# Co-evolutionary Biclustering for microRNA Expression Profiles Analysis

**Soo-Jin Kim**[1, 2] **Je-Gun Joung**[1, 2] **Byoung-Tak Zhang**[1, 2, 3]
[1] *Graduate Program in Bioinformatics, Seoul National University, Seoul, Korea*
[2] *Center for Bioinformation Technology, Seoul National University, Seoul, Korea*
[3] *School of Computer Science and Engineering, Seoul National University, Seoul, Korea*
*Email: sjkim@bi.snu.ac.kr; jgjoung@bi.snu.ac.kr; btzhang@bi.snu.ac.kr*

**ABSTRACT**: Small non-coding RNAs, known as microRNAs (miRNAs), have critical functions across various biological processes. They can play important regulatory roles in animals and plants via the RNA-interference pathway by targeting mRNAs for the cleavage or translational repression. This regulatory process also can be involved in the cancer development and progression. However, the specific function of most human miRNAs is unknown. Recently, the miRNA expression profiling method using high-throughput microarray technology has allowed us to study the functional roles of miRNAs in the systemic analysis. In this paper, we propose the algorithm to cluster functionally coherent miRNAs effectively. It is based on the biclustering method that clusters simultaneously the rows and columns of an expression matrix. We apply our algorithm to miRNA expression profiles dataset relate to human cancer. The experimental results show that the proposed algorithm can find some biclusters of the biologically significant miRNAs, which have co-regulated expression patterns. Furthermore, for the validation of the biclusters, we performed Gene Ontology (GO) analysis. It evaluate that the target genes of clustered miRNAs indeed are closely related in the specific biological process, such as gene regulations involved in the cellular process and the nervous system development.

## 1 INTRODUCION

MicroRNAs (miRNAs) are endogenous 21-22 nt RNAs that can play important regulatory roles in animals, plants and viruses. They interact with target mRNAs at specific sites to induce cleavage of the message or inhibit translation [1]. Each miRNA has the potential to bind to many different transcripts and down-regulate protein expression of multiple target genes by binding to the mRNA transcripts. Also, they are reported that related nearly to the cancer development and progression [2]. Now, hundreds of different miRNAs have been identified in complex eukaryotes. However, the cellular function of most mammalian miRNAs is unknown [3]. So, it is necessary to understand their biological mechanism that miRNAs are analyzed functionally. According to taking notice of miRNAs importance in various biological processes, recently, their expression profiling methods using high-throughput microarray technology have developed and allowed us to study the functional roles of miRNAs in systemic analysis.

In general, clustering analysis [4][5] is a first step to elucidating the function of genes in expression data. In this manner, we can perform clustering of miRNAs. It is assume that functionally related miRNAs may be co-expressed and co-regulated. Especially, the biclustering is a useful method, because it can identify groups of miRNAs that exhibit a coherent pattern across a specific subset of samples. Several miRNAs group response to specific stimuli that presents only in certain samples. Also, the biclustering has proved method to find interesting patterns in microarray expression data, for different biological samples. Recently, various biclustering algorithms have been introduced in microarray data analysis [6][7][8][9]. However, most algorithms focus on finding possible biclusters of not miRNAs but mRNAs.

In this paper, we propose the biclustering algorithm to cluster effectively functionally coherent miRNAs. It makes use of the concept of co-evolutionary learning. Co-evolutionary learning evolves two populations with the context of each other [10][11][12]. Our algorithm (Co-evolutionary biclustering algorithm, CBA) maintains two populations for a set of miRNA and sample and leads them toward the minimum of objective function. And, it is important to consider the issue of the evolution of interdependent subcomponents in the biclustering [13]. The CBA can be reflected upon interdependent subcomponents, since it in terms of an evolutionary computation permits the interaction between miRNAs and samples. So, it can search on coupled landscapes. Furthermore, we use the global statistical information extracted from a current population to generate offspring. This is based on the idea underlying the estimation of distribution algorithms (EDAs) that have better search ability than that of general population based evolutionary algorithms [14].

And, in order to validate biclusters, we performed Gene Ontology (GO) analysis of genes targeted by clustered miRNAs. Since the miRNA affects regulatory process in the biological mechanism by binding to specific mRNAs. It is evaluated that the roles of these genes indeed are closely related in specific biological process, such as gene regulation involved in the cellular process and the nervous system development. These results supported the quality of found biclusters.

The paper is organized as follows. In section 2, the biclustering problem is defined and our algorithm is described. Experimental results are reported in section 3. Discussion is drawn in section 4.

## 2 METHODS

### 2.1 Biclustering of miRNA expression data

As yielding miRNA expression data, the focus of the miRNAs functional study has shifted on mining the groups as opposed to individual ones that exhibit a similar expression across a wide range of samples. General clustering methods including *k*-means and hierarchical clustering [4] [5] aim at discovering miRNA groups through the similarity measure over all samples. They may fail to uncover miRNA sets involving not all samples but several samples. Often, miRNAs involving in the same pathway are activated in response to specific stimuli that present only in certain samples. In this case, biclustering of in the miRNA expression data is a useful method, since it can identify groups of miRNAs that exhibit a coherent pattern across a subset of samples, not over all samples. We are able to obtain functionally closely related miRNA groups which exhibit a coherent pattern.

Let $M = \{m_1, m_2, \ldots, m_{Nm}\}$ be the set of miRNAs and $S = \{s_1, s_2, \ldots, s_{Ns}\}$ be the set of samples such as different cell lines including some cancer types. The data can be viewed as an $N_m \times N_s$ matrix of real values $E$ (Figure 1). Each entry $e_{ij}$ in matrix is the expression level of a miRNA $m_i$ under a specific sample $s_j$. A bicluter is defined on this miRNA expression profiles matrix.

A bicluster corresponds to a submatrix that exhibits coherence in rows and columns. Let $I$ denotes the set of indices of the rows in a row cluster and $J$ denote the set of indices of the columns in a column cluster. Thus, a bicluster is a matrix, denoted as $(I, J)$, where $I$ and $J$ are the set of miRNAs (rows) and samples (columns), respectively, and $| I | \leq | N_m |$ and $| J | \leq | N_s |$. Our goal is to minimize the following mean squared residue (MSR) score [7]:

$$H_{IJ} = \frac{1}{| I || J |} \sum_{i \in I, j \in J} h_{ij}^2 , \qquad (1)$$

where the residue of an element $e_{ij}$ in bicluter determined by index sets $I$ and $J$ is defined as

$$h_{ij} = e_{ij} - e_{iJ} - e_{Ij} + e_{IJ} , \qquad (2)$$

and

$$e_{iJ} = \frac{\sum_{j \in J} e_{ij}}{| J |} , \quad e_{Ij} = \frac{\sum_{i \in I} e_{ij}}{| I |} , \quad e_{IJ} = \frac{\sum_{i \in I, j \in J} e_{ij}}{| I || J |} , \qquad (3)$$

$e_{iJ}$ is the mean of the entries in row $i$ whose column indices are in $J$, $e_{Ij}$ is the mean of the entries in column $j$ whose row indices are in $I$, and $e_{IJ}$ is the mean of all the entries in the bicluter.

In order to find a bicluster, we designed the fitness function by employing the MSR score. Our algorithm minimizes the MSR score. It is the variance of the set of all elements in the bicluster, adding to the mean differences between row values and the mean difference between column values. Thus, the lower MSR score means the stronger coherence. Finally, the bicluster makes miRNAs hold coherent trends under specific set of samples.

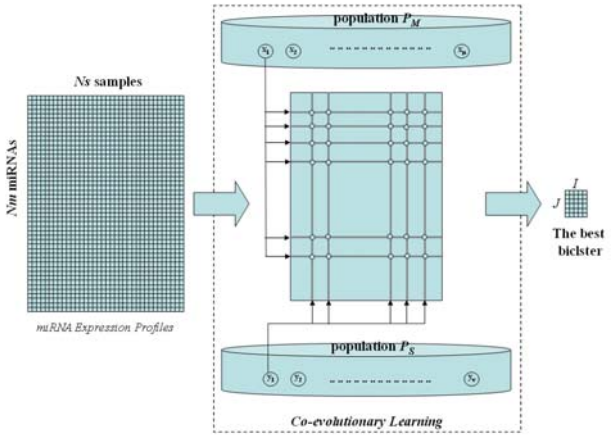


Figure 1. The schematic diagram of co-evolutionary biclustering algorithm (CBA) for miRNA expression profiles analysis.

### 2.2 The co-evolutionary algorithm for biclustering

In this section, we describe in detail the algorithm. Figure 1 presents the schematic diagram of our algorithm on the miRNA expression analysis. The algorithm co-evolves two populations for a set of miRNA and sample. These miRNAs and samples correspond to the row and column in the expression data of Figure 1, respectively. Each individual contains selected indices from the miRNA or sample set. The fitness function is measured by combining individuals of two populations. Then, two populations are updated based on statistical information extracted from previous them. As generation goes, probabilities of miRNAs and samples are changed by a cooperative co-evolutionary manner of two populations.

#### 2.2.1 Representation

Our algorithm has two populations for a set of miRNA and sample. The population $P_M$ for the miRNA set consists of $\{x_1, \ldots, x_\mu\}$ and the population $P_C$ for the sample set consists of $\{y_1, \ldots, y_v\}$. Here, each individual $x_i$ contains indices of several $w_x$ miRNAs among $\{m_1, m_2, \ldots, m_{Nm}\}$ and each individual $c_i$ contains indices of several $w_y$ samples among $\{s_1, s_2, \ldots, s_{Ns}\}$. The combination of $w_x$ and $w_y$ represents the volume size of the bicluster.

#### 2.2.2 Algorithm

Initial populations of $Pop_M$ and $Pop_S$ are created randomly. Each individual is measured by the fitness function. Then, individuals with best fitness are selected and the probabilities $P_M = \{p_1, p_2, \ldots, p_{Nm}\}$ and $P_S = \{p_1, p_2, \ldots, p_{Ns}\}$ are updated. Next, the populations are generated based on these current probability vectors.

**Algorithm 1** Co-evolutionary biclustering

*Initialize*

$Pop_M(0) := \{x_1(0), \ldots, x_\mu(0)\}$,

$Pop_S(0) := \{y_1(0), \ldots, y_v(0)\}$;

Evaluation

$Pop_M(0) : \{\Phi(x_1(0)), \ldots, \Phi(x_\mu(0))\}$,

$Pop_S(0) : \{\Phi(y_1(0)), \ldots, \Phi(y_v(0))\}$;

**While** ( t < $gen_{max}$ ) **do**
*Selection*:
$Pop`_M(t) := s(Pop_M(t))$,
$Pop`_S(t) := s(Pop_S(t))$;
*Probability Update*:
$P_M(t+1) := u(P_M(t))$,
$P_S(t+1) := u(P_S(t))$;
*Generation*:
$Pop_M(t+1) := \pi(P_M(t+1))$,
$Pop_S(t+1) := \pi(P_S(t+1))$;
*Evaluation*:
$Pop_M(t+1) : \{\Phi(x_1(t+1)), ..., \Phi(x_\mu(t+1))\}$,
$Pop_S(t+1) : \{\Phi(y_1(t+1)), ..., \Phi(y_\upsilon(t+1))\}\}$; $t := t+1$;
**end.**

### 2.2.3 Initialization and Update of the Probability

The next population $Pop(t+1)$ is generated by sampling a vector of probabilities. An initial vector has a uniform distribution.

$$Pop_M(t+1) \leftarrow \pi(p_1, p_2, ..., p_{Nm})$$
$$Pop_S(t+1) \leftarrow \pi(p_1, p_2, ..., p_{Ns}) \quad (4)$$

Here, each probability for a set of miRNA and sample is updated by the following equations:

$$p_i = (1-\alpha)p_i + \alpha \frac{f_{m_i}}{\sum_{k=1}^{B_m} f_{m_k}}, i \in I$$
$$p_j = (1-\beta)p_j + \beta \frac{f_{s_j}}{\sum_{k=1}^{B_s} f_{s_k}}, j \in J \quad (5)$$

where $\alpha$ and $\beta$ are parameters for controlling updates and $f$ is the frequency of selected miRNA and sample. In each generation, the best individuals ($B$) are selected, and each probability is updated by considering the fraction of each miRNA or sample.

### 2.2.4 Fitness function

The fitness function is defined as the mean squared residue (MSR) scores of the biclusters. The fitness of the individual for the miRNA set is an average MSR, when it is combined with individuals for the sample set. The fitness of individual for sample set is measured in same way. The following fitness functions for a set of miRNA and sample are used:

$$\Phi(x_i) = \frac{\sum_{k=1}^{Z_m} H^B_{x_i y_k}}{Z_m}, \quad \Phi(y_j) = \frac{\sum_{k=1}^{Z_s} H^B_{x_k y_j}}{Z_s} \quad (6)$$

$H^B$ means the lowest score of MSR that represents a high homogeneous bicluster. $Z_m$ and $Z_s$ are the number of the best individuals for the miRNA set and the sample set, respectively.

If the individual has good performance by combining with the opposite best individuals, it may have a high fitness. This measurement of fitness is efficient at extracting the global structure. The number of the best individuals $Z_m$ and $Z_s$ can be controlled as the generation goes by

$$Z \begin{cases} Z\gamma & if \quad Z \geq 1, \\ 1 & otherwise, \end{cases} \quad (7)$$

| Parameters | setting |
|---|---|
| Num. of $Pop_M$ | 1200 |
| Num. of $Pop_S$ | 600 |
| Num. of generations | 300 |
| Initial $Z_m$, $Z_s$ | 600, 360 |
| $\gamma$(miRNA, sample) | 0.96, 0.8 |
| $\alpha$, $\beta$ | 0.8, 0.8 |
| $w_x$, $w_y$ | 15, 10 |

Table 1. Experimental parameter setting

where $\gamma \in (0, 1)$. The large value of $\gamma$ gives rise to decreasing $Z$ quickly. It means that it has a high rate on converting the global search to the local search is high.

## 2.3 Preparation of miRNAs expression dataset

We applied the CBA to a microarray dataset, which contains expression profiles of miRNAs in human. The experimental dataset was obtained from Volinia *et al.* [2]. This dataset consists of an expression matrix of 381 miRNAs (rows) and 155 samples (columns). In detail, these samples consist of 115 primary tumors and 40 normal tissues, and were obtained from three solid tumors, which were gastric carcinoma, prostate carcinoma and endocrine pancreatic tumor.

The normalization process was executed as follows. First of all, expression values of replicated miRNAs were averaged and log transformed. Then, it was performed by adjusting to a median in each chip. And, negative values converted to 0.001 in each sample, since expressed values are subtracted background from intensity in cDNA microarray data, and they were considered as errors.

# 3 RESULTS

## 3.1 Experimental parameters setup

The experimental parameters were described in Table 1. The sizes of two populations were set differently, because the number of miRNAs is greater than that of samples. The maximum number of generation was 300. Initial fraction of the best individuals for a fitness measure was 0.3 and the decreasing rates of the best individuals were 0.96 and 0.8. The parameters for probabilities $\alpha$ and $\beta$ were 0.8, respectively. The parameters $w_x$ and $w_y$, the size of bicluster, were fixed 15 and 10, respectively.

## 3.2 Identification of coherent miRNAs block

The CBA was executed on the miRNA expression dataset. Figure 2 shows the biclustering results. Biclusters found by our algorithm contain highly homogeneous elements. The expression of miRNAs presents a similar pattern in specific samples. This means that clustered miRNAs may be functionally related in specific samples.

Also, CBA is able to detect miRNAs which are activated in response to specific stimuli, since it calculates similarity over selected samples. On the contrary, it is hard to find such miRNAs by other clustering methods because of the similarity measure over overall samples.

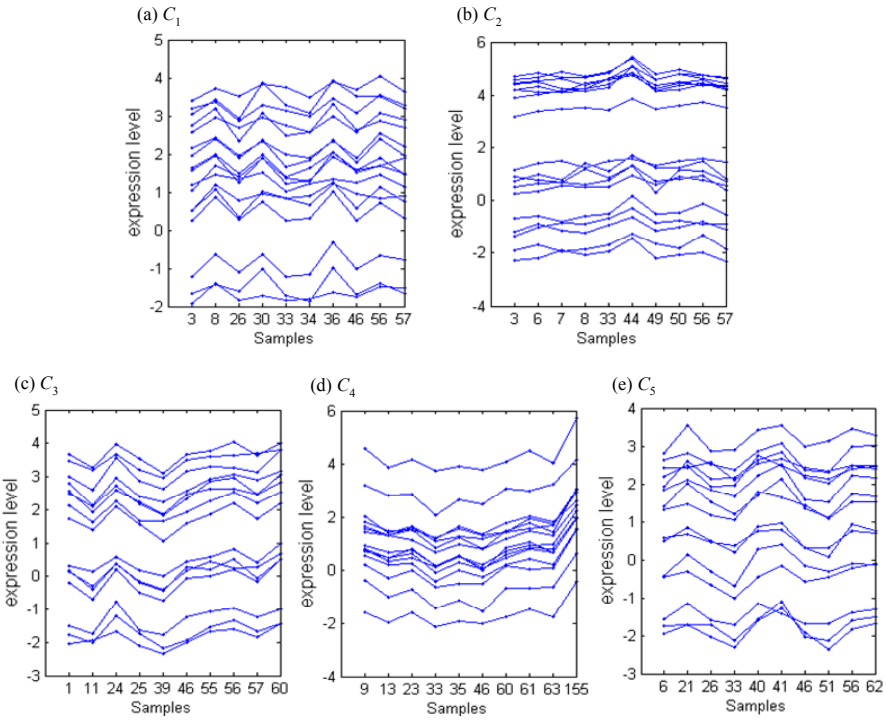


Figure 2. The result of solutions from the miRNA expression dataset. Our algorithm found five biclusters ($C_1$, $C_2$, $C_3$, $C_4$ and $C_5$) containing highly homogeneous elements. The aspects of miRNA expression in each bicluster have similar patterns. And, the *x*-axis and *y*-axis represent the expression levels of selected miRNAs and selected samples, respectively.

From this result, we notice that our algorithm can effectively cluster functionally coherent miRNAs. The reason lies in more elaborate searching of a global solution space by evolutionary learning.

### 3.3 Validation with Gene Ontology

We performed GO analysis using target genes of clustered miRNAs in order to validate biclusters. Using GO (The Gene Ontology Consortium or GO project, 2000) has become a standard way to validate the functional coherence of genes in a list. This project aims to develop three structured, controlled vocabularies that describe gene products in terms of their associated biological processes (BP), cellular components (CC), and molecular functions (MF), in a species-independent manner. To determine whether any GO terms annotate genes at a frequency greater than would be expected by chance. Typically, this type of validation is accompanied by a statistical significance analysis. In order to this purpose, we extracted target genes of clustered miRNAs in each bicluster by miRBASE [15]. The analysis using target genes can be biologically significant, since miRNAs determine their functions of target genes in a specific biological context.

Figure 3 shows the annotation of the genes in each bicluster, $C_1$ and $C_2$ with the terms in BP category of GO. It presents how these terms are related in the GO directed acyclic graph (dag). We observed that the abundant terms are related to the nervous development and the cellular process in GO dag, respectively.

Furthermore, we used the tool GO::TermFinder [16] to find significantly over-represented GO terms. This tool calculates a *p*-value calculated by the hypergeometric distribution and also performs the multiple comparison correction. Table 2 shows the shared GO terms used to describe miRNA target genes of each bicluster. We display ed only the significantly shared terms with a adjust *p*-value less than 0.05. Also, when multiple hierarchical terms were involved in the same group of genes, we reported only the most significant terms. For example, for bicluster $C_1$, we find significant genes involved in development of nervous system and morphogenesis. And, the result of bicluster $C_2$ shows that finding genes relate to the cellular and biological process. These results indicate that CBA can find potentially biologically significant biclusters.

## 4 DISCUSSION

We proposed a co-evolutionary biclustering algorithm (CBA) that clusters simultaneously rows and columns of an expression matrix based on co-evolutionary learning. Co-evolutionary learning employs the probabilistic search based on EDA. We applied our algorithm to miRNA expression profiles dataset. It provides an efficient procedure for the discovery of miRNA expression patterns on specific samples, since the algorithm tried to decompose the task by co-evolutionary learning in complex problem of large-scale matrix. Accordingly, it can find biclusters which contain highly homogeneous elements. The clustered miRNA of in a bicluster is functionally closely related that it exhibits a

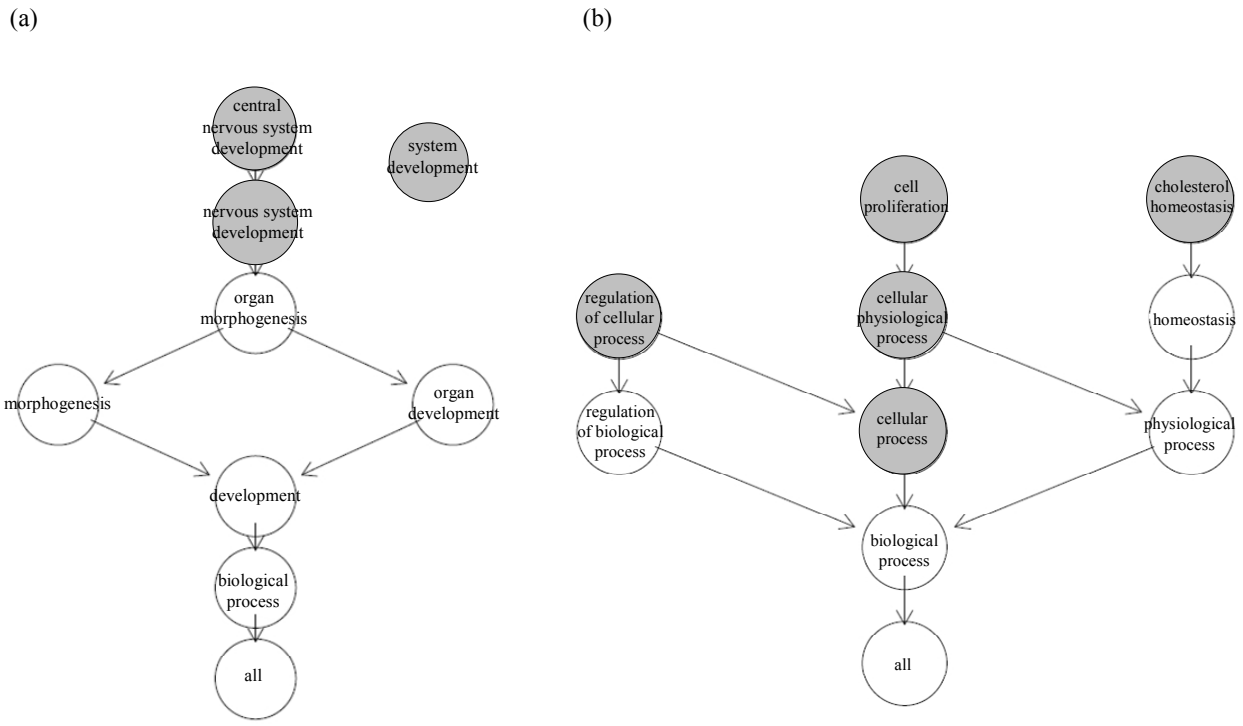


Figure 3. The annotation with GO terms. The GO dag shows the relationship among the GO terms. The colored nodes indicate terms of $p$-value < 0.05. The GO dags correspond to bicluster $C_1$ (a) and $C_2$ (b), respectively.

Table 2. The significantly shared Gene Ontology terms (BP, CC and MF) for target genes in different biclusters.

| Cluster (size) | Gene Ontology Term | Annotated genes | Adjust p-value |
|---|---|---|---|
| $C_1$ (10x17) | BP: nervous system development | HES1, EFNB2, UBE3A, PTEN | 0.0002 |
| | BP: central nervous system development | UBE3A, PTEN | 0.0003 |
| | BP: system development | UBE3A, PTEN, HES1, EFNB2, UBE3A, PTEN | 0.0003 |
| $C_2$ (10x18) | BP: cellular physiological process | CPNE3, RBMS1, SPC18, CTGF, AP1M2, EYA4, G3BP, LAMC1, GNAI3, MYH9, MTMR6, PECI, CDK6, LASS2, PTBP1, SP1, VAMP3, BTG3, ELF4, DFFB, CDK4, CEBPA, IFRD2, RYK, NFIC, CAV1, SMAD5, PLP2, TLN1, APEX2, UHRF1, PLOD3, NEK9, DCTD, NEK6, LITAF, ARH, UBE3A, AHR, PLDN, STX10, TEAD1, PGM1, ABHD5, PTEN | 0.02 |
| | BP: cholesterol homeostasis | CAV1, ARH | 0.02 |
| | BP: regulation of biological process | ELF4, CDK4, CEBPA, RBMS1, NFIC, SMAD5, CTGF, UHRF1, EYA4, LAMC1, GNAI3, CDK6, NEK6, LITAF, LASS2, ARH, AHR, SP1, TEAD1, BTG3, SOS2, PTEN | 0.02 |
| | BP: cellular process | CPNE3, RBMS1, SPC18, CTGF, AP1M2, EYA4, G3BP, LAMC1, GNAI3, MYH9, MTMR6, PECI, CDK6, DVL2, CD59, LASS2, PTBP1, SP1, VAMP3, BTG3, CAV1, SMAD5, PLP2, EFNB2, PODXL, TLN1, APEX2, UHRF1, PLOD3, NEK9, DCTD, NEK6, LITAF, ARH, UBE3A, AHR, PLDN, STX10, TEAD1, PGM1, ABHD5, PTEN | 0.03 |
| | BP: cell proliferation | ELF4, LAMC1, CDK4, CDK6, BTG3, IFRD2, PTEN | 0.03 |
| | CC: intracellular part | CPNE3, RBMS1, SPC18, AP1M2, EYA4, G3BP, MYH9, PECI, STOM, DVL2, LASS2, PTBP1, SP1, BTG3, ELF4, DFFB, LMNB1, CEBPA, NFIC, CAV1, SMAD5, PLP2, TLN1, APEX2, UHRF1, PLOD3, NEK6, AHR, ARH, UBE3A, STX10, TEAD1, PGM1, PTEN | 0.02 |
| | CC: endomembrane system | STX10, LMNB1, PLP2, AP1M2, CAV1, ARH | 0.02 |
| | MF: protein binding | ELF4, DFFB, CDK4, CEBPA, CAV1, CTGF, BDNF, PLP2, EFNB2, TLN1, UHRF1, G3BP, LAMC1, PLOD3, MYH9, CDK6, DVL2, CD59, PLDN, ARH, AHR, PTBP1, SP1, TEAD1, SOS2, PTEN | 0.002 |

coherent pattern. The result shows that it offers biologically significant biclusters through searching global solution space. In future work, to search more biologically significant clusters, we plan to develop additional new techniques for miRNA analysis. We expect that this study will provide an application based on the evolutionary computation to extract meaningful patterns from a dataset of matrix form generated in various fields including miRNA.

## ACKNOWLEDGEMENT

This work was supported by the Korean Ministry of Science and Technology (MOST) through National Research Lab (NRL) project.

## REFERENCES

[1] D. P. Bartel, *et al*., MicroRNA: genomics, biogenesis, mechanism, and function, *Cell*, 116: 281--297, 2004.

[2] V. Stefano, *et al*., A microRNAs expression signature of human solid tumors define cancer gene targets, *PNAS*, 103: 2257--2261, 2006.

[3] J. Bino, *et al*., Human microRNA targets, *PLoS Biol*., .2(11): e363, 2004.

[4] M. B. Eisen., *et al*., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci*., 95(25): 14863--14868, 1998.

[5] R. Herwig, *et al*., Large-scale Clustering of cDNA-fingerprinting data, *Genome Research*, 9(11): 1093--1105, 1999.

[6] S. C. Madeira, *et al*., Biclustering algorithms for biological data analysis: a survey, *IEEE/ACM Transaction on Computational Biology and Bioinformatics*, 1: 24--45, 2004.

[7] Y. Cheng, *et al*., Biclustering of expression data, In *Proceedings ISMB*, 93--103, 2000.

[8] S. Bergmann, *et al*., Iterative signature algorithm for the analysis of large scalegene expression data, *Phys. Rev. E. Stat. Nonlin Foft Matter Phys*., 67: 0319201--18, 2003.

[9] J. Yang, *et al*., Enhanced biclustering on expression data, In *Proceedings of the 3rd IEEE Conference on Bioinformatics and Bioengineering*, 321--327, 2003.

[10] D. W. Hillis, Co-evolving parasites improve simulated evolution in an optimization procedure, *Physica*, 42: 228--234, 1990.

[11] R. Axelrod, The evolution of strategies in the iterated Prisoner's Dilemma, In *L. Davis (Ed.), Genetic Algorithms and Simulated Annealing, London: Pitman Publishing*, 32--41, 1987.

[12] N. A. Barricelli, Numerical testing of evolution theories, Part I: theoretical introduction and basic tests, *Acta Biotheoretica*, 16: 69--98, 1962.

[13] M. Petter, *et al*., Cooperative co-evolution: an architecture for evolving coadapted subcomponents, *Evolutionary Computation*, 8: 1--9, 2000.

[14] P. Larranaga, *et al*., Estimation of Distribution Algorithms, A New Tool for Evolutionary Computation, Kluwer Academic Publishers, 2001.

[15] G. J. Sam, *et al*., miRBASE: microRNA sequences, targets and gene nomenclature, *Nucleic Acids Research*, 34: 140--144, 2006.

[16] E. I, Boytel, *et al*., GO ::TermFider-open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, 20: 371--3715, 2004.