Classification of the Risk Types of Human Papillomavirus by Decision Trees

Seong-Bae Park, Sohyun Hwang, and Byoung-Tak Zhang

School of Computer Science and Engineering
Seoul National University
Seoul 151-744, Korea
{sbpark,shhwang,btzhang}@bi.snu.ac.kr

Abstract. The high-risk type of Human Papillomavirus (HPV) is the main etiologic factor of cervical cancer, which is a leading cause of cancer deaths in women worldwide. Therefore, classifying the risk type of HPVs is very useful and necessary to the daignosis and remedy of cervical cancer. In this paper, we classify the risk type of 72 HPVs and predict the risk type of 4 HPVs of which type is unknown. As a machine learning method to classify them, we use decision trees. According to the experimental results, it shows about 81.14% of accuracy.

1 Introduction

Cervical cancer is a leading cause of cancer deaths in women worldwide. Since the main etiologic factor for cervical cancer is known as high-risk Human Papillomavirus (HPV) infection [5], it is now largely a preventable disease. There are more than 100 types of HPV that are specific for epithelial cells including skin, respiratory mucosa, and the genital tract. Genital tract HPV types are classified by their relative malignant potential into low-, and high-risk types [4]. The common, unifying oncogenic feature of the vast majority of cervical cancers is the presence of high-risk HPV. Therefore, the most important thing for diagnosis and therapy is discriminating whether patients have the high-risk HPVs and what HPV types are highly risky.

One way to discriminate the risk types of HPVs is using a text mining technique. Since a great number of research results on HPV have been already reported in biomedical journals [3], they can be used as a source of discriminating HPV risk types. Since there are a number of research results on HPV and cervical cancer, the textual data about them can be easily obtained. In this paper, we use the textual information describing the characteristics of various HPV types as document data, and decision trees as a learning algorithm to classify the their risk types. One advantage of this work is usefulness for designing the DNA-chip, diagnosing the presence of HPV in cervical cancer patients. Since there are about 100 HPV types, making the DNA chip needs to choose some dangerous ones related with cervical cancer among them. Therefore, this result classifying the risk of HPVs can be a big help to save time to understand information about HPV and cervical cancer from many papers and to choose the HPV types used for DNA chip.

J. Liu et al. (Eds.): IDEAL 2003, LNCS 2690, pp. 540–544, 2003.

<definition>
Human papillomavirus type 43 (HPV-43) E6 region.
</definition>
<source>
Human papillomavirus type 43 DNA recovered from a vulvar biopsy with hyperplasia.
</source>
<comment>
HPV-43 was classified by Lorincz et. al [435] as a "low-risk" virus. Prevalence studies indicate that HPV-44 and HPV-43 have been found in 4% of cervical intraepithelial neoplasms, but in none of the 56 cervical cancers tested. During an anlaysis of approximately 100 anogenital tissue samples, two new HPV types, HPV-43 and HPV-44, were identified. The complete genome of HPV-43 was recovered from a vulvar biopsy and cloned into bacteriophage lambda. The biopsy was taken from a woman living in the Detroit Michigan area. The DNA consisted of two fragments: a 6.3 kb BamHI fragment and a 2.9 kb HindIII fragment. The total quantity of unique DNA was 7.6 kb. Only the E6 region of the cloned sample has been sequenced, although all positions of the ORFs have been deduced and are consistent with the organization of DNA from HPV-6b. A possible feature of HPV types associated with malignant lesions is the potential to produce a different E6 protein by alternative splicing. This potential has been found in types HPV-16, HPV-18, and HPV-31. HPV-43 has both the potential E6 splice donor site at nt 233 and the potential splice acceptor at nt 413.
</comment>

Fig. 1. An example description of HPV43 from HPV sequence database.

2 Dataset

In this paper, we use *the HPV Sequence Database* in Los Alamos National Laboratory as a dataset. This papillomavirus database is an extension of the HPV compendiums published in 1994–1997 and provides the complete list of 'papillomavirus types and hosts' and the records for each unique papillomavirus type. An example of the data made from this database is given in Figure 1. This example is for HPV43 and consists of three parts: `definition`, `source`, and `comment`. The `definition` indicates the HPV type, the `source` explains where the information for this HPV is obtained, and the `comment` gives the explanation for this HPV. In the all experiments below, we used only `comment`. The comment for a HPV type can be considered as a document in text classification. Therefore, each HPV type is represented as a vector of which elements are $tf \cdot idf$ values. When we stemmed the documents using the Porter's algorithm and removed words from the stop-list, the size of vocabulary is just 1,434. Thus, each document is represented as a 1,434-dimensional vector.

To measure the performance of the results in the experiments below, we manually classified HPV risk types using the 1997 version of HPV compendium and the comment in the records above. The classifying procedure is as follows. First, we divided roughly HPV types by the groups in the compendium. Second,

Table 1. The manually classified risk types of HPVs.

Type	Risk	Type	Risk	Type	Risk	Type	Risk
HPV1	Low	HPV2	Low	HPV3	Low	HPV4	Low
HPV5	Low	HPV6	Low	HPV7	Low	HPV8	Low
HPV9	Low	HPV10	Low	HPV11	Low	HPV12	Low
HPV13	Low	HPV14	Low	HPV15	Low	HPV16	High
HPV17	Low	HPV18	High	HPV19	Low	HPV20	Low
HPV21	Low	HPV22	Low	HPV23	Low	HPV24	Low
HPV25	Low	HPV26	Don't Know	HPV27	Low	HPV28	Low
HPV29	Low	HPV30	Low	HPV31	High	HPV32	Low
HPV33	High	HPV34	Low	HPV35	High	HPV36	Low
HPV37	Low	HPV38	Low	HPV39	High	HPV40	Low
HPV41	Low	HPV42	Low	HPV43	Low	HPV44	Low
HPV45	High	HPV47	Low	HPV48	Low	HPV49	Low
HPV50	Low	HPV51	High	HPV52	High	HPV53	Low
HPV54	Don't Know	HPV55	Low	HPV56	High	HPV57	Don't Know
HPV58	High	HPV59	High	HPV60	Low	HPV61	High
HPV62	High	HPV63	Low	HPV64	Low	HPV65	Low
HPV66	High	HPV67	High	HPV68	High	HPV69	Low
HPV70	Don't Know	HPV72	High	HPV73	Low	HPV74	Low
HPV75	Low	HPV76	Low	HPV77	Low	HPV80	Low

Table 2. The performance of decision trees.

Trial	**Accuracy** (%)
1	86.7
2	86.7
3	66.7
4	73.3
5	92.3
Average	81.14 ± 10.68

if the group is skin-related or cutaneous, the members of the group are classified into low-risk type. Third, if the group is known to be cervical cancer-related HPV, the members of the group are classified into high-risk type. Lastly, we used the comment of HPV types to classify some types difficult to be classified. Table 1 shows the summarized classification of HPVs according to its risk. "Don't know"s in this table are the ones that can not be classified by above knowledge.

3 Experimental Results

Since we have only 76 HPV types and the explanation of each HPV is relatively short, *5-fold cross validation* is used to determine the performance of decision trees. That is, in each experiment, we used 58 examples as a training set, and the remaining 15 as a test set. As a learning algorithm of decision trees, Quinlan's C4.5 release 8 is used.

Table 2 shows the performance of decision trees. The average accuracy is $81.14 \pm 10.68\%$. Among the misclassified HPV types, 9 low-risk HPVs are classified as high-risk, and 5 high-risk HPVs are classified as low-risk. Four (HPV13, HPV14, HPV30, and HPV40) of 9 low-risk HPVs that are misclassied as high-risk have a potential to cause a laryngeal cancer, though they are not directly

Table 3. The misclassifed HPVs.

HPV Type	Real Type	Predicted Type	Characteristics
HPV2	Low	High	normal wart
HPV13	Low	High	oral infection, some progress to cancer
HPV14	Low	High	some progress to cancer
HPV18	High	Low	
HPV30	Low	High	laryngeal cancer
HPV40	Low	High	laryngeal cancer
HPV42	Low	High	genital wart
HPV43	Low	High	
HPV44	Low	High	
HPV53	Low	High	
HPV56	High	Low	
HPV59	High	Low	
HPV62	High	Low	mid-high-risk
HPV72	High	Low	mid-high-risk

Table 4. Risk type of the HPVs whose risk type is known as 'Don't Know'.

Trial	HPV26	HPV54	HPV57	HPV70	Accuracy (%)
1	Low	Low	Low	Low	75
2	Low	Low	High	Low	50
3	Low	Low	High	Low	50
4	Low	Low	Low	Low	75
5	High	Low	High	Low	25

related with cervical cancer. And, three (HPV59, HPV62, and HPV72) of 5 high-risk HPVs which are misclassified as low-risk are the ones which some clinical researchers classify them as mid-high-risk types. That is, they do not have many expression about cervical cancer, though we classify them as high-risk to make a problem simple. Therefore, if we exclude these 7 cases, the errors are only 5 low-risk HPVs (HPV2, HPV42, HPV43, HPV44, and HPV53) and 2 high-risk HPVs (HPV18 and HPV56). Table 3 summarizes these errors.

One point we should emphasize in classifying the risk-type of HPVs is that false negatives are far more important than false positives. That is, it is no problem to misclassify low-risk HPVs as high-risk, but it is fatal to misclassify high-risk HPVs as low-risk. In false negative case, dangerous HPVs are missed, and there is no further chance to detect cervical cancer by them. Because only 2 are false negatives in our experiments, the results are reliable.

Four HPVs in Table 1 are classified as "Don't Know" since their risk type is not certain. According to the other research results [1,2], their risk types are as follows.

Type	HPV26	HPV54	HPV57	HPV70
Risk	Low	Low	Low	High

When we classify their risk types by the decision trees, we obtain $55 \pm 20.92\%$ of accuracy. Table 4 summarizes this results. Especially, HPV70 is not correctly

classified at all. This is because the comment for HPV70 does not describe its risk but explains that it is found at the cervix of patients and its sequence is analyzed.

4 Conclusions

In this paper, we classified the risk type of HPVs by decision trees. The accuracy is about 82%, which is a little bit higher than expected. But, if we are to use this result as not reference material but basic information before biomedical experiments, we need higher accuracy. In addition, it is no problem to classify low-risk HPVs as high-risk, but it is fatal to classify high-risk HPVs as low-risk. If we have about 20% of errors, we come to have too many misclassified high-risk HPVs. If the DNA-chips are designed only based on this result, some of them will cause misdiagnosis. Therefore, we suggest this result as reference material to design DNA-chips.

Because the virus is easy to mutate in order to survive within the host, most viruses have various types like HPVs and each type has different effect on serious diseases. Therefore, it is required to keep studying on classifying mutants of viruses like HPVs according to their relation with diseases. In case of HPV, it is an easy work because there are a number of research results and databases. But we need to study further how to classify the risk type of viruses when there is not such data available.

Acknowledgements. This research was supported by the Korean Ministry of Education under the BK21-IT Program, by BrainTech and NRL programs sponsored by the Korean Ministry of Science and Technology.

References

1. S. Chan, S. Chew, K. Egawa, E. Grussendorf-Conen, Y. Honda, A. Rubben, K. Tan and H. Bernard. Phylogenetic Analysis of the Human Papillomavirus Type 2 (HPV-2), HPV-27, and HPV-57 Group, Which is Associated with Common Warts. *Virology*, 239, pp. 296–302, 1997.
2. M. Favre, D. Kremsdorf, S. Jablonska, S. Obalek, G. Pehau-Arnaudet, O. Croissant, and G. Orth. Two New Human Papillomavirus Types (HPV54 and 55) Characterized from Genital Tumours Illustrate the Plurality of Genital HPVs. *International Journal of Cancer*, 45, pp. 40–46, 1990.
3. H. Furumoto and M. Irahara. Human Papilloma Virus (HPV) and Cervical Cancer. *The Jounral of Medical Investigation*, 49(3-4), pp. 124–133, 2002.
4. M. Janicek and H. Averette. Cervical Cancer: Prevention, Diagnosis, and Therapeutics. *Cancer Journal for Clinicians*, 51, pp. 92–114, 2001.
5. M. Schiffman, H. Bauer, R. Hoover, A. Glass, D. Cadell, B. Rush, D. Scott, M. Sherman, R. Kurman and S. Wacholder. Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasis. *Jounral of the National Cancer Institute*, 85, pp. 958–964, 1993.