ELSEVIER

Computers in Biology and Medicine 36 (2006) 656–667

Computers in Biology and Medicine

www.intl.elsevierhealth.com/journals/cobm

Protein sequence-based risk classification for human papillomaviruses

Je-Gun Joung[a, b], Sok June O[c], Byoung-Tak Zhang[a, b, d, *]

[a] *Graduate Program in Bioinformatics, Seoul National University, Seoul 151-742, Republic of Korea*
[b] *Center for Bioinformation Technology (CBIT), Seoul National University, Seoul 151-742, Republic of Korea*
[c] *Department of Pharmacology and Pharmacogenomics Research Center, Inje University College of Medicine, Busan 614-735, Republic of Korea*
[d] *School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Republic of Korea*

Received 20 November 2003; accepted 12 April 2004

Abstract

Human papillomaviruses (HPVs) are small DNA tumor viruses which infect epithelial tissues and induce hyperproliferative lesions. Infection by high-risk genital HPVs is associated with the development of anogenital cancers. Classification of risk types is important in understanding the mechanisms in infection and in developing novel instruments for medical examination such as DNA microarrays. The sequence-based classification methods are useful in classifying risk types by considering residues in conserved positions. In this paper, we present a machine learning approach to the classification of HPV risk types by using the protein sequences. Our approach is based on the hidden Markov model and the kernel method. The former searches informative subsequence positions and the latter computes efficiently to classify protein sequences. In the experiments, the classifier predicted four unknown HPV types exactly. An additional result shows that the kernel-based classifiers learned with more informative subsequences outperform the classifiers learned with the whole sequence or random subsequences.

Keywords: Human papillomavirus; Machine learning; Kernel methods; Hidden Markov models; Sequence classification

* Corresponding author. School of Computer Science and Engineering, Seoul National University, Seoul 151-742, Republic of Korea. Tel.: +82 2 8801847; fax: +82 2 8572240.
E-mail address: btzhang@bi.snu.ac.kr (B.-T. Zhang).

0010-4825/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiomed.2004.04.007

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 women. HPVs have a double-stranded DNA genome of approximately 8000 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. On the other hand, high-risk viral types are associated with high-grade cervical lesions and cancers [4].

The most urgent and important aspect 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 models (HMMs) and the kernel method. 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 [5–7]. Whereas HMM is a generative model, the kernel-based classifier is a discriminant model. Ultimately, the proposed method uses the generative model to obtain an easily distinguishable sequence source and a discriminant model to maximize classification ability.

The presented kernel-based classifier includes the string kernel that deals with HPV 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 in analyzing 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 [8–11].

Kernel-based learning methods use an implicit mapping of the input data into a high-dimensional feature space defined by a kernel function, i.e. a function returning the inner product between the images of two data points in the feature space. The learning then takes place in the feature space, providing the learning algorithm to be entirely rewritten so that the data points appear only inside the dot products with other points. Several linear algorithms can be formulated in this way, for classification, regression, and clustering. The most typical example of kernel-based systems is the support vector machines (SVMs) that implement linear classification. 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 the kernel and HMM embedded system to classify HPV genotype is described. Then the kernel method for HPV sequence analysis is presented. In Section 3, the experimental results are provided by the proposed method applied to HPV sequence data sets. In Section 4, some possible applications using our method are described. Concluding remarks and directions on further research are given in Section 5.

2. Data sets and methods

2.1. Data source

The data set was extracted from the HPV sequence database at Los Alamos National Laboratory (LANL). Fig. 1 shows an example of the HPV type 31 among many types. 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 obtained 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 of total types were labeled as high-risk types [12]. The rest were labeled as low-risk types.

2.2. Data pre-processing using HMM

The overall process to classify HPV risk types is presented in Fig. 2. 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 HPV sequences are aligned by a multiple alignment tool such as Clustal W [13]. Second, they are divided into positive and negative sequences, and then an HMM is constructed from positive segments. Each segment is the subsequence that is a window of size w and is aligned over the same position by using the multiple alignment tool. Third, the

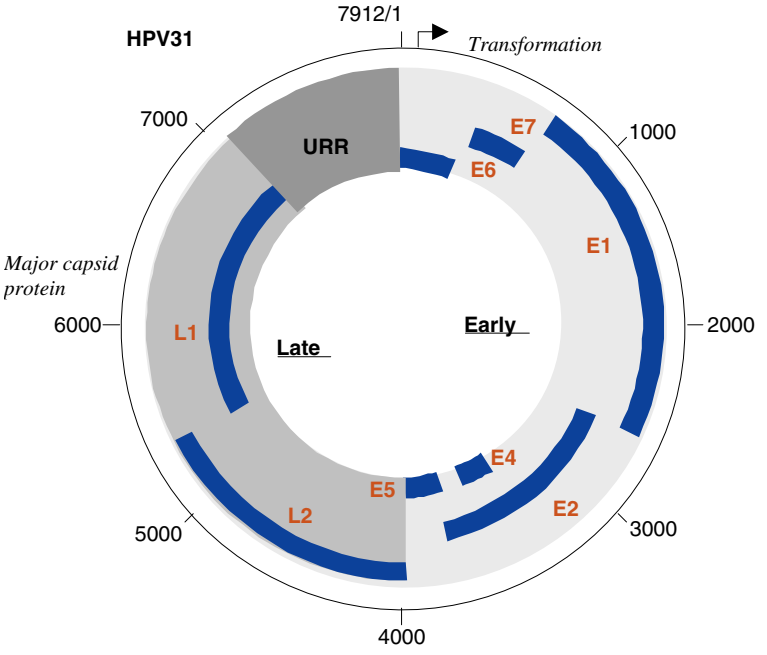

Fig. 1. The structure of the HPV-31 genome: HPV is the 8 kb double-stranded DNA genome. The oncoproteins E6 and E7 form complexes with host tumor suppressor proteins p53 and Rb, respectively, inactivating them and disrupting cell-cycle control.

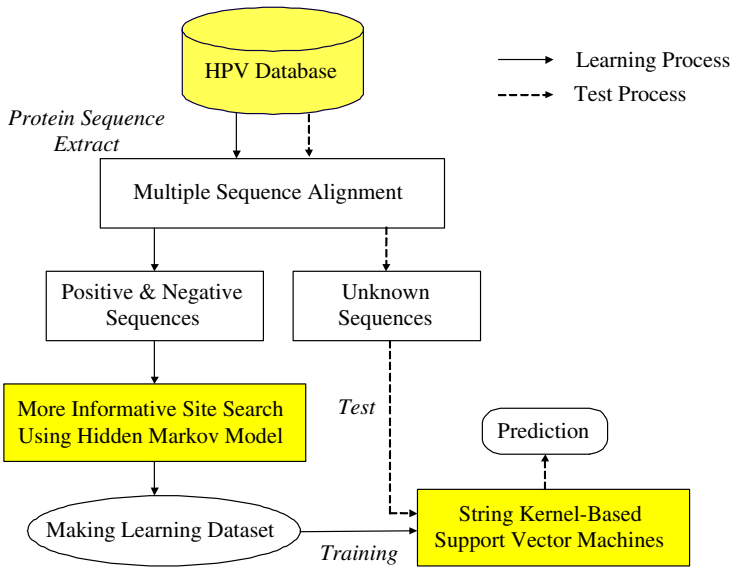

Fig. 2. Overview of the HPV risk classification procedure. The data pre-processing contains Hidden Markov models in order to find more informative subsequence regions. Then, a string kernel-based SVM learns these subsequence data sets. Given an unknown sequence, its risk type is predicted using the kernel SVM.

log-likelihoods of positive and negative segments are calculated from the HMM model. Fourth, the 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 is extracted from subsequences that have a high score.

The biological sequence analysis has developed a reasonably successful solution using HMMs. HMMs are a statistical sequence-comparison technique. 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.

2.3. Kernel function

After the data-preprocessing, the string kernel-based SVM is trained on the HPV sequence data set and tested on the unknown sequences. Here, the main module is the mismatch-spectrum kernel that is a function of SVM. The mismatch-spectrum kernel is a new string kernel that was used to detect remote homology detection [9]. It is very simple and efficient to compute because it is based on occurrences of possible subsequences. In order to capture significant information from the sequence data, mismatch-spectrum kernels use the spectrum. The mismatch-spectrum kernel is an 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 $\mathscr{A}$ of size $|\mathscr{A}| = l$ and the vector space is the l^k-dimensional vectors indexed by the set of k-length subsequences (k-mers) from $\mathscr{A}$.

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

$$\Phi_{(k)}(x) = (\phi_\alpha(x))_{\alpha \in \mathscr{A}^k}, \tag{1}$$

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

$$K_{(k)}(x, y) = \langle \Phi_{(k)}(x), \Phi_{(k)}(y) \rangle. \tag{2}$$

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.

A more biologically realistic kernel is the model allowing mismatch in k-mer subsequences. A fixed k-mer subsequence of amino acids is defined as $\alpha = a_1 a_2 \ldots a_k$. Here each a_i is a character in $\mathscr{A}$. The (k, m)-neighborhood generated by α is the set of all k-length sequences β that differ from α by at most m mismatches. This set is denoted by $N_{(k,m)}(\alpha)$. For k-mer and m mismatch, the feature map $\Phi_{(k,m)}$ is defined as follows:

$$\Phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))_{\beta \in \mathscr{A}^k}, \tag{3}$$

where $\phi_\beta(\alpha) = 1$ if β belongs to $N_{(k,m)}(\alpha)$, and $\phi_\beta(\alpha) = 0$ otherwise.

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

$$\Phi_{(k,m)}(x) = \sum_{k\text{-mers } \alpha \text{ in } x} \Phi_{(k,m)}(\alpha). \tag{4}$$

The β-coordinate of $\Phi_{(k,m)}(x)$ is just a count of all instances of the k-mer β occurring m mismatches in x. If Eq. (2) is extended, then the (k, m)-mismatch kernel $K_{(m,k)}$ is the inner product in the feature space of feature vectors:

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

The mismatch kernel is used in combination with the SVM. Fig. 3 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.

2.4. Support vector machines

Support vector machines were developed by Vapnik for classification of data based on a trained model [14]. Recently, they have found several applications in biological data analysis. These applications contain gene classification from microarray [15,16], translation initiation site recognition in DNA [17] and identifying splicing sites in eukaryotic RNA [18].

Given a kernel and a set of labeled 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 unlabeled example is then predicted to be positive or negative depending on the position of its image in the feature space relative to the linear boundary.

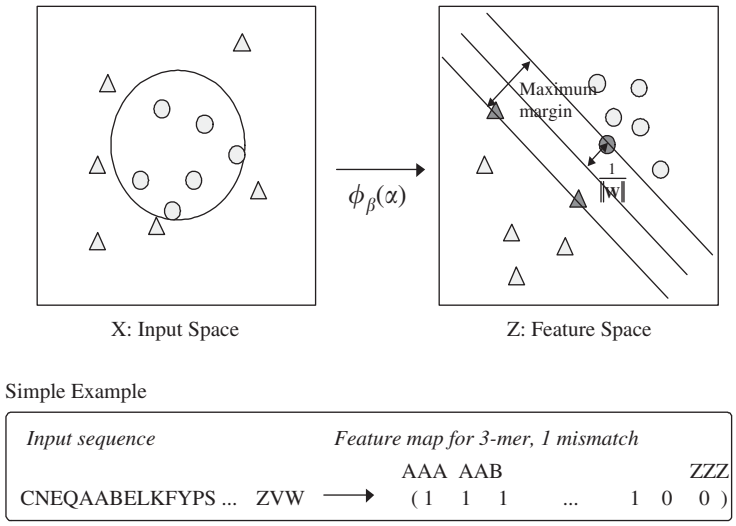

Fig. 3. The maximal margin classifier (or SVM) learns a linear discriminant function in a high-dimensional feature so that the hyperplane optimally separates with maximum margin. For the mismatch-spectrum kernel, the feature map $\phi_\beta(\alpha)$ is given from input space into a high-dimensional feature space (vector space). The feature map is indexed by all possible k-mers.

SVMs learn non-linear discriminant functions in an input space. This is achieved by learning a linear discriminant function in a high-dimensional feature space. A feature mapping ϕ 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 $1/\|\mathbf{w}\|$. Among normalized hyperplanes, SVMs find the maximal margin hyperplane that has the maximal margin.

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

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

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

where parameters α_i are called *Lagrange multipliers*. The parameters $(\mathbf{w}, b)$ are determined by the optimal α_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 the following parameters:

$$f^*(\mathbf{z}) = \sum_{i=1}^{n} \alpha_i^* y_i \langle \mathbf{z}_i, \mathbf{z} \rangle + b. \tag{8}$$

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

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 α_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 looking up k-mer scores [9].

3. Results

3.1. Searching informative subsequences

Fig. 4 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 the zinc-binding region of E6 [19]. 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

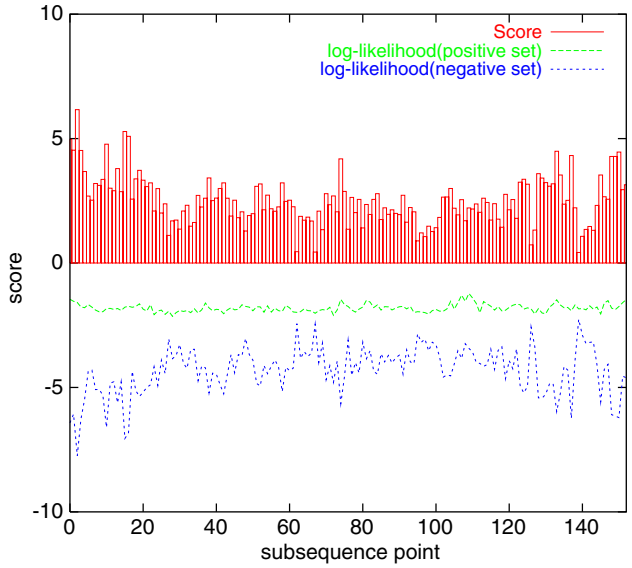

Fig. 4. The scores of subsequences through HMM learning for E6. Points 3, 17, 75, 138 and 150 show high scores. These points may possibly play an important role in the tumor-related suppression or activation.

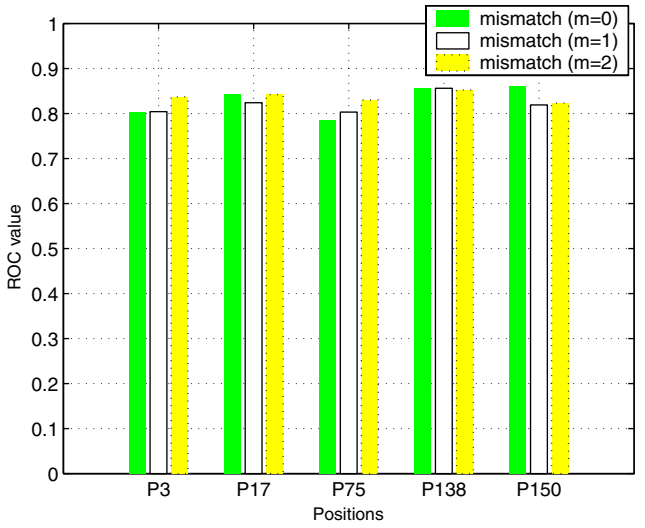

Fig. 5. The ROC values of subsequences that present high scores by HMM. The point 138 indicates the highly conserved sequence position so that the highly conserved regions are associated with the classification performance.

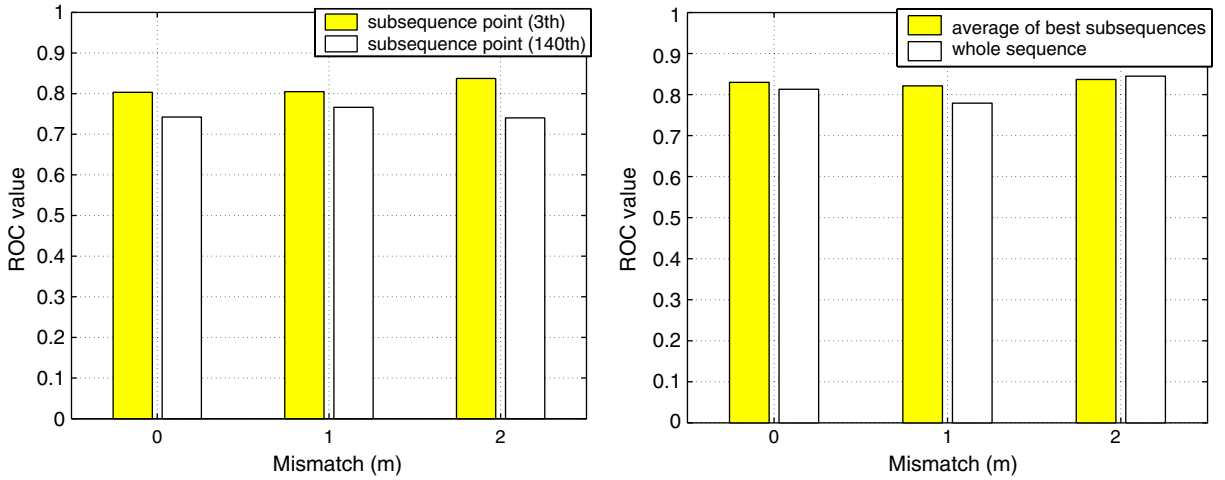

Fig. 6. Comparison of ROC values of various data sets. Subsequences with high score vs. subsequences with low score (left) and the whole sequence vs. average of informative subsequences (right). When the classifier learns the informative subsequences, the accuracy becomes better.

suppressor protein [20]. Moreover, E6 from high-risk HPV binds p53 with a higher affinity than that from low-risk HPV, and mediates the degradation of p53 through the ubiquitin-dependent system.

3.2. *Prediction performance of subsequences*

Fig. 5 shows the ROC (receiver-operating characteristic) [21] values of subsequences that present high scores in Fig. 4. An ROC represents the joint values of the true-positive ratio (sensitivity) and false positive

Table 1
Comparison between the manually tagged answer (Man.) and the string kernel-based prediction (Class)

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

ratio (1-specificity). Each bar represents an average value after 100 runs. The size of k-mers is 4. In 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.

Fig. 6 demonstrates a comparison of the ROC values of various data sets. The left figure shows a comparison between the best and worst informative subsequences. When the classifier learns informative

subsequences, the accuracy is better. The left figure is the result that measures the accuracy for the whole sequence and the average of the informative subsequences. The kernel method using the informative subsequences outperforms for two mismatches ($m = 0, 1$).

3.3. Classification of risk types

Table 1 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 [12] 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 cannot be determined, and there are three unknown types. The prediction is the result of leave-one-out cross-validation.

The most interesting fact is that the classifier predicted high-risks for HPV70. According to the previous study on HPV [22], the document contains that HPV70 was also detected in genital intraepithelial neoplasia from one patient. This is very important result because the classifier in this paper provides the probability of 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 sure to make a mistake. HPV type 53 was detected in genital specimens of the 16 of the patients [23]. 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 the E7 or L1 gene.

4. Discussion

Classification of risk types is important in understanding the mechanisms in infection and in developing novel instruments for medical examination such as DNA chips. To design a genotyping DNA microarray, probe selection is one of the important tasks. Many viruses, as well as HPV can be easily detected by DNA microarrays. The probe selection problem is to determine probe sequences from target sequences. It is based on the criteria of specificity, melting temperature, and secondary structure stability. Satisfying these criteria is very difficult. Our approach can provide useful information about the informative sites to select probes. In other words, it can catch specificity to classify high-risk and low-risk viral infections. It can be used as a prior process for probe selection.

For a better performance, there is a method to extend our approach. As an alternative, the input sequence data could be combined with information of the protein structure. In principle, amino acid sequences for proteins contain sufficient information to determine their structure. Furthermore, the structure for proteins is closely related to function. The secondary structure for each HPV is easily predicted by a suitable tool. Use of the secondary structure can possibly affect HPV classification.

5. Summary

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. In the prediction strategy, when the classifier learns informative subsequences, the accuracy is better. The informative subsequences could indicate the highly conserved regions.

In addition to this result, we predicted the risk type for all types via leave-one-out cross-validation. The most interesting question is 'what is the risk type of HPV70'. This paper provides the probability of 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. For more accurate prediction, the input sequence data could be combined with information of the protein structure.

Acknowledgements

This research was supported by NRL, Systems Biology (M10309000002-03B5000-00110), and BK21.

References

[1] H. Pfister, J. Krubke, W. Dietrich, T. Iftner, P.G. Fuchs, Classification of the papilomavirues-mapping the genome, Ciba Found. Symp. 120 (1986) 3–22.
[2] H. zur Hausen, Papillomaviruses causing cancer: evasion from host-cell control in early events in carcinogenesis, J. Nat. Cancer Inst. 92 (2000) 690–698.
[3] IARC Monographs on the Evaluation of the Carcinogenic Risks to Humans, IARC Scientific Publications, Lyon, France, 1995.
[4] M.F. Janicek, H.E. Averette, Cervical cancer: prevention, diagnosis, and therapeutics, Cancer J. Clin. 51 (2001) 92–114.
[5] R. Haghey, A. Krogh, Hidden Markov models for sequence analysis: extension and analysis of the basic method, CABIOS 12 (1996) 95–107.
[6] P. Baldi, Y. Chauvin, et al., Hidden Markov models of biological primary sequence information, PNAS 91 (1994) 1059–1063.
[7] S. Eddy, Multiple alignment using hidden Markov models, ISMB 95 (1995) 114–120.
[8] C. Leslie, E. Eskin, W. Noble, The spectrum kernel: a string kernel for SVM protein classification, in: Proceedings of the Pacific Symposium on Biocomputing, 2002, pp. 564–575.
[9] C. Leslie, E. Eskin, J. Weston, W. Noble, Mismatch string kernels for SVM protein classification, Neural Information Processing Systems (NIPS) 15 (2003) 1441–1448.
[10] J.-P. Vert, Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings, in: Proceedings of the Pacific Symposium on Biocomputing, 2002, pp. 649–660.
[11] T. Jaakkola, M. Diekhans, D. Haussler, A discriminative framework for detecting remote protein homologies, J. Comput. Biol. (2000).
[12] N. Muñoz, F.X. Bosch, et al., Epidemiologic classification of human papillomavirus types associated with cervical cancer, N. Engl. J. Med. 348 (2003) 518–527.
[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. 22 (1994) 4673–4680.
[14] V.N. Vapnik, Statistical Learning Theory, Springer, New York, 1998.

[15] M.P.S. Brown, W.N. Grund, et al., Knowledge-based analysis of microarray gene expression data by using support vector machines, Proc. Natl. Acad. Sci. 97 (2000) 262–267.
[16] T.S. Furey, N. Duffy, et al., Support vector machines classification and validation of cancer tissues samples using microarray expression data, Bioinformatics 16 (2000) 906–914.
[17] A. Zien, G. Ratsch, et al., Engineering support vector machine kernels that recognize translation initiation sites, Bioinformatics 16 (2000) 799–807.
[18] Y.F. Sun, X.D. Fan, Y.D. Li, Identifying splicing sites in eukaryotic RNA: support vector machine approach, Comput. Biol. Med. 33 (2003) 17–29.
[19] 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. 319 (1996) 229–239.
[20] 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) (2000) 1189–1203.
[21] F. Schoonjans, C. Depuydt, F. Comhaire, Presentation of receiver-operating characteristic (ROC) plots (Letter to the Editor), Clin. Chem. 42 (1996) 986–987.
[22] 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) (1996) 738–744.
[23] 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) (2001) 198–204.

Je-Gun Joung has studied computer science and received his M.S. degree in bioinformatics from Seoul National University in 2004. Currently, he is a Ph.D. student at the Graduate Program in Bioinformatics and is working at the Center for Bioinformation Technology (CBIT) of Institute of Computer Technology, Seoul National University. His research interests lie in biological datamining, bionetwork analysis using machine learning, and evolutionary computing.

Sok June O received his Ph.D. in agricultural biotechnology from Seoul National University in 2000. He has worked as R&D section head of the Center for Bioinformation Technology (CBIT), Seoul National University. At present, he is a full-time instructor in the Department of Pharmacology and Pharmacogenomics Research Center, College of Medicine, Inje University, Busan, Korea. His research interests include biological information processing, biomedical informatics, and biomolecular computing.

Byoung-Tak Zhang received his Ph.D. in computer science from University of Bonn, Germany in 1992. He is currently an associate professor of School of Computer Science and Engineering at Seoul National University, and directs the Biointelligence Laboratory and the Center for Bioinformation Technology (CBIT). His research interests include probabilistic models of learning and evolution, biomolecular computing, and molecular learning/evolvable machines.