

## BICLUSTERING ALGORITHM OF DIFFERENTIAL CO-EXPRESSION FOR GENE DATA

Ran Jin<sup>1,3</sup>, Chunhai Kou<sup>2</sup> and Ruijuan Liu<sup>3</sup>

<sup>1</sup> School of Computer Science and Information Technology, Zhejiang Wanli University;

<sup>2</sup> School of Information Science and Technology, Donghua University;

<sup>3</sup> School of Science, Donghua University.

Email: ran.jin@163.com

### ABSTRACT

The gene expression dataset has the characteristic of small number of samples, large number of genes and many genes related to noise and useless genes, it is necessary to work out effective method to find genes related to disease. The differential co-expression bicluster is helpful to find genes related to cancer and biological process of aging. This paper first puts forward a new dispersed method of multi-valued sample based on the rough set, and then proposed the DCECluster biclustering algorithm. It extended nodes with the strategy of depth priority, and added several pruning steps to mine maximal differential co-expression biclusters effectively. The result of experiment shows DCECluster is more efficient than current algorithms.

**Keywords:** Bicluster, Differential gene co-expression, Discretize, Weighted graph, Pruning strategy.

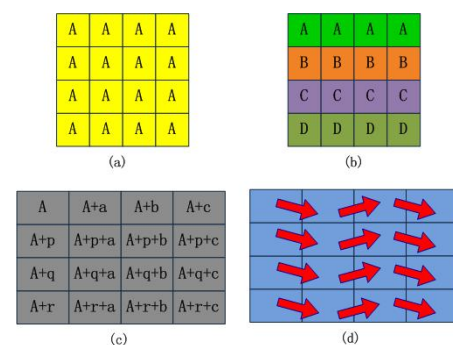
### 1. INTRODUCTION

In recent years, the microarray has been a new high throughput test technology which can simultaneously test thousands of gene data's expression level in the meantime. So it has been a useful tool to research the interrelation of genes in post-gene era. However, how to give a reasonable analysis of the large amount of experimental data brought by this technology and how to reveal the biographic knowledge underneath the gene data have been a bottleneck of effectively using the technology.

Cluster Analysis is a common way to research the microarray. By the Cluster Analysis, it can effectively predict the function of unknown gene and protein. The purpose of traditional Cluster Analysis is to divide the gene according to the gene expression level and classify the genes which have the similar expression pattern into one kind by various mathematical models. Based on this, we can find related genes and research on the function of genes. However, the traditional Cluster Analysis deals with the whole sample sets and have some defects: (a) Incapable to reach Cross Clus. From the biological perspectives, one gene can be involved in many bio-processes or bio-functions which mean genes can exist in various cluster results but the traditional Cluster Analysis can't realize the Cross Cluster. (b) Incapable to reach the partial spatial clustering. Two genes can have the similar expression patterns under some conditions while the traditional cluster analysis can't reach this kind of clustering under partial conditions. (c) Incapable to find the relationship between clusters [1]. Therefore, bicluster emerges as the times require and it has been a widely used excavation method in microarray technology.

According to the characteristics of gene expression in bicluster, Madeira and other scholars divide the bicluster

into the following types [2], as the Fig. 1 shows: (a) The bicluster whose expression have no difference in overall situation. (b) The bicluster which expression have no difference in a line or row.(c) The bicluster having coherent variation which means differences between lines and rows are constant. (d) The bicluster having coherent variation tend to the same variation trend.



**Figure 1.** Four types of bicluster

Differential co-expression is another popular method in microarray technology which can recognize the bicluster which have the differential co-expression. That is to say, the gene groups have strong coherence in one gene group but have no coherence in another gene group. Differential co-expression is helpful to find the genes related to the cancer and bio-process of growing old [3]. Therefore it is necessary to work out the data of differential co-expression.

With the deep research, many differential co-expression bicluster have been put forward. FANG [4] and some scholars put forward the algorithm of SDC which is based on the ideas of Apriori. In this way, she cleverly defines a differentiated support level to meet the anti-monotonicity.

By taking this measure, it greatly improves the productivity of algorithm by reducing the amount of candidate item sets and simultaneously works out its biological mode. However, the SDC has the following defects: (a) In order to revise the modes which are not suitable to the support level, a large amount of data have been saved. It not only increases the complexity of accumulation but also reduce the internal storage's coefficient of utilization. (b) The definition of differential co-expression's support level coming from algorithm will result in some ignorance of differential co-expression's data. ODIBAT [5] and some scholars put forward the "DiBiCLUS" biclustering algorithm which directly adopts the Cluster method by working out the two data to meet differential co-expression. They also mentioned that this kind of algorithm has some dangers of losing some information: one group of genes may have positive expression and negative expression under different experimental conditions and the algorithm can only save the expression which has the most samples. However, "DiBiCLUS" is a kind of dispersion to the original data which may result in the lack of information. WANG [6] and some scholars have put forward the "DRCluster" which defines the samples' support level and they also put forward the new constant differentiated bicluster based on this. This algorithm is developed on the weight chart which is of high efficiency but have defects of soft-real definition for differential co-expression bicluster and inconspicuous effects about the data worked out. XIE [3] has put forward a new definition—MiSupport for the differential co-expression and a developed algorithm—MiCluster based on differential weight chart. From two real microarray dataset, we can work out the data of differential co-expression bicluster. However, because the results worked out are few and fragmented and the rid definition of MiSupport, the recognition rate of differential bicluster is not high. LI [7] and some scholars have put forward some algorithms which adopt the method that is putting the depth at first and avoiding remaining a large amount of candidate nodes in internal storage. In this way, the utilization rate of internal storage has been enhanced. By adopting the "revised" policy, the efficiency has been improved. However, the algorithm has taken no consideration of the filtration on noise data and dispersion on microarray data.

To work out more effective differential co-expression bicluster, this article put forward a new algorithm—DCECluster for a new differential co-expression bicluster. The algorithm adopts a new support level of differential co-expression and turns the dataset into a gene differential co-expression weight chart. In the end, by adopting the effective revising policy to revise the useless nodes in search process, the most differential co-expression bi-cluster can be worked out. The tasks of this article are as following:

(1) To make up for the defects that the ordinary gene data is dispersed in "0,1" or "0,1,-1", they put forward a dispersed method of multi-valued data based on rough set.

(2) They put forward a new support level of differential co-expression.

(3) According to the relationship of gene co-expression, it guarantees the effectiveness of working out the largest bi-cluster by adopting the revising policy and turning the gene sample data into weight chart of differential co-expression.

(4) By using the two dataset, this article proves the superiority of bicluster by two experiments.

## 2. PROBLEM DESCRIPTIONS AND RELATED DEFINITIONS

**Definition 1** Gene expression data sets.  $D = G \times S$ , in it: D represents the set of all genes, S represents the set of all experimental time points. Each element  $D_{ij}$  in the matrix represents the true value of the expression levels of the gene at the time point  $S_j$ . Table 1 is an example of a gene expression matrix.

**Table 1.** A gene expression dataset

	$S_1$	$S_2$	$S_3$
$G_1$	5.78	2.45	-3.91
$G_2$	-4.57	1.41	5.55
$G_3$	6.11	3.65	-2.82

**Definition 2** Attribute discretization method of multi-valued samples based on rough sets. Divide time point attribute set into a number of different Intersect or phase sequences, while obtaining results in a breakpoint and using a new heuristic algorithm to determine the breakpoint subset of the result based on the conditional attribute importance and extreme algorithm. Algorithm is divided into three steps: (a) Determine the set of candidate breakpoints for each successive attribute discretization; (b) determine the breakpoint subset of the result; (c) discrete the properties of the gene expression data sets using the breakpoint subset of the result. Among them, the extreme algorithm is based on the Boolean logic and Rough. It is used to solve a minimum breakpoint sets. Each time the extreme algorithm takes the most important breakpoint. The importance of the breakpoint is measured by the number of 1 in each column. The more the number of 1 is, the higher the importance of the breakpoint is.

**Algorithm 1** result-dots

**Input:** Sample properties  $S = \{s_1, s_2, \dots, s_n\}$ , each set of sample properties  $S_m = \{S_m(g_1), S_m(g_d)\}$ , each breakpoint set of sample properties  $S_{s_i} = \{s_{11}, s_{12}, \dots, s_{1m_i}\}$ , ...,  $S_{s_m} = \{s_{m1}, s_{m2}, \dots, s_{mm_m}\}$ . Gene expresses data collection  $D$ .

**Output:** Breakpoint sets of each sample properties.

for i=1:m

Z=length( $S_{s_i}$ );

For j=1:Z

If

$(\{S_{ij}, S_{i(j+1)}\} \subseteq \bigcap \min(S_i(g_{t_1}), S_i(g_{t_2})), \max(S_i(g_{t_1}), S_i(g_{t_2})))$

$D_j^{S_i}(g_{t_1}, g_{t_2}) = 1$ ;

Else  $D_j^{S_i}(g_{t_1}, g_{t_2}) = 0$ ;

end

end

end

Constructing the new gene expression dataset  $D^*$  according to  $D_j^{S_i}(g_{t_1}, g_{t_2})$  and sorting for the attribute importance of gene expression dataset according to each set  $S_m$  of sample properties. This discretized method is more precise compared with the traditions, because traditional

method only discretizes data into two values 0 and 1, or three values 0, 1 and -1. This new method can better reflect the real data, which can decrease the noise influence and the useful information loss. Table 2 is an example for discretizing using definition 2.

**Table 2.** The gene expression dataset after discretizing

	$S_1$	$S_2$	$S_3$	$S_4$
$G_1$	4	2	-1	-3
$G_2$	-4	1	0	-5
$G_3$	5	3	-2	-2
$G_4$	-1	0	0	-4

(a) Dataset A

	$S_1$	$S_2$	$S_3$	$S_4$
$G_1$	0	0	-3	0
$G_2$	4	0	5	0
$G_3$	3	0	-2	2
$G_4$	-1	0	2	0

(b) Dataset B

### 3. CASE STUDY

**Definition 3** Bicluster. A bicluster consists of a row set and a column set, row set can be seen as all gene set  $S$ , and column set can be seen as all conditions set  $G$ . Bicluster can be defined as  $B = (B_G, B_S)$ , where,  $B_G \subset G, B_S \subset S$ .

**Definition 4** Relationship of gene co-expression. The relationship between gene  $G_1$  and  $G_2$  has three conditions:

(1) If  $|G_1| = |G_2|$ , and  $G_1 * G_2 > 0$ , then the relationship between gene  $G_1$  and  $G_2$  is positive co-expression, namely promote each other.

(2) If  $|G_1| = |G_2|$ , and  $G_1 * G_2 < 0$ , then the relationship between gene  $G_1$  and  $G_2$  is negative co-expression, namely restrain each other.

(3) Neither promoting nor inhibition, the relationship between  $G_1$  and  $G_2$  is non co-expression.

**Definition 5** Close degree of differential co-expression.  $X = \{x_1, x_2, \dots, x_m\}$  and  $Y = \{y_1, y_2, \dots, y_m\}$  are two gene set.

$$CCF(\alpha) = \frac{\sum_{i \in \{A, B\}} w(u_i)w(v_i)(x_i - m_{X,k})(y_i - m_{Y,k})}{\sqrt{\left\{ \sum_{i \in \{A, B\}} w^2(u_i)(x_i - m_{X,k})^2 \right\} \left\{ \sum_{i \in \{A, B\}} w^2(v_i)(y_i - m_{Y,k})^2 \right\}}}$$

$$\text{Where, } w(\alpha) = \begin{cases} (1-\alpha^2)^2 & |\alpha| \leq 1 \\ 0 & \text{otherwise} \end{cases}, \text{ and } m_{X,k} \text{ is the median}$$

of  $X$  in dataset  $k$ .

**Definition 6** The maximum bicluster. Suppose  $B$  is all bicluster set probably in gene expression dataset  $D$ , where,  $Q = W \times P \subseteq E$  is a bicluster. If  $Q$  is the maximum bicluster, then it must satisfy following condition: there does not exist another bicluster  $A = L \times H$ , so  $Q \subseteq A$ .

## 4. DCECLUSTER ALGORITHMS

### 4.1 Constructing weighted graph of gene differential co-expression

The number of large scale gene is far more than that of the experimental conditions (samples) in gene expression dataset.

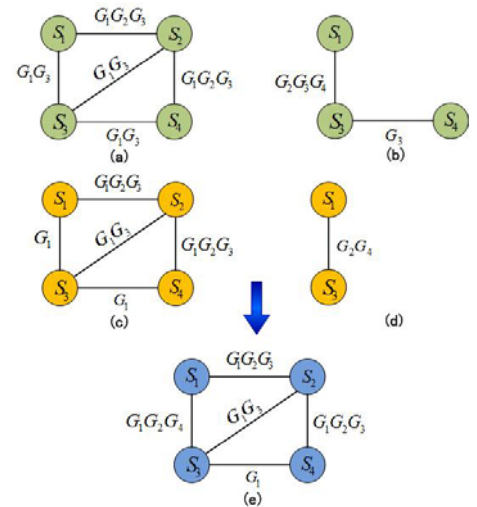
This proposed algorithm adopts the differential weighted graph based on experimental conditions in order to mine the maximum differential bicluster effectively.

**Definition 7** Weighted graph on gene differential samples. Suppose that  $G = (V, E, W)$  is a weighted undirected graph, and it represents a weighted graph. Where, vertex  $v$  ( $A$ ) represent the experimental conditions in dataset A. It represents that there exists a gene set of differential co-expression between two vertexes if they are connected by edge  $e_{ij}$ . The weight  $w_{ij}$  on each edge denotes the gene set of differential co-expression between vertexes  $v_i$  and  $v_j$ . In weight, including not only solving difference from dataset A to B, but also from B to A, and can avoid secondary search, promote efficiency. According to article [8], the definition of differential graph from gene dataset A to B as follows:

$$G_{sub(A-B)} = \{(u_i, v_j, w_{ij}) | [(u_i \in v(A)) \cap (v_j \in v(A)) \cap (u_i \in v(A)) \cap (v_j \in v(B)) \cap (w_{ij} = w_{ij}(A) - w_{ij}(B))] \cup [(u_i \in v(A)) \cap (v_j \in v(A)) \cap (u_i \in v(B)) \cap (v_j \notin v(B)) \cap (w_{ij} = w_{ij}(A))]\}$$

Then, the definition of weighted graph on differential samples between gene dataset A and B as follows:

$$G_{dc(AB)} = G_{sub(A-B)} \cup G_{sub(B-A)}$$



**Figure 2.** The generation process of weighted graph on gene differential samples

According to definition 4 and 7 told above, we construct the weighted graph on gene differential samples for table 2, as Figure 2 shows. In Figure 2, Figure (a) and Figure (b) are respectively the weighted graph of gene dataset A and B, Figure (c) represents the differential graph from dataset A to B, and Figure (d) is from B to A, Figure (e) is the final weighted graph based on gene differential samples.

#### 4.2 Bicluster mining of gene differential co-expression using pruning strategy

This algorithm can acquire the maximum bicluster by generating candidate sample set and pruning strategy. Where, it is acquired through sample extension and level advancement for generating candidate sample set.

**Definition 8** Generating method of the candidate set. Suppose that  $S_i(G)$  is the gene number belongs to dataset sample  $S_i$ , gene differential co-expression bicluster is  $S_i S_{i+1} \dots S_{j-1} S_j$ ,  $lessgene$  is the least threshold of gene number in bicluster,  $r_{less}$  denotes the least row threshold and  $c_{less}$  denotes the least column threshold,  $S$  is the all sample set in dataset, and  $G$  denotes the all gene set in dataset. If  $S_h$  is the candidate sample and  $S_h \in S$ , and then  $|S_i S_{i+1} \dots S_{j-1} S_j(G) \cap S_i S_{i+1} \dots S_{j-1} S_h(G)|$  can be satisfied.

**Theorem 1** Pruning strategy 1. If the gene sequence of sample  $S_i$  is empty in gene sample tree, then, it is unnecessary to generate sub-nodes for  $S_i$ .

**Theorem 2** Pruning strategy 2. Suppose  $i < j$ , and if the gene sequence is exactly the same for sample node A on  $i$  th layer and B on  $j$  th in the procedure of generating sample gene tree, then it is unnecessary to generate sub-nodes for A.

**Inference 1** If the gene sequence of new nodes between sample node A and B through intersection operation is the same with A, then A can be replaced by C.

#### 4.3 The flow of DCECluster algorithm

Gene sample dataset can be discretized using effective method in algorithm, and generate weighted graph based on gene differential samples, remove the irrigated gene samples with mining, at the same time, use the strategy of depth prior search and pruning to mine the maximum differential co-expression bicluster.

Figure2 is the weighted graph based on gene differential samples. The differential genome of dataset A to B can be represented as  $+G_i G_{i+1} \dots G_j$ , and The differential genome of dataset B to A can be represented as  $-G_i G_{i+1} \dots G_j$ . As Figure 2(c) shown, the genome between node 1 and 2 can be represented as  $+G_1 G_2 G_3$ , while in Figure 2(d), the genome between node 1 and 3 can be represented as  $-G_2 G_4$ . After constructing the weighted graph on gene differential samples, using the method of generating candidate sample and pruning strategies, the following is DCECluster algorithm.

#### Algorithm 2 DCECluster algorithm

**Input:** gene expression dataset A and B, threshold is  $\alpha$ , the least gene number in differential bicluster is minige, the least number of samples is minisam, the weighted graph on

gene differential samples is DCSWG and candidate bicluster is P.

**Output:** the satisfied maximum differential bicluster.

```

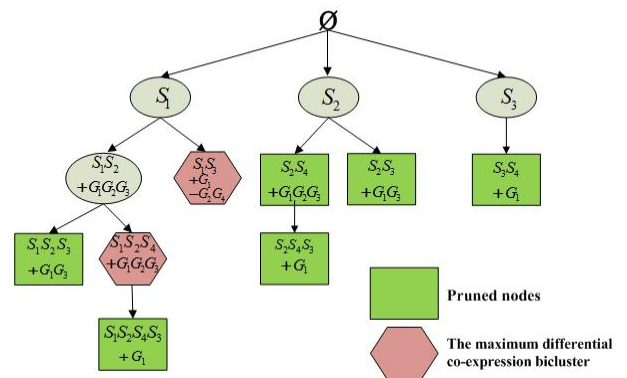
Initialization.P=NULL, DCSWG=NULL;
DCECluster(P);
i) if P=NULL
    then scan dataset A and B, create the weighed graph DCSWG;
    end if
ii) scan the weighed graph DCSWG and acquire the candidate sample set C;
    if C=NULL and both the samples and genes of P can satisfy the minisam and minige
        then output P
    end if
iii) if one node  $C_i$  conform to pruning strategy
        then cut the node
    else if both the samples and genes of P can satisfy the minisam and minige
        then output P
    end if
iv)while each node  $C_i$  belong to sample C
    add each  $C_i$  and take related genes in P with genes in C to conduct intersecting operations
    if both the samples and genes of P can satisfy the minisam and minige
        then add them to P
    else delete the node
    end if
end while

```

#### 4.4 Analysis of time complexity

Suppose that gene number is  $g$  in experiment, sample number is  $t$  and generally  $g \gg t$ . DCECluster algorithm consists of generating weighted graph on gene differential samples and search pruning:

(a) Generating weighted graph. There is one edge between two vertexes for undirected weighted graph and  $C_t^2$  edges for undirected weighted graph having  $t$  vertexes, meanwhile, the weight on each edge is genome sequence.  $g$  genes have to be attended when constructing each edge, so the time complexity of generating weighted graph is  $O(g.C_t^2)$  in the case of the undirected weighted graph possesses the maximum edge number.



**Figure 3.** The procedure of DCECluster algorithm



(b) Search and pruning. In this paper, both depth prior search and pruning method are used, and there are  $2^t$  elements at most on the structure tree of solution space for the collection having  $t$  vertexes. Searching the structure tree of solution space using depth prior method, and using the pruning function to avoid the ineffective search, the time complexity mentioned above is  $O(t \cdot 2^t)$ .

All told, the time complexity of DCECluster algorithm is  $O(t \cdot 2^t + g \cdot C_t^2)$ . The  $2^t$  elements on the structure tree of solution space need to compute each gene if the weighted graph on gene differential samples is not generated, namely, the time of complexity is  $O(g \cdot 2^t)$  to mine the maximum bicluster using non-weighted graph. We can see from above comparison, the superiority of DCECluster algorithm is not obvious even worth than the general algorithm when the number of genes  $g$  is very small or close to the number of samples  $t$ , but it has strong superiority when  $g$  reaches to large scale or mass level.

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

This paper compares the DCECluster algorithm with MiCluster algorithm [3], SDC algorithm [4] and DRCluster [6] algorithm in efficiency and effectiveness.

In order to evaluate the effectiveness of the algorithm, in the first part experiment, we adopt six different dataset from table 3 to mine biclusters using different algorithms, and compare the number of biclusters that they can find respectively. The data in table 3 is from different biological field. In the second part experiment, we adopt the gene expression data AGEMAP [9] to compare the efficiency of different algorithms. AGEMAP records the changes in the level of gene expression of little mice in different age, and concludes 8932 gene expression. At the same time, AGEMAP also concludes 16896 cDNA from sixteen organizations, which has five male and female mice respectively in one month, six month, sixteen month and twenty-four month age. In this paper, we mainly analysis three organizations including hippocampus, heart and sex organ. We can find the genes correlated with age or senescence through the analysis.

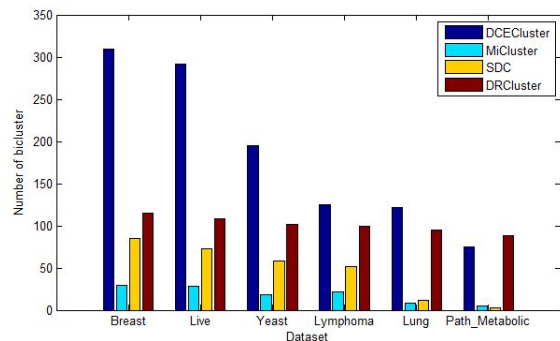
The hardware environment in this experiment as follows: Intel core quad 2.8GHZ, 8G memory. Software environment

as follows: windows7, java programming language and Eclipse.

**Table 3.** Six dataset in the first part experiment

Dataset	Gene number	Sample number
Breast	22283	286
Live	10200	203
Yeast	2993	173
Lymphoma	4026	96
Lung	54837	79
Path_Metabolic	734	69

### 5.1 Effectiveness of algorithm

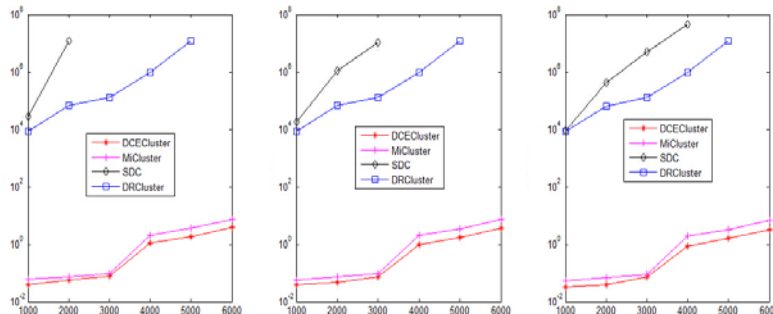


**Figure 4.** Comparison of algorithm effectiveness

Figure 4 shows that the mined number of biclusters for four algorithms in six dataset. It can be seen from Figure 4, the number of biclusters found by MiCluster algorithm is the least, this result mainly is related with the defined differential support. DCECluster algorithm can mine the most clustering number and the most abundant, because it adopts effective support, differential weighted graph and effective pruning strategy.

### 5.2 The comparison of algorithm's running time

To value on the algorithm's effectiveness, we select different scaled data from AGEMAP gene expression dataset(the scale of gene are 1000, 2000, 3000, 4000, 5000, 6000).We test the four types of bicluster under differential support level.



**Figure 5.** Running time of different algorithms

From the Figure 5, the speed of "DCECluster" and "MiCluster" which adopt differential weight chart is faster than the other two algorithms. Because "SDC" takes no effective measures to avoid the candidate nodes or the

expanding of intermediate results, the depletion of internal storage is severe. When the support level is 0.1,the number of genes increase from 1000 to 2000,the running time tends to straight up. But when the number of genes is over 2000,it

is incapable to get the exact running time because of long time enforcement of procedure. When the support level reach to 0.3 and 0.5, the effectiveness of SDC has been greatly improved and get the ideal running time at gene of 3000 and 4000. That is to say, for SDC, the efficiency of algorithm has been improved by increasing the differential support level and shortening the running time. DRCluster also adopts a search policy which is of higher efficiency than SDC. But when the gene is over 5000, the procedure also can't get the exact running time. Although MiCluster can work out a little data of bicluster, it is of high efficiency and close to the algorithm of DCECluster. With the increase of genes, the running time of DCECluster is gently increased and the DCECluster's expanding performance is good which is not only suitable for ordinary scaled data but for working out the large scaled data.

## 6. CONCLUSIONS

The gene expression data gained from the DNA can simultaneously test thousands of gene expression level and work out an effective way to analyze the genes related to diseases. Because the gene expression dataset has the characteristic of small number of samples, large number of genes and many genes related to noise and useless genes, it is necessary to work out effective method to find genes related to disease. The differential co-expression bicluster is helpful to find genes related to cancer and biological process of aging. This paper starts with the data of cluster analysis and gene expression, it vividly analyze the basic concept and main classification of bicluster by giving explanation on the current study situation of differential co-expression bicluster. For the defects of current differential co-expression, this article puts forward a new dispersed method of multi-valued sample based on the rough set and DCECluster which is a method of working out the maximum bicluster based on the gene differential co-expression weight chart and searching revising policy. This algorithm firstly turns the dispersed dataset into gene differential sample chart and effectively reduces the unconcerned genes. It gives a new definition of differential support level and revises the bicluster by effectively using the search policy and revising policy to revise the candidate bicluster. By testing the effectiveness and efficiency of four types of differential co-expression bicluster, it shows the advantages of fast running, saving internal storage, working out large amount of bicluster. The next study on large-scaled gene data will developed in the following aspects:

a) This article mainly deals with the differential co-expression relationship of two genes and study on the differential co-expression relationship among various genes.

b) The dispersed method and revising policy of gene data need to be optimized.

c) We can establish the social network between people and study on the mechanism of cyberspace and social community. In this way, we can analyze the relationship of genes and find out the pathogeny and biological function of disease by working out the differential co-expression data to make up the gene network. In this way, we can infuse new energy for life science.

## ACKNOWLEDGEMENTS

This work was supported by the Science and Technology Research Program of Zhejiang Province under grant No. 2011C21036, and by the Shanghai Natural Science Foundation under grant No.10ZR1400100, and by the National Undergraduate Training Programs for Innovation and Entrepreneurship under grant No. 201410876011, and by the new-shoot Talents Program of Zhejiang Province under grant No.2014R419020.

## REFERENCES

1. Tao Xu, Xuequn Shang, Mijing Yang, et al., Bicluster Algorithm on Discrete Time-Series Gene Expression Data [J], *Application Research of Computers*, 2013, 30(12): 3551-3567 (in Chinese).
2. Madeira S. C., Oliveira A. L., Biclustering Algorithms for Biological Data Analysis: A Survey [J], *IEEE/ACM Trans. on Computational Biology and Biology and Bioinformatics*, 2004, 1(1): 24-45.
3. Huabo Xie, Xuequn Shang, Miao Wang. Differential Co-Expression Relative Constant Row Bicluster Mining Algorithm [J], *Journal of Computer Applications*, 2013, 33(8): 2188-2193 (in Chinese).
4. Fang Gang, Kuang Rui, Pandey G., et al., Subspace Differential Coexpression Analysis: Problem Definition and a General Approach [C], *Proc. of the 15th Pacific Symposium on Biocomputing*, 2003: 145-156.
5. Odibat O., Reddy C. K., Giroux C. N., Differential Biclustering for Gene Expression Analysis [C], *Proceedings of the ACM Conference on Bioinformatics and Computational Biology*, New York: ACM, 2010: 275-284.
6. Wang M., Shang X. Q., Li X. Y., et al., Efficient Mining Differential Co-Expression Constant Row Bicluster in Real-Valued Gene Expression Datasets [J], *Applied Mathematics & Information Sciences*, 2013, 7(2): 587-598.
7. Xiaoyuan Li, Xuequn Shang, Miao Wang, DiCluster Approach: Effective Mining Differential Co-expression Bicluster in Gene Expression Data [J], *Application Research of Computers*, 2012, 29(11): 4087-4092 (in Chinese).
8. Jingni Diao, Xuequn Shang, Miao Wang, Differential Biclustering Algorithm of Microarray Dataset Based on Weighted Graph [J], *Application Research of Computers*, 2011, 28(1): 49-53 (in Chinese).
9. Southworth L. K., Owen A. B., Kim S. K., Aging Mice Show a Decreasing Correlation Of Gene [J], *PLoS Genetics*, 2009, 5(12): 1-13.