2009 International Symposium on Nonlinear Theory and its Applications
NOLTA'09, Sapporo, Japan, October 18-21, 2009

Detection of Differentially Expressed Gene Modules in Humana Cancer

Terufumi Inoue, Tomomasa Nagashima and Yoshifumi Okada*

Computer Science and Systems Engineering, Muroran Institute of Technology,
27-1, Mizumoto-cho, Muroran 050-8585, Japan
Email: okada@csse.muroran-it.ac.jp

Abstract– Identifying biologically useful genes from massive gene expression data produced by DNA microarray experiments is a crucial issue in bioinformatics and clinical area. Recent studies on gene module discovery have shown substantial usefulness for identifying genetic subtypes in single disease class, but the extension to different disease classes has remained to unsolved. In this paper, we propose a new method to discover differentially expressed gene modules from two class dataset. The proposed method is applied to breast cancer and leukemia datasets, and the biological functions of the extracted modules are evaluated by functional enrichment analysis. As a result, we show that our method can extract genes reflecting known biological functions compared to a traditional approach.

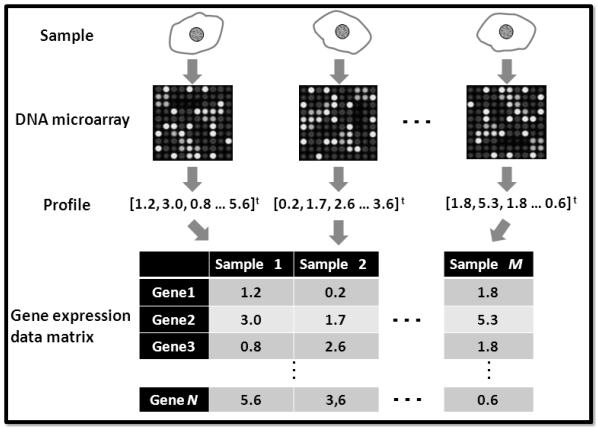


Figure 1: Gene expression datasets obtained from DNA microarray experiments

1. Introduction

DNA microarray technology has enabled us to measure expressions of thousands of genes simultaneously under a certain condition and has yielded various biological applications such as functional analysis of genes or identification of up- and down-expressed genes in complex diseases like cancer. An important step of microarray data analysis is to identify groups of genes with similar expression patterns across multiple samples (e.g., normal/disease cells) in a gene expression dataset. Although traditional clustering algorithms like hierarchical clustering provide natural solutions to this problem, it is bound by a limitation that they use all dimensions of samples to compare pair of genes even if these have relevance only in a subset of samples.

On the other hand, a new clustering technique called biclustering has focused on finding gene expression modules ("modules" for short) with a locally similar expression pattern across subset of samples. We previously developed an exhaustive and efficient biclustering algorithm (BiModule) for module search, and reported that it can well reflect known biological functions compared to other salient methods [1]. Existing methods including BiModule have targeted single class dataset but have not been applied to multiple classes so far.

The aim of this study is to provide a new method for identifying differentially expressed modules between different two-classes in a gene expression dataset. In the proposed method, specificity score for each module is defined to rank according to expression difference between classes. It is expected that such discriminative modules in different classes would become candidates of genetic biomarkers in disease diagnosis. In this study, this method is applied to two public cancer datasets and its performance is evaluated through functional enrichment analysis of obtained modules.

2. DNA microarray data and gene expression module

2.1 Gene expression dataset

Figure 1 illustrates a gene expression dataset obtained by multiple DNA microarray experiments in a single class. A single DNA microarray generates expression values of thousands of genes simultaneously for a single sample (e.g., a normal/disease cell). Each point on a DNA microarray indicates a gene and its intensity represents the expression level. The set of expression values obtained from a single DNA microarray is called a gene expression profile. In multiple samples, the gene expression profiles are arranged in the form of a matrix in which each row and each column respectively correspond to a gene and a sample, and each element is an expression value of a gene.

2.2 Module extraction by biclustering

Figure 2 depicts module extraction from gene expression dataset in a single class. A module is defined as a subset of genes with a common expression pattern over subset of samples. In many past studies, it has been shown that genes composing a module play a biologically important

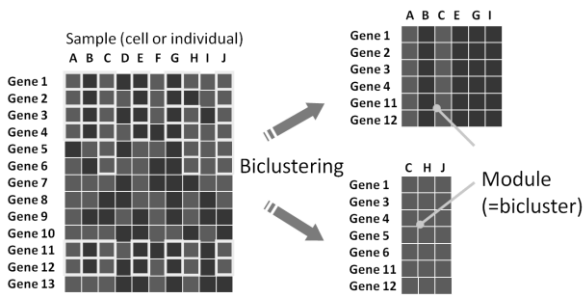


Figure2: Module extraction from gene expression

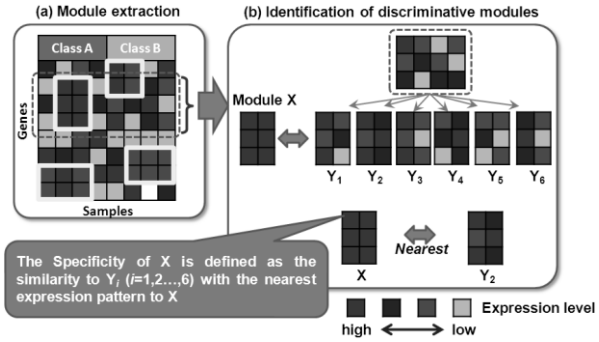


Figure3: The procedure of our method

role, participating in an identical genetic pathway. Biclustering can identify various sample subgroups having different co-expressed genes from single class but has not been extended to discriminative module extraction from two-classes as proposed in this study.

3. Materials and Method

3.1. Outline

Motivated by the issue as described in Section 2.2, we develop a new method for identifying discriminative modules between different classes. Figure 3 shows the outline of this method. First, we search for modules exhaustively from respective classes by using a biclustering technique called BiModule (Fig. 3a). Next, the extracted modules are scored and ranked based on their specificities representing the discriminative powers between classes (Fig. 3b).

3.2. Module extraction from each class

To extract modules from each class, we utilize the high-performance biclustering tool called BiModule. Biclustering typically requires high computational complexity due to combinatorial searches for both of genes and samples, whereas BiModule can exhaustively search for maximal modules from discretized expression data in real time based on a closed itemset mining algorithm called LCM [2]. BiModule shows the highest enrichment of gene function sets as well as the fastest running time among five salient algorithms in yeast data and human cell/tissues data. This tool requires a discretization bins and the minimum size of modules as the input parameters. In this study, we use 7 as the discretization bins, and set 10 and 4 as the minimum number of genes and samples, respectively. In this method, BiModule are applied separately to each class as illustrated in Fig. 3a.

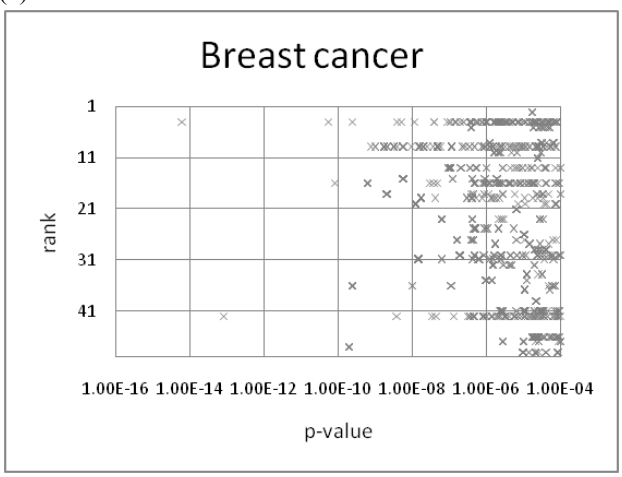
3.3. Identification of discriminative modules

As the candidates of discriminative modules, first, we pick up only the constant modules in which all discretized values have identical signs as depicted in Fig. 2b. Discriminative modules between classes are selected from those constant modules. Here we define the specificity score that represents the discrimination power between classes. The specificities of the constant modules are calculated in each class separately. Hereinafter the targeting class and another class are respectively referred to as class A and class B, where the targeting class means the class in which we perform the specificity calculation. We consider calculating the specificity of a constant module X in class A. First, in class B, we enumerate all combinations of modules Y_i (i=1,2,..,C) in the same genes and the same size of samples as the X as illustrated in Fig. 2b. Next the specificity S of the X is calculated by the following expression:

$$S = \min_{1 \le i \le C} \operatorname{sgn}(m) \frac{\log s_X}{\log s_{Y_i}}$$
$$\operatorname{sgn}(m) = \begin{cases} 1, & m > 0 \\ 0, & m = 0 \\ -1, & m < 0 \end{cases} \quad (1)$$
$$m = -(m_X \times m_{Y_i})$$

where S_X and S_{Yi} are respectively the standard deviations of the discretized values for the X and the Y_i, and m_X and m_{Yi} are respectively the mean values of the discretized values for the X and the Y_i. The larger specificity of the X is the larger expression difference between classes is. The specificity calculation is performed for every constant module X in class A. These modules are ranked in descending order of their specificities. The specificity calculation in class B is performed in the same manner as class A. Finally, discriminative modules of each class are extracted by setting a threshold to the rank orders of the specificities.

(a) Breast cancer



(b) Leukemia

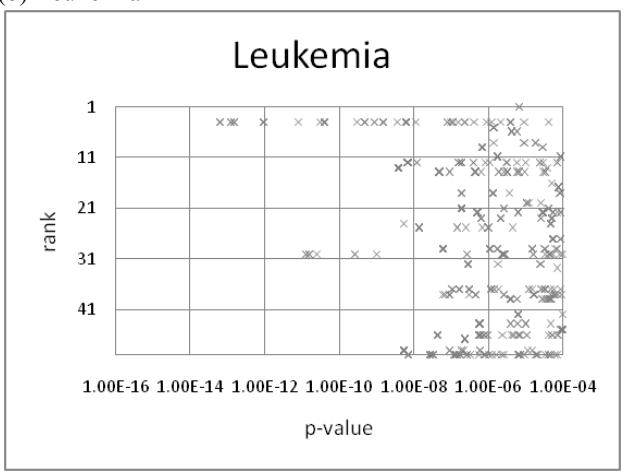


Figure4. Correlation between the specificity scores vs. module ranking

4. Experiments

4.1. Datasets

To evaluate the usefulness of our method, we use the two-class gene expression datasets, leukemia [3] and breast cancer [4]. The breakdown of each dataset is as follows:

- Leukemia : 12,582 genes
  class1 (ALL): 24 samples
  class2 (AML): 28 samples
- Breast cancer : 7,129 genes
  class1 (positive) 25 samples
  class2 (negative) 24 samples

4.2. Functional analysis of discriminative modules

We evaluate if the genes composing the extracted discriminative modules (called module genes below) reflect properly known biological functions. In this study, the functions of the module genes are identified by using a functional enrichment analysis tool called GeneCoDis [5, 6]. GeneCoDis provides a statistical probability (*p*-value) that a certain biological function occurs *x*-times by chance in a given list of genes. This tool enables us to find statistically significant biological functions for the four functional themes, gene function (GO: Gene Ontology), molecular interaction (KEGG: KEGG pathway), motif sequence (IPM: InterPro Motifs) and transcription factor (TF).

4.3. Evaluation

The evaluation is conducted on the following two:

1) Correlation between specificity scores and module ranking.
2) Comparison with a traditional method.

To examine the first one, we extract the top 50 discriminative modules in descending order of the specificities and then generate the distribution for the *p*-values of the significant functions found in their module genes. In the second one, we compare our method with the t-test-based approach (called t-test approach below) that has been widely used in differentially expressed gene analysis.

5. Results and Discussion

5.1. Correlation between the specificity scores and module ranking

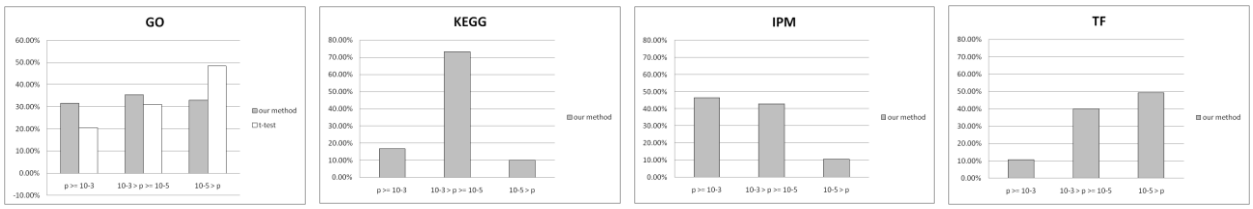
Figure 4 show the *p*-values judged to be significant functions ($p<0.0001$) in the respective rank orders of specificities for the breast cancer (Fig. 4a) and leukemia datasets (Fig. 4b). In this figure, the *p*-values for the four functional themes are plotted all together. The horizontal and vertical axes show the *p*-value and the rank order of specificity, respectively. From these two figures, we can see that discriminative modules with larger specificities are characterized by the more significant functions. This result suggests that specificity in our method well reflects known biological functions.

5.2. Comparison with a traditional method

Subsequently, we compare our method with t-test approach. The t-test approach used here consists of the following steps; first, t-test is applied for each gene and only genes with smaller *p*-values than a certain significant level are selected. Next, these selected genes are grouped into gene clusters showing similar expression patterns by using a hierarchical clustering (HCL). After that, we utilize the cluster boundary discovery tool ASIAN [7] to obtain the optimal cluster separation. Finally, functional enrichment analysis for each cluster is conducted by GeneCoDis.

The significant functions of discriminative modules are compared to those of the clusters generated by the t-test

(a) Breast cancer



(b) Leukemia

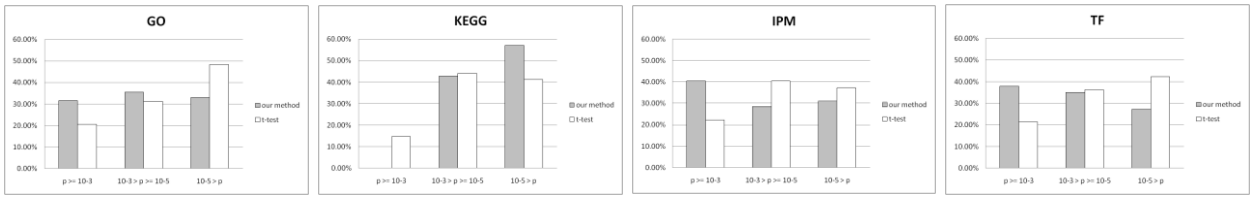


Figure5. Relative frequency distribution of *p*-value in four functional themes

approach. The comparison is performed using the relative frequency distributions of *p*-values (< 0.001) for the four functional themes in Section 4.2. Figure 5 shows the results for breast cancer (Fig. 5a) and leukemia datasets (Fig. 5b), where the relative frequency (%) is depicted with three ranges of the *p*-values. The horizontal and vertical axes are respectively the *p*-value and the relative frequency, and gray and white bars show the results for our method and the t-test approach.

In the breast cancer dataset (Fig. 5a), the discriminative modules obtained by our method are characterized by significant functions in all of the themes. In contrast, the clusters in the t-test approach include no significant functions except for GO functions. As for the leukemia dataset (Fig. 5b), we cannot observe obvious differences between the two approaches although the both exhibit significant functions in all themes. From the two figures, however, we can see that our method shows better results than the t-test approach in the KEGG functions. Namely, this suggests that our method outperforms the t-test approach in discovery of genes interacting within the actual living cells.

6. Conclusions

In this paper, we proposed a new method for extracting differentially expressed gene modules from two-class gene expression dataset and applied it to breast cancer and leukemia datasets. The results of functional enrichment analysis for extracted discriminative modules revealed that our method can extract genes well-reflecting known biological functions compared to the traditional t-test approach. We expect that our method becomes a promising tool for identifying candidates of gene biomarkers for various intractable diseases like cancer.

We however have not provided any definition of the critical value (the threshold) in the ranking of the specificities. Thus the top 50 discriminative modules used in this study might include indifferent modules between classes. In the future work, we will develop a method to detect automatically the threshold value for specificity. In addition, we will extend the method to a new classification tool based on the discriminative modules.

Acknowledgments

This work was supported by Grant-in-Aid for Young Scientists (B) No.21700233 from MEXT Japan and A Research for Promoting Technological Seeds from JSTA.

References

[1] Okada Y, Fujibuchi W, and Horton P: A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm, IPSJ Trans. on Bioinformatics 48 (2007), no. SIG5(TBIO2), pp.39-48.

[2] Uno, T., Kiyomi, M. and Arimura, H., “Lcm ver.3: Collaboration of array, bitmap and prefix tree for frequent itemset mining”, Open Source Data Mining Workshop on Frequent Pattern Mining Implementations 2005, 2005.

[3] S.A. Armstrong et al. : MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia, 2001.

[4] M. West et al. : Predicting the clinical status of human breast cancer by using gene expression profiles, 2001.

[5] Carmona-Saez P, Chagoyen M, Tirado F, Carazo JM, Pascual-Montano A: GENECODIS: A web-based tool for finding significant concurrent annotations in gene lists. *Genome Biology 2007 8(1):R3*

[6]Nogales-Cadenas R, Carmona-Saez P, Vazquez M, Vicente C, Yang X, Tirado F, Carazo JM, Pascual-Montano A: GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Research 2009; doi: 10.1093/nar/gkp416*

[7] Horimoto, K. and Toh, H. (2001) Bioinformatics 17:1143-1151, Statistical estimation of cluster boundaries in gene expression profile data.