Genetic Mining of DNA Sequence Structures for Effective Classification of the Risk Types of Human Papillomavirus (HPV)

Jae-Hong Eom[1], Seong-Bae Park[2], and Byoung-Tak Zhang[1]

[1] Biointelligence Lab., School of Computer Science and Engineering,
Seoul National University, Seoul 151-744, South Korea
{jheom,btzhang}@bi.snu.ac.kr
[2] Language & Information Processing Lab., Dept. of Computer Engineering,
Kyungpook National University, Daegu 702-701, South Korea
seongbae@knu.ac.kr

Abstract. Human papillomavirus (HPV) is considered to be the most common sexually transmitted disease and the infection of HPV is known as the major factor for cervical cancer. There are more than 100 types in HPV and each HPV has two risk types, low and high. In particular, high risk type HPV is known to the most important factors in medical judgment. Thus, the classifying the risk type of HPV is very important to the treat of cervical cancer. In this paper, we present a machine learning approach to mine the structure of HPV DNA sequence for effective classification of the HPV risk types. We learn the most informative subsequence segment sets and its weights with genetic algorithm to classify the risk types of each HPV. To resolve the problem of computational complexity of genetic algorithm we use distributed intelligent data engineering platform based on active grid concept called "IDEA@Home." The proposed genetic mining method, with the described platform, shows about 85.6% classification accuracy with relatively fast mining speed.

1 Introduction

Cervical cancer is a leading cause of cancer deaths in women worldwide. It is well established that persistent infection with the Human Papillomavirus (HPV) is associated cervical cancer. Large studies have shown that HPV is present in up to 90% of cervical cancers [1][2]. Since the main etiologic factor for cervical cancer is known as high-risk Human Papillomavirus infection [3], it is now largely a preventable disease [4]. This Human Papillomavirus is one of the most common sexually transmitted diseases and the infection of HPV is still known as the major factor for cervical cancer. This HPV is a 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 by their relative malignant potential as low-, intermediate-, and high-risk types [5]. The common, unifying oncogenic feature of the vast majority of cervical cancers is the presence of HPV, especially high-risk type HPV.

N.R. Pal et al. (Eds.): ICONIP 2004, LNCS 3316, pp. 1334–1343, 2004.

Since the HPV classification is important in medical judgments and it is becoming more important, there have been many approaches to classify the risk types of HPVs. Bosch *et al*. [1] investigated whether the association between HPV infection and cervical cancer is consistent worldwide in the context of geographic variation in the distribution of HPV types. Burk *et al*. [6] inspected the risk factors for HPV infection in 604 young college women and they detected various factors of HPV infection (e.g., age, ethnicity, number of lifetime male vaginal sex partners, etc.) through L1 consensus primer polymerase chain reaction and Southern blot hybridization. Park *et al*. [4] used text mining technique to discriminate the risk types of HPVs and they predicted the risk types of several HPVs whose risk types were have been unknown. Muñoz *et al*. [7] classified the risk types with practical experiments based on risk factor analysis. They collected real data from 1,900 cervical cancer patients and analyzed it by PCR (polymerase chain reaction) based assays.

Practical analysis with experiment is the most accurate analysis process. However, it is not easy to conduct every experimental trial when we have many cases to analyze. One alternative for this problem is exploiting computational power to the analysis. The "systems biology" is the field which makes use of this approach to analyze and understand biological phenomena through systems approaches.

In this paper, we present a machine learning approach to mine the structure of HPV DNA sequence for effective classification of the HPV risk types. We learn the most informative subsequence segments and its weights with genetic algorithm to classify the risk types of each HPV. Also we use new data engineering platform based on active grid computing, called "IDEA@Home," to alleviate the problem of computational complexity of genetic algorithm [8].

The remainder of the paper is organized as follows. In Section 2, we simply describe the concept of the distributed intelligent data engineering platform which was used for proposed mining algorithm. Section 3 represents the genetic mining method for learning substructure and its weights of HPV DNA subsequences. Section 4 presents the experimental results. Finally, Section 5 draws conclusions.

2 IDEA@Home: Intelligent Data Engineering Platform

As the data size is becoming increasingly large, more powerful computational ability is needed. One possible alternative for this problem is utilizing distributed computing and there have been many attempts to employ this approach for analyzing high dimensional data which called "Grid Computing." Grid computing is distributed computing, in which a network of computers taps into a main computer server that stores software and data.

One of the most famous grid computing projects is the SETI@Home which aims to analyze radio telescope data to search extraterrestrial intelligence. Grid computing for biological analysis also started. The FightAIDS@Home by the Olson laboratory at the Scripps Research Institute is the first biomedical distributed computing project based on grid computing to discover new drugs, using the growing knowledge of the structural biology of AIDS. The Folding@Home from Stanford University is designed to understand protein folding and related diseases. The "Screensaver Lifesaver" projects from research group of W. Graham Richards has developed and ap-

plied computational methods for drug design, molecular similarity analysis and protein structure prediction, and performed simulations of enzyme reaction mechanisms, DNA recognition, and lipid bilayers [9][10].

But, many conventional grid systems are designed to solve only domain-specific problems. Usually they have one master node controlling the whole network and it distributes all information needed for computation. However, the platform IDEA@Home proposed by Eom *et al*. [8] allows *n* master nodes and each master divides the entire network into *n* sub-networks. The platform also supports computational flexibility through the concept of operation object and data object. The operation object represents the task-specific operation which will be applied to the data object for the analysis and the data object presents the task-specific data. This data object also includes some constraint information which will be referenced by operation object [8].

The GA based method proposed in this paper, actually, is suitable for both stand-alone computing platform and distributed computing platforms. But, in the stand alone computing platform, general approach of GA require lots of computing time for the analysis of relatively big and complex data. And the data from biological domain usually have these complex characteristics. Thus the grid platform, such as "IDEA@Home," is somewhat useful for these kinds of fields.

In this paper, since the proposed genetic mining method requires relatively heavy computation, we use the IDEA@Home grid computing platform for computational efficiency.

3 Genetic Mining of HPV Sequence Structures

In this paper, we use genetic algorithms (GAs) for mining discriminative sequence sets of HPVs and classifying the risk types of HPVs. Although there are many fast methods (e.g., sequence alignment, SVM and NN, etc.) and GAs are still somewhat time-consuming, GAs are easily implementable in distributed fashion and they also have great potentials for improved search performance of solution space search. Thus, here we use GAs for sequence structure mining (other method will be considered in the future works).

Encoding

Genetic algorithm is a probabilistic search method based on the mechanism of natural selection and genetics [11]. Genetic search is characterized by the fact that *N* potential solutions for an optimization problem simultaneously sample the search space. These solutions are called *individuals* or *chromosomes* and denoted as $J_i \in \mathbf{J}$, where $\mathbf{J}$ represents the space of all possible individuals. The *population* $\vec{J} = \{ J_1, J_2, \ldots, J_N \} \in \mathbf{J}^N$ is modified according to the natural evolutionary process with predefined modification rates (i.e., crossover and mutation rates). After the initial population is generated arbitrarily, selection ω: $\mathbf{J}^N \Rightarrow \mathbf{J}^N$ and variation Ξ: $\mathbf{J}^N \Rightarrow \mathbf{J}^N$ are applied in a loop until some termination criterion is satisfied. Each run of the loop is called generation [12].

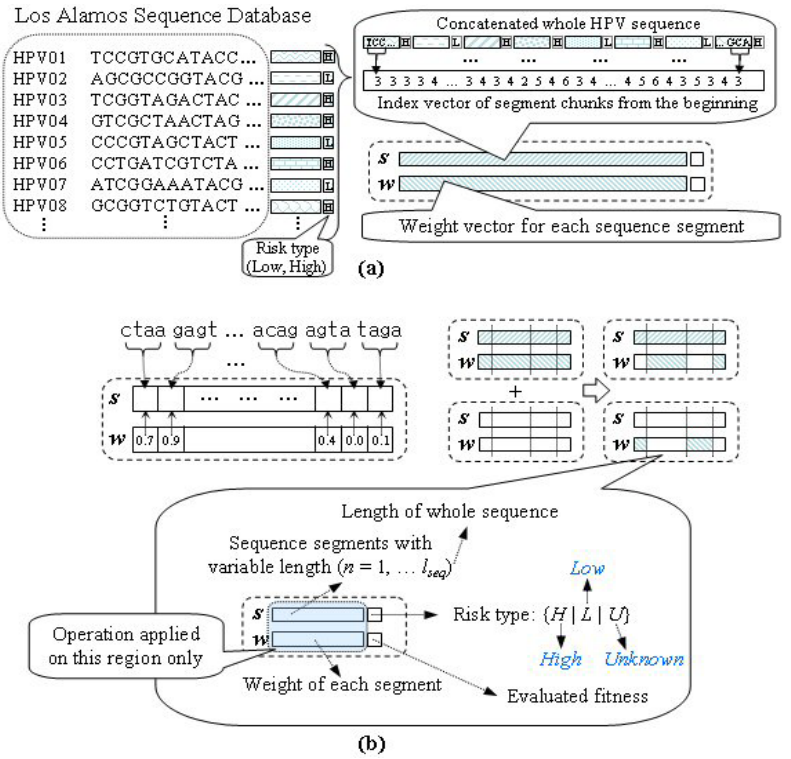

Fig. 1. The encoding scheme of individual chromosome on HPV sequences. Figure (a) shows the encoding scheme of each individual. Each individual in the population has different index vector and its corresponding weight vector. ***S*** is a vector of subsequences and ***W*** is a vector of weights of subsequences in the segment vector ***S***. An encoding example is described in (b). In this example, all segment length is equally set to 3. Figure (b) shows the schematic working mechanism of 3-point crossover.

The chromosome for HPV classification is defined as a pair of two vectors: a vector for subsequence segments and the other for their weights. The weight of chromosome is a real number ranging from 0.0 to 1.0 and the weight represents the importance of corresponding segment when classifying HPVs. This chromosome is represented as:

$$\mathbf{J} = \begin{pmatrix} j_{s_1}, j_{s_2}, \ldots, j_{s_i}, \ldots, j_{s_L} \\ j_{w_1}, j_{w_2}, \ldots, j_{w_i}, \ldots, j_{w_L} \end{pmatrix} \tag{1}$$

where j_{si} denotes the i-th subsequence segment of DNA sequence of HPV*nn* and j_{wi} denotes the weight of corresponding subsequence segment. '*nn*' corresponds to the index of HPV. L is the number of segments to be considered for classification and is determined by dividing the length of sequence n by the length of segment ρ. Figure 1 shows the representation of chromosome and the mechanism of simple n-point crossover (n=3). Generally, GAs use crossover and mutation as variation functions Ξ: $\mathbf{J}^N \Rightarrow \mathbf{J}^N$. For crossover, only the fractions of weights are switched in alternative way

based on n-cutting points for selected chromosome pairs. Cutting points are chosen with crossover probability p_c and the uniform random variable sampled again for each point. For mutation, the weight of randomly selected position of chromosome is changed with mutation probability p_m ranging from 0.0 to 1.0. These two operations maintain the diversity of the population.

In this paper, each chromosome represents a subsequence segment with various length and its weights. This is for classifying the corresponding HPV to high or low risk types. The index vector of segment chunks in an individual represents the granularity of sub sequences. For example, an index vector "3 3 4" represents that the chromosome is constructed with successive subsequences with length 3, 3, and 4. In addition, each offspring represents the ordering information of sequence segments through the weights of each segment. A weight represents consideration priority or the importance of a segment for the classification of HPV risk types.

Fitness Function

The fitness function f: $\mathbf{J} \Rightarrow \mathbf{R}$ measures the fitness of a chromosome in terms of classification performance. In this paper, we evaluate the fitness of each offspring with its classification accuracy, which is defined as following:

$$Accuracy = \frac{a+d}{a+b+c+d} \cdot 100\% \tag{2}$$

where a, b, c and d are defined in Table 1. In the proposed approach, the fitness function represents the classification accuracy of each offspring. Answer set was obtained from Los Alamos sequence database [13]. The risk type described in this database is used as an answer set to calculate the classification accuracy. The calculation of the classification accuracy in Equation 2 is based on this information. That is; "answer should be…" are measured according to this answer and the "test results" are the classification result according to the encoding of each individual.

The tag of risk types of each individual is assigned at training step to find the most informative subsequence segments in the whole sequence and to evaluate the fitness value. The estimated fitness value is stored in another tag and this value is used when we select appropriate offspring in the selection procedure of GAs.

The overall procedure of proposed genetic mining method is described in Figure 2. The genetic mining procedure described in Figure 2 is a stand-alone system version.

4 Experimental Results

Datasets

In this paper, we use the HPV sequence database in Los Alamos National Laboratory as a dataset [13]. This database includes HPV compendiums published in 1994–1997 and provides the complete list of 'papillomavirus types and hosts' and the records for each unique papillomavirus types.

Table 1. The contingency table to evaluate the classification performance.

		Test results	
		Low	*High*
Risk type	Answer should be *Low*	*a*	*b*
	Answer should be *High*	*c*	*d*

Do following procedures with given parameters:
- Chromosome set $\mathbf{J} = \{ J_1, J_2, \ldots, J_N \}$
- Crossover probability p_c, Mutation probability p_m
- Number of maximum generation *gmax*
- Number of population to select in each generation *M*

1. **Set** all chromosomes with randomly generated initial values.
   - Initialize sequence index vectors.
   - Initialize sequence weight vectors.
2. **Repeat** Step 2.1 to 2.3 for all i, $i = 1$ to *gmax*.
   2.1 **Evaluate** all chromosomes by fitness function *f*.
   2.2 **Repeat** Step 2.2.1 to 2.2.3 for all j, $j = 1$ to *M*.
      2.2.1 **Select** two chromosomes J_a and J_b.
      2.2.2 **Set** offspring[j] = Crossover (J_a, J_b).
      2.2.3 **Set** offspring[j] = Mutation (offspring[j]).
   2.3 **Replace** *M* chromosomes by offspring generated from Step 2.2.
3. **Return** the optimal chromosome: $\mathbf{J}_{opt}$

Fig. 2. The general procedure of genetic mining. Step 2 represents one generation of generational GAs and it repeats until the population satisfies some predefined convergence criterion.

Settings

To measure the fitness of each individual in genetic mining we use the table of manually classified HPVs as a correct answer, which was previously constructed by Park *et al*. [4] (Table 2) and we used the parameters for genetic mining described in Table 3. The HPVs with "don't know" types are classified according to the discovered subsequence segments and its weights learned from the remaining 72 HPVs. The classification accuracy is measured by Equation 2. This experimental result was calculated through 10-fold cross validation test. For cross validation, the whole dataset is divided into disjoint 10 bins and used as train and test datasets (leave-one-out validation was used; 9 as train, 1 as test dataset).

Results

Muñoz *et al*. classified the types of HPV based on epidemiologic classification method [7]. They pooled data from more than 1,900 patients who have cervical cancer. They detected HPV DNA and assigned type by polymerase-chain-reaction-based

Table 2. The manually classified risk types of 76 HPVs. The "D/K" mean "don't know." There are 18 HPVs with high risk types and 4 HPVs with "don't know" types. This classification of total 76 HPVs is based on the 1997 version of HPV compendium. The classifications of HPV [46, 71, 78, 79] are missing from the table due to the lack of its research data (Park *et al.* [4]).

HPVs	Type	HPVs	Type	HPVs	Type	HPVs	Type
HPV01	Low	HPV20	Low	HPV39	High	HPV59	High
HPV02	Low	HPV21	Low	HPV40	Low	HPV60	Low
HPV03	Low	HPV22	Low	HPV41	Low	HPV61	High
HPV04	Low	HPV23	Low	HPV42	Low	HPV62	High
HPV05	Low	HPV24	Low	HPV43	Low	HPV63	Low
HPV06	Low	HPV25	Low	HPV44	Low	HPV64	Low
HPV07	Low	*HPV26*	D/K	HPV45	High	HPV65	Low
HPV08	Low	HPV27	Low	HPV47	Low	HPV66	High
HPV09	Low	HPV28	Low	HPV48	Low	HPV67	High
HPV10	Low	HPV29	Low	HPV49	Low	HPV68	High
HPV11	Low	HPV30	Low	HPV50	Low	HPV69	Low
HPV12	Low	HPV31	High	HPV51	High	*HPV70*	D/K
HPV13	Low	HPV32	Low	HPV52	High	HPV72	High
HPV14	Low	HPV33	High	HPV53	Low	HPV73	Low
HPV15	Low	HPV34	Low	*HPV54*	D/K	HPV74	Low
HPV16	High	HPV35	High	HPV55	Low	HPV75	Low
HPV17	Low	HPV36	Low	HPV56	High	HPV76	Low
HPV18	High	HPV37	Low	*HPV57*	D/K	HPV77	Low
HPV19	Low	HPV38	Low	HPV58	High	HPV80	Low

Table 3. Parameters for genetic mining.

	Parameters				
	Pop. size N	Mutation rate P_m	Crossover rate P_c	Replace rate M	Segment length (ρ)
Value	2,000	0.3	0.5	50%	4

assays which are considered as a relatively accurate detection method in current literatures. In their experiment, both HPV54 and HPV70 were classified as low risk types. From our genetic mining method, we also classified these two HPVs as low risk types. This result is identical to the result of Muñoz *et al*. for HPV [54, 70]. Currently, there are insufficient research results on the HPVs which have undetermined risk types, HPV [26, 54, 57, 70], to decide whether the classification result is correct or not. But, from the experimental result of Muñoz *et al.* in comparison with the result of Park *et al.*, we can conclude that our genetic mining method has more ability than previous text mining method of Park *et al*. [4] for capturing the genetic and the structural sequence characteristics of HPVs when we assume that the results of Muñoz *et al.* are correct.

Table 4. Predicted risk types of the HPVs whose risk types are known as "don't know." Note that the HPV70 is classifyed as low risk type which was classified as high-risk type in the previous research result of Park *et al*. [4].

Type	HPV26	HPV54	HPV57	HPV70
Risk	Low	Low	Low	Low

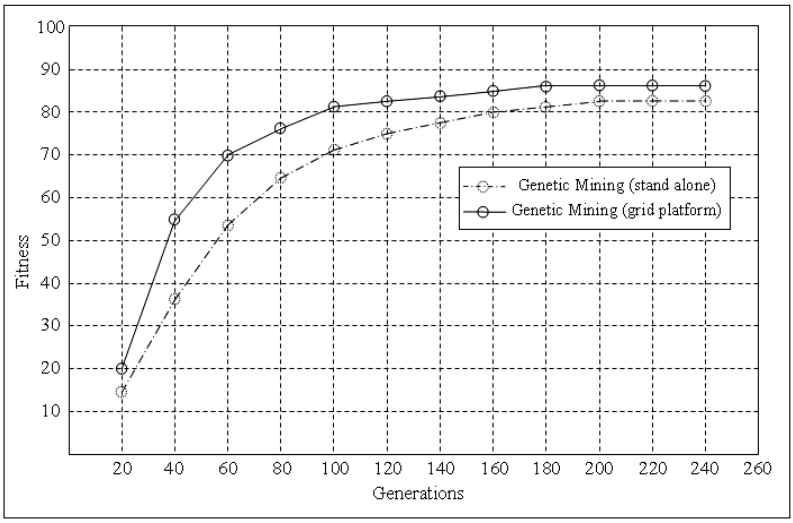

Fig. 3. The average fitness of genetic mining of HPV sequence structures. This figure shows the tendency that genetic mining on the grid platform converges relatively faster than the genetic mining on a single platform (represented as "stand alone"). The maximum fitness of genetic mining on the grid platform was 85.6%. The fitness means the classification accuracy of HPV risk types on sequence data. Whole experiment was conducted five times and averaged for each plot of the graph.

The proposed genetic mining method showed about 85.6% classification accuracy for 72 HPVs in Table 2. This result outperforms that of decision tree based text mining (it was about 81.14%) of Park *et al*. [4] by 4.5%. The population was converged to this accuracy at least within 250 generations which takes about 8.1 hours on described grid platform. Figure 3 shows the fitness convergence graph of proposed genetic mining method on both stand-alone and grid based platform. The graph of genetic mining on the grid platform shows relatively fast converge tendency.

To evaluate the relative success of the proposed genetic mining approach, we compared the classification performance with the result of our previous works [4] which use decision tree as a HPV classification method. And we also implemented Naïve Bayes classifier as a baseline model to compare the classification results. These classification results are described in Table 5. Both classifications with GA in stand-alone system and with GA in grid platform show improved classification accuracy than the previous decision tree based classification result. These GA based method outperform the baseline model, Naïve Bayes, by about 6 to 8 percents.

Table 5. Accuracy of the baseline model (Naïve Bayes), decision tree, and proposed approaches. Decision tree was used in the text ming approach of Part *et al.* [4] as a base classifier. Genetic mining based approches outperformed other two methods in term of classifiaction accuracy.

Method	Naïve Bayes	Decision Tree	GA (stand alone)	GA (GRID)
Accuracy	77.18	81.14	83.93	85.64

5 Conclusions

In this paper, we proposed genetic mining method to classify the risk types of HPVs. The proposed method achieved the improvement in both classification accuracy and processing time (the processing time is compared to the stand-alone system). In particular, the classification results of the HPVs whose risk types were known as "don't know" show coincidence with the recent HPV classification results based on practical experiments. This result based on the genetic mining is different from the previous classification results based on decision tree on the text description of each HPV. Thus, we can conclude that the proposed method is appropriate to analyze biological sequences for classification. Also, the proposed platform cut down the computing time from 22.9 (in stand-alone system) hours to just 8.1 hours. Also, the convergence time of GA is relatively faster in the grid platform than in stand-alone system. We assume that this is maybe due to the 'randomness' in applying genetic operators in distributed platform and also due to relatively powerful computational power. More detailed investigation on this issue will be conducted in the future works.

Moreover, HPVs have many mutants and classification on the risk types of these mutated HPVs is important for medical remedy. Thus, study on the new method for classifying these mutants and study on the more efficient encoding scheme to exploit the proposed genetic mining method with theoretical analysis remain as future works.

Acknowledgements

This research was supported by the Korean Ministry of Science and Technology under the NRL Program and the Systems Biology Program.

References

1. Bosch, F.X., Manos, M.M., Muñoz, N., Sherman, M., Jansen, A.M., Peto, J., Schiffman, M.H., Moreno, V., Kurman, R., Shah, K.V.: Prevalence of human papillomavirus in cervical cancer: a worldwide perspective. *J Natl Cancer Inst* **87** (11) (1995) 796-802.
2. Furumoto, H. and Irahara, M.: Human Papillomavirus (HPV) and Cervical Cancer. *Journal of Medical Investigation* **49** (2002) 124-33.
3. Schiffman, M., Bauer, H., Hoover, R., Glass, A., Cadell, D., Rush, B., Scott, D., Sherman, M., Kurman, R., and Wacholder, S.: Epidemiologic evidence showing that Human Papillomavirus infection causes most cervical intraepithelial neoplasis. *Journal of the National Cancer Institute* **85** (1993) 958-64.
4. Park, S.-B., Hwang, S.-H., and Zhang, B.-T.: Classification of the risk types of human papillomavirus by decision trees. In *Proceedings of the 4th International Conference on Intelligent Data Engineering and Automated Learning* (2003) 540–44.
5. Janicek, M.F. and Averette, H.E.: Cervical cancer: prevention, diagnosis, and therapeutics. *Cancer Journals for Clinicians* **51** (2001) 92-114.
6. Burk, R.D., Ho, G.Y., Beardsley, L., Lempa, M., Peters, M., and Bierman, R.: Sexual behavior and partner characteristics are the predominant risk factors for genital human papillomavirus infection in young women. *J Infect Dis*. **174**(4) (1996) 679-89.

7. Muñoz, N., Bosch, F.X., Sanjosé, S., Herrero, R., Castellsagué, X., Shah, K.V., Snijders, P.J.F., and Meijer, C.J.L.M.: Epidemiologic classification of human papillomavirus types associated with cervical cancer. *The New England Journal of Medicine* **348**(6) (2003) 518-27.
8. Eom, J.-H. and Zhang, B.-T.: IDEA@home: The flexible active grid computing platform based on P2P and network segmentation. *Technical Report BI-04-01*, School of Computer Sci.&Eng., Seoul National Univ., Seoul, Korea, February (2004).
9. Richards, W.G.: Virtual screening using grid computing: the screensaver project. *Nature Reviews Drug Discovery* **1** (2002) 551-55.
10. Davies, E.K., Glick, M., Harrison, K.N., and Richards, W.G.: Pattern recognition and massively distributed computing. *Journal of Computational Chemistry* **23**(16) (2002) 1544-50.
11. Bäck, T., Evolutionary algorithms in theory and practice, *Oxford University Press*. (1996)
12. Kim, S. and Zhang B.-T.: Genetic mining of HTML structures for effective web-document retrieval. *Applied Intelligence* **18** (2003) 243–56.
13. The HPV sequence database in Los Alamos laboratory. http://hpv-web.lanl.gov/stdgen/virus/ hpv/index.html.