A Forecasting System for Lung Cancer Sensitivities Using SNP Data

Myung-Chun Ryoo*, Sang-Jin Kim*, Chang-Hyeon Park**
*Dept. of Computer Engineering, KyungWoon University
**Dept. of Computer Engineering, YeungNam University, Korea

Abstract

SNP (Single Nucleotide Polymorphism) refers to the difference in a base pair existed in DNAs of individuals. Each of it appears per 1,000 bases in human genome and it enables each gene to defer in functions, interacts with each other to make different shapes of humans, and produces different disease sensitivities. In this paper, we propose a system to forecast lung cancer sensitivities using SNP data related with the lung cancer. A lung cancer sensitivity forecasting model is also constructed through analysis of genetic and non-genetic factors for squamous cell carcinomas, adeno carcinomas, and small cell carcinomas that may frequently appear in Korean. The proposed system with the model gives the probabilities of the onset of lung cancers in the experimental subjects.

1. Introduction

According to recent data announced by Korea National Statistical Office, the lung cancer is one of the cancers whose mortality rate has been increased most rapidly over the last decade. It is known that 28.4 persons per 100,000 people (21.1%) died by the lung cancer in 2005. The lung cancer is classified into non-small cell carcinomas and small cell carcinomas according to the size and shape of the cancer cells. The non-small cell carcinomas takes about 80~85% of the lung cancer and is also classified into squamous cell carcinomas, adeno carcinomas, large cell carcinomas, and so on. Among them, the squamous cell carcinomas is the most well-taken lung cancer to Korean people. It takes about 60% of men's lung cancer and 25% of women's lung cancer in Korea. The adeno carcinoma, which is the next well-taken lung cancer, takes about 18% of men's lung cancer and 50% of women's lung cancer.

Although smoking is the main cause of the lung cancer, only a few smokers catch the lung cancer. It means that the genetic or epigenetic factor plays an important role in determining the lung cancer sensitivity of a person.

In this paper, we propose a forecasting system for the lung cancer sensitivity. The proposed system forecasts the onset possibility of the lung cancer through the analysis of the genes related to the onset of lung cancer.

2. Related studies

2.1 SNP

Genetic differences between individuals are resulted from differences between the DNA base sequences of individuals and the differences in base sequences are classified into variations and polymorphism [1]. The polymorphism gives rise to minute differences in phenotype. In case that the occurrence frequency of the polymorphism is 1% or more of population, it is defined as the polymorphous genotype. The SNP is one of the polymorphisms existed in the human genome, where the only 0.1% of the bases are different in a base sequence. The SNP occupies about 90% of the polymorphisms. It can be clinically used to forecast sensitivities of diseases, drug efficacies, and drug adverse effects. So it may be utilized in the customized medicine which implements the diagnoses and the treatment strategies suitable to the genetic characteristics of each individual [2].

2.2 Genes related with lung cancer

The genetic characteristics of individuals show considerable differences depending on human races as

** Corresponding Author : Chang-Hyeon Park

well as individuals. In addition, whether a certain polymorphism exists or not, the frequency of polymorphism, and the relationship between different polymorphisms are very different according to human races. Accordingly, a program for the diagnoses of lung cancer sensitivities of Korean must be based on the analysis materials of the clinical data extracted from Korean. The genetic sensitivity of the lung cancer is determined by various polymorphisms existed in various genes rather than by a certain polymorphism existed in a certain gene. The lung cancer occurs through continuous damages of genes by carcinogens. That is, the lung cancer occurs by the polymorphism of genes related with the lung cancer such as the genes concerned with the activation and detoxification of carcinogens, with the innate or acquired changes, and with the cell death, cell cycle regulation, and restoration of damaged genes. In this paper, we use seven kinds of genes related with the lung cancer.

2.3 Classifying methods of SNP data

2.3.1 Conventional classifying methods.

The linkage analysis and the association study have been used to analyze SNP data [3]. The linkage analysis calculates the correlation between data by calculating recombined values using mutation traits and their following molecule markers. In order to implement the linkage analysis, association maps, multiple pedigree patterns, information on genetic signs, and data on the characteristic status of family members are required. The association study is one of the most efficient classification methods. It gives the correlation between the genetic polymorphism and its related diseases by comparing genetic differences between a group of patients and its corresponding group of normal people.

2.3.2 Pattern recognition methods.

Most pattern recognitions are implemented through the pattern classification, which classifies given data into groups of similar characteristics. As typical pattern recognition methods, there are statistical pattern recognition, k-NN (k-nearest neighbor), SVM (support vector machine), and etc [4, 5]. The statistical pattern recognition assigns each characteristic vector to each class by using the extracted characteristic values. This method can induce some rules of pattern classification based on statistical averages so that can reduce the probability of misclassification. In the k-NN, each pattern is classified into its nearest class. This method adopts the calculation of the distance between a pattern and the reference pattern instead of the calculation of the distinguishing function using the information on the distribution of patterns [6]. The SVM classifies patterns by maximizing the minimum distance between the supporting vector and the hyperplane in the vector space [7].

3. The architecture of the proposed system

Figure 1 shows the architecture of the proposed sensitivity forecasting system. As shown in figure 1, the proposed system consists of SNP analysis module, lung cancer forecasting module, test data input module, and test result reporting module. It manages the database through a server which can store and retrieve the personal information, the genetic information, and the non-genetic factors of each subject. One can register and retrieve his information using the client program of the system connected with networks. We developed the programs of the proposed system using the Microsoft VC++ and used the MDB of MS Access as the database.

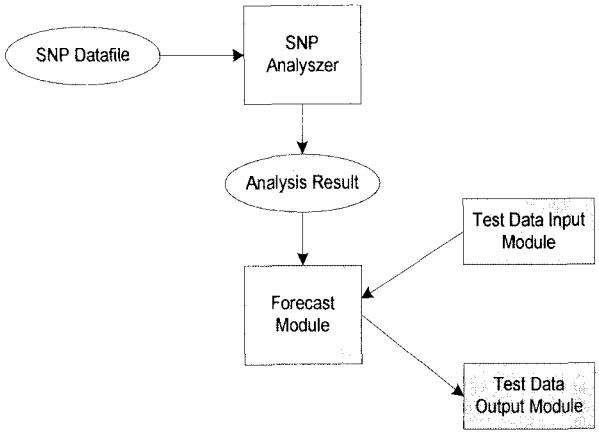

Figure 1. The architecture of the proposed sensitivity forecasting system.

3.1 SNP data

Table 1 shows the composition of the patient group-the control group. The clinical data of a group of 432 patients with lung cancers and those of 432 normal people were used as the control group. The patient group was composed of 210 patients with squamous cell carcinomas (48.6%), 141 patients with adeno carcinomas (32.6%), 73 patients with small cell carcinomas (16.9%), and 8 patients with large cell carcinomas (1.9%). In the control group, 432 subjects were randomly selected from the healthy adults, who had received normal diagnosis, by one to one matching

with the patient group according to the genders and the ages.

Table 1. The composition of the Patient group-the control group

Factor	Patient group	Control group
age	61.6	60.9
gender	352:80	352:80
smoking	317	229
non-smoking	39	98
no experience in smoking	76	105
period of smoking	39.9	34.4

All the people in the patient and control groups were Korean who resides in Daegu Metropolitan City or near the city. The ages, genders, smoking histories, and past disease histories of the people in the patient and control group were obtained through interview of records. In the factor of smoking, the average number of cigarettes smoked in a day and the periods of smoking of the subjects were included. In the factor of non-smoking, subjects who did not smoke for at least one year before their diagnoses were included in the case of the patient group and those who did not smoke for at least one year before the date of the survey were included in the control group.

3.2 Analysis of SNP data and the model for forecasting lung cancer sensitivities

To construct the model for forecasting the lung cancer sensitivity, we first performed statistical analysis for the data from the seven kinds of SNP known to be related to the lung cancer among the genes of the 432 lung cancer patients and the 432 normal people. We then construct the model based on the results of the analysis.

Figure 2 shows the proposed lung cancer sensitivity forecasting algorithm used in the model. As shown in figure 2, the proposed forecasting algorithm consists of the evaluation of the sensitivity using only genetic factors ($t = 1$) and that including non-genetic factors ($t = 2$). The non-genetic factors include the ages, the genders, the periods of smoking, and the periods of non-smoking. The kinds of lung cancers included in the lung cancer sensitivity algorithm are the squamous cell carcinomas ($c = 1$), the adeno carcinomas ($c = 2$), and the small cell carcinomas ($c = 3$). The forecasted values were calculated differently for each kind of the lung cancer. The lung cancer sensitivities for the seven kinds of SNP ($g = 1 \sim 7$) seemed to be associated with the occurrence of lung cancer were forecasted by the algorithm.

if $t = 1$ **then**

$F_{t,c} = 0$

else

for $i = 1$ **to 4 do**

$F_{t,c} = F_{t,c} + Fact_{c,i} + \log IFact_i$

end for

$d = 0$

for $g = 1$ **to 7 do**

$r = Result_g$

$d = Data_{t,c,g,r} + F_{t,c}$

end for

$G_{t,c} = \exp(d)$

$$\text{Prob}_{t,c} = \frac{G_{t,c}}{1 + G_{t,c}}$$

Figure 2. The proposed lung cancer sensitivity forecasting algorithm.

The proposed algorithm uses the three kinds of tables. The tables include the three constant tables (Base, Data, Fact) defined by constants, the two input tables (Result, IFact) prepared by the information of the study subjects, and the result table (Prob) prepared by calculation formulae.

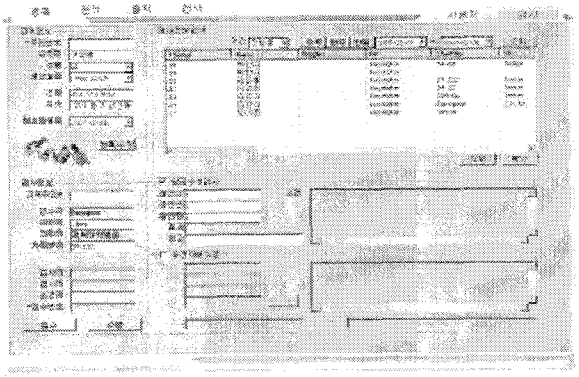

Figure 3. The screen into which the information of an experimental subject is entered.

3.3 Test data input module

Figure 3 shows the data input module. As shown in figure 3, the information related to cancer sensitivity such as the age, the gender, the smoking history, and gene information of an experimental subject who wants to get the forecasting of the lung cancer sensitivity is entered through the input interface.

3.4 Test result reporting module

Figure 4 shows the analysis result screen. In the screen, the search result of the experimental subject, the gene type, and the genes are displayed. The analysis results of the lung cancer sensitivity appear as the seven types of genes. The red points indicate the high risk and the blue ones low risk. 2~3 high risk of genes are usually found in each subject.

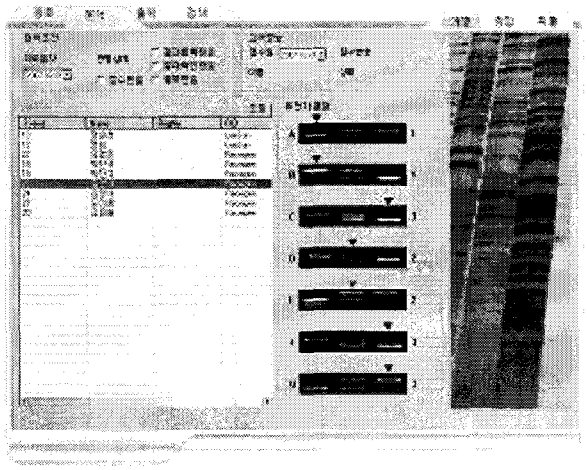

Figure 4. Analysis result screen.

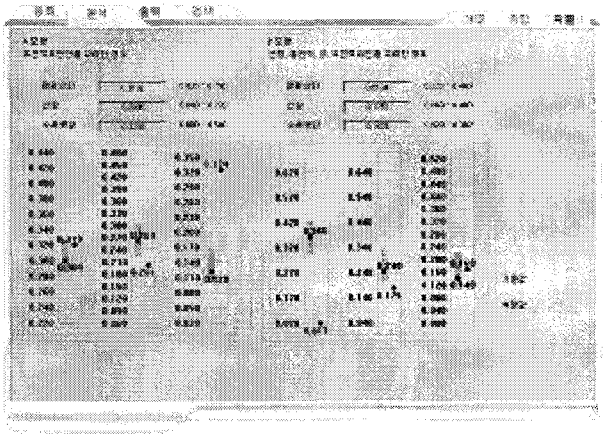

Figure 5. The screen of the analysis of the risk of onset of lung cancers.

Figure 5 shows the analysis result on the risk for the lung cancer tissue types. In figure 5, the results for the case considering only the genetic factors and those including the non-genetic factors are displayed. The reference points indicate the thresholds of dividing the low risk groups and the high risk groups. If the result value of each examination is greater than its threshold, the point is displayed in red color. Otherwise, it is displayed in blue color.

4. Conclusion

Due to the development of the genetics, the scope of genetic diseases is being widened. So the genes associated with the onset and progress of many diseases such as cancers, hypertension, diabetes, and etc are being found out.

In this paper, a system to forecast the lung cancer sensitivity has been proposed. The proposed system can forecast the risk of lung cancer onsets using SNP data. A lung cancer sensitivity forecasting model has been constructed through analysis of the genetic factors and the non-genetic factors for the squamous cell carcinomas, the adeno carcinomas, and the small cell carcinomas that may frequently appear in Koreans. Through the lung cancer sensitivity forecasting model, we could obtain the results of the probabilities of the onsets of lung cancers in the experimental subjects.

It is needed that an optimization method such as the simulated annealing method is added to the proposed system to enhance the accuracy of the lung cancer sensitivity forecasting. We leave it as a further study.

5. References

[1] J. I. Bell, "Single nucleotide polymorphism disease gene mapping," Arthritis Research, vol. 4, pp. s273-s278, Apr. 2002.

[2] A. J. Brookes, The essence of SNPs, Gene, 1999.

[3] N. J. Risch, "Searching for genetic determinants in the new millennium," Nature, 405(6788), pp. 847-856, June 2000.

[4] V. N. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

[5] J. A. Cruz et al., "Applications of machine learning in cancer prediction and prognosis," Cancer Informatics, pp. 59-77, 2006.

[6] "k-nearest neighbor algorithm", http://en.wikipedia.org/wiki/Nearest_neighbor_(pattern_recognition)

[7] T. Joachims, Learning to Classify Text Using Support Vector Machines. Dissertation, Kluwer, 2002.