A Novel Cancer Classifier based on Differentially Expressed Gene Network

Jaegyoon Ahn  
Department of Computer Science  
Yonsei University  
3rd Engineering Bldg. 533-1, Shinchon-dong, Seodaemun-gu, Seoul, Korea  
00822-2123-7757  
ajk@cs.yonsei.ac.kr

Youngmi Yoon  
Information Technology Department  
Gachon Univ. of Medicine & Science  
1108 Gachon-Kwan, Yongsu-dong, Yonsu-gu, Incheon, Korea  
008232-820-4393  
ymyoon@gachon.ac.kr

Sanghyun Park  
Department of Computer Science  
Yonsei University  
3rd Engineering Bldg. 520, Shinchon-dong, Seodaemun-gu, Seoul, Korea  
00822-2123-5714  
sanghyun@cs.yonsei.ac.kr

ABSTRACT
It is fundamental and essential to elucidate how cancer-related genes interact with each other. In this study, we build two undirected graphs: one is a graph consisting of edges only observed in tumor samples, and the other is a graph consisting of edges only observed in normal samples. We apply a genetic algorithm for searching sub-networks of these genetic networks. Those gene sub-networks identify new cancer-related genes that might be related with previously known cancer-related genes, and also show a higher accuracy in classifying tumor and normal samples than the current methods.

Categories and Subject Descriptors
H.2.8 [Database Management]: Data mining; J.3 [Life and Medical Sciences]: Biology and Genetics

General Terms
Algorithms

Keywords
Cancer Classification, Microarray, Genetic algorithm

1. INTRODUCTION
It is important to develop cancer classification methods based on microarray experiments. Especially, elucidating how cancer-related genes interact with each other is more fundamental and essential. Using the microarray data, implementing a classifier in the form of gene sub-network gives a hint to understand the interaction of cancer-related genes.

Large numbers of cancer classification methods based on microarray data use various machine learning techniques which extract cancer-related genes by examining gene expression profiles which are differentially expressed in cancer tissues, classify a new sample with these genes. These showed that the machine learning methods are effectively applied to cancer classification [1-6]. These machine learning classification methods generally filter individual marker genes out, and use them collectively as a classifier without considering gene-gene interactions. We propose a novel method by adopting genetic network which is recently recognized as a model to describe a complex biological occurrences and diseases such as cancers. In this study we consider two kinds of undirected graphs: one is a graph consisting of edges only observed in tumor samples, and the other is a graph consisting of edges only observed in normal samples.

The search space for sub-networks that can differentiate tumor versus normal in the complicated and massive gene network is extremely large. This study applies a genetic algorithm for efficient search. Consequently, this study identifies a classifier with minimally 18 genes, and also exhibits a high accuracy rate when is applied to prostate cancer microarray data. Moreover, the resulting classifier includes new cancer-related genes that might be related with previously known cancer-related genes.

2. ALGORITHM
2.1 Constructing the gene network
We construct the differentially expressed gene network which is built up with genes whose expression values show significant difference between tumor and normal samples. The definitions for the differentially expressed gene network are as follows.

Definition 1 Tumor and Normal edge: Let \( \exp(A) \) be expression value of gene \( A \). For genes \( A \) and \( B \), edge \((A, B)\) is defined as tumor edge, if \( \exp(A) > \exp(B) \) on all the samples in tumor sample set \( T \) and on less than \( p\% \) samples in normal sample set \( N \). On the contrary, edge \((A, B)\) is defined as normal edge, if \( \exp(A) < \exp(B) \) on all the samples in \( N \) and on less than \( p\% \) samples in \( T \).

Definition 2 Differentially expressed gene network: Differentially expressed gene network is defined as a
network whose nodes are genes, and edges are tumor or normal edges.

We abbreviate the differentially expressed gene network as gene network hereafter. The gene network can be built through identifying the tumor and normal edges from every gene pairs of microarray data.

2.2 Making classifier using genetic algorithm

The tumor classifier is composed of genes in the sub-network of the gene network. One tumor or normal edge cannot classify at most p% of normal or tumor samples. Good classifier is composed of best combination of edges which can classify most of samples. One of the effective methods to search such combination of edges is using a genetic algorithm (GA). In this study, the chromosome is simply a set of genes. Each gene in the chromosome is connected with other genes in the chromosome, and these edges are all of the same type (tumor or normal edge). The chromosome can represent a sub-graph of the gene network.

Definition 3 TClassifier and NClassifier: If all the edges of the sub-network are tumor edges, the sub-network is defined as TClassifier. Likewise, if all the edges of the sub-network are normal edges, then this sub-network is defined as NClassifier.

TClassifier and NClassifier are evolved separately. The initial generation is composed of randomly selected tumor or normal edges. Those edges are selected to have k genes each. To select the chromosomes to reproduce offspring, we adopt roulette wheel sampling strategy whose selection probability is proportional to its fitness. To define the fitness function, we define necessary concepts.

Definition 4 PCCt(G) and PCCn(G): Let ev(G, s) be expression values of genes in gene set G on sample s. Given gene set G and sample pair (s1, s2), Pearson’s Correlation Coefficient (PCC) between ev(G, s1) and ev(G, s2) can be calculated. PCCt(G) and PCCn(G) is defined as the average of PCCs for all possible sample pairs in tumor and normal sample set, respectively.

Definition 5 f1(G) and f2(G): f1(G) and f2(G) are fitness functions for TClassifier and NClassifier respectively, f1(G) = PCCt(G) – PCCn(G), f2(G) = PCCn(G) - PCCt(G).

2.3 Prediction of unknown sample

Let Gt and Gn be the gene set of selected TClassifier and NClassifier, respectively. Given unknown sample s, we calculate Ct and Cn by following formula.

\[ C_t = \frac{\sum \text{PCC}(ev(G_t, s_t), ev(G_t, s))}{\text{number of tumor samples}}, \quad C_n = \frac{\sum \text{PCC}(ev(G_n, s_n), ev(G_n, s))}{\text{number of normal samples}}, \]

where sample st is each sample in the tumor sample set and sn is each sample in the normal samples set. The class label of an unknown sample s is predicted as tumor if \( C_t \geq C_n \) and predicted as normal if \( C_t < C_n \).

3. EXPERIMENTAL RESULT

3.1 Gene network and the classifier

To construct the differentially expressed gene network, we used Affymetrix microarray data [7] with 12600 probes (8828 gene symbols), 50 normal samples and 52 prostate tumor samples. Resulting gene network in Figure 2 has 365 tumor edges, 692 normal edges and 1021 unique genes. In figure 2, we can observe that the normal edges are widely distributed while the tumor edges are relatively congregated each other. This observation implies that the normal cell can lose many functions when it changes to the tumor cell. We can also observe that many genes are distributed to be topologically clustered. This observation supports the existing researches saying that genes which are involved in same function can be clustered together on gene network.

3.2 Performance test

Firstly, we performed 10-fold cross validation with k=3~15 and p=50~80%, and get optimal parameter k=12 and p=0.6.
The result showed that accuracy was high as long as \( k > 8 \) and \( p < 0.7 \). For independent test, we built a classifier using these optimal parameters, as in Figure 2, and then measured the accuracy using two independent microarray datasets [9, 10]. Table 1 shows the accuracy, sensitivity and specificity, compared with other algorithms. Results of our algorithm are average of ten independent tests. We also performed 10-fold cross validation to find optimal parameters for the comparison algorithms. Note that the results of SVM, Random Forest and Naïve Bayesian networks are average of ten independent tests. We also measured the accuracy using two independent microarray datasets [9, 10]. Table 1 shows the accuracy, sensitivity and specificity, compared with other algorithms. Results of our algorithm are average of ten independent tests. We also performed 10-fold cross validation to find optimal parameters for the comparison algorithms. Note that the results of SVM, Random Forest and Naïve Bayesian networks are average of ten independent tests. We also measured the accuracy using two independent microarray datasets [9, 10].

Table 1. Comparison with other algorithms

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Accuracy (%)</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>98.47</td>
<td>99.36</td>
<td>95.00</td>
</tr>
<tr>
<td>SVM</td>
<td>93.22</td>
<td>93.62</td>
<td>91.67</td>
</tr>
<tr>
<td>Random Forest</td>
<td>93.22</td>
<td>93.62</td>
<td>91.67</td>
</tr>
<tr>
<td>Naïve Bayesian Network</td>
<td>94.91</td>
<td>95.74</td>
<td>91.67</td>
</tr>
<tr>
<td>k-TSP [5]</td>
<td>81.36</td>
<td>93.62</td>
<td>33.33</td>
</tr>
<tr>
<td>Shah et al. [6]</td>
<td>55.88</td>
<td>52.00</td>
<td>66.67</td>
</tr>
</tbody>
</table>

3.3 Genes related with prostate cancer

Among 10 sets of classifiers built in 3.2, the number of genes that are included in more than 3 TClassifiers is 12. Among those, EIF3H, S100A4, and FGFR3 have been disclosed to be related with prostate cancer. Also, the number of genes that are included in more than 3 NClassifiers is 10. Among those, PSMD9, FPR1, NCAM1, YBX1, and SLC19A1 have been reported to be related with prostate cancer.

In all the tumor samples, expression level of EIF3H is greater than that of S100A4, while such relation was not found in 46% of normal samples. From this observation, we can infer that NOL7 is a good candidate tumor gene.

4. CONCLUSION

This study exhibited that the sub-network of the differentially expressed gene network can be a very effective prostate tumor classifier. The classifier has higher accuracy rate than others, and can be used in clinical setting since it consists of relatively smaller number of genes. The differentially expressed gene network can be used in various cancer related studies. For example, we can expect that if this network is combined and analyzed along with gene regulatory information, the set of causal genes to cancer can be more accurately clarified.

5. ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea funded by Korea Government under Grant 2007-2003965.

6. REFERENCES