Human Papillomavirus Risk Type Classification from Protein Sequences Using Support Vector Machines

Sun Kim and Byoung-Tak Zhang

Biointelligence Laboratory,
School of Computer Science and Engineering,
Seoul National University, Seoul 151-744, South Korea
{skim, btzhang}@bi.snu.ac.kr

Abstract. Infection by the human papillomavirus (HPV) is associated with the development of cervical cancer. HPV can be classified to high- and low-risk type according to its malignant potential, and detection of the risk type is important to understand the mechanisms and diagnose potential patients. In this paper, we classify the HPV protein sequences by support vector machines. A string kernel is introduced to discriminate HPV protein sequences. The kernel emphasizes amino acids pairs with a distance. In the experiments, our approach is compared with previous methods in accuracy and F1-score, and it has showed better performance. Also, the prediction results for unknown HPV types are presented.

1 Introduction

The cervical cancer is a leading cause of cancer death among women worldwide. Epidemiologic studies have shown that the association of genital human papillomavirus (HPV) with cervical cancer is strong, independent of other risk factors [1]. HPV infection causes virtually all cases of cervical cancer because certain high-risk HPVs develop cancer even though most cases of HPV are low-risk and rarely develop into cancer. Especially, high-risk HPV types could induce more than 95% of cervical cancer in woman.

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, or the genital tract. And the genital tract HPV types are classified into two or three types such as low-, intermediate-, and high-risk types by their relative malignant potential [2]. The common, unifying oncogenic feature of the vast majority of cervical cancers is the presence of HPV, especially high-risk type HPV [3]. Thus the risk type detection of HPVs have become one of the most essential procedures in cervical cancer remedy. Currently, the HPV risk types are still manually classified by experts, and there is no deterministic method to expect the risk type for unknown or new HPVs.

Since the HPV classification is important in medical judgments, there have been many epidemiological and experimental studies to identify HPV risk types

F. Rothlauf et al. (Eds.): EvoWorkshops 2006, LNCS 3907, pp. 57–66, 2006.

[3]. Polymerase chain reaction (PCR) is a sensitive technique for the detection of very small amounts of HPV's nucleic acids in clinical specimens. It has been used in most epidemiological studies that have evaluated the role of these viruses in cervical cancer causation [4]. Bosch et al. [1] investigated epidemiological characteristic that whether the association between HPV infection and cervical cancer is consistent worldwide in the context of geographic variation. Burk et al. [5] inspected the risk factors for HPV infection in 604 young college women through examining social relationship and detected various factors of HPV infection with L1 consensus primer PCR and Southern blot hybridization. Muñoz et al. [6] classified the HPV risk types with epidemiological experiments based on risk factor analysis. They pooled real data from 1,918 cervical cancer patients and analyzed it by PCR based assays.

Detection of HPV risk types can be a protein function prediction even though functions are described at many levels, ranging from biochemical function to biological processes and pathways, all the way up to the organ or organism level [7]. Many approaches for protein function prediction are based on similarity search between proteins with known function. The similarity among proteins can be defined in a multitude of ways [8]: sequence alignment, structure match by common surface clefts or binding sites, common chemical features, or certain motifs comparison. However, none of the existing prediction systems can guarantee generally good performance. Thus it is required to develop classification methods for HPV risk types. Eom et al. [9] presented a sequence comparison method for HPV classification. They use DNA sequences to discriminate risk types based on genetic algorithms. Joung et al. [10] combined with several methods for the risk type prediction from protein sequences. Protein sequences are first aligned, and the subsequences in high-risk HPVs against low-risk HPVs are selected by hidden Markov models. Then a support vector machine is used to determine the risk types. The main drawback of this paper is that the method is biased by one sequence pattern. Alternatively, biomedical literature can be used to predict HPV risk types [11]. But, text mining approaches have the limitation for prediction capability because they only depend on texts to capture the classification evidence, and the obvious keywords such as 'high' tend to be appeared in the literature explicitly.

In this paper, we propose a method to classify HPV risk types using protein sequences. Our approach is based on support vector machines (SVM) to discriminate low- and high-risk types and a string kernel is introduced to deal with protein sequences. The string kernel first maps to the space consisting of all subsequences of amino acids pair. A RBF kernel is then used for nonlinear mapping into a higher dimensional space and similarity calculation. Especially, the proposed kernel only uses amino acids of both ends in k-length subsequences to improve the classification performance. It is motivated by the assumption that amino acids pairs with certain distance affects the HPV's biological function, i.e. risk type, more than consecutive amino acids. The experimental results show that our approach provides better performance than previous approaches in accuracy and F1-score.

Our work addresses how to classify HPV risk types from protein sequences by SVM approaches, which can provide a guide to determine unknown or new HPVs. The paper is organized as follows. In Section 2, we explain the SVM method for classification. Then the string kernel for HPV protein sequence is presented in Section 3. In Section 4, we present the experimental results and draw conclusions in Section 5.

2 Support Vector Machine Classifiers

We use support vector machines to classify HPV risk types. A string kernel-based SVM is trained on HPV protein sequences and tested on unknown sequences. Support vector machines have been developed by Vapnik to give robust performance for classification and regression problems in noisy, complex data [12]. It has been widely used from text categorization to bioinformatics in recent days. When it is used for classification problem, a kernel and a set of labeled vectors, which is marked to positive or negative class are given. The kernel functions introduce nonlinear features in 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 $S = \{\mathbf{x}_i, y_i\}_{i=1}^n$, is 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 $sgn(\langle \mathbf{w}, \Phi(\mathbf{x}) \rangle + b)$. 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.

To find the optimal hyperplane, it can be formulated as the following problem:

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{w}\|^2 \tag{1}$$

$$\text{subject to} \quad y_i(\langle \mathbf{w}, \Phi(\mathbf{x}_i) \rangle + b) \geq 1, \ i = 1, \ldots, n. \tag{2}$$

By introducing Lagrange multipliers $\alpha_i \geq 0, \ i = 1, \ldots, n$, we get the following dual optimization problem:

$$\text{maximize} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle \tag{3}$$

$$\text{subject to} \quad \alpha_i \geq 0, \ i = 1, \ldots, n, \tag{4}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0. \tag{5}$$

By solving this dual problem, one obtains optimal solution $\alpha_i, 1 \leq i \leq n$, which needs to determine the parameters $(\mathbf{w}, b)$. For the solution $\alpha_i, \ldots, \alpha_n$, the

nonlinear decision function $f(\mathbf{x})$ is expressed in terms of the following parameters:

$$f(\mathbf{x}) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}_i)\rangle + b\right) \tag{6}$$

$$= \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b\right). \tag{7}$$

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

3 Kernel Function

For HPV protein classification, we introduce a string kernel based on the spectrum kernel method. The spectrum kernel was used to detect remote homology detection [13][14]. The input space $\mathcal{X}$ consists of all finite length sequences of characters from an alphabet $\mathcal{A}$ of size $|\mathcal{A}| = l$ ($l = 20$ for amino acids). Given a number $k \geq 1$, 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 $\mathcal{A}$. The k-spectrum feature map $\Phi_k(x)$ from $\mathcal{X}$ to $\mathbb{R}^{l^k}$ can be defined as:

$$\Phi_k(x) = (\phi_a(x))_{a \in \mathcal{A}^k}. \tag{8}$$

where $\phi_a(x)$ = number of occurrences of a occurs 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:

$$K_k^s(x_i, x_j) = \langle \Phi_k(x_i), \Phi_k(x_j)\rangle. \tag{9}$$

To fit in with HPV risk type classification, we want to modify the spectrum kernel. Proteins are linear chains of amino acids, which are made during the process of translation, and it is called primary structure. The natural shape of proteins are not such as straight lines, rather 3-dimensional structures formed by protein folding, which is a consequence of the primary structure. The structure of a similar homologous sequence can be helpful to identify the tertiary structure of the given sequence. Here, we assume that the amino acids pair with certain distance affect HPV's risk type function more than consecutive amino acids according to its 3-dimensional structure property, and the HPV risk types can be identified by the amino acids pair with a fixed distance, which mostly influence on risk type decision. This assumption is somewhat rough, but it can be useful for relatively short and α helix-dominant sequences.

Under the assumption, we want to define a string kernel, the gap-spectrum kernel based on k-spectrum. For a fixed k-mer $a = a_1 a_2 \ldots a_k$, $a_i \in \mathcal{A}$, 2-length sequence $\beta = a_1 a_k$, $\beta \in \mathcal{A}^2$. β indicates the amino acids pair with (k-2) gap. The feature map $\Psi_k(x)$ is defined as:

Table 1. The manually classified risk types of 72 HPVs

Type	Class	Type	Class	Type	Class	Type	Class
HPV1	Low	HPV20	Low	HPV38	Low	HPV57	?
HPV2	Low	HPV21	Low	HPV39	High	HPV58	High
HPV3	Low	HPV22	Low	HPV40	Low	HPV59	High
HPV4	Low	HPV23	Low	HPV41	Low	HPV60	Low
HPV5	Low	HPV24	Low	HPV42	Low	HPV61	High
HPV6	Low	HPV25	Low	HPV43	Low	HPV63	Low
HPV7	Low	HPV26	?	HPV44	Low	HPV65	Low
HPV8	Low	HPV27	Low	HPV45	High	HPV66	High
HPV9	Low	HPV28	Low	HPV47	Low	HPV67	High
HPV10	Low	HPV29	Low	HPV48	Low	HPV68	High
HPV11	Low	HPV30	Low	HPV49	Low	HPV70	?
HPV12	Low	HPV31	High	HPV50	Low	HPV72	High
HPV13	Low	HPV32	Low	HPV51	High	HPV73	Low
HPV15	Low	HPV33	High	HPV52	High	HPV74	Low
HPV16	High	HPV34	Low	HPV53	Low	HPV75	Low
HPV17	Low	HPV35	High	HPV54	?	HPV76	Low
HPV18	High	HPV36	Low	HPV55	Low	HPV77	Low
HPV19	Low	HPV37	Low	HPV56	High	HPV80	Low

$$\Psi_k(x) = (\phi_\beta(x))_{\beta \in \mathcal{A}^2}. \tag{10}$$

where $\phi_\beta(x) =$ number of occurrences of β occurs in x. Furthermore a nonlinear kernel function, RBF kernel is appended to increase the discrimination ability between HPV risk types. By closure properties of kernels [15], the gap-spectrum kernel is defined as follows:

$$K_k(x_i, x_j) = K'(\Psi_k(x_i), \Psi_k(x_j)) \tag{11}$$

$$= \exp\left(-\gamma \|\Psi_k(x_i) - \Psi_k(x_j)\|^2\right). \tag{12}$$

where $\gamma > 0$. This string kernel is used in combination with the SVM explained in Section 2.

4 Experimental Results

4.1 Data Set

In this paper, we use the HPV sequence database in Los Alamos National Laboratory (LANL) [16], and total 72 types of HPV are used for experiments. The risk types of HPVs were determined based on the HPV compendium (1997). If a HPV belongs to skin-related or cutaneous groups, the HPV is classified into low-risk type. On the other hand, a HPV is classified as a high-risk if it is known to be high-risk type for cervical cancer. The comments in LANL database are used to decide risk types for some HPVs, which are difficult to be classified. Seventeen sequences out of 72 HPVs were classified as high-risk types (16, 18,

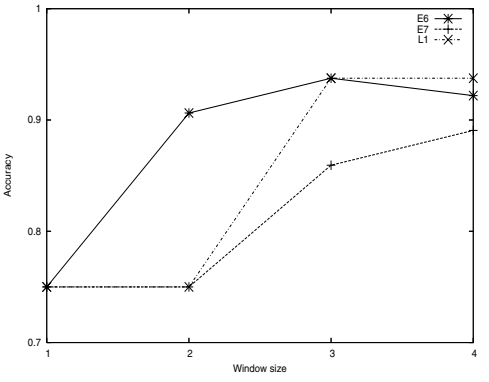

Fig. 1. SVM classification performance by window size

31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 61, 66, 67, 68, and 72), and others were classified as low-risk types. Table 1 shows the HPV types and their classified risk type. The symbol '?' in the table denotes unknown risk type that cannot be determined.

Since several proteins can be applied to discriminate HPVs, we have evaluated the classification accuracy using the SVM with RBF kernel to determine the gene products to be used for the experiments. The input data is the normalized frequency vector by sliding window method. It has been performed to decide the most informative protein among gene products for HPV risk type classification. Figure 1 depicts the accuracy changes by window size for E6, E7, and L1 proteins. The accuracy is the result of leave-one-out cross-validation. It indicates that the accuracy using E6 protein is mostly higher than using E7 and L1 proteins. However, the overall accuracy gets high by increasing window size for all proteins because the HPV sequences are relatively short and unique patterns are more generated when window size is long. That is, the learners overfit protein sequences for long window size. 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 [16][17]. This process plays an important role in the development of cervical cancer. For these reasons, we have chosen E6 protein sequences corresponding to the 72 HPVs.

4.2 Evaluation Measure

For the HPV prediction, it is important to get high-risk HPVs as many as possible, although a few low-risk HPVs are misclassified, hence we evaluate the system performance using F1-score rather than Matthews correlation coefficient. F1-score is a performance measure usually used for information retrieval systems, and it is effective to evaluate how well the classifier did when it assigned classes such as high-risk type.

Table 2. The contingency table to evaluate the classification performance

		Risk type answer	
		High	*Low*
Prediction result	*High*	a	b
	Low	c	d

When binary scales are used for both answer and prediction, a contingency table is established showing how the data set is divided by these two measures (Table 2). By the table, the classification performance is measured as follows:

$$precision = \frac{a}{a+b} \cdot 100\%$$
$$recall = \frac{a}{a+c} \cdot 100\%$$
$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall}.$$

4.3 HPV Classification

We have tested the gap-spectrum SVM method using E6 protein sequences. Leave-one-out cross-validation is used to determine the classification performance. Figure 2 shows the accuracy changes according to k given in Equation (12). The graph shows the accuracy has the highest performance (97.22%) when $k = 4$, and the performance is decreases as it gets more or less k. $k = 4$ means that we only use the amino acids pairs which have two gaps between them for HPV classification. $k = 2$ is exactly same as the SVM using RBF kernel with 2-spectrum method, and the accuracy is 94.44% for $k = 2$. Even though it gives higher score than other methods as shown in Figure 3, the kernel methods with $k > 2$ still gives better performance. As a result, the amino acids pair with a distance can provide more evidence than consecutive amino acids to discriminate low- and high-risk HPV proteins.

The final classification performance in accuracy and F1-score is given in Figure 3. It compares with previous results using SVM approaches based on sequence alignment and text mining approaches. The SVM method which utilizes alignment information has been reported in [10]. AdaCost and naïve Bayes are text mining methods using HPV literature data, which have been reported in [11]. Our approach shows 97.22% of accuracy and 95.00% of F1-score, while previous SVM method shows 93.15% of accuracy and 85.71% of F1-score. For text-based classification, the AdaCost method shows 93.05% of accuracy and 86.49% of F1-score, and the naïve Bayes method shows 81.94% of accuracy and 63.64% of F1-score. Additionally, the accuracy obtained from the DNA sequence-based method [9] is 85.64%. It is interesting that it gets relatively higher score in F1-score than in accuracy. F1-score is related with the number of high-risk HPVs found by classifiers, while accuracy is related with the number of HPVs which is correctly classified. Therefore, F1-score is more important than accuracy in this task.

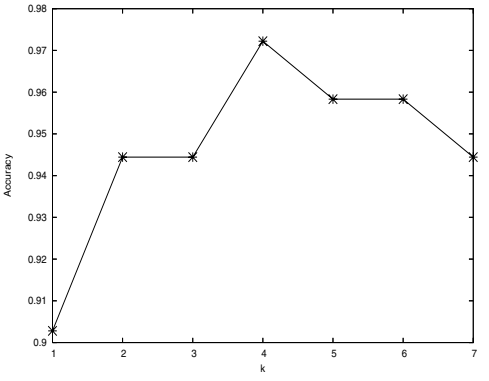

Fig. 2. Accuracy changes by the gap-spectrum kernel

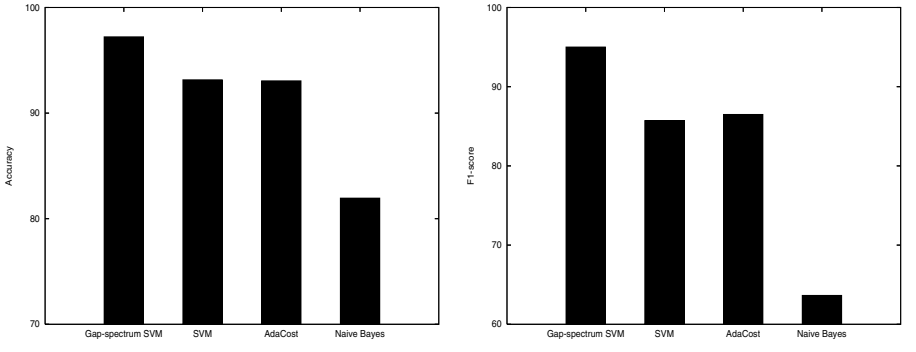

Fig. 3. The performance comparison of proposed approaches and previous classification methods

Text mining approaches only depend on the clues from text sentences. If the text documents are unavailable for unknown HPVs, there is no way to classify them, whereas the sequence-based classification does not need to use any additional information except sequence itself.

Table 3 shows the risk type prediction for HPVs marked as unknown in Table 1. HPV26, HPV54, HPV57, and HPV70 are predicted as high-, low-, low-, and high-risk, respectively. The prediction results for HPV26 and HPV54 are identical to the one in Muñoz et al. [6], and we assume that their results are correct because it is based on epidemiologic classification from over 1,900 patients. For HPV70, there are different decisions for the risk type according to previous research [6][18][19], and the risk type of HPV57 cannot be decided yet because of insufficient previous works. By the prediction results, we can conclude our approach provides certain probability for whether unknown HPVs are high-risk or not.

Table 3. Predicted risk type for unknown HPVs

Type	HPV26	HPV54	HPV57	HPV70
Risk	High	Low	Low	High

5 Conclusion

We have presented a machine learning approach to classify HPV risk types. Our method uses the SVM classifier with the gap-spectrum kernel based on k-spectrum methods. The proposed kernel is designed to emphasize amino acids pair with a fixed distance, which can be suitable for relatively short and α helix-dominant sequences. For the experiments, the performance has been measured based on leave-one-out cross-validation. According to experimental results, amino acids pair with a fixed distance provides good performance to discriminate HPV proteins by its risk. Especially, it is important not to have false negatives as many as possible in this task. Therefore F1-score is important because it considers both precision and recall based on high-risk type. Our approach shows significant improvement in F1-score as compared with previous methods, and the prediction for unknown HPV types has given promising results. We can conclude that the relationship between amino acids with $k = 4$ supports important role to divide low- and high-risk function in HPV E6 proteins.

In this paper, we consider all protein subsequences equally. Even though SVMs naturally detect the important factors in a high-dimensional space, it is necessary to analyze what components are more informative for HPV risk types. Also, protein structure or biological literature information can be combined with this method for more accurate prediction. Thus, study on exploring efficient analysis method remains as future works.

Acknowledgement

This work was supported by the Korea Ministry of Science and Technology through National Research Lab (NRL) project and the Ministry of Education and Human Resources Development under the BK21-IT Program. The ICT at the Seoul National University provided research facilities for this study.

References

[1] Bosch, F. X., Manos, M. M., et al.: Prevalence of Human Papillomavirus in Cervical Cancer: a Worldwide Perspective. *Journal of the National Cancer Institute* **87** (1995) 796–802.

[2] Janicek, M. F. and Averette, H. E.: Cervical Cancer: Prevention, Diagnosis, and Therapeutics. *Cancer Journals for Clinicians* **51** (2001) 92–114.

[3] Furumoto, H. and Irahara, M.: Human Papillomavirus (HPV) and Cervical Cancer. *Journal of Medical Investigation* **49** (2002) 124–133.

[4] Centurioni, M. G., Puppo, A., et al.: Prevalence of Human Papillomavirus Cervical Infection in an Italian Asymptomatic Population. *BMC Infectious Diseases* **5**(77) (2005).
[5] Burk, R. D., Ho, G. Y., et al.: Sexual Behavior and Partner Characteristics Are the Predominant Risk Factors for Genital Human Papillomavirus Infection in Young Women. *The Journal of Infectious Diseases* **174** (1996) 679–689.
[6] Muñoz, N., Bosch, F.X., et al.: Epidemiologic Classification of Human Papillomavirus Types Associated with Cervical Cancer. *New England Journal of Medicine* **348** (2003) 518–527.
[7] Watson, J. D., Laskowski, R. A., and Thornton, J. M.: Predicting Protein Function from Sequence and Structural Data. *Current Opinion in Structural Biology* **15** (2005) 275–284.
[8] Borgwardt, K. M., Ong, C. S., et al.: Protein Function Prediction via Graph Kernels. In *Proceedings of Thirteenth International Conference on Intelligenc Systems for Molecular Biology* (2005) 47–56.
[9] Eom, J.-H., Park, S.-B., and Zhang, B.-T.: 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) 1334–1343.
[10] Joung, J.-G., O, S.-J, and Zhang, B.-T.: 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) 723–731.
[11] Park, S.-B., Hwang, S., and Zhang, B.-T.: Mining the Risk Types of Human Papillomavirus (HPV) by AdaCost. In *Proceedings of the 14th International Conference on Database and Expert Systems Applications* (2003) 403–412.
[12] Vapnik, V. N.: Statistical Learning Theory. Springer (1998).
[13] Leslie, C., Eskin, E., and Noble, W. S.: The Spectrum Kernel: A String Kernel for SVM Protein Classification. In *Proceedings of the Pacific Symposium on Biocomputing* (2002) 564–575.
[14] Leslie, C., Eskin, E., et al.: Mismatch String Kernels for Discriminative Protein Classification. *Bioinformatics* **20**(4) (2004) 467–476.
[15] Shawe-Taylor, J. and Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004).
[16] The HPV sequence database in Los Alamos laboratory, http://hpv-web.lanl.gov/stdgen/virus/hpv/index.html.
[17] Pillai, M., Lakshmi, S., et al.: High-Risk Human Papillomavirus Infection and E6 Protein Expression in Lesions of the Uterine Cervix. *Pathobiology* **66** (1998) 240–246.
[18] Longuet, M., Beaudenon, S., and Orth, G.: Two Novel Genital Human Papillomavirus (HPV) Types, HPV68 and HPV70, Related to the Potentially Oncogenic HPV39. *Journal of Clinical Microbiology* **34** (1996) 738–744.
[19] Meyer, T., Arndt, R., et al.: Association of Rare Human Papillomavirus Types with Genital Premalignant and Malignant Lesions. *The Journal of Infectious Diseases* **178** (1998) 252–255.