

# Interrelated Two-way Clustering: An Unsupervised Approach for Gene Expression Data Analysis

Chun Tang, Li Zhang and Aidong Zhang  
Department of Computer Science and Engineering  
State University of New York at Buffalo  
Buffalo, NY 14260  
{chuntang, lizhang, azhang}@cse.buffalo.edu

Murali Ramanathan  
Department of Pharmaceutical Sciences  
State University of New York at Buffalo  
Buffalo, NY 14260  
murali@acsu.buffalo.edu

## Abstract

*DNA arrays can be used to measure the expression levels of thousands of genes simultaneously. Currently most research focuses on the interpretation of the meaning of the data. However, majority methods are supervised-based, less attention has been paid on unsupervised approaches which is important when domain knowledge is incomplete or hard to obtain. In this paper, we present a new framework for unsupervised analysis of gene expression data which applies an interrelated two-way clustering approach on the gene expression matrices. The goal of clustering is to find important gene patterns and perform cluster discovery on samples. The advantage of this approach is that we can dynamically use the relationships between the groups of the genes and samples while iteratively clustering through both gene-dimension and sample-dimension. We illustrate the method on gene expression data from a study of multiple sclerosis patients. The experiments demonstrate the effectiveness of this approach.*

## 1 Introduction

DNA microarray technology permits rapid, large-scale screening for patterns of gene expression and gives a simultaneous, semi-quantitative readouts on the level of expression of thousands of genes [18, 15, 24, 27, 36, 20, 21, 8, 30]. This new technology also gives rise to a challenge: to interpret the meaning of this immense amount of biological information usually formatted in numerical matrices. To meet the challenge, various methods have been developed using both traditional and innovative techniques to extract, analyze and visualize gene expression data generated from DNA microarrays.

A key step in the analysis of gene expression data is to detect groups that manifest similar expression patterns. Be-

sides, clustering gene expression will reduce complexity, facilitate interpretation, avoid redundancy and curb the noise.

Information in gene expression matrices is special in that it can be studied in two dimensions [2]: analyzing expression profiles of genes by comparing rows in the expression matrix [25, 23, 22, 3, 33, 10, 6, 26] and analyzing expression profiles of samples by comparing columns in the matrix [11, 32, 9]. While most researchers focus on either gene dimension or sample dimension, in a few occasions, sample clustering has been combined with gene clustering. Alon et al. [34] have applied a partitioning-based clustering algorithm to study 6500 genes of 40 tumor and 22 normal colon tissues for clustering both genes and samples. Getz et al. [12] present a method applied on colon cancer and leukemia data set. By identifying subsets of the genes and samples such that when one of these is used to cluster the other, stable and significant partitions emerge. They call it coupled two-way clustering.

Although many different methods for data clustering have been proposed, two major paradigms can be identified: supervised clustering and unsupervised clustering. The supervised approach assumes that for some (or all) profiles there is additional information attached, such as functional classes for the genes, or diseased/normal attributes for the samples. Having this information, a typical task is to build a classifier to predict the labels from the expression profile. Brown et al. [23] have applied various supervised learning algorithms on six functional classes of yeast genes using gene expression matrices from 79 samples. Golub et al. [32] used neighborhood analysis to construct class predictors for samples. They built a weighted vote classifier based on 38 training samples and applied it on a collection of 34 new samples. Hastie et al. [33] proposed a tree harvesting method for supervised learning from gene expression data to discover genes that have strong effects on their own as well as genes that interact with others. Our group has developed a maximum entropy approach to classify gene array data sets [29]. We used part of pre-known classes of sam-

ples as training set and applied the maximum entropy model to generate an optimal pattern model which can be used on new samples.

Unsupervised approaches assume little or no prior knowledge. The goal of such approaches is to partition the set of samples or genes into statistically meaningful classes [1]. A typical example of unsupervised data analysis is to find groups of co-regulated genes or related samples. Currently most of the research focuses on the supervised analysis, relatively less attention has been paid to unsupervised approaches in gene expression data analysis which is important when domain knowledge is incomplete or hard to obtain [31, 37]. The hierarchical [22, 16, 5] and K-means clustering algorithms [14, 28] as well as self-organizing maps [26] are major unsupervised clustering methods which have applied to various gene array datasets.

In this paper, we present an interrelated two-way clustering approach for unsupervised analysis of gene expression data. Unlike previous work mentioned above, in which genes and samples were clustered either independently or both dimensions being reduced, our approach is to dynamically use the relationships between the groups of the genes and samples while iteratively clustering through both gene-dimension and sample-dimension to extract important genes and classify samples simultaneously.

We have applied the method to a data set on multiple sclerosis patients collected by the Neurology and Pharmaceutical Sciences departments in our university (Multiple sclerosis (MS) is a chronic, relapsing, inflammatory disease and interferon- $\beta$  (IFN- $\beta$ ) has been the most important treatment for the MS disease for the last decade [35]). In particular, we perform class discovery on the healthy control, MS and IFN-treated samples based on the data collected from the DNA microarray experiments. The gene expression levels are measured by the intensity levels of the corresponding array spots. The experiments demonstrate the effectiveness of this approach.

This paper is organized as follows. Section 2 introduces our approach. Section 3 presents the experimental results on multiple sclerosis data set. And finally, the conclusion is provided in Section 4.

## 2 Interrelated Two-way Clustering

### 2.1 Motivation

Gene expression data are matrices where rows represent genes, columns represent samples such as tissues or experimental conditions, and numbers in each cell characterizes the expression level of a particular gene in a particular sample. Let  $G = \{g_1, \dots, g_i, \dots, g_n\}$  be the set of all genes,  $S = \{s_1, \dots, s_j, \dots, s_m\}$  be the set of all samples, and  $w_{i,j}$  be the intensity value associated with each gene  $g_i$

and sample  $s_j$  in the matrix. Thus the gene expression matrix  $W = \{w_{i,j} | 1 \leq i \leq n, 1 \leq j \leq m\}$  has  $n$  rows (gene vectors) and  $m$  columns (sample vectors).

Based on the gene expression matrix  $W$  which usually has thousands of rows and less than a hundred of columns ( $n \gg m$ ), a common problem is: can we effectively cluster the samples with similar properties using the genes automatically?

Note that for the unsupervised analysis, the previous knowledge and the training data are not available. Also, because  $n \gg m$ , the dimension of sample vectors is much higher than the number of samples, it is very hard to get good result by directly using traditional clustering algorithms [14, 13] for classifying samples on such a high dimensional space.

To achieve better class discovery on samples, we should first try to lower the vector space into a relatively small one, which means reduce the number of genes (equals to the dimension of sample vectors) to a smaller dimension and then perform clustering. If the dimension of sample vectors is still too large, we continue to reduce until it reaches a reasonable level on which clustering algorithms can work effectively and efficiently. However, the dimension reduction is non-trivial.

In recognizing the above problems, we propose a general framework for the unsupervised gene expression data analysis. In this framework, an interrelated two-way clustering approach as well as a pre-processing procedure is applied on the gene expression matrix  $W$ , and the goal of clustering is to find important gene patterns and to perform class discovery on samples simultaneously. To be more specific, we have two goals:

- (1) Find a subset of genes, usually called important genes, which are highly related to the experiment conditions. This can also be considered as the gene dimension reduction.

- (2) Cluster the samples into different groups. According to the most popular experimental platforms, the number of different groups is usually two, for example, diseased samples and control samples.

These two goals are actually two sides of one coin. If we can find important genes, then it is relatively easy to use traditional clustering algorithms to cluster samples because the sample vectors' dimension is reduced to a reasonable level (usually around 100). On the other hand, if we can correctly cluster the samples, important genes can be found by sorting all genes using similarity scores such as correlation coefficient [32, 4] with patterns according to the cluster results.

One of the advantages of our approach is that we can dynamically use the relationships between the groups of the genes and samples while iteratively clustering through both gene-dimension and sample-dimension. In doing iterative

clustering, reducing gene-dimension will improve the accuracy of class discovery, which in turn will guide further gene-dimension reduction.

## 2.2 Pre-processing of Data

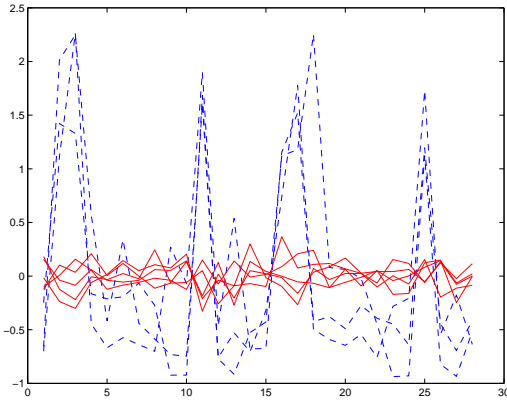
In the gene expression matrix, different genes have different ranges of intensity values. The intensity values alone may not have significant meaning, but the relative values are more intrinsic. So we first normalize the original gene intensity values into relative values [19, 38].

Our general formula is

$$w'_{i,j} = \frac{w_{i,j} - \mu_i}{\mu_i}, \quad \text{where } \mu_i = \frac{\sum_{j=1}^m w_{i,j}}{m}; \quad (1)$$

$w'_{i,j}$  denotes normalized intensity value for gene  $i$  of sample  $j$ ,  $w_{i,j}$  represents the original intensity value for gene  $i$  of sample  $j$ ,  $m$  is the number of samples, and  $\mu_i$  is the mean of the intensity values for gene  $i$  over all samples.

Notice that among thousands of genes, not all of them have the same contribution in distinguishing the classes. Actually, some genes have little contribution. We need to remove those genes which have little reaction to the experiment condition. We believe genes whose intensity values keep invariant or change very little belong to this class (Figure 1 shows an example of gene distributions).



**Figure 1. Genes intensity value distributions after normalization. Horizontal axis represents samples. Each polygonal line indicates a gene changing level varies among samples. The red-solid lines represent gene intensity values which vary little through all samples, and the blue-dash lines represent gene intensity values which vary much among samples.**

Let's assume we have  $n$  genes and  $m$  samples. We de-

note each gene vector (after normalization) as

$$g_i = (w'_{i,1}, w'_{i,2}, \dots, w'_{i,m}), \quad (2)$$

where  $i = 1, 2, \dots, n$  for each gene. We use *vector-cosine* between each gene vector and a pre-defined stable pattern  $E$  to test whether a gene intensity value varies much among samples. The pattern can be denoted as  $E = (e_1, e_2, \dots, e_m)$ , where all  $e_i$  are equal.

$$\cos(\theta) = \frac{\langle \vec{g}_i, \vec{E} \rangle}{\|\vec{g}_i\| \cdot \|\vec{E}\|} = \frac{\sum_{j=1}^m w'_{i,j} \times e_j}{\sqrt{\sum_{j=1}^m w'^2_{i,j}} \times \sqrt{\sum_{j=1}^m e^2_j}}, \quad (3)$$

where  $\theta$  is the *angle* between two vectors  $\vec{g}_i$  and  $\vec{E}$  in  $m$ -dimensional space. If the two vector patterns are more similar, the vector-cosine will be closer to 1. The extreme case is that when two vectors are parallel, the vector-cosine value is 1. On the other hand, vector-cosine value of two perpendicular vectors is 0. After calculating vector-cosine values, we can choose a threshold to remove genes matching pattern  $E$  (those genes' vector-cosine values with  $E$  are higher than the threshold, which means these genes change little during the experiment). Usually we can remove twenty to thirty percent of genes by this step, thus facilitating clustering in the next stage.

## 2.3 Interrelated Two-way Clustering

To perform two-way clustering, a distance measure to be used during the clustering procedure should be carefully chosen. One commonly used distance is the *Euclidean* distance. But for gene data, patterns' similarity seems more important than their spatial distance [32, 4]. So we choose *correlation coefficient* [17] which measures the strength of the linear relationship between two vectors. This measure has the advantage of calculating similarity depending only on the pattern but not on the absolute magnitude of the spatial vector. The formula of correlation coefficient between two vectors  $X = (x_1, x_2, \dots, x_k)$  and  $Y = (y_1, y_2, \dots, y_k)$  is:

$$\rho_{x,y} = \frac{k(\sum_{i=1}^k x_i \times y_i) - (\sum_{i=1}^k x_i) \times (\sum_{i=1}^k y_i)}{\sqrt{\left[ k \sum_{i=1}^k x_i^2 - (\sum_{i=1}^k x_i)^2 \right] \left[ k \sum_{i=1}^k y_i^2 - (\sum_{i=1}^k y_i)^2 \right]}} \quad (4)$$

where  $k$  is the length of vector  $X$  and  $Y$ .

Then we cluster genes as well as samples. Dynamic relationship between gene clustering and sample clustering is used to reduce the vector space of samples into a reasonable level and perform class discovery. Our approach, illustrated in Figure 3, is an iterative procedure based on  $G$  with  $n_1$

genes after pre-processing. Within each iteration there are five main steps:

*Step 1: clustering in the gene dimension.* The task of this step is to cluster  $n_1$  genes into  $k$  groups, denoted as  $G_i$  ( $1 \leq i \leq k$ ), each of which is an exclusive subset of  $G$ . The clustering method can be any method for which we can give the cluster number, such as K-means or SOM [13, 14].

*Step 2: clustering in the sample dimension.* Based on each group  $G_i$  ( $1 \leq i \leq k$ ), we independently cluster samples into two clusters (according to the most popular experimental conditions [2]), represented by  $S_{i,a}$  and  $S_{i,b}$ .

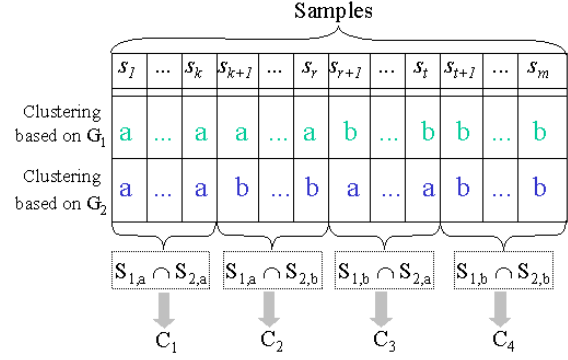
*Step 3: clustering results combination.* This step combines the clustering results of the *step 1* and *step 2*. Without loss of the generality, let  $k = 2$ . Then the samples can be divided into four groups:

- $C_1$  (all samples clustered into  $S_{1,a}$  based on  $G_1$  and clustered into  $S_{2,a}$  based on  $G_2$ );
- $C_2$  (all samples clustered into  $S_{1,a}$  based on  $G_1$  and clustered into  $S_{2,b}$  based on  $G_2$ );
- $C_3$  (all samples clustered into  $S_{1,b}$  based on  $G_1$  and clustered into  $S_{2,a}$  based on  $G_2$ );
- $C_4$  (all samples clustered into  $S_{1,b}$  based on  $G_1$  and clustered into  $S_{2,b}$  based on  $G_2$ ).

Figure 2 illustrates the results of this combination. If  $k = 3$ , there will be 8 possible sample groups. In general, the number of possible sample groups equals  $2^k$ . Usually  $k$  is set to be 2 to reduce the computational complexity.

*Step 4: finding heterogeneous groups.* Among the sample groups  $C_1, C_2, C_3, C_4$ , we choose two distinct groups  $C_s$  and  $C_t$  ( $1 \leq s, t \leq 4$ ) which satisfy the following condition: for  $\forall u \in C_s, \forall v \in C_t$ , where  $u$  and  $v$  are samples, if  $u \in S_{i,r_1}, v \in S_{i,r_2}$ , then  $r_1 \neq r_2$  ( $r_1, r_2 \in \{a, b\}$ ) for all  $i$  ( $1 \leq i \leq k$ ). We call  $(C_s, C_t)$  *heterogeneous group*. For example,  $(C_1, C_4)$  is such a heterogeneous group (when  $k = 2$ ) because all samples in group  $C_1$  are clustered into  $S_{i,a}$  ( $1 \leq i \leq k$ ), while all samples in group  $C_4$  are clustered into  $S_{i,b}$  ( $1 \leq i \leq k$ ). For the same reason,  $(C_2, C_3)$  is another heterogeneous group.

*Step 5: sorting and reducing.* For each heterogeneous group, for example,  $(C_1, C_4)$ , two patterns  $(0, 0, \dots, 0, 1, 1, \dots, 1)$  and  $(1, 1, \dots, 1, 0, 0, \dots, 0)$  are introduced. The pattern  $(0, 0, \dots, 0, 1, 1, \dots, 1)$  includes  $|C_1|$  (number of samples in group  $C_1$ ) zeros followed by  $|C_4|$  (number of samples in group  $C_4$ ) one's. Similarly,  $(1, 1, \dots, 1, 0, 0, \dots, 0)$  includes  $|C_1|$  one's followed by  $|C_4|$  zeros. For each pattern, we use it to calculate vector-cosine defined in Equation (3) with each gene vector, then sort all genes according to the similarity values in descending order, and keep the first one third of the sorted gene sequence by cutting off the other two thirds of the gene sequence. By merging the



**Figure 2. Clustering results combination when  $k = 2$ .**  $s_1, s_2, \dots, s_m$  in the first line represent samples. The second and third lines show cluster results on samples based on gene groups  $G_1$  or  $G_2$  independently. In each case, samples are clustered into two groups, which are marked as “a” or “b”. We use green color (second line) to represent cluster results based on  $G_1$  and blue color (third line) for results based on  $G_2$ . By combination, four possible sample groups are generated:  $C_1$  includes samples marked as “a” based on  $G_1$  and marked as “a” based on  $G_2$ ;  $C_2$  includes samples marked as “a” based on  $G_1$  and marked as “b” based on  $G_2$ ;  $C_3$  includes samples marked as “b” based on  $G_1$  and marked as “a” based on  $G_2$ ;  $C_4$  includes samples marked as “b” based on  $G_1$  and marked as “b” based on  $G_2$ .

remaining sorted gene sequences from two patterns, we obtain the reduced gene sequence  $G'$  where at least one third of the genes in  $G$  are cut off.

Similarly, for the other heterogeneous group  $(C_2, C_3)$ , another reduced gene sequence  $G''$  is generated. Now the problem is which gene subset should be chosen for the next iteration,  $G'$  or  $G''$ ? The semantic meaning behind it is to select a heterogeneous group which is a better representation for the original distribution of samples because  $G'$  and  $G''$  are generated based on the corresponding heterogeneous groups. Here we use the cross-validation method [32] to evaluate each group. In each heterogeneous group, first choose one sample, then use the remaining samples of this group to select important genes, and predict the class of the withheld samples. The process is repeated for each sample, and the cumulative error rate is calculated. When the heterogeneous group which has lower error rate is found, its corresponding reduced gene sequence is selected as  $\hat{G}$  with  $n_2$  genes for the next iteration.

After *Step 5*, the gene number is reduced from  $n_1$  to  $n_2$ .

The above steps can be repeated by clustering  $n_2$  genes, and so on. The iteration will be terminated until the termination conditions are satisfied.

## 2.4 Termination Condition

To explain the termination condition, we first define the *occupancy ratio* between samples in heterogeneous groups and all samples, let  $\chi$  denote all heterogeneous groups:

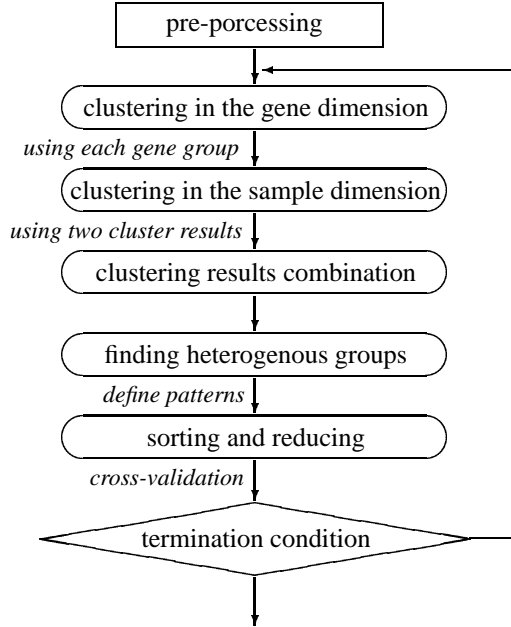
$$Occratio = Max(\{ \frac{|C_i| + |C_j|}{m} \}), \quad (5)$$

where  $(C_i, C_j) \in \chi$  ( $1 \leq i, j \leq 2^k$ ),  $m$  is the total number of the samples,  $|C_i|$  is the number of samples in  $C_i$ .

Thus if  $k = 2$ , the occupancy ratio will be:

$$Occratio = Max(\frac{|C_1| + |C_4|}{m}, \frac{|C_2| + |C_3|}{m}). \quad (6)$$

Because the sum of the number of samples in all heterogeneous groups is equal to  $m$ , the minimum value of *Occratio* is 0.5. If the gene clustering results based on  $G_1$  and  $G_2$  are the same, then either  $C_1 \cup C_4 = S$  ( $S$  is the set of all samples) or  $C_2 \cup C_3 = S$ , in this case  $\hat{G}$  (the remaining genes) is good enough for sample clustering. Note that under such optimal condition, the *Occratio* value will reach the maximum value 1.



**Figure 3.** The structure of Interrelated Two-way Clustering.

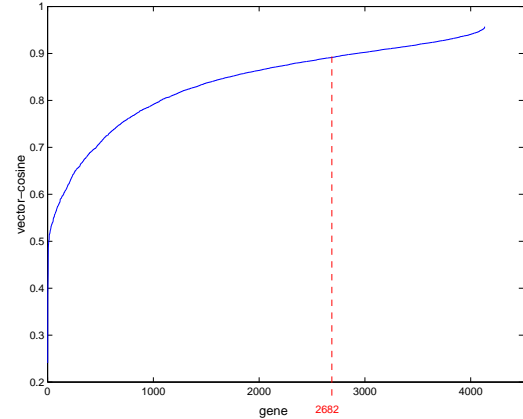
*Occratio* value can be used as one of the termination conditions for the iteration. If the *Occratio* value reaches 1,

we can stop the iteration. However, since the optimal condition is hard to reach, usually the iteration can be stopped when the *Occratio* value reaches a threshold such as 0.9, meaning samples cluster result at step 2 based on  $G_1$  and  $G_2$  are quite similar. Sometimes after many iterations, the *Occratio* value still cannot reach the threshold, but the remaining gene number ( $n_2$ ) is very small (for example, 100). This also can be used as termination condition.

The whole procedure of interrelated two-way clustering is presented in Figure 3.

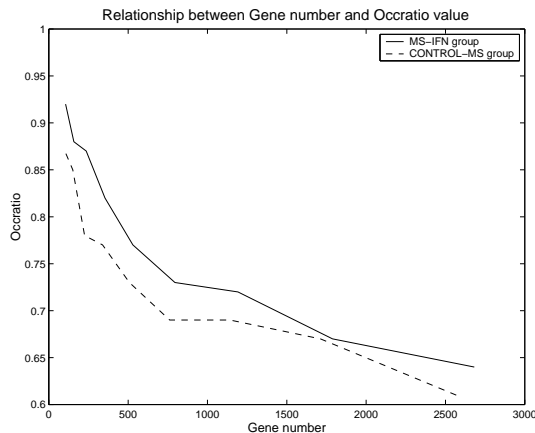
## 3 Experimental Results

The experiments are based on two data sets on multiple sclerosis patients: the MS\_IFN group and the CONTROL\_MS group. The MS\_IFN group contains 28 samples while the CONTROL\_MS group contains 30 samples. Each sample consists 4132 genes. We perform the interrelated two-way clustering approach for unsupervised classification separately on each group. To test the performance of our approach, we choose these two datasets in which the ground-truth is already known, that is, in the MS\_IFN group, there are 14 MS samples and 14 IFN samples, and in CONTROL\_MS group, there are 15 control samples and 15 MS samples. We only use this ground-truth to evaluate our experimental results.



**Figure 4.** Distribution of genes' vector-cosine calculated from Equation (3). Horizontal axis represents samples, vertical axis means vector-cosine value. Samples are sorted in an ascending order, where we choose threshold 0.89 to reduce 4132 genes to 2682.

During the data pre-processing procedure, by sorting genes using vector-cosine calculated from Equation (3), we choose threshold 0.89 (See Figure 4), then remove genes for which vector-cosine with pattern  $E$  is higher than that



**Figure 5. Relationship between Gene number and Occratio value.**

threshold, which means the gene intensity values vary little among the samples. 1450 genes are removed from 4132. As the result, 2682 genes are left.

K-means clustering method is used during the interrelated two-way process, and correlation coefficient (Equation (4)) is used as the distance measure.

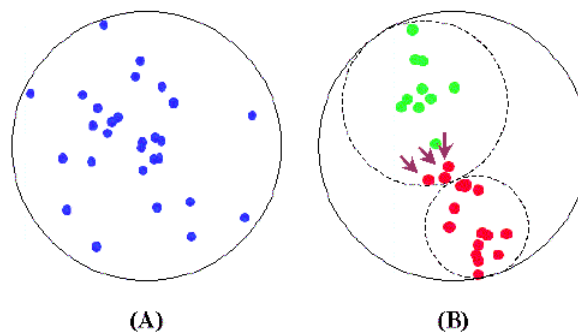
During the iterative process, after each iteration the remaining genes are traced together with the occupancy ratio for the heterogeneous groups of the two dataset (See Figure 5). One observation is that in our experiment, while gene-dimension is reduced, the *Occratio* value increases, the samples cluster results based on  $G_1$  and  $G_2$  become more similar.

On the MS\_IFN group, after nine iterations, the *Occratio* value reaches 0.92. We reduce 2682 genes to 100 genes and cluster samples into two group: 11 samples in group one, which are all correctly classified to samples having MS disease. Another 17 samples are in group two, in which 14 of them is IFN treated, but another 3 were in the wrong group.

In Figure 6, we use a liner mapping function [7] which maps the  $n$ -dimension vectors into two dimensions to show the samples' distribution before and after our approach for the MS\_IFN group.

Similarly, for the CONTROL\_MS group, we reduce 1474 genes with the same threshold as the MS\_IFN group in the pre-processing step, and use the remaining 2658 genes to perform the interrelated two-way cluster. The result is 8 samples being incorrectly classified out of 30 samples.

For the purpose of comparison, we also directly perform K-means clustering method and self-organizing maps on both the MS\_IFN and CONTROL\_MS group data after normalization but without any gene-dimension deduction. Figure 7 lists the sample clustering accuracy rate achieved



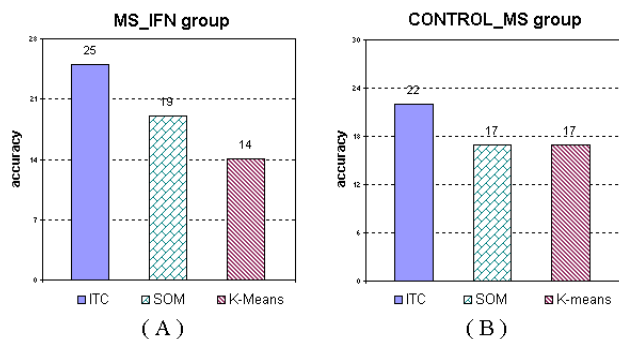
**Figure 6. Approach applying on the MS\_IFN group. (A) Shows the original 28 samples' distribution, Each point represents a sample, which is a mapping from the sample's 4132 genes intensity vectors. There is no obvious cluster border as we see. (B) Shows the same 28 samples distribution after using our approach. We reduce 4132 genes to 100 genes. So each sample is a 100-dimension vector. The green and red colors show the cluster result using our approach, while two dash circles indicate the real sample cluster and three arrows point out the incorrectly classified samples.**

by these methods. From this figure, we can see that using our approach, the accuracy of class discovery is higher than those of traditional methods, which illustrates the effectiveness of the interrelated two-way clustering method on such high dimensional gene data.

## 4 Conclusion

In this paper, we have presented a new framework for the unsupervised analysis of gene expression data. In this framework, an interrelated two-way clustering method is developed and applied on the gene expression matrices transformed from the raw microarray data. We were able to find important gene patterns and to perform class discovery on samples simultaneously. It has the advantage of dynamically using the relationships between the groups of the genes and samples while iteratively clustering through both gene-dimension and sample-dimension. In doing iterative clustering, reducing gene-dimension will benefit the accuracy improvement of class discovery, which in turn will guide further gene-dimension reduction.

In particular, we used the above approach to distinguish the healthy control, MS and IFN-treated samples based on the data collected from DNA microarray experiments. From



**Figure 7. Comparison of accuracy rate achieved by interrelated two-way clustering (ITC), self-organizing maps (SOM) and K-means clustering methods. (A) Shows clustering results on the MS\_IFN group which include 28 samples. (B) Shows clustering results on the CONTROL\_MS group which include 30 samples.**

our experiments, we demonstrated that this approach is a promising approach to be used for unsupervised analysis of gene array data sets.

## References

- [1] A. Ben-Dor, N. Friedman, and Z. Yakhini. Class discovery in gene expression data. In *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB 2001)*, 2001.
- [2] Alvis Brazma and Jaak Vilo. Minireview: Gene expression data analysis. *Federation of European Biochemical societies*, 480:17–24, June 2000.
- [3] Amir Ben-Dor, Ron Shamir and Zohar Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
- [4] Anna Jorgensen. Clustering excipient near infrared spectra using different chemometric methods. Technical report, Dept. of Pharmacy, University of Helsinki, 2000.
- [5] Ash A. Alizadeh, Michael B. Eisen, R. Eric Davis, Chi Ma, Izidore S. Lossos, Adreas RosenWald, Jennifer C. Boldrick, Hajeer Sabet, Truc Tran, Xin Yu, John I. Powell, Liming Yang, Gerald E. Marti et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, Vol.403:503–511, February 2000.
- [6] Charles M. Perou, Stefanie S. Jeffrey, Matt Van De Rijn, Christia A. Rees, Michael B. Eisen, Douglas T. Ross, Alexander Pergamenschikov, Cheryl F. Williams, Shirley X. Zhu, Jeffrey C. F. Lee, Deval Lashkari, Dari Shalon, Pat rick O. Brown, and David Bostein. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA*, Vol. 96(16):9212–9217, August 1999.
- [7] D. Bhadra and A. Garg. An interactive visual framework for detecting clusters of a multidimensional dataset. Technical Report 2001-03, Dept. of Computer Science and Engineering, University at Buffalo, NY., 2001.
- [8] D. Shalon, S.J. Smith, P.O. Brown. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6:639–645, 1996.
- [9] Donna K. Slonim, Pablo Tamayo, Jill P. Mesirov, Todd R. Golub, and Eric S. Lander. Class Prediction and Discovery Using Gene Expression Data. In *RECOMB 2000: Proceedings of the Fifth Annual International Conference on Computational Biology*. ACM Press, 2000.
- [10] Elisabetta Manduchi, Gregory R. Grant, Steven E. McKenzie, G. Christian Overton, Saul Surrey and Christian J. Stoeckert Jr. Generation of patterns form gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, Vol. 16(8):685–698, 2000.
- [11] Francisco Azuaje Department. Making genome expression data meaningful: Prediction and discovery of classes of cancer through a connectionist learning approach, 2000.
- [12] Gad Getz, Erel Levine and Eytan Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, Vol. 97(22):12079–12084, October 2000.
- [13] J. Hartigan and M. Wong. Algorithm AS136: a k-means clustering algorithms. *Applied Statistics*, 28:100–108, 1979.
- [14] Hartigan J.A. *Clustering Algorithm*. John Wiley and Sons, New York., 1975.
- [15] J. DeRisi, L. Penland, P.O. Brown, M.L. Bittner, P.S. Meltzer, M. Ray, Y. Chen, Y.A. Su, J.M. Trent. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.
- [16] Javier Herrero, Alfonso Valencia, and Joaquin Dopazo. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.
- [17] Jay L. Devore. *Probability and Statistics for Engineering and Sciences*. Brook/Cole Publishing Company, 1991.
- [18] J.J. Chen, R. Wu, P.C. Yang, J.Y. Huang, Y.P. Sher, M.H. Han, W.C. Kao, P.J. Lee, T.F. Chiu, F. Chang, Y.W. Chu, C.W. Wu, K. Peck. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics*, 51:313–324, 1998.
- [19] Johannes Schuchhardt, Dieter Beule, Arif Malik, Eryc Wol-ski, Holger Eickhoff, Hans Lehrach and Hanspeter Herzl. Normalization strategies for cDNA microarrays. *Nucleic Acids Research*, Vol. 28(10), 2000.
- [20] M. Schena, D. Shalon, R.W. Davis, P.O. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.
- [21] Mark Schena, Dari Shalon, Renu Heller, Andrew Chai, Patrick O. Brown, and Ronald W. Davis. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA*, Vol. 93(20):10614–10619, October 1996.
- [22] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95:14863–14868, 1998.

- [23] Michael P. S. Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Sugnet, Terrence S. Furey, Manuel Ares and Jr.David Haussler. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sci.*, 97(1):262–267, January 2000.
- [24] O. Ermolaeva, M. Rastogi, K.D. Pruitt, G.D. Schuler, M.L. Bittner, Y. Chen, R. Simon, P. Meltzer, J.M. Trent, M.S. Boguski. Data management and analysis for gene expression arrays. *Nature Genetics*, 20:19–23, 1998.
- [25] Orly Alter, Patrick O. Brown and David Bostein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, Vol. 97(18):10101–10106, August 2000.
- [26] Pablo Tamayo, Donna Solni, Jill Mesirov, Qing Zhu, Sutsak Kitareewan, Ethan Dmitrovsky, Eric S. Lander and Todd R. Golub. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, Vol. 96(6):2907–2912, March 1999.
- [27] R.A. Heller, M. Schena, A. Chai, D. Shalon, T. Bedilion, J. Gilmore, D.E. Woolley, R.W. Davis. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA*, 94:2150–2155, 1997.
- [28] S. Tavazoie, D. Hughes, M.J. Campbell, R.J. Cho and G.M. Church. Systematic determination of genetic network architecture. *Nature Genet*, pages 281–285, 1999.
- [29] Shumei Jiang, Chun Tang, Li Zhang and Aidong Zhang, Murali Ramanathan. A maximum entropy approach to classifying gene array data sets. In *Proc. of Workshop on Data mining for genomics, First SIAM International Conference on Data Mining*, 2001.
- [30] S.M. Welford, J. Gregg, E. Chen, D. Garrison, P.H. Sorensen, C.T. Denny, S.F. Nelson. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Research*, 26:3059–3065, 1998.
- [31] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D., Futcher B. . Exploring the metabolic and genetic control of gene expression on a genomic scale. *Mol. Biol. Cell*, page 3273, 1998.
- [32] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, D.D. Bloomfield and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, Vol. 286(15):531–537, October 1999.
- [33] Trevor Hastie, Robert Tibshirani, David Boststein and Patrick Brown. Supervised harvesting of expression trees. *Genome Biology*, Vol. 2(1):0003.1–0003.12, January 2001.
- [34] U. Alon, N. Barkai, D.A. Notterman, K.Gish, S. Ybarra, D. Mack and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide array. *Proc. Natl. Acad. Sci. USA*, Vol. 96(12):6745–6750, June 1999.
- [35] V. Yong, S. Chabot, Q. Stuve and G. Williams. Interferon beta in the treatment of multiple sclerosis: mechanisms of action. *Neurology*, 51:682–689, 1998.
- [36] V.R. Iyer, M.B. Eisen, D.T. Ross, G. Schuler, T. Moore, J.C.F. Lee, J.M. Trent, L.M. Staudt, Jr. J. Hudson, M.S. Boguski, D. Lashkari, D. Shalon, D. Botstein, P.O. Brown. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.
- [37] Y Barash and N Friedman. Context-specific bayesian clustering for gene expression data. *Bioinformatics, RECOM01*, 2001.
- [38] Yang Y.H., Dudoit S., Luu P. and Speed T. P. Normalization for cDNA Microarray Data. In *Proceedings of SPIE BiOS 2001*, San Jose, California, January 2001.