DSS: A biclustering method to identify diverse and state specific gene modules in gene expression data

Jungrim Kim*, Yunku Yeu*, Jeongwoo Kim*, Youngmi Yoon§ and Sanghyun Park*†
*Department of Computer Science, Yonsei University
Seoul, South Korea
Email: see http://delab.yonsei.ac.kr/eng/member
§Department of Computer Science, Gachon University
Seongnam, South Korea
Email: ymyoon@gachon.ac.kr
†Corresponding Author

Abstract—**The biclustering method is a useful co-clustering technique to identify biologically relevant gene modules. In this paper, we propose a novel method to find not only functionally-related gene modules but also state specific gene modules by applying a genetic algorithm to gene expression data. To identify these gene modules, the proposed method finds biclusters in which genes are statistically overexpressed or under expressed, and are differentially-expressed in the samples in the bicluster compared to the samples not in the bicluster. In addition, we improve the genetic algorithm by adding a selection pool for preserving the diversity of the population. The resulting gene modules exhibit better performances than comparative methods in the GO (Gene Ontology) term enrichment test and an analysis connection between gene modules and disease. This is especially the case with gene modules that receive the highest score in the breast cancer dataset; they are closely linked to the ribosome pathway. Recent studies show that dysregulation of ribosome biogenesis is associated with breast tumor progression.**

I. Introduction

Due to the availability of large amounts of biological data, there have been many research studies for identifying new biologically valuable knowledge. Clustering is a technique for finding a gene module that shows a similar expression pattern across the set of all samples. Although it is a very useful technique for identifying relevant knowledge, it is difficult to find a disease-related gene module because disease-related gene modules do not affect the whole process of the disease progression [28]. As an example, subtypes of a heterogeneous disease like cancer are characterized by distinct genetic alteration. To overcome this limitation, a biclustering technique can be used. Biclustering is a co-clustering technique which allows simultaneous clustering of the genes and samples in order to find a gene module which shows crucial expression on specific samples. It requires more complicated calculations than the one-way clustering technique; accordingly, it has problems with time complexity. Thus, most of the research uses a heuristic method or probability approximation for finding gene modules.

Since Cheng and Church [31] have introduced a biclustering method to analyze gene expression data, many researchers have used biclustering methods [6], [7], [17], [23], [27] for analyzing gene expression data. To find biclusters, the CC (Cheng and Church) method calculates the mean squared residue score of candidate biclusters that exhibit an additive pattern. If this score is close to zero, it means that the bicluster becomes optimized. The order-preserving submatrix (OPSM) method [1] finds order-preserving submatrices (bicluster). Order-preserving submatrices have a permutation pattern where the columns of matrices are in non-decreasing order. The Iterative Signature Algorithm (ISA) [22] finds cis-regulatory biclusters of which the gene expression is high (at the gene and the sample) by giving a high weight to the gene and sample. The Debi (Differentially Expressed BIclusters) method [2] finds a bicluster that exhibits a statistical difference in the expression between the samples in the bicluster and the samples not in the bicluster, using a frequent item set approach. QUBIC (QUalitative BIClustering) algorithm [12] is a method which identifies biclusters efficiently with 'scaling patterns' utilizing a graph technique. Chakraborty et al [3] use a genetic algorithm to find biclusters without the threshold of the maximum allowable dissimilarity in gene expression data. Nepomuceno et al [18] present a scatter search approach to find biclusters based on linear correlations.

In this paper, we propose a new biclustering method which aims to find not only functionally-related gene modules, but also state specific gene modules. To find state specific gene modules, the proposed method finds biclusters where gene modules are statistically overexpressed or under expressed. Besides, gene modules in biclusters are differentially expressed between the samples in the bicluster and the samples not in the bicluster. Figure 1 shows an example of the biclusters that we want to find. In this figure, gene expression data are z-scored in each sample. Rows represent genes and columns represent samples. Colors of rectangles represent the degree of the gene expression. The more that gene expressions are statistically high, the closer this color is to green; the more that gene expressions are statistically low, the closer this color is to red. In addition, bolded rectangles represent biclusters. As shown in Figure 1, we find biclusters according to the following two rules: i) genes in a bicluster should be statistically coexpressed

978-1-5090-1897-0/16/$31.00 ©2016 IEEE

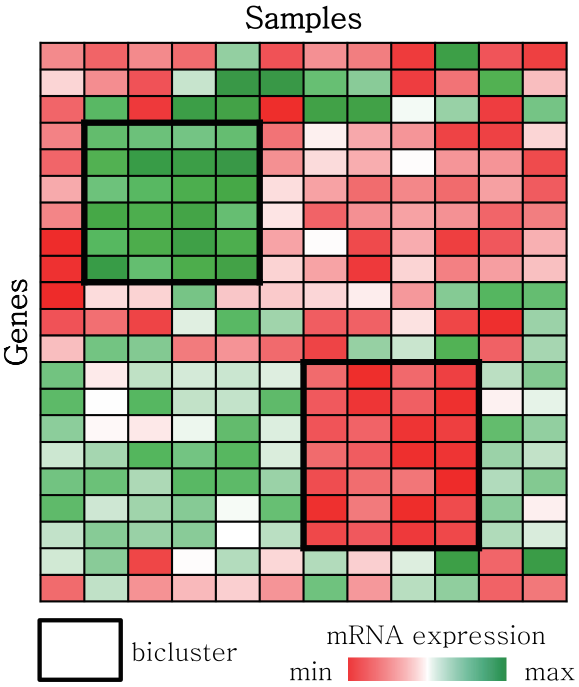


Fig. 1. An example of biclusters generated by the proposed method.

TABLE I
SUMMARY OF GENE EXPRESSION DATA

Dataset	Rows	Columns	Description
GDS181	12,559	84	gene expression of the normal physiological state profiles from various tissues, organs and cell lines.
GDS1027	15,866	154	Sulfur mustard effect on lungs: dose response and time course.
GDS3715	12,422	110	Insulin effect on human skeletal muscle.
GDS3716	22,215	42	Breast cancer: histologically normal breast epithelium.

TABLE II
SUMMARY OF PREPROCESSED GENE EXPRESSION DATA

	GDS181	GDS1027	GDS3715	GDS3716
Genes	3,177	13,152	9,351	13,965
Samples	84	154	110	42

in the samples in the bicluster and ii) genes in a bicluster should be differentially expressed between the samples in the bicluster and the other samples. We assume that statistically overexpressed or under expressed gene modules are more likely to play an important role in samples because it may occur problem in biological process like gene dysregulation. Additionally, if these overexpressed or under expressed gene modules are differentially expressed compared to the samples not in the bicluster, we can assume that they have a stronger relation with the specific states in samples. For example, if gene modules have these characteristics in disease samples, it could have a stronger relation with the progression of the disease. Because a disease-related gene module does not affect the whole process of the disease, it could represent a more specific state of the disease.

To develop the proposed method, we use a genetic algorithm to overcome the time complexity problem. A genetic algorithm [16] is a technique which mimics the processes of natural evolution such as mutation, inheritance, selection and crossover to find a local optimal solution. Because it mimics the process of natural evolution, it may incur a loss of diversity problem and it is hard to retain a variety of biclusters. With the loss of diversity problem, modules in a population tend to have a similar composition of genes during the algorithm's execution. This problem can occur when a few highly-scored gene modules survive and the next population is constructed from these few modules. To overcome this problem, we improve the genetic algorithm by adding a selection pool to give various biclusters a chance to be trained, which helps to solve the loss of diversity problem. Figure 2 shows the difference between the original genetic algorithm and the improved genetic algorithm processes.

This paper is organized as follows. Section 2 introduces the proposed DSS (Diverse and State Specific) biclustering method. In Section 3, we present our results. We conclude our work by discussing the implications in Section 4.

II. METHOD

A. Data preprocessing

We downloaded the gene expression data sets GDS181, GDS1027, GDS3715, and GDS3716 [8], [11], [25], [29], [30], [33] from the gene Expression Omnibus (GEO) database [5] to use. The description of each set of gene expression data is summarized in Table 1. We begin by preprocessing the gene expression data. First, we removed rows that have a null value in the gene expression data. After that, we combine rows that belong to the same gene. To combine rows, we calculate the average gene expression of the genes in each column. Table 2 shows the preprocessed gene expression data. Finally, Data set GDS181 has many rows that have a null value compared to the other data sets. Hence, a relatively large amount of rows from GDS181 are removed. Then, we apply z-scoring to the gene expression data based on the samples by the Equation (1).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

x : mRNA expression of a gene in a sample.
μ : mean of the mRNA expression in a sample.
σ : standard deviation of the mRNA expression in a sample.

After that, if the expression values were higher than 2, we assumed that expression values were highly expressed

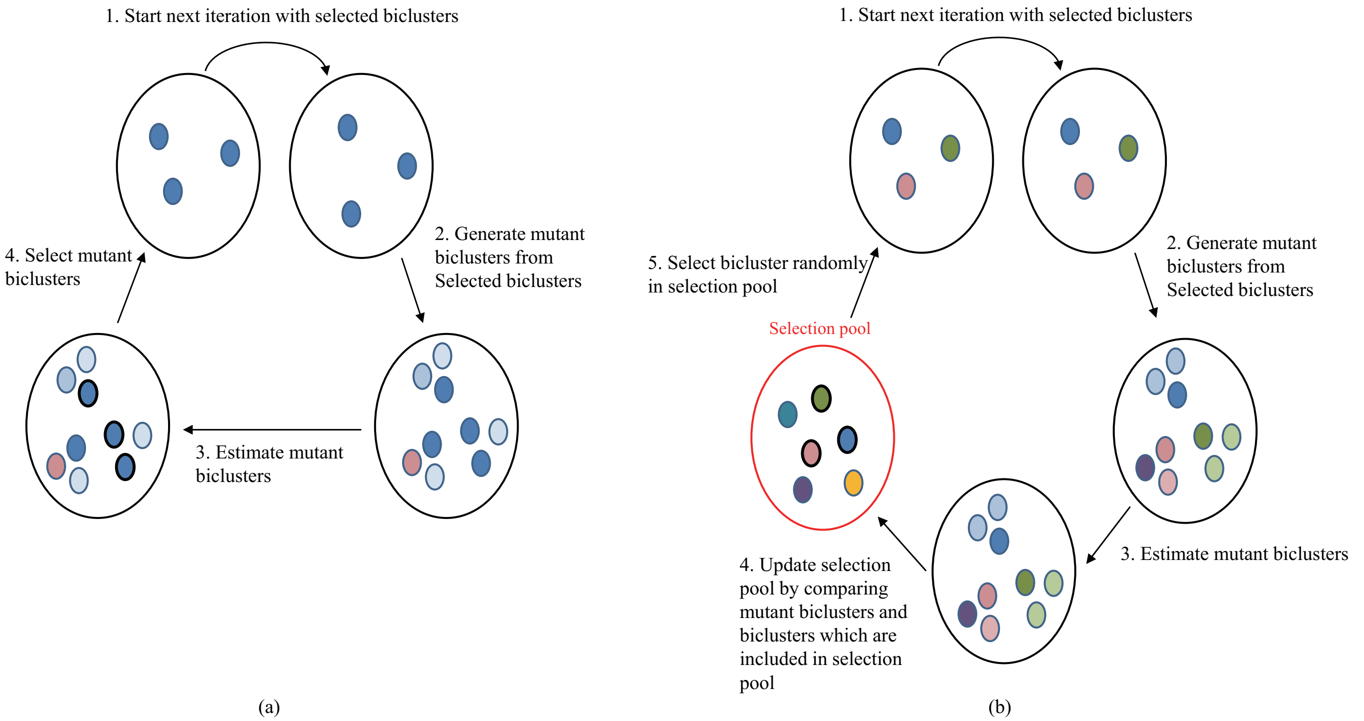

1. Start next iteration with selected biclusters
2. Generate mutant biclusters from Selected biclusters
3. Estimate mutant biclusters
4. Select mutant biclusters
(a)
1. Start next iteration with selected biclusters
2. Generate mutant biclusters from Selected biclusters
3. Estimate mutant biclusters
4. Update selection pool by comparing mutant biclusters and biclusters which are included in selection pool
5. Select bicluster randomly in selection pool
Selection pool
(b)

Fig. 2. The difference between (a) the original genetic algorithm and (b) the improved genetic algorithm. The improved genetic algorithm has a selection pool. The original genetic algorithm starts with biclusters that are generated from the previous iteration and it causes a loss of diversity problem. However, the improved genetic algorithm starts with biclusters that are randomly selected from a selection pool. It can give various biclusters a chance to be trained and find non-overlapping biclusters (biclusters that are generated from the previous iteration are used only for updating the selection pool).

enough to be converted to a value of 2. Likewise, if the expression values were lower than -2, we assumed that it was under expressed enough to be converted to a value of -2. The proposed method gives a score to the bicluster in proportion to the absolute value of the gene expression value. If we don't convert the expression value, the proposed method may find only a small number of biclusters, which includes genes with very high or very low expression. Further, it is hard to say if these biclusters have a more significant meaning. Thus, we preprocess the gene expression data as above.

B. DSS biclustering

In this paper, we proposed a new biclustering method that uses a genetic algorithm. It aims to find gene modules that are not only statistically overexpressed or under expressed, but also differentially expressed between the sample in the bicluster and the samples not in the bicluster. Although our method is developed based on a genetic algorithm, it has an important difference; our method has a selection pool to overcome hereditary traits compared to the original genetic algorithm. Figure 3 shows a system overview of our method. It has six steps: 1) generating initial biclusters; 2) estimating biclusters; 3) updating the selection pool; 4) selecting biclusters; 5) modifying biclusters; and 6) integrating biclusters. To generate biclusters, we generate initial bicluster sets and then repeat steps 2) - 5) a total of N times. After that, we integrate the biclusters that are included in the selection pool to make a more valuable bicluster in step 6. Further details are described below.

1) Generating initial biclusters: Generating the initial sets is a process for initializing the population. We heuristically generate 1,000 random biclusters that contain 16 genes and 3 samples. These biclusters are used in our proposed method.

2) Estimating biclusters: In the Estimating Biclusters process, we score biclusters to use in the updating selection pool step. Algorithm 1 shows the implementation of bicluster estimation. To estimate bicluster D, we calculated sum_{in} and sum_{out} from gene g which is included in the bicluster. sum_{in} is a summation of the mRNA gene expression in samples that are in biclusters and sum_{out} is a summation of a gene gene expression in samples that are not in the biclusters. After calculating sum_{in} and sum_{out}, we calculate avg_{in} and avg_{out} [Lines #1-10]. We then calculate the absolute value of $(avg_{in} * avg_{in}) + |(avg_{in} * (avg_{in} - avg_{out})|$ for g's score (=$gene_{score}$) [Line #11]. The $gene_{score}$ is affected by g's expression value and the difference between g's expression value in the sample in the bicluster and the samples not in the bicluster. Finally, we can obtain bicluster D's score (=D_{score}), which is a summation of all of the gene's scores in D. From this scoring method; we want to find biclusters according to the following two rules as we mentioned in introduction section: i) genes in a bicluster should be statistically coexpressed in the samples in the bicluster and ii) genes in a bicluster should be differentially expressed between the samples in the bicluster

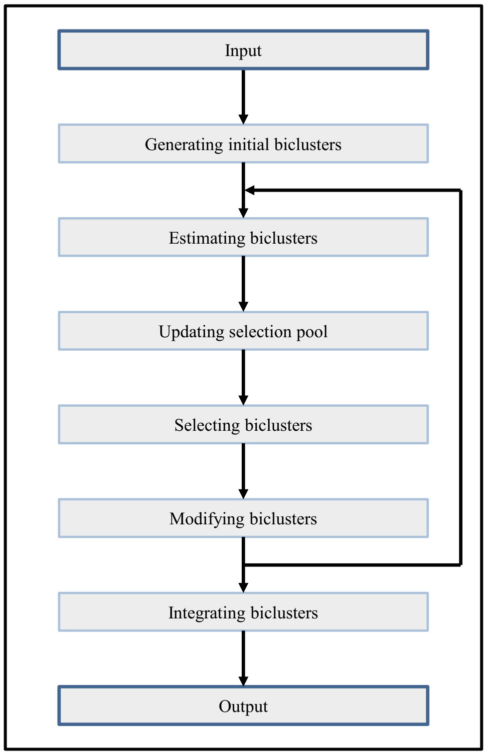


Fig. 3. System overview

Algorithm 1: Estimating biclusters

Input: bicluster B which has genes and samples

Output: score of bicluster B_{score}

Notation

sum_{in} : the summation of a gene expression in a bicluster

sum_{out} : the summation of a gene expression not in a bicluster

avg_{in} : the average of a gene expression in a bicluster

avg_{out} : the average of a gene expression not in a bicluster

$gene_{score}$: the score of a gene

```
FOR each gene g in bicluster B
    $sum_{in} = sum_{out} = 0$
    FOR each sample s in gene expression data
        IF the s in B
            $sum_{in} = sum_{in}$ +g's expression in s
        ELSE
            $sum_{out} = sum_{out}$ + g's expression in s
        ENDIF
    ENDFOR
    $avg_{in} = |sum_{in}|$ / The number of samples in B
    $avg_{out} = |sum_{out}|$ / The number of samples not in B
    $gene_{score} = (avg_{in} * avg_{in}) + (avg_{in} * (|avg_{in} - avg_{out}|)$
    $B_{score} = B_{score} + gene_{score}$
ENDFOR
```

and the other samples.

Algorithm 2: Updating the selection pool

Input: candidate bicluster B which has genes and samples, score of bicluster B_{score} , selection pool S which has biclusters

Output: update selection pool S

Notation

threshold: The threshold of overlap rate between B and B'

```
1: FOR each bicluster B' in selection pool S
2:      FOR each gene g in candidate bicluster B
3:          IF the g in B'
4:              overlap_count++
5:          ENDIF
6:      ENDFOR
7:      IF overlap_count > threshold * gene, set size of B'
8:          IF $B_{score} > B'_{score}$
9:              update S by changing B' to B
10:         ENDIF
11:     ELSE
12:         Insert B into S
13:     ENDIF
14:ENDFOR
```

3) Updating the selection pool: In the step for updating the selection pool, we compare the gene set of a candidate bicluster that received a higher score than the threshold in the previous step and a gene set of biclusters in the selection pool, and decide whether to include a new bicluster in the pool and update the pool accordingly. In this step, we want to solve the loss of diversity problem between biclusters and keep diverse biclusters which have the potential to be functionally-related gene modules. Algorithm 2 shows the implementation of updating the selection pool. We compare the gene set of a candidate bicluster and every gene set of biclusters in the selection pool [Lines #2-6]. If more than r % of the genes between a candidate bicluster and a selection pool bicluster overlap, which is the threshold based on B, we compare the two biclusters scores. Then, if the score of the candidate bicluster is higher than the score of the bicluster in the selection pool, we replace the selection pool bicluster with the candidate bicluster [Lines #7-10]. If less than r % of the genes overlap between a candidate bicluster and a selection pool bicluster, we insert the candidate bicluster into the selection pool [Lines #11-14]. In this step, the selection pool gives various biclusters a chance to be trained. Additionally, for filtering overlapping bicluster, we heuristically set a threshold of r=20, which showed the best performance.

4) Selecting biclusters: Selecting biclusters is a process that selects biclusters to be used in the next iteration. In this process, we select k biclusters by repeating the selection method k times. The selection method selects a bicluster that is included in the selection pool with a probability of p and selects a bicluster that is included in the initial sets with a probability of (1-p) to ensure the diversity of biclusters.

5) Modifying biclusters: Modifying biclusters is a process that modifies biclusters for diversity. Genes and samples between two biclusters are switched and mutated by randomly deleting and adding samples and genes from biclusters. At first, we additionally make $2 * {}_k\mathrm{C}_2$ new biclusters by switching randomly selected genes and samples from k selected biclusters. In other words, the crossover process is applied for every possible pair of biclusters selected in 4) and then two biclusters that underwent crossover are generated. As a result, we get k^2 $(= 2 * {}_k\mathrm{C}_2 + k)$ biclusters which include the originally selected biclusters. Additionally, we make $4 * k^2$ biclusters by randomly deleting and adding samples and genes to the k^2 biclusters in the previous step, and finally we obtain $5 * k^2$ biclusters.

6) Integrating biclusters: In the integrating biclusters process, we integrate biclusters that are included in the selection pool to generate a larger gene module. Firstly, we filter biclusters that are not trained enough in the selection pool based on the threshold, and check whether there are some biclusters that have the same sample set. If two or more biclusters have the same sample set, these biclusters are integrated into one bicluster.

III. Result

For our experiments, we used a Windows 7 operating system with an Intel Core i5-3470, 3.2 GHz, 16.00 GB RAM machine and we implemented our algorithm using the Java language. We used k=10 and p=0.9 in section 2.B.4, which showed the best performance in the proposed method.

A. Comparison of gene ontology terms

Table 3 represents the number of gene modules that are found from existing and proposed biclustering methods in the gene expression datasets GDS181, GDS1027, GDS3715, and GDS3716. Rows represent biclustering methods and columns represent gene expression datasets. For a comparison experiment, we use the BicAT_v2.22 tool [4] which generates biclusters of existing methods, including OPSM, ISA, and CC from gene expression data. Other biclusters are acquired from applications which are provided by various authors in their papers. The proposed method finds the largest number of biclusters in GDS1027 and GDS3715. On the other hand, it finds the smallest number of biclusters in GDS181 because the proposed method is greatly influenced by the size of the gene expression data. The size of the gene expression data is the number of rows multiplied by the number of columns. Because we essentially find genes which are statistically overexpressed or under expressed. And, the number of such overexpressed or under expressed genes is directly proportional to the size of the gene expression data. As seen in table 2, the size of GDS181 is smaller than other data, and then it cause this result.

TABLE III
THE NUMBER OF GENE MODULES FROM BICLUSTERING METHODS

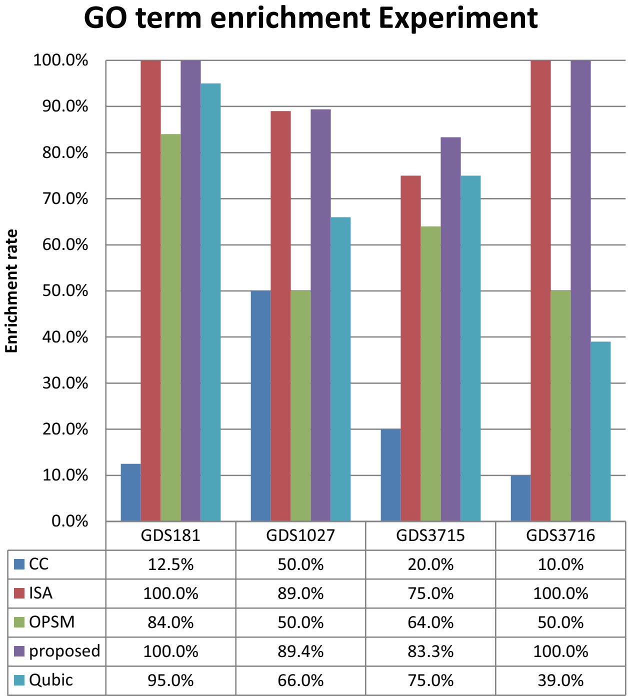
	GDS181	GDS1027	GDS3715	GDS3716
CC	8	10	10	10
ISA	28	46	8	9
OPSM	13	6	14	12
QUBIC	100	6	4	53
Proposed	8	57	30	15

To identify the functional association of the gene modules that are found by the proposed method, we analyzed the gene modules using the FuncAssociate tool [10]. The FuncAssociate tool is a GO term enrichment algorithm which takes a gene module as the input and then shows enriched GO terms that the gene module shares, along with the p-value. We use a default parameter, which is provided by the FuncAssociate tool; it includes evidence codes, p-value cut-offs, etc. GO terms represent genes and gene product attributes in terms of their related biological processes, cellular elements, and molecular functions. Figure 4 shows the GO term enrichment experiment result using GDS181, GDS1027, GDS3715, and GDS3716. GDS181, GDS3715, and GDS3716 are gene expression profiles from Homo sapiens tissues, and GDS1027 is the gene expression profile from Rattus norvegicus tissues. Each bar represents the comparison method. The proposed method showed a better performance than other comparison methods. This indicates that the proposed method is better at finding functionally-related biclusters than other comparison methods.

In addition, we analyzed the gene modules that receive the highest score using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7. DAVID [14], [15] is a tool that provides a comprehensive set of functional annotation tools. We mapped the gene module onto the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway [19], [20] using DAVID in GDS3716. GDS3716 is the gene expression profile of prophylactic mastectomy patients and breast cancer patients. The results indicated that our gene module is closely linked to the ribosome KEGG pathway. The p-value, which represents the significance of the gene module over the KEGG pathway, is calculated using DAVID v6.7. The calculated p-value is $6.7e^{18}$. Many recent studies have reported that dysregulation of ribosome biogenesis is associated with breast tumor progression [24], [26].

B. Analysis connection between gene module and disease

In this section, we build a pathway to understand the roles of gene modules which are found in breast cancer samples by the proposed method. The pathway is built based on the KEGG pathway and we add associations supported by existing studies. Figure 5 shows relationships between some of the found gene modules (#4, 9,11) and the cancer pathway. According to existing studies [9], [13], [32], RPL5, 11 and RPS7 bind to MDM2, which is known as a breast cancer gene, and it affects TP53 accumulation. RAC1, included in the found gene module #11, activates JNK and the TP53 and the JNK indirectly activate, evading the apoptosis process which is a hallmark of cancer. Moreover, RPL11 in module #9 binds MYC [21] and HSP activates KLK3, and then it indirectly



	GDS181	GDS1027	GDS3715	GDS3716
CC	12.5%	50.0%	20.0%	10.0%
ISA	100.0%	89.0%	75.0%	100.0%
OPSM	84.0%	50.0%	64.0%	50.0%
proposed	100.0%	89.4%	83.3%	100.0%
Qubic	95.0%	66.0%	75.0%	39.0%

Fig. 4. GO term enrichment result. The x-axis represents data sets, and the y-axis represents the enrichment rate, which is the proportion of biclusters for which the adjusted p-value is less than or equal to 0.05.

affects the proliferation process which is included in the cancer KEGG pathway. Summing up, there are a total of 3 gene modules and 6 genes that are included in this pathway.

IV. Conclusion

In this paper, we propose the DSS(Diverse and State Specific) method to find not only functionally-related gene modules, but also state specific gene modules using an improved genetic algorithm. Because a genetic algorithm mimics the process of natural selection, it may incur a loss of diversity problem. Therefore, we add a selection pool to the genetic algorithm to solve this problem. The experimental results show that the proposed method has a better performance than the existing bicluster algorithm for finding functionally-related and disease-related gene modules. Our gene modules are closely linked to ribosome proteins, and recent research has reported that dysregulation of ribosome biogenesis is associated with breast cancer. From this research, we build a breast cancer pathway based on the KEGG pathway and associations supported by existing studies to understand the roles of gene modules. In this paper, we proved that our method performs well for finding functionally-related gene modules and state specific gene modules. Furthermore, the proposed method can help to improve disease pathways more intricately.

Acknowledgment

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2015R1A2A1A05001845).

References

[1] 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.

[2] Akdes Serin, Martin Vingron *DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach*, Algorithms for molecular biology, vol. 6(18), 2011.

[3] Anupam Chakraborty, Hitashyam Maka, *Biclustering of Gene Expression Data Using Genetic Algorithm*, IEEE Symposium on computational intelligence in Bioinformatics and Computational Biology, pp.18, 2005.

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

[5] 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 setsupdate*, Nucleic Acids Res, vol.41(Database issue), pp.D991-D995, 2013.

[6] Bozdag D, Kumar A, Catalyurek UV, *Comparative analysis of biclustering algorithms*, In: Proceedings of 1st ACM, International Conference Bioinformatics and Computational Biology, pp.265274, 2010.

[7] C. Cano, L. Adarve, J. Lopez, A. Blanco, *Possibilistic approach for biclustering microarray data*, Computers in Biology and Medicine, vol.37, pp.1426-1436, 2007.

[8] Dillman JF 3rd, Phillips CS, Dorsch LM, Croxton MD et al., *Genomic analysis of rodent pulmonary tissue following bis-(2-chloroethyl) sulfide exposure*, Chem Res Toxicol, vol.18(1), pp.28-34, 2005.

[9] Everardo Macias, Aiwen Jin, Chad Deisenroth, Krishna Bhat, Hua Mao, Mikael S. Lindstrm, Yanping Zhang, *An ARF-Independent c-MYC-Activated Tumor Suppression Pathway Mediated by Ribosomal Protein-Mdm2 Interaction*, Cancer Cell, vol.18(3), pp.231-243, 2010.

[10] 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.

[11] Graham K, de las Morenas A, Tripathi A, King C et al., *Gene expression in histologically normal epithelium from breast cancer patients and from cancer-free prophylactic mastectomy patients shares a similar profile*, Br J Cancer, vol.102(8), pp.1284-1293, 2010.

[12] Guojun Li, Qin Ma, Haibao Tang, Andrew H. Paterson and Ying Xu, *QUBIC: a qualitative biclustering algorithm for analyses of gene expression data*, Nucleic Acids Research, vol. 37 pp.e101, 2009.

[13] HF Horn and KH Vousden, *Cooperation between the ribosomal proteins L5 and L11 in the p53 pathway*, Oncogene, vol.27, pp.5774-5784, 2008.

[14] Huang DW, Sherman BT, Lempicki RA., *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*, Nucleic Acids Res, vol.37(1), pp.1-13, 2009.

[15] Huang DW, Sherman BT, Lempicki RA., *Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources*, Nature Protoc, vol.4(1), pp.44-57, 2009.

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

[17] Jaegyoon Ahn, Youngmi Yoon, Sanghyun Park, *Noise-robust algorithm for identifying functionally associated biclusters from gene expression data*, Information Sciences, vol. 181 pp.435-449, 2011.

[18] Juan A. Nepomuceno, Alicia Troncoso, Jesus S. Aguilar-Ruiz, *Scatter search-based identification of local patterns with positive andnegative correlations in gene expression data*, Applied Soft Computing, vol. 35, pp.637-651, 2015.

[19] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M, *Data, information, knowledge and principle: back to metabolism in KEGG*, Nucleic Acids Res, vol.42, pp.D199-D205, 2014.

[20] Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M, *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Res, vol.28, pp.27-30, 2000.

[21] Mu-Shui Dai, Hua Lu, *Crosstalk between c-Myc and ribosome in ribosomal biogenesis and cancer*, Journal of Cellular Biochemistry, vol.105(3), pp.670-677, 2008.

[22] 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, vol. 67(3 Pt 1) pp. 03190201-18, 2003.

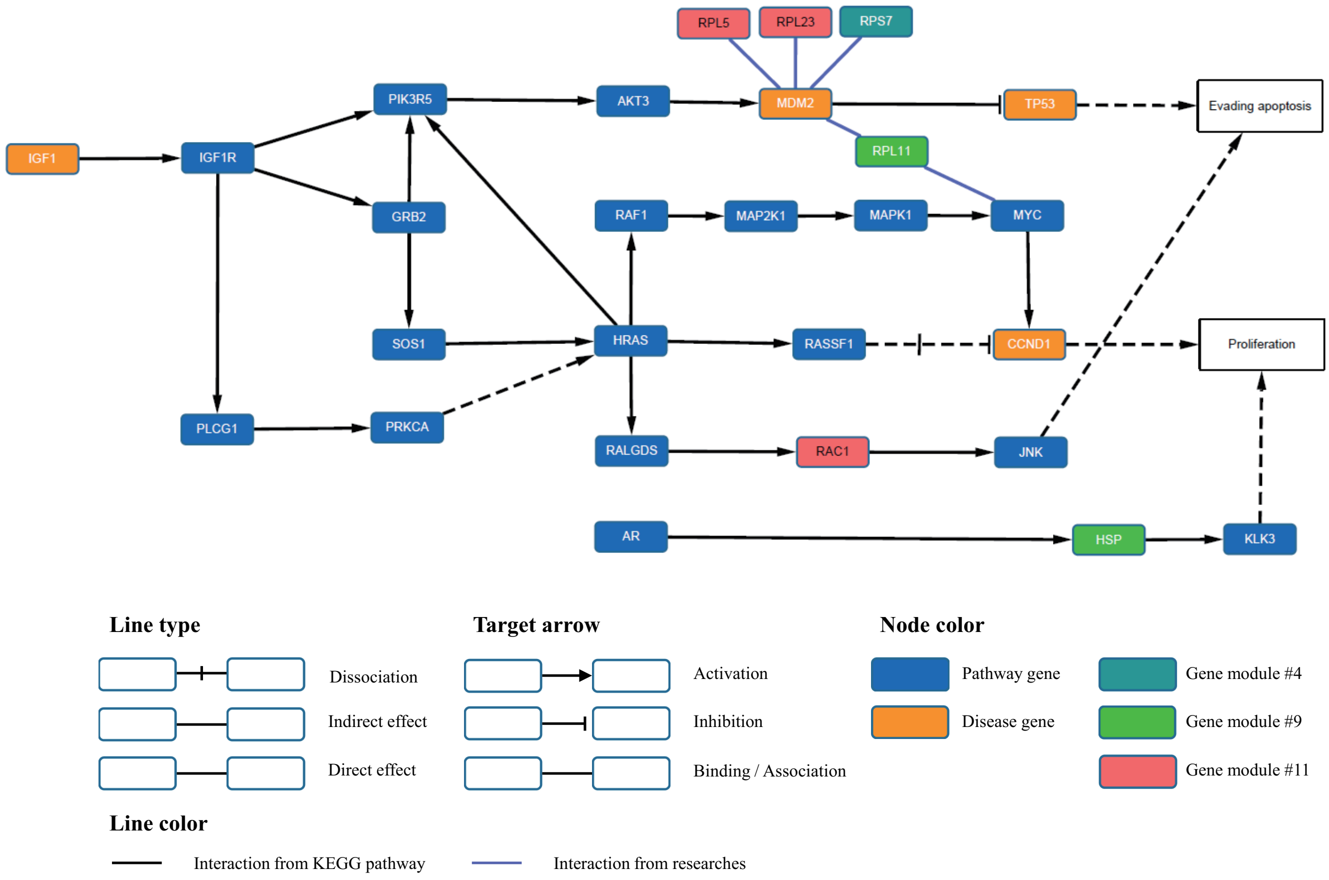


Fig. 5. The breast cancer pathway built based on the KEGG pathway and associations supported by existing studies.

[23] Shahreen Kasim, Safaai Deris, Razib M. Othman, *Multi-stage filtering for improving confidence level and determining dominant clusters in clustering algorithms of gene expression data*, Computers in Biology and Medicine, vol.43(9), pp.1120-1133, 2013.

[24] Stephane Belin., Anne Beghin., Eduardo Solano-Gonzaez., Laurent Bezin,, Stephanie Brunet-Manquat., Julien Textoris., Anne-Catherine Prats., Hichem C. Mertani., Charles Dumontet., Jean-Jacques Diaz., *Dysregulation of Ribosome Biogenesis and Translational Capacity Is Associated with Tumor Progression of Human Breast Cancer Cells*, PLOS One, vol.4(9), p.e7147, 2009.

[25] Su AI, Cooke MP, Ching KA, Hakak Y et al., *Large-scale analysis of the human and mouse transcriptomes*, Proc Natl Acad Sci U S A, vol.99(7), pp.4465-4470, 2002.

[26] Swagat Ray., Rebecca Johnston., David C. Campbell., Shannen Nugent., Simon S. McDade., David Waugh., Konstantin I. Panov., *Androgens and estrogens stimulate ribosome biogenesis in prostate and breast cancer cells in receptor dependent manner*, Gene, vol.526(1), pp.46-53, 2013.

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

[28] 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.

[29] Wu X, Patki A, Lara-Castro C, Cui X et al., *Genes and biochemical pathways in human skeletal muscle affecting resting energy expenditure and fuel partitioning*, Journal of Applied Physiology, vol.110(3), pp.746-755, 2011.

[30] Wu X, Wang J, Cui X, Maianu L et al., *The effect of insulin on expression of genes and biochemical pathways in human skeletal muscle*, Endocrine, vol.31(1), pp.5-17, 2007.

[31] 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.

[32] Yanping Zhang, Gabrielle White Wolf, Krishna Bhat, Aiwen Jin, Theresa Allio, William A. Burkhart and Yue Xiong, *Ribosomal Protein L11 Negatively Regulates Oncoprotein MDM2 and Mediates a p53-Dependent Ribosomal-Stress Checkpoint Pathway*, Molecular and Cellular Biology, vol.23(23), pp.8902-8912, 2003.

[33] Yu X, Griffith WC, Hanspers K, Dillman JF 3rd et al., *A system-based approach to interpret dose- and time-dependent microarray data: quantitative integration of gene ontology analysis for risk assessment*, Toxicol Sci, vol.92(2), pp.560-577, 2006.