Discovering phenotype specific gene module using a novel biclustering algorithm in colorectal cancer

Jungrim Kim
Department of Computer Science
Yonsei University
Seoul, Korea
Kimgogo02@cs.yonsei.ac.kr

Jeagyoon Ahn
Department of Integrative Biology and Physiology
University of California, Los Angeles
Los Angeles, CA
jgahn@ucla.edu

Youngmi Yoon
Department of Computer Engineering
Gachon University
Gyeonggi-do, Korea
ymyoon@gachon.ac.kr

Yunku Yeu
Department of Computer Science
Yonsei Unviersity
Seoul, Korea
yyk@cs.yonsei.ac.kr

Sanghyun Park*
Department of Computer Science
Yonsei Unviersity
Seoul, Korea
sanghyun@cs.yonsei.ac.kr

***Abstract*— Gene clustering is a method for finding gene sets which are related to the same biological processes or molecular function. In order to find these gene sets, previous studies have clustered genes which showed similar mRNA expression or a specific expression pattern in a (sub) sample set. However, for two contrasting groups of samples, it is not easy to identify gene sets which show significant expression pattern in only one group using current gene clustering methods. Existing biclustering methods use only one group (disease) of samples. It is hard to identify disease specific biclusters which are differentially expressed in the disease although those methods can find biclusters which have specific expression pattern. Here, we proposed a novel method using a genetic algorithm in gene expression data, in order to find gene sets which can represent specific subtype of cancer. Proposed method finds gene sets which have statistically differential mRNA expression on two contrasting samples and fraction of cancer samples. The resulting gene modules share higher number of GO (Gene Ontology) terms related to a specific disease than gene modules identified by current algorithms. We also identify that when we integrate protein-protein interaction data with gene expression data of colorectal cancer samples, proposed method can find more functionally related gene sets.**

Keywords— Gene module; Biclustering; Microarray; Genetic Algorithm

I. INTRODUCTION

Disease related genes do not affect the whole process of the disease progress and subtypes of heterogeneous disease like cancer are characterized by distinct genetic alteration [1]. Therefore, finding gene sets whose expression patterns reflect characteristics of cancer subtypes would be helpful to understand cancer more comprehensively. For identifying such genes, individual samples as well as genes should be considered during search process, since many genes show high variance in their expression even on samples of same group. In this study, we use a bicluster method to find gene module. Bicluster method is a kind of co-clustering technique, which clusters genes and samples at the same time in order to find gene sets which show significant expression on specific samples. Finding biclusters is NP-hard problem and most of bicluster algorithms use heuristic method or probabilistic approximation.

Since Cheng and Church [2] have started to use bicluster method for analyzing gene expression data at first, many bicluster methods [2-7] have been introduced for analyzing gene expression data. CC [2], OPSM [3], ISA [4] are most typical bicluster methods. CC takes finding biclusters as optimization problem. It calculates mean squared residue score of candidate biclusters for finding optimized biclusters which has additive pattern. Additionally, if the mean squared residue score of a bicluster is close to 0, the bicluster becomes optimized. OPSM method is the bicluster method which finds order preserving sub matrixes (bicluster). In order preserving sub matrix, there exists a permutation of columns and the permutation is arranged in non-decreasing pattern. For example, it finds biclusters of which mRNA expression is order preserved sample-wise as an output. ISA method gives high weight to gene and sample, if the mRNA expression is high at the gene and the sample, and consequently finds cis-regulatory

*: corresponding author

bicluster as a result. In this paper, we compare our method with CC, OPSM and ISA method which we introduced.

Although numerous researches were introduced after Cheng and Church for gene expression data, current bicluster methods primarily aim at finding Order preserving patterns and Additive or Multiplicative patterns. However, it is not sure that there is a correlation between these patterns and specific cancer subtypes because current bicluster method uses only disease (control) samples without control (disease) samples. And, actually, we observed that there is much overlap of genes between two contrasting biclusters which are founded on disease samples and control samples. Because of this reason, our study aims at finding pattern which can reflect specific subtype of disease and we name this pattern phenotype specific pattern.

In this study, gene sets which have phenotype specific patterns have statistically very high or low gene expression in a fraction of disease samples and control samples. And, we assume that, it can represent a specific aspect of a disease. For example, driver genes can regulate other disease genes which are involved in early state of a cancer, and these genes can have similar expression patterns only on subsets of disease samples.

To find these phenotype specific gene sets, we propose PSBF (Phenotype Specific Bicluster Finder), in order to find gene sets which represent specific disease aspect of disease. Proposed method analysis disease and control samples and finds biclusters in which gene sets are statistically high or low gene expressed in specific disease samples compared to control and fraction of disease samples as shown in Fig 1. For example, gene a, b and c's mRNA expression are statistically high or low in tumor condition 2, 3 and 5. Because gene a, b and c are differentially expressed compared to normal samples, it can be considered as disease specific gene set. Also, this gene set is differentially expressed compared to tumor sample 1, 4 therefore it can be considered as phenotype specific gene set. In other words, gene a, b and c can represents specific subtype of cancer in which gene set is differentially expressed compared to control and fraction of tumor samples. At this table, if the value is higher than '2', it means statistically high mRNA expression, and if the value is lower than '-2', it means statistically low mRNA expression.

	Tumor 1	Tumor 2	Tumor 3	Tumor 4	Tumor 5	Normal 1	Normal 2
…							
Gene a	0.14	2.23	2.34	0.56	1.98	-0.12	0.34
…							
Gene b	-0.87	-1.88	-2.13	0.34	-2.33	0.17	-0.62
…							
Gene c	0.81	3.02	2.29	0.22	2.87	0.54	0.89
…							

Fig. 1. An example of gene set with statically different mRNA expression.

Additionally, we apply weight to the gene set depending on the distance between genes in protein-protein interaction (PPI) network, in order to select more related gene sets. And, we develop new method using genetic algorithm for solving NP problem of biclustering. Genetic algorithm is developed by John Holland [8] firstly and it is derived from the process of natural evolution. It contains genetic technique such as inheritance, mutation, selection and cross over and it is usually used to find NP hard problem's optimal solution.

II. METHOD

A. Data preprocessing

We use GSE24514 gene microarray dataset [9] which includes 15 normal tissue samples and 34 colorectal tumor samples from GEO (Gene Expression Omnibus) [10]. We assume that gene expression of a sample follows a normal distribution, and accordingly apply z-scoring to the gene expression values.

B. Generating Protein distance matrix

We download protein interaction pair data from I2D version 2 (Interologous Interaction Database) [11] and we remove duplicated interaction and self-interaction pair. Finally, we get 306,419 protein interaction pairs which include 13,207 proteins. Based on this protein interaction data, we build an undirected protein interaction graph where all of its edges have 1 as weight. Then, we apply Floyd-Warsall algorithm [12] to this graph to make a protein distance matrix. It is a 13207 X 13207 matrix of which row and column means gene, and each value means shortest distance between the gene in a row and the gene in a column. Afterward, this matrix is used to get the weight of bicluster during process for pruning result sets in PSBF (Phenotype Specific Bicluster Finder).

C. Phenotype Specific Bicluster

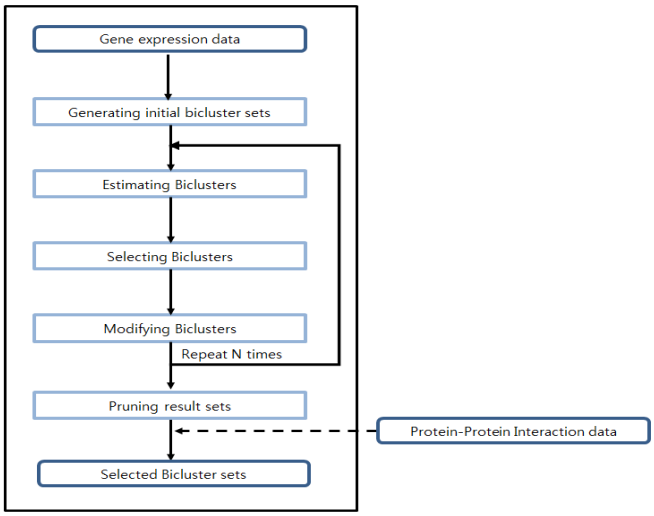

Fig. 2. System overview.

In this paper, we developed a biclustering method based on genetic algorithm to find gene sets which have statistically high or low mRNA expression in specific tumor samples. Figure 2 shows system overview of proposed algorithm. We design PSBF method based on genetic algorithm, and it has 5 main processes 1) generating initial bi-cluster sets 2) estimating biclusters 3) selecting biclusters 4) modifying biclusters 5) pruning result sets. Proposed method repeats process 2) ~ 4) N times to find biclusters and select best biclusters which satisfy threshold in process 5.

1) Generating initial biclusters: Firstly, we generate 1000 random biclusters as initial population for the genetic algorithm. Each bicluster contains 4 ~ 8 genes and samples randomly for being used in the method.

2) Estimating Biclusters : Estimating Biclusters is a process for scoring bicluster to be used in the selection stage. Figure 3 shows bicluster scoring implementation. Bicluster D stores n sample IDs and m gene IDs from modification process. From $gene_i$ which is included in D, we calculated sum_{in} which is a summation of mRNA expression value in all of samples which are included in D and sum_{out} which is a summation of mRNA expression in all of samples which are not included in D. And then, we calculate avg_{in} and avg_{out} by dividing sum_{in} and sum_{out} by number of samples [Line #1-*12*]. After that we calculate absolute value of avg_{in} * $(avg_{in} - avg_{out})$ for $gene_i$'s score (=$gene_{score}$) [Line #13]. The $gene_{score}$ is affected by $gene_i$'s expression value and the difference of $gene_i$'s expression value between a fraction of disease samples and other samples. After all of gene's scores in D are calculated, finally, we can get bicluster D's score (=D_{score}) which is a summation of all of gene's score in D.

INPUT : bi-cluster D which has m gene ID and n sample ID
OUTPUT : score of bi-cluster D.
PARAMETER :
$|M|$: The number of gene in microarray
$|N|$: The number of sample in microarray
MA: M X N microarray
$MA_{z-score(gene)}$ = microarray which is applied to z-score formulation based on gene(row)
$MA_{z-score(sample)}$ = microarray which is applied to z-score formulation based on sample(column)
$gene_i$: i th gene in microarray
$sample_j$: j th sample in microarray
sum_{in}: summation of $gene_i$ expression in bicluster D
sum_{out}: summation of $gene_i$ expression out of bicluster D
avg_{in}: average of $gene_i$ expression in bicluster D
avg_{out}: average of $gene_i$ expression out of bicluster D
METHOD :
1: **FOREACH** $gene_i \in D$ **do**
2: $sum_{in} = 0$
3: $sum_{out} = 0$
4: **FOR** j = 0 **to** N **do**
5: **IF** $sample_j \in D$ **THEN**
6: $sum_{in} = sum_{in} + MA_{z-score(gene)}[i][j] + MA_{z-score(sample)}[i][j]$
7: **ELSE**
8: $sum_{out} = sum_{out} + MA_{z-score(gene)}[i][j] + MA_{z-score(sample)}[i][j]$
9: **ENDIF**
10: **ENDFOR**
11: $avg_{in} = |sum_{in}| / n$
12: $avg_{out} = |sum_{out}| / (N - n)$
13: $gene_{score} = (avg_{in} * |avg_{in} - avg_{out}|)$
14: $D_{score} = D_{score} + gene_{score}$
15: **ENDIF**
16: **FOREACHEND**

Fig. 3. Process of estimating biclusters.

3) Selecting biclusters: This process is for selecting biclusters which are inherited to the next generation. In this experiment, we select t biclusters which will be used in the modifying process by repeating the process t times. At this process a bicluster which is from top k scored biclusters of the previous generation is selected with probability of p. And the rest of the biclusters are selected with probability of (1-p), for securing diversity of biclusters. We use t=10, p=0.9, and k=10 which shows best performance.

4) Modifying biclusters: Modifying Biclusters is a process to modify bicluster for diversity. There are switching genes and samples between two biclusters, and mutation by randomly deleting and adding samples and genes from biclusters. At first, we additionally make 2 * ${}_tC_2$ new biclusters by switching randomly selected genes and samples from t selected biclusters. In other words, the crossover process is applied for every possible pairs of biclusters selected in 3) then 2 crossovered biclusters are generated. As a result, we get t^2 (= 2 * ${}_tC_2$ + t) biclusters which include the originally selected biclusters. Additionally, we make 4 * t^2 biclusters by randomly deleting and adding samples and genes to the t^2 biclusters in the previous step then finally we get 5 * t^2 bicluster. In other word, we get 500 bicluster which result from t=10.

5) Pruning result set: This process filters the gene sets of the selected biclusters. The genetic algorithm has a hereditary trait so it makes unequal distribution of gene's composition included in the biclusters. Therefore, it is necessary to filter similar biclusters to avoid generating redundant outputs. In order to filter the similar biclusters, we sort biclusters by its score and we select high-quality biclusters which have higher score than threshold. At that time, if there is more than 20 % similar bicluster to the high-quality biclusters, the latter biclusters in the sorted list are removed.

$$D'_{score} = D_{score} / \textit{gene set's average distance} \quad (1)$$

Additionally, if the PSBF method uses PPI data together, these bicluster's score are calculated again like (1) above. In this formulation, we calculate gene set's average distance between all of genes included in a bicluster from protein distance matrix. And, we calculate bicluster's score (= D'_{score}) again.

III. Results

Experimental environment is that, we used a Windows 7 operating system on an Intel Core i5-3470, 3.2 GHz, 3.69 GB RAM machine and we have implemented our algorithm using the java language.

A. Comprison of Gene Ontology terms

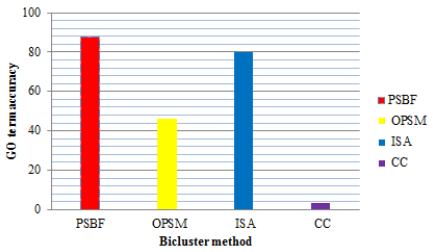

Fig. 4. GO term validation results using FuncAssociate.

Functional association of the genes which are in a bicluster identified by PSBF is tested using GO database searching. Gene Ontology is a database which defines GO terms representing genes and gene products properties of various organisms hierarchically, and we use FuncAssociate [13] for GO term verification. FuncAssociate is a tool which takes gene sets as input and then gives GO terms that the gene set shares and p-value as output. For experiment, we use biclusters which have only 4 genes, and are ranked in the top 25 percent of bicluster score for verification. For comparison

with existing algorithms, other biclusters are obtained from BicAT_v2.22 [14]. The BicAT is a tool which takes gene expression data as input and then gives biclusters which are got from different methods including OPSM, ISA and CC method as outputs. Figure 4 shows the Go Term validation results using FuncAssociate. X-axis represents algorithm, and y-axis represents the proportion of biclusters significantly enriched with significance level of 0.00005. The proportion of PSBF is highest with 83%, which indicates that the biclusters of PSBF are more functionally related.

B. Phenotype Specific Bicluster Finder integrated with protein interation data

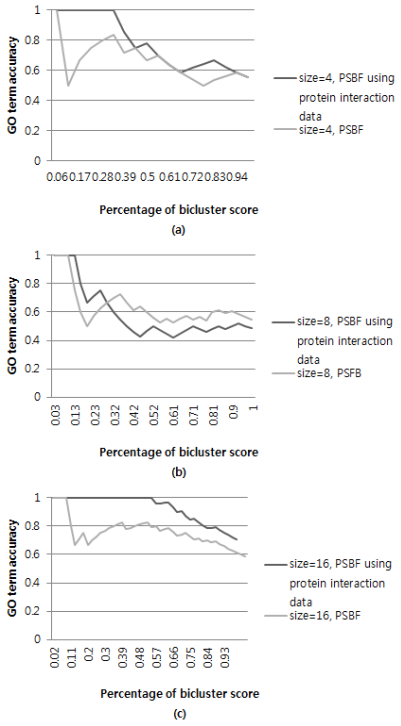

Fig. 5. The GO term accuracy of PSBF using protein interaction with mRNA expression data or not

Our method generates functionally more related biclusters if when we add protein interaction data. We identify that our method generates functionally more related biclusters when we added protein interaction data in various size of gene set. We use biclusters of which gene set have 4, 8 and 16 genes and we checked whether protein interaction data affect GO term accuracy or not. Figure 5 shows GO term accuracy of biclusters when PSBF does incorporate PPI and when it does not incorporate PPI. Y-axis shows GO term accuracy and x-axis shows percentage of bicluster score which are used in experiment. For example, if x is 0.1, it means that we use only biclusters which are ranked in top 10 percent at bicluster score. From experiment result, it shows that bicluster score and GO term accuracy are directly proportional to each other so it represents that our experiment is designed well.

IV. CONCLUSION

In this paper, we suggest PSBF algorithm to find phenotype specific gene modules. In result, PSBF algorithm was proved to be better existing bicluster algorithms in finding gene modules which are closely related to each other functionally. Also in a process of finding biclusters by PSBF algorithm, the performance was found to be improved when using PPI data along with gene expression data.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2012R1A2A1A01010775).

REFERENCES

[1] Toshinori Hinoue, Daniel J. Weisenberger, Christopher P.E. Lange, Hui Shen, Hyang-Min Byun, David Van Den Berg, Simeen Malik, Fei Pan,Houtan Noushmehr, Cornelis M. van Dijk, Rob A.E.M. Tollenaar, and Peter W. Laird, "Genome-scale analysis of aberrant DNA methylation in colorectal cancer," Genome Res, vol. 22, pp. 271–282, 2012

[2] Y. Cheng and G. M. Church, "Biclustering of expression data," In Proc. of the International Conference on Intelligent Systems for Molecular Biology, pp. 93–103, 2000

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini, "discovering local structure in gene expression data: The order-preserving sub matrix problem," In Proc. International Conference on Computational Biology, pp. 49–57, 2002

[4] S. Bergmann, J. Ihmels, and N. Barkai, "Iterative signature algorithm for the analysis of largescale gene expression data," *Phys Rev E Stat Nonlin Soft Matter Phys*, 67(3 Pt 1):03190201–18. 2003

[5] Bozdag˘ D, Kumar A, Catalyurek UV, "Comparative analysis of biclustering algorithms," In: Proceedings of 1st ACM, International Conference Bioinformatics and Computational Biology, pp. 265–274, 2010

[6] TANAY, A., SHARAN, R., AND SHAMIR, R, "Biclustering algorithms: A survey," In *Handbook of Computational Molecular Biology*, S. Aluru, Ed, Chapman & Hall, 2006,

[7] Jaegyoon Ahn, Youngmi Yoon, Sanghyun Park, "Noise-robust algorithm for identifying functionally associated biclusters from gene expression data," Elsevier, vol. 181, pp. 435~449, , 2011

[8] J. H. Holland, "Adaptation in Natural and Artificial Systems," MIT Press, Cambridge, MA , 1992

[9] Pia Alhopuro, Heli Sammalkorpi, Iina Niittymki, Mia Biström, Anniina Raitila, Juha Saharinen, Kari Nousiainen, Heli J. Lehtonen, Elina Heliövaara, Jani Puhakka, Sari Tuupanen, Sónia Sousa, Raquel Seruca, Ana M. Ferreira, Robert M. W. Hofstra, Jukka-Pekka Mecklin, Heikki Jrvinen, Ari Ristimki, Torben F. Ø rntoft, Sampsa Hautaniemi, Diego Arango, Auli Karhu and Lauri A. Aaltonen, "Candidate driver genes in microsatellite-unstable colorectal cancer," International Journal of Cancer vol. 130, pp. 1558–1566, 2012

[10] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A. "NCBI GEO: archive for functional genomics data sets—update," Nucleic Acids Res, 41(Database issue):D991-5, Jan 2013.

[11] Brown, K.R., and Jurisica, I, "Unequal evolutionary conservation of human protein interactions in interologous networks," Genome Biology, vol. 8, R95, 2007

[12] Robert W. Floyd, "Algorithm 97: shortest path," Communications of the ACM, vol. 5 p. 345, 1962.

[13] G.F. Berriz, O.D. King, B. Bryant, C. Sander, F.P. Roth, "Characterizing gene sets with FuncAssociate," Bioinformatics, vol. 19, pp. 2502-2504 , 2003.

[14] Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., and E. Zitzler. "BicAT: a biclustering analysis toolbox," Bioinformatics, vol. 22, pp. 1282-1283, 2006.