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, Yonsu-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 identify cancer-related genes, and to develop cancer classification methods using microarray experiments. Especially, elucidating how cancer-related genes interact with each other is more fundamental and essential. Using the microarray data, implementing a cancer 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.

ACM-BCB 2010, Niagara Falls, NY, USA

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

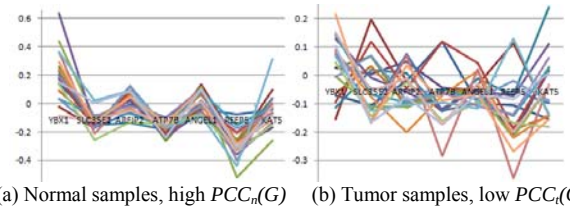
The tumor classifier is composed of genes in the sub-network of the gene network. One tumor or normal edge cannot classify at most *p*% of normal or tumor samples. Good classifier is composed of best combination of edges which can classify most of samples. One of the effective methods to search such 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 (s_1, s_2), Pearson's Correlation Coefficient (PCC) between $ev(G, s_1)$ and $ev(G, s_2)$ 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 ***$f_1(G)$ and $f_2(G)$*** **:** $f_1(G)$ and $f_2(G)$ are fitness functions for *TClassifier* and *NClassifier* respectively, $f_1(G) = PCCt(G) – PCCn(G)$, $f_2(G) = PCCn(G) - PCCt(G)$.



(a) Normal samples, high $PCC_n(G)$ (b) Tumor samples, low $PCC_t(G)$

Figure 1. Example of *NClassifier* with good fitness.

Figure 1 shows the example of *NClassifier* with 7 genes. Figure 1-(a) and (b) are gene expression graphs of randomly selected 20 normal and tumor samples, respectively. We can see that *NClassifier* in figure 1 has high $f_2(G)$, and shows definite difference in gene expression patterns. We can also think of *TClassifier* in opposite way.

To give variation to the offspring, crossover and mutation are used. We use 1-point crossover with two selected chromosomes. If two chromosomes have common genes, the number of genes of both chromosomes may not be conserved, because chromosomes do not allow duplicated genes. In that case, we include randomly selected genes which share *tumor* and *normal edge* with conserved genes, in case of *TClassifier* and *NClassifier*, respectively.

In case of mutation, genes in the chromosome are selected by given mutation rate, and replaced with randomly chosen genes which share *tumor* and *normal edge* with conserved genes, in case of *TClassifier* and *NClassifier*, respectively.

We used population = 30, generations = 50, crossover rate = 0.3 and mutation rate = 0.03 for the experiments in chapter 3.

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.

$$Ct = \frac{\sum PCC(ev(Gt, st_i), ev(Gt, s))}{\text{number of tumor samples}}, Cn = \frac{\sum PCC(ev(Gn, sn_i), ev(Gn, s))}{\text{number of normal samples}}$$

, where sample st_i is each sample in the tumor sample set and sn_i is each sample in the normal samples set. The class label of an unknown sample *s* is predicted as tumor if $Ct \geq Cn$, and predicted as normal if $Ct < Cn$.

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.

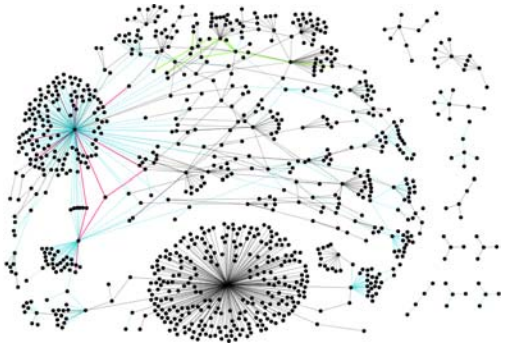


Figure 2. Differentially expressed gene network constructed with normal and prostate tumor samples, with p=60%. Blue and red edges indicate *tumor edges*, and black and green edges indicate *normal edges*. *TClassifier* and *NClassifier* with 12 genes each are indicated with red and green edges, respectively. Visualization was performed with Cytoscape [8].

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 Network use whole 8828 genes.

Table 1. Comparison with other algorithms

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

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 44% of normal samples. This supports the existing studies [11, 12], and also gives new insight that the relationship of those two genes might affect the cancer. NOL7 which has not been cited as a cancer gene is shown 6 times in *TClassifiers*, and reveals a similar trend with S100A4, in that the expression level of EIF3H is greater than NOL7 in all the tumor samples, 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

[1] Pirooznia M, Yang J Y, Yang M Q and Deng Y, “A comparative study of different machine learning methods on microarray gene expression data”, BMC Genomics 2008, 9 Suppl 1:S13.

[2] Duan K B, Rajapakse J C, Wang H and Azuaje F, “Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data”, IEEE Transactions on Nanobioscience, vol. 4, no. 3, pp. 228-234, 2005.

[3] Pan F, Wang B and Perrizzo W, “Comprehensive vertical sample-based k-NN/LSVM classification for gene expression analysis”, Journal of Biomedical Informatics, vol. 37, pp. 241–249, 2004.

[4] Diaz-Uriarte R and Alvarez de Andres S, “Gene selection and classification of microarray data using random forest”, BMC Bioinformatics, vol. 7, no.13, 2006.

[5] Tan A, Naiman D, Xu L, Winslow R and Geman D, “Simple decision rules for classifying human Cancers from gene expression profiles”, Bioinformatics, vol. 21, pp. 3896-3904, 2005.

[6] Shah S and Kusiak A, "Cancer gene search with data-mining and genetic algorithms", Computers in Biology and Medicine, vol. 37, Issue. 2, pp. 251-261, 2007.

[7] Singh D, Febbo P G, Ross K, Jackson D G, Manola J and Ladd C, “Gene expression correlates of clinical prostate Cancer behavior”, Cancer Cell, vol. 1, pp. 203-209, 2002.

[8] Shannon P, Markiel A, Ozier O, Baliga N S, Wang J T, Ramage D, Amin N, Schwikowski B and Ideker T, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks", Genome Research, 2003, 13: 2498-2504.

[9] Welsh J B, Sapinoso L M, Su A I, Kern S G, Wang-Rodriguez J and Moskaluk C A, “Analysis of gene expression identifies candidate markers and pharmacological targets in prostate Cancer”, Cancer Research, vol. 61, pp. 5974-5978, 2001.

[10] LaTulippe E, Satagopan J, Smith A, Scher H, Scardino P and Reuter V, “Comprehensive gene expression analysis of prostate Cancer reveals distinct transcriptional programs associated with metastatic disease”, Cancer Research, vol. 62, pp. 4499-4506, 2002.

[11] Savinainen K J, Linja M J, Saramäki O R, Tammela T L, Chang G T, Brinkmann A O and Visakorpi T, "Expression and copy number analysis of TRPS1, EIF3S3 and MYC genes in breast and prostate cancer", British Journal of Cancer, 2004 Mar 8; 90(5):1041-6.

[12] Sherbet G V, "Metastasis promoter S100A4 is a potentially valuable molecular target for cancer therapy", Cancer Letters, 2009 Jul 18; 280(1):15-30.