secondary structures to discriminate HPV risk types. A neural network-based method, Jnet [10] first predicts the protein secondary structure, then ensembled SVMs with string kernels are used to determine the risk types from the extracted subsequences. The string kernel maps given input sequences to the space consisting of all subsequences of amino acid pairs. In particular, the string kernel only uses amino acids of both ends in k-length subsequences to capture a specific structural effect, which is motivated by the assumption that amino acid pairs with certain distances affect the HPV biological function, i.e., risk type, in different ways. The risk type is determined by a voting scheme from the outputs of all SVM classifiers. It gives the confidence level by providing three classes, high-, possible high-, and low-risk type.

2. Materials and methods

2.1. Data

In this paper, we use the HPV database from the Los Alamos National Laboratory (LANL), and a total of 72 types of HPV are used for experiments. The HPV risk types were manually determined based on the HPV compendium (1997). If an HPV belongs to a skin-related or cutaneous group, the HPV is classified as a low-risk type. On the other hand, an HPV is classified as a high-risk type if it is known to be a high-risk type for cervical cancer. The comments in the LANL database are used to decide the risk types for some HPVs that are difficult to classify. Seventeen sequences out of 72 HPVs were classified as high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 61, 66, 67, 68, and 72), and others were classified as low-risk types. Four HPVs (26, 54, 57, and 70) that could not be determined remain as unknown risk types. Fig. 1 shows an example of the HPV type 16, which is one of high-risk HPVs.

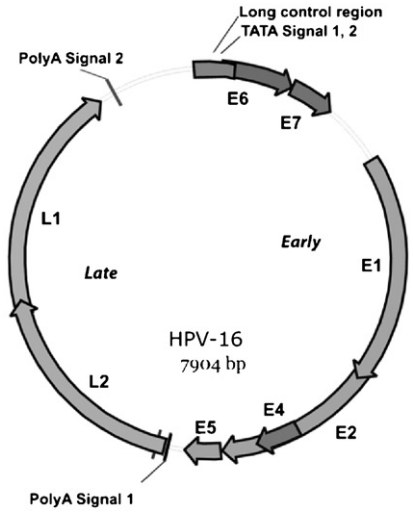

Fig. 1. The structure of HPV 16 genome. The E6 protein in high-risk HPVs can bind to and inactivate the tumor suppressor gene products.

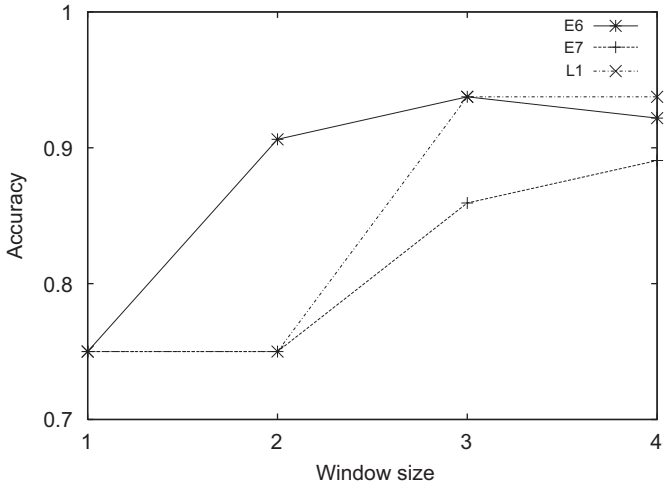

Fig. 2. SVM classification performance for E6, E7, and L1 gene products. Classification accuracies are measured according to window size, and it shows that E6 proteins are the most informative to discriminate HPV risk types.

2.2. Gene product selection for HPV risk type classification

High-risk HPV types can be distinguished from other HPV types based on the structure and function of the E6, E7, and L1 gene products. The late-region gene L1 that encodes a major capsid protein appears to be the most polymorphic, and sequence polymorphism exists in the early genes E6 and E7 [11]. These sequence differences endow individual HPV types with varying degrees of oncogenic transformation potential, and they can be exploited for developing type-specific molecular tests.

Here, we empirically evaluate E6, E7, and L1 gene products to obtain the most informative protein with high classification performance. We have measured the classification accuracy using a string kernel-based SVM classifier. The SVM classifier is same as the one used in our ensemble approach. The only difference is that the k-spectrum kernel is used instead of the gap-spectrum kernel. Hence, this empirical analysis can be compatible with the ensembled SVMs. The parameter k in the k-spectrum kernel is the number of consecutive amino acids, i.e., window size. The SVM classifier, the k-spectrum kernel, and the gap-spectrum kernel are explained in later sections.

Fig. 2 depicts the accuracy changes by the window size. The results are obtained by leave-one-out cross-validation, which is common for evaluating small data problems. The figure indicates that the accuracy using the E6 protein is mostly higher than the one using E7 and L1 proteins. We also find that using E7 protein is a less effective way to discriminate HPV risk types, based on sequence similarities. However, the overall accuracy gets high scores by increasing window size for all proteins because the HPV sequences are relatively short, and unique patterns are more frequently generated when window size becomes larger. That is, the learners overfit protein sequences for larger windows. Viral early proteins E6 and E7 are known for inducing immortalization and transformation in rodent and human cell types. E6 proteins produced by the high-risk HPV types can bind to and inactivate the tumor suppressor protein, thus facilitating tumor progression [12,13]. This process plays an important role in the development of cervical cancer. For these reasons and the empirical results, we have chosen E6 proteins corresponding to the 72 HPVs.

2.3. Protein secondary structure prediction

In our approach, we divide HPV sequences according to the secondary structural elements because it helps to analyze the functional

roles of proteins. Note that the function of a protein depends on its structure. Even though the structure is mostly determined by amino acid sequences, secondary structure is known to facilitate the prediction of protein functions [14,15].

The secondary structure of a protein defines the general three-dimensional form of local regions and may include regions of α helices, β sheets, or segments of the chain that assume no stable shape. Being able to predict accurately secondary structural elements along the sequence provides a good starting point toward elucidating three-dimensional structure. There have been many prediction methods for protein secondary structure [16–19], and neural network-based systems generally show good performance.

The tool, Jnet, is a consensus method using neural network ensembles [10]. We assume that Jnet can give more precise results, but in conservative ways because it is based on a consensus of the results obtained from different aspects. The prediction by Jnet is the definition of each residue into either α helix, β sheet, or other types of structure. In Jnet, a network ensemble consists of two artificial neural networks. The process of assessing the prediction methods results in different neural network ensembles that are trained with different alignment data. Each of the networks is combined and an average is taken for each predicted state.

Our method takes only two types, α helices and other types, to make input subsequences for ensembled SVMs since the β sheets predicted by the consensus method are too short to be used.

2.4. SVM ensemble approach

Using the protein subsequences classified by their secondary structures, PPI risk type classification is performed by an SVM ensemble method. Our ensemble approach is motivated by the intuitive idea that the risk type prediction might be improved by combining the outputs of individual classifiers. The theoretical framework of ensemble averaging is related to bias and variance in statistics [20]. The ensemble machine has the advantages of avoiding low bias and high variance, i.e., overfitting, and reducing the generalization error rather than using a single classifier.

For classifying HPV proteins, our approach uses ensembled SVMs, which is the combination of string kernel-based SVMs. The individual SVMs are trained on HPV protein subsequences of either α helices or other types, and tested on unknown subsequences. A simple string kernel the so-called gap-spectrum kernel is introduced to apply proteins to SVMs directly. The string kernel will only consider amino acid pairs with a fixed gap k, thus we can obtain different information according to k. The overall process has the following three steps (Fig. 3):

(1) Secondary structure prediction by Jnet from HPV protein sequences.
(2) SVM learning with a string kernel parameter k for each structural element.
(3) HPV risk type prediction through the SVM ensemble.

In this paper, we have taken two structural categories, α helices and other types as previously described, and set the kernel parameter k into 2, 3, and 4, respectively, for each structure. Therefore, the SVM ensemble consists of six SVM classifiers. HPV risk types are finally determined to high-, possible high-, or low-risk by the voting scheme based on the outputs from the SVM classifiers.

2.5. Gap-spectrum kernels

Here, we introduce a simple string kernel based on the spectrum kernel method [21,22], which has been used to detect remote homology detection. The input space $\mathscr{X}$ consists of all finite length sequences of characters from an alphabet $\mathscr{A}$ of size $|\mathscr{A}|=l$ ($l=20$ for amino acids). Given a number $k \geqslant 1$, the k-spectrum of a protein sequence is the set of all possible k-length subsequences (k-mers) that it contains. The feature map is indexed by all possible subsequences a of length k from $\mathscr{A}$. The k-spectrum feature map $\Phi_k(x)$ from $\mathscr{X}$ to $\mathbb{R}^{l^k}$ can be defined as

$$\Phi_k(x) = (\phi_a(x))_{a \in \mathscr{A}^k}, \tag{1}$$

where $\phi_a(x)$ is the number of occurrences of a that occur in x. Thus the k-spectrum kernel function $K^s(x_i, x_j)$ for two sequences x_i and x_j is obtained by taking the inner product in feature space:

$$K_k^s(x_i, x_j) = \langle \Phi_k(x_i), \Phi_k(x_j) \rangle. \tag{2}$$

HPV proteins are relatively short, so that evolutionary variation such as insertion and deletion cannot be sensitive to the predictions comparing long proteins. For instance, E6 gene products from HPVs have less than 170 bp. We assume that an amino acid pair with a certain distance affects the HPV risk type in a specific way according to its three-dimensional structural properties, and the HPV risk types can be identified by considering the amino acid pairs which mostly influence the risk type decision.

Under the above assumption, we define the gap-spectrum kernel [23] based on the k-spectrum. The gap-spectrum kernel handles neighbor or distant amino acid pairs by adjusting the parameter k. For a fixed k-mer $a = a_1 a_2 \ldots a_k$, $a_i \in \mathscr{A}$, 2-length sequence $\beta = a_1 a_k$, $\beta \in \mathscr{A}^2$. β indicates the amino acid pair with a $(k-2)$ gap. The feature map $\Psi_k(x)$ is defined as

$$\Psi_k(x) = (\phi_\beta(x))_{\beta \in \mathscr{A}^2}, \tag{3}$$

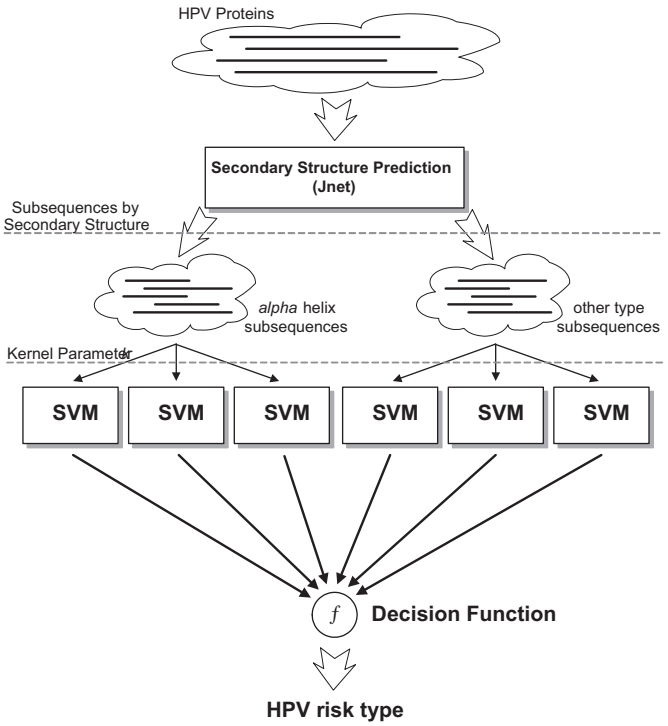

Fig. 3. Overview of the SVM ensemble approach for HPV risk type classification. Protein sequences are divided into two subsequences, α helices and other types by Jnet. Then, the subsequences are entered into the SVM classifiers. Each SVM classifier differently handles the subsequences by the kernel parameter k. k indicates a fixed distance between amino acid pairs to be considered for classification. The outputs from the SVM classifiers are gathered, and calculated to produce a HPV risk type.

where $\phi_\beta(x)$ is the number of occurrences of β that occur in x. Furthermore, for a nonlinear kernel function, the RBF kernel is appended to increase the discriminative ability between HPV risk types. By closure properties of kernels [24], the gap-spectrum kernel is defined as follows:

$$K_k(x_i, x_j) = K'(\Psi_k(x_i), \Psi_k(x_j)) \tag{4}$$

$$= \exp(-\gamma \| \Psi_k(x_i) - \Psi_k(x_j) \|^2), \tag{5}$$

where $\gamma > 0$. The gap-spectrum kernel may have some problems in accepting some variants since proteins are handled in the same manner over whole sequences. However, it is compensated by using the ensembled SVMs in which various gaps are simultaneously handled.

2.6. SVM classifiers

SVMs have been developed by Vapnik to give robust performance for classification and regression problems [25]. Kernel functions introduce nonlinear features in the hypothesis space without explicitly requiring nonlinear algorithms. SVMs learn a linear decision boundary in the feature space mapped by the kernel in order to separate the data into two classes.

For a feature mapping ϕ, the training data $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$ are mapped into the feature space $\Phi(S) = \{\Phi(\mathbf{x}_i), y_i\}_{i=1}^n$. In the feature space, SVMs learn the hyperplane $f = \langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b$, $\mathbf{w} \in \mathbb{R}^N$, $b \in R$, and the decision is made by $\mathrm{sgn}(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b)$. The decision boundary is the hyperplane $f = 0$ and its margin is $1/\|\mathbf{w}\|$. SVMs find a hyperplane that has the maximal margin from each class among normalized hyperplanes.

Finding the optimal hyperplane for nonseparable data can be formulated as the following optimization problem:

$$\text{maximize} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \tag{6}$$

$$\text{subject to} \quad \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leqslant \alpha_i \leqslant C, \quad i = 1, \ldots, n. \tag{7}$$

By solving this problem, one obtains an optimal solution, Lagrange multiplier α_i, $1 \leqslant i \leqslant n$, which needs to determine the parameters, $\mathbf{w}$ and b. For the solution $\alpha_i, \ldots, \alpha_n$, the decision function $f(\mathbf{x})$ is expressed in terms of the following parameters:

$$f(\mathbf{x}) = \mathrm{sgn}\left(\sum_{i=1}^n \alpha_i y_i \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i) \rangle + b \right). \tag{8}$$

We can work on a high-dimensional feature space by using kernel functions, and any kernel function K satisfying Mercer's condition can be used.

In this paper, we use the gap-spectrum kernel for the kernel function K. The ensembled SVMs were implemented by using LIBSVM [26], a freely available software package based on the sequential minimal optimization (SMO) method.

2.7. HPV risk type decision

After individual SVM classifiers are trained for the assigned sequences, a final HPV type is given from the SVM outputs by a voting scheme, i.e., an SVM ensemble calculates the weighted sum $F = \sum_{i=1}^m \omega_i f_i(\mathbf{x})$, where $\omega_i > 0$ and f_i is the output of an SVM classifier. For $\varepsilon_1, \varepsilon_2 > 0$, the HPV risk type is then predicted as follows:

$$Output = \begin{cases} \text{High} & \text{if } F > \varepsilon_1, \\ \text{Possible high} & \text{if } \varepsilon_2 < F \leqslant \varepsilon_1, \\ \text{Low} & \text{otherwise.} \end{cases} \tag{9}$$

For experiments, we have simply set the decision weight ω to uniform values, i.e., $\omega_i = 1$, and the decision boundaries have been set to $\varepsilon_1 = 4$, $\varepsilon_2 = 2$.

3. Results and discussion

3.1. Evaluation measure

In the HPV risk type classification, it is important to detect as many high-risk HPVs as possible, although a few low-risk HPVs are misclassified, hence we use both accuracy and F1-score as performance evaluation measure.

When binary scales are used for both answer and prediction, a contingency table is established showing how the dataset is divided by these two measures (Table1). Using the table, the accuracy and F1-score are calculated as follows:

$$accuracy = \frac{a+d}{a+b+c+d} \cdot 100\%,$$

$$F1\text{-}score = \frac{2 \cdot PPV \cdot specificity}{PPV + specificity} \cdot 100\%,$$

where $PPV = a/(a+b)$ and $specificity = a/(a+c)$. PPV is the abbreviation for positive predictive value.

3.2. HPV classification

Table 2 shows the comparison of the manually tagged answer and the results from our approach using the SVM ensemble. Leave-one-out cross-validation was applied to determine the prediction performance for all experimental results. We classify HPV types into three classes, high-, possible high-, and low-risk as presented in Eq. (9). In other words, this gives three confidence levels for the high-risk type. For instance, 'possible high' means that a HPV is predicted as a high-risk type, but with low confidence.

The prediction performance in accuracy and F1-score is given in Table 3. The performance results using mismatch kernels have been reported in Joung et al. [7]. The linear kernel method is the same as the SVM classifier with $k = 1$ in the gap-spectrum kernel. The BLAST predictions are obtained from taking the majority class between the most similar five HPVs, a slight modification of the k-nearest neighbor method. The ensembled SVMs reach 94.12% accuracy and 88.89% F1-score, while the mismatch kernel shows 92.70% accuracy and 85.70% F1-score. The linear kernel shows 90.28% accuracy and 83.72% F1-score. For BLAST, 91.18% accuracy and 88.24% F1-score are given. They all show good performance, but the ensembled SVMs get better results than other sequence-based methods, on average a 3.5% improvement in the dataset. In particular, BLAST, mismatch kernels, and linear kernels can produce incorrect results with noisy data, whereas the ensembled SVMs can be more consistent because it combines the results of individual classifiers.

The AdaCost [27] and naïve Bayes [28] in Table 3 are the performance results using biomedical literature, which are reported in Park et al. [9]. Even though text-based classification has an advantage in having explicit keywords in the documents, this does not provide superior results compared with sequence-based methods. In particular, text-mining approaches only depend on the evidence obtained

Table 1
The contingency table to evaluate the classification performance.

		HPV risk type	
		High	Low
Prediction Result	High	a	b
	Low	c	d

Table 2
Comparison of the manually classified risk types (Man) and the prediction results using the proposed approach (Out).

Type	Man	Out	Type	Man	Out	Type	Man	Out
HPV1	Low	Low	HPV25	Low	Low	HPV50	Low	Low
HPV2	Low	Low	HPV27	Low	Low	HPV51	High	High
HPV3	Low	Low	HPV28	Low	Low	HPV52	High	High
HPV4	Low	Low	HPV29	Low	Low	HPV53	Low	PH
HPV5	Low	Low	HPV30	Low	High	HPV55	Low	Low
HPV6	Low	Low	HPV31	High	High	HPV56	High	PH
HPV7	Low	Low	HPV32	Low	Low	HPV58	High	High
HPV8	Low	Low	HPV33	High	High	HPV59	High	High
HPV9	Low	Low	HPV34	Low	Low	HPV60	Low	Low
HPV10	Low	Low	HPV35	High	High	HPV61	High	PH
HPV11	Low	Low	HPV36	Low	Low	HPV63	Low	Low
HPV12	Low	Low	HPV37	Low	Low	HPV65	Low	Low
HPV13	Low	Low	HPV38	Low	Low	HPV66	High	Low
HPV15	Low	Low	HPV39	High	High	HPV67	High	High
HPV16	High	High	HPV40	Low	Low	HPV68	High	High
HPV17	Low	Low	HPV41	Low	Low	HPV72	High	PH
HPV18	High	High	HPV42	Low	Low	HPV73	Low	PH
HPV19	Low	Low	HPV43	Low	Low	HPV74	Low	Low
HPV20	Low	Low	HPV44	Low	Low	HPV75	Low	Low
HPV21	Low	Low	HPV45	High	High	HPV76	Low	Low
HPV22	Low	Low	HPV47	Low	Low	HPV77	Low	Low
HPV23	Low	Low	HPV48	Low	Low	HPV80	Low	Low
HPV24	Low	Low	HPV49	Low	Low			

'PH' means 'possible high'.

Table 3
The performance comparison of the proposed approach and other approaches based on manually tagged answers in Table 2.

Method	Protein sequence-based classification				Text-based classification	
	Ensemble	Mismatch	Linear	BLAST	AdaCost	Naïve Bayes
Accuracy	94.12	92.70	90.28	91.18	93.05	81.94
F1-score	88.89	85.70	83.72	88.24	86.49	63.64

'Ensemble', 'Mismatch', and 'Linear' mean ensembled SVMs, mismatch kernels, and linear kernels, respectively.

Table 4
Comparison of the manually classified risk types (Manual) and the result from Muñoz et al. (Muñoz) for the HPVs predicted as 'possible high' in our approach.

Type	Manual	Muñoz
HPV53	Low	Possible high
HPV56	High	High
HPV61	High	Low
HPV72	High	Low
HPV73	Low	High

from the literature. If the documents are unavailable for unknown HPVs, it is not possible to classify them.

While manually tagged answers are only based on the literature in the LANL database, we can refer to recent research for additional information on HPV risk types. Muñoz et al. [5] suggest a reliable risk type classification based on epidemiologic studies of over 1900 patients. In most cases, the HPV risk types follow the same decision between both manually tagged answers and the results in Muñoz et al. However, there exist a few differences in HPV risk types. Table 4 shows the comparison of the manually tagged answers from the LANL database and the decisions from Muñoz et al. for the HPVs predicted as 'possible high' by our method. Four HPV types (53, 61, 72, and 73) except HPV56 are predicted with opposing risk by these two sources. This means that the four HPVs have low confidence for the high-risk type, and the SVM ensemble exactly predicts the risk types as 'possible high'.

Fig. 4 depicts the performance comparison of the SVM ensemble and the mismatch kernel on manually tagged answers and modified answers. The modified answers are the corrected version of the manually tagged answers in Table 2, by adding new high-risk types (Table 4) reported in Muñoz et al. [5]. The figure shows that our approach using protein secondary structure and amino acid pair extraction outperforms the previous SVM method, which does not utilize the protein structural information. As a result, we can conclude that it is critical to use protein structures to get more accurate HPV risk type prediction.

It would be interesting to analyze the outputs from the SVM classifiers so that we can infer which factor has a certain function that predicts the high-risk type. Table 5 presents all SVM predictions for the HPVs classified as high-risk types. According to the table, only α helices have the important function in determining correct decisions even though α helices and other types are all used to produce final decisions in the experiments. In particular, the risk types can be obtained with good performance using SVM classifiers with $k=2$ and 4 in the regions of α helices. $k=2$ means two consecutive amino acids and $k=4$ means an amino acid pair with two gaps. Because an α helix is spiral shaped, there is more chance of interaction between amino acids with the same interval. In addition, HPV E6 proteins are relatively short, so that any variations such as evolutionary or amino acid order changes do not create serious pitfalls in the proposed approach. Here, the i th amino acids may interact with $(i+3)$th amino acids in the α helices of E6 proteins, which significantly affects the decision of the high-risk type. In conclusion, it might that the regions of α helices in E6 are important in predicting whether HPVs are high risk or not regardless of the imperfect secondary structure prediction.

E6 protein produced by the high-risk HPV types binds to host p53 causing inactivation of its function through a mechanism of ubiquitin-dependent degradation. However, the structure of E6

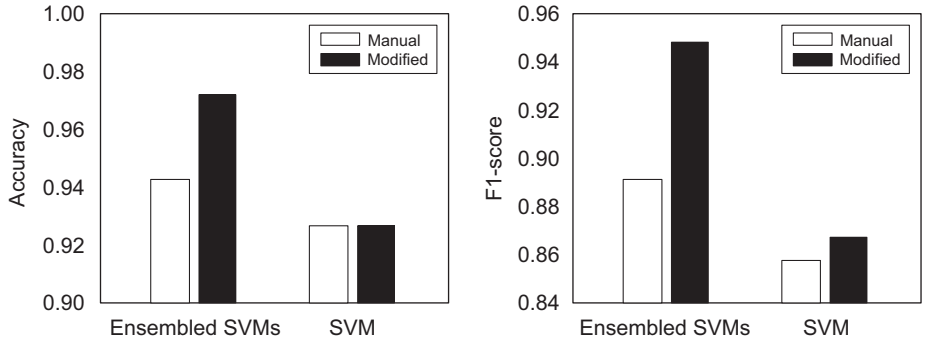

Fig. 4. The performance comparison of the proposed method and the previous SVM approach from Joung et al. on manually tagged answers (Manual) and modified answers (Modified). For the modified answers, some HPV risk types were corrected according to the results of Muñoz et al. as shown in Table 4.

Table 5
SVM predictions for the HPVs manually classified as high-risk types.

Type	$k=2$		$k=3$		$k=4$		Output
	α helix	Others	α helix	Others	α helix	Others	
HPV16	High	High	High	High	High	High	High
HPV18	High	High	High	High	High	High	High
HPV31	High	High	High	High	High	High	High
HPV33	High	High	High	High	High	High	High
HPV35	High	High	High	High	High	High	High
HPV39	High	High	High	High	High	High	High
HPV45	High	High	High	High	High	High	High
HPV51	High	High	Low	High	High	High	High
HPV52	High	High	High	High	High	High	High
HPV56	High	High	Low	High	Low	High	Possible high
HPV58	High	High	High	High	High	High	High
HPV59	High	High	Low	High	High	High	High
HPV61	High	Low	Low	High	Low	High	Possible high
HPV66	High	Low	Low	Low	Low	Low	Low
HPV67	High	High	High	High	High	High	High
HPV68	High	High	High	High	High	High	High
HPV72	High	High	Low	Low	Low	High	Possible high

protein has not been fully solved and only a predicted model can be applied [29]. HPV E6 proteins do not act as high-risk types independently, rather associated with p53 tumor suppressor and the ubiquitin ligase E6-AP, then α helices might form part of the binding regions, i.e., zinc-finger proteins according to the experimental results. There have been previous reports that the α helix in certain domains is related to viral infection and oncogenic transformation [30–32]. It is also known that many HPV16 E6 binding proteins, including E6-AP, paxillin, E6-BP, and IRF-3, contain a conserved α-helical domain and presumably interact with similar E6 sequences [12,33]. The HPV16 is the most prevalent high-risk HPV type and we realize that the α-helical feature can also be common for other high-risk HPVs.

3.3. Prediction for unknown HPV types

One of the most important issues in this task is to predict unknown, but potentially high-risk HPV types. We have evaluated the HPVs marked as unknown from the LANL database, and the results are presented in Table 6. HPV26, HPV54, HPV57, and HPV70 are predicted as high-, low-, low-, and high-risk types, respectively. The prediction results for HPV26 and HPV54 are identical to the one in Muñoz et al. [5]. However, there have been different predictive decisions for HPV70 according to previous reports [5,34,35], and the risk type of HPV57 cannot be predicted with accuracy because of insufficient published work. This shows that the SVM ensemble method can provide a guideline for the investigation of potentially high-risk HPVs.

Table 6
Predicted risk type for the HPVs, whose types are unknown.

Type	Risk
HPV26	High
HPV54	Low
HPV57	Low
HPV70	High

4. Summary

We have proposed an ML approach to classify HPV risk types that are closely related to cervical caner. The proposed method uses the secondary structure information using Jnet, and for each structural element, gapped amino acid pairs are considered to reflect the interactions between amino acids by the gap-spectrum kernel. The ensembled SVMs give confidence levels for the high-risk type, following performance improvement. Even though the ensemble method is computationally expensive, it provides intuitive explanations of how results are obtained.

By performing leave-one-out cross-validation, the experimental results show that the SVM ensemble utilizing E6 protein structures improves the HPV classification performance more than other methods. In particular, we have discovered that α helical regions can have an important role in oncogenic transformation in high-risk HPVs. We also simulated predictions for unknown HPV types, which provide promising results. Even though E6 structure is not fully elucidated,

our approach would provide prior knowledge for drug design for cervical cancer prophylaxis.

Conflict of interest statement

None declared.

Acknowledgments

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the National Research Laboratory Program funded by the Ministry of Science and Technology (No. M10400000349-06J0000-34910).

References

[1] E.-K. Yim, J.-S. Park, Role of proteomics in translational research in cervical cancer, Expert Rev. Proteomics 3 (1) (2006) 21–36.
[2] F.X. Bosch, M.M. Manos, N. Muñoz, M. Sherman, A.M. Jansen, J. Peto, M.H. Schiffman, V. Moreno, R. Kurman, K.V. Shah, Prevalence of human papillomavirus in cervical cancer: a worldwide perspective, J. Natl. Cancer Inst. 87 (1995) 796–802.
[3] M.F. Janicek, H.E. Averette, Cervical cancer: prevention, diagnosis, and therapeutics, CA Cancer J. Clin. 51 (2001) 92–114.
[4] H. Furumoto, M. Irahara, Human papillomavirus (HPV) and cervical cancer, J. Med. Invest. 49 (2002) 124–133.
[5] N. Muñoz, F.X. Bosch, S. de Sanjose, R. Herrero, X. Castellsague, K.V. Shah, P.J. Snijders, C.J. Meijer, Epidemiologic classification of human papillomavirus types associated with cervical cancer, N. Engl. J. Med. 348 (2003) 518–527.
[6] J.-H. Eom, S.-B. Park, B.-T. Zhang, Genetic mining of DNA sequence structures for effective classification of the risk types of human papillomavirus (HPV), in: Proceedings of the 11th International Conference on Neural Information Processing, 2004, pp. 1334–1343.
[7] J.-G. Joung, S.-J. O, 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, 2004, pp. 723–731.
[8] J.-G. Joung, S.-J. O, B.-T. Zhang, Protein sequence-based risk classification for human papillomaviruses, Comput. Biol. Med. 36 (2006) 656–667.
[9] S.-B. Park, S. Hwang, B.-T. Zhang, Mining the risk types of human papillomavirus (HPV) by AdaCost, in: Proceedings of the 14th International Conference on Database and Expert Systems Applications, 2003, pp. 403–412.
[10] J.A. Cuff, G.J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, Proteins 40 (2000) 502–511.
[11] R.A. Hubbard, Human papillomavirus testing methods, Arch. Pathol. Lab. Med. 127 (8) (2003) 940–945.
[12] K. Münger, A. Baldwin, K.M. Edwards, H. Hayakawa, C.L. Nguyen, M. Owens, M. Grace, K. Huh, Mechanisms of human papillomavirus-induced oncogenesis, J. Virol. 78 (21) (2004) 11451–11460.
[13] M.R. Pillai, S. Lakshmi, S. Sreekala, T.G. Devi, P.G. Jayaprakash, T.N. Rajalakshmi, C.G. Devi, M.K. Nair, M.B. Nair, High-risk human papillomavirus infection and E6 protein expression in lesions of the uterine cervix, Pathobiology 66 (1998) 240–246.
[14] B. Rost, Review: protein secondary structure prediction continues to rise, J. Struct. Biol. 134 (2001) 204–218.
[15] Z. Aydin, Y. Altunbasak, M. Borodovsky, Protein secondary structure prediction for a single-sequence using hidden semi-Markov models, BMC Bioinform. 7 (2006) 178.
[16] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (2) (1999) 195–202.
[17] D.G. Kneller, F.E. Cohen, R. Langridge, Improvements in protein secondary structure prediction by an enhanced neural network, J. Mol. Biol. 214 (1) (1990) 171–182.
[18] J. Martin, G. Letellier, A. Marin, J.F. Taly, A.G. de Brevern, J.F. Gibrat, Protein secondary structure assignment revisited: a detailed analysis of different assignment methods, BMC Struct. Biol. 5 (17) (2005).
[19] A.A. Salamov, V.V. Solovyev, Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments, J. Mol. Biol. 247 (1) (1995) 11–15.
[20] S. Geman, E. Bienenstock, R. Doursat, Neural networks and the bias/variance dilemma, Neural Comput. 4 (1992) 1–58.
[21] C. Leslie, E. Eskin, W.S. Noble, The spectrum kernel: a string kernel for SVM protein classification, in: Proceedings of the Pacific Symposium on Biocomputing, 2002, pp. 564–575.
[22] C. Leslie, E. Eskin, A. Cohen, J. Weston, W.S. Noble, Mismatch string kernels for discriminative protein classification, Bioinformatics 20 (4) (2004) 467–476.
[23] S. Kim, B.-T. Zhang, Human papillomavirus risk type classification from protein sequences using support vector machines, in: Proceedings of the European Workshop on Evolutionary Computation and Machine Learning in Bioinformatics, 2006, pp. 57–66.
[24] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, Cambridge, 2004.
[25] V.N. Vapnik, Statistical Learning Theory, Springer, Berlin, 1998.
[26] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001, Software available at ⟨http://www.csie.ntu.edu.tw/~cjlin/libsvm⟩.
[27] W. Fan, S. Stolfo, J. Zhang, P. Chan, AdaCost: misclassification cost-sensitive boosting, in: Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 97–105.
[28] Y.-H. Kim, S.-Y. Hahn, B.-T. Zhang, Text filtering by boosting naive Bayes classifiers, in: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000, pp. 168–175.
[29] S.-W. Hung, J.-K. Hwang, F. Tseng, J.-M. Chang, C.-C. Chen, C.-C. Chieng, Molecular dynamics simulation of the enhancement of cobra cardiotoxin and E6 protein binding on mixed self-assembled monolayer molecules, Nanotechnology 17 (2006) S8–S13.
[30] V. Brass, E. Bieck, R. Montserret, B. Wölk, J.A. Hellings, H.E. Blum, F. Penin, D. Moradpour, An amino-terminal amphipathic α-helix mediates membrane association of the hepatitis C virus nonstructural protein 5A, J. Biol. Chem. 277 (10) (2002) 8130–8139.
[31] K. Briknárová, F. Nasertorabi, M.L. Havert, E. Eggleston, D.W. Hoyt, C. Li, A.J. Olson, K. Vuori, K.R. Ely, The serine-rich domain from Crk-associated substrate ($p130^{cas}$) is a four-helix bundle, J. Biol. Chem. 280 (23) (2005) 21908–21914.
[32] J.S. Orlando, D.A. Ornelles, An arginine-faced amphipathic alpha helix is required for adenovirus type 5 e4orf6 protein function, J. Virol. 73 (6) (1999) 4600–4610.
[33] M. Nguyen, S. Song, A. Liem, E. Androphy, Y. Liu, P.F. Lambert, A mutant of human papillomavirus type 16 E6 deficient in binding α-helix partners displays reduced oncogenic potential in vivo, J. Virol. 76 (24) (2002) 13039–13048.
[34] 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 (1996) 738–744.
[35] T. Meyer, R. Arndt, E. Christophers, E.R. Beckmann, S. Schroder, L. Gissmann, E. Stockfleth, Association of rare human papillomavirus types with genital premalignant and malignant lesions, J. Infect. Dis. 178 (1998) 252–255.

Sun Kim received the B.S. degree in Computer Science from Soongsil University, Seoul, Korea, in 1999 and the M.S. degree in Computer Science and Engineering from Seoul National University (SNU), Seoul, Korea, in 2001. He is currently working towards the Ph.D. degree at the School of Computer Science and Engineering, SNU. His research interests include text mining, bioinformatics, and evolutionary computation.

Jeongmi Kim received the B.S. degree in Biology from Ewha Woman University, Seoul, Korea, in 1986, the M.S. degree in Epidemiology from Seoul National University, Seoul, Korea, in 1989, and the Ph.D. degree in Molecular Toxicology from University of Text at Austin, USA. She was a Postdoctorial Associate at the Division of Toxicology, Massachusetts Institute of Technology (MIT), USA from 1996 to 1997. She was a Senior Research Scientist at the Division of Cancer Research, NIH, Korea from 1997 to 2000. She was a Director at the Microarray Centerin Biomedlab Co., Korea from 2000 to 2002. She was a Director at the Division of Diagnostic Business, ISU Chemical Co. Ltd./ISU ABXIS Co. Ltd., Korea from 2002 to 2007. She is currently the CEO at the VetAll Laboratories, Korea. His research interest includes molecular diagnostics.

Byoung-Tak Zhang received the B.S. and M.S. degrees in Computer Science and Engineering from Seoul National University (SNU), Seoul, Korea, in 1986 and 1988, respectively, and the Ph.D. degree in Computer Science from University of Bonn, Bonn, Germany, in 1992. He is a Professor at the School of Computer Science and Engineering and of the Graduate Programs in Bioinformatics, Brain Science, and Cognitive Science at Seoul National University (SNU), and directs the Biointelligence Laboratory and the Center for Bioinformation Technology (CBIT). Prior to joining SNU, he had been a Research Associate at the German National Research Center for Information Technology (GMD) from 1992 to 1995. He had been a Visiting Professor at the Computer Science and Artificial Intelligence Laboratory (CSAIL), Massachusetts Institute of Technology (MIT), Cambridge, from August 2003 to August 2004. His research interests include probabilistic models of learning and evolution, biomolecular/DNA computing, and molecular learning/evolvable machines. Dr. Zhang serves as an Associate Editor of the IEEE Transactions on Evolutionary Computation, Advances in Natural Computation, and Genomics and Informatics. He is on the Editorial Board of Genetic Programming and Evolvable Machines and Applied Soft Computing.