

Interrelated Two-way Clustering and Its Application on Gene Expression Data

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

Abstract

Microarray technologies are capable of simultaneously measuring the signals for thousands of messenger RNAs and large numbers of proteins from single samples. Arrays are now widely used in basic biomedical research for mRNA expression profiling and are increasingly being used to explore patterns of gene expression in clinical research. Most research has focused on the interpretation of the meaning of the microarray data which are transformed into gene expression matrices where usually the rows represent genes, the columns represent various samples. Clustering samples can be done by analyzing and eliminating of irrelevant genes. However, majority methods are supervised (or assisted by domain knowledge), less attention has been paid on unsupervised approaches which are important when little domain knowledge is available. 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 identify important genes and perform cluster discovery on samples. The advantage of this approach is that we can dynamically manipulate the relationship between the gene clusters and sample groups while conducting an iterative clustering through both of them. The performance of the proposed method with various gene expression data sets is also illustrated.

1 Introduction

Bioinformatics is defined as conceptualizing biology in terms of molecules and applying informatics techniques to understand and organize the information associated with these molecules [34]. Knowledge of the spectrum of genes expressed at a given time or under certain conditions proves instrumental to understand the working of a living cell [55].

Microarray technologies are capable of simultaneously measuring the signals for thousands of messenger RNAs and large numbers of proteins from single samples. Figure 1 illustrates a typical microarray experiment whose outcome are scanned images. Arrays are now widely used in basic biomedical research for mRNA expression profiling and are increasingly being used to explore patterns of gene expression in clinical research [10, 28, 44, 45, 46, 57]. The customary approach in array analysis is to obtain data from fluorescence scanners or phosphorimagers and to analyze the array images using dedicated, custom image analysis software, usually provided by the array manufacturer. Minimally, these software identify spots and analyze spot intensities, map spots to genes, and condition the data. The normalized results are exported as

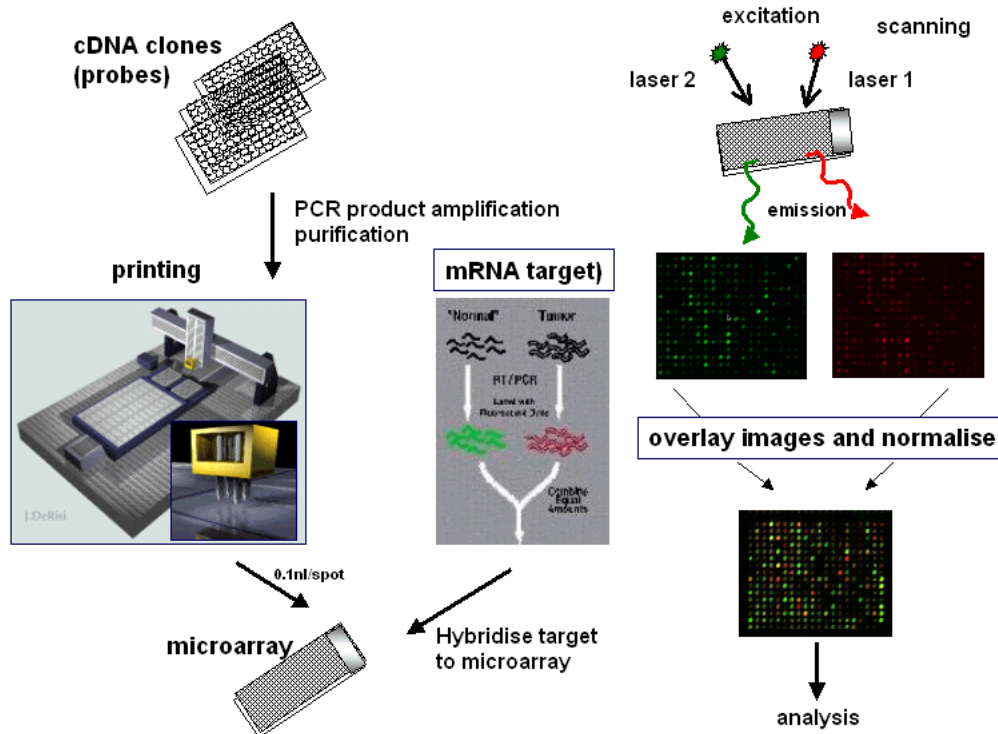


Figure 1: cDNA microarray experiment, courtesy of IPAM: Institute for Pure and Applied Mathematics, UCLA [48]. The raw microarray data are images, each may contain over 5000 genes. The images can then be transformed into numeric gene expression matrices for further analysis.

flat tables to other software where a typical preliminary analysis may involve exploratory cluster analysis, biostatistical analysis and bioinformatics research for interesting genes [12, 13, 16, 25].

The raw microarray data are images which can then be transformed into gene expression matrices where usually the rows represent genes, the columns represent various samples. The numeric value in each cell characterizes the expression level of the particular gene in a particular sample. Innovative techniques to efficiently and effectively analyze these fast growing gene data are required, which will have a significant impact on the field of bioinformatics. But the high-dimensionality and size of array-derived data poses challenging problems in both computational and biomedical research, and the difficult task ahead is converting genomic data into knowledge. Various methods have been developed using both traditional and innovative techniques to extract, analyze, and visualize gene expression data generated from DNA microarrays.

The many data-clustering methods which have been proposed fall into two major categories: supervised clustering and unsupervised clustering. The supervised approach assumes that additional information is attached to some (or all) data, for example, samples are labeled as diseased vs. normal. Using this information, a classifier can be constructed to predict the labels from the expression profile. The major supervised clustering methods include neighborhood analysis [19], the support vector machine [11, 17, 41], the tree harvesting method [24], the decision tree method [63], statistical approaches such as the maximum-entropy model [30], and a variety of ranking-based methods [6, 33, 40, 53, 54, 38]. Unsupervised approaches assume little or no prior knowledge. The goal of such approaches is to partition the dataset into statistically meaningful classes [6]. A typical example of unsupervised data analysis is to find groups of co-regulated genes or related sam-

ples. 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 in a context where little domain knowledge is available [5, 49]. The hierarchical clustering method [1, 15, 26, 9, 65, 29, 56, 58, 36, 20], the k-means clustering algorithms [21, 22, 52, 61] and the self-organizing maps [19, 32, 50, 27, 37] are the major unsupervised clustering methods which have been commonly applied to various data sets.

Information in gene expression data can then be studied in two angles [10]: analyzing expression profiles of genes by comparing rows in the expression matrix [3, 11, 15, 7, 24, 35, 42, 50] and analyzing expression profiles of samples by comparing columns in the matrix [4, 19, 47]. While most researchers focus on either genes or samples, in a few occasions, sample clustering has been combined with gene clustering. Alon et al. [2] proposed a partitioning-based algorithm in which genes and samples were clustered independently. Getz et al. [18] proposed a coupled two-way clustering method to identify subsets of both genes and samples. Xing et al. [59] proposed a clustering method called CLIFF which iteratively uses sample partitions as a reference to filter genes. None of these approaches offers a definitive solution to the fundamental challenge of detecting meaningful patterns in the samples while pruning out irrelevant genes in a context where little domain knowledge is available.

In this paper, we will introduce 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 data being reduced, our approach is to delineate the relationships between gene clusters and sample partitions while conducting an iterative search for sample patterns and detecting significant genes of empirical interest. This iterative framework incorporates a variety of improved techniques extended from the previous work [51]. The performance of the proposed method will be illustrated in the context of various data sets.

The remainder of this paper is organized as follows. Section 2 introduces the motivation and framework while the algorithm to implement the framework is presented in Section 3. Experimental results appear in Section 4 and concluding remarks in Section 5.

2 Motivation

Let $\widehat{\mathbf{M}} = \{m_{i,j} | i \in I, j \in J\}$ represent the original gene expression matrix, where $I = \{1, 2, \dots, n_g\}$ represents the labels of the genes and $J = \{1, 2, \dots, n_s\}$ represents the labels of the samples. Clustering can be used to group genes that manifest similar expression patterns for a set of samples [3, 7, 11, 15, 24, 35, 42, 50]. This view considers the $N = n_g$ genes as objects to be clustered, each represented by its expression profile, as a point in a $D = n_s$ dimensional space, measured over all of the samples [18]. Another type of clustering is to cluster samples into homogeneous groups which may correspond to particular macroscopic phenotypes, such as clinical syndromes or cancer types [4, 19, 47]. In this instance, the $N = n_s$ samples are viewed as the objects to be clustered, with the levels of expression of n_g genes playing the role of the features, representing each sample as a point in a $D = n_g$ dimensional space.

Sample clustering presents interesting but also very challenging problems. In typical microarray data sets, the sample space and gene space are of very different dimensionality, for example, $10^1 \sim 10^2$ samples versus $10^3 \sim 10^4$ genes. Clustering on the original high dimensional data is not guaranteed to capture a meaningful partition corresponding to empirical interest because [59]:

1. A gene expression matrix is usually generated according to some actual empirical interest, like diseased vs. healthy condition for samples. But the same set of samples may also display gender, age, or

other variability.

2. Microarrays are not typically task-specific and most of the genes are not necessarily of interest. Sample-pattern detection is subject to interference from the large number of irrelevant or redundant genes which should be pruned out or filtered when clustering samples.
3. For unsupervised analysis, uncertainty about which genes are relevant makes it difficult to construct an informative gene space to detect real sample partition.

Thus how to select the significant genes which contribute to the clustering of the samples and reveal the empirical interest pattern of the samples are very important in the data analyzing procedure. These two tasks are actually interconnected. Once the important genes are identified, the dimensions of the data will be efficiently reduced so to allow conventional clustering algorithms to be used to cluster samples. Conversely, once the salient sample patterns have been found, genes can be sorted for importance using similarity scores, such as correlation coefficient with the pattern. In general, if either an accurate sample partition or a set of significant genes is known, the other can then be easily obtained by supervised approaches [19, 30, 31]. With unsupervised clustering, however, factors such as the sparsity of data, the high dimensionality of the gene space, and the high percentage of irrelevant or redundant genes make it very difficult either to classify samples or pick out substantial genes in a context where little domain knowledge is available.

To address these problems, we propose the interrelated two-way clustering framework for unsupervised gene expression data analysis which is illustrated in Figure 2.

The goal involves two interrelated tasks: detection of meaningful patterns within the samples and selection of those significant genes which contribute to the samples' empirical pattern. To be more specific, they are:

- To select a subset of genes, usually called important genes, which are highly associated with the samples experimental distributions. This can also be considered as genes filtering.
- To 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 health control samples.

Since the volume of genes is large and no information regarding the actual partition of the samples assumed to be available, we cannot directly identify the sample patterns or significant genes. Rather, these goals must be gradually approached. First, we use the relationships of sample clusters and gene groups thus discovered to post a partial or approximate pattern. We then use this pattern to direct the elimination of irrelevant genes. In turn, the remaining meaningful genes will guide further sample pattern detection. Thus, we can formulate the problem of pattern discovery in the original data var an interplay between approximate partition detection and irrelevant gene shaving. Because of the complexity of the matrix, this procedure usually requires several iterations to achieve satisfactory results.

The criterion for terminating the series of iterations is determined by evaluating the quality of the sample partition. This is achieved in the "class validation" phase by assigning certain statistical measures to the selected genes and the related sample partition. When a stable and significant pattern of samples emerges, the iteration stops, and the selected genes with the related sample pattern become the final result of the process.

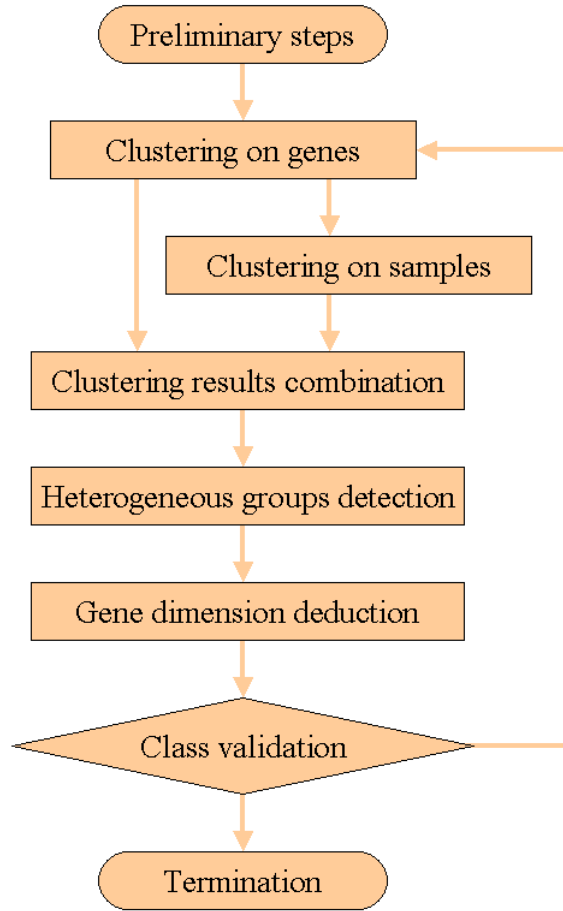


Figure 2: Framework of interrelated two-way clustering.

3 Interrelated Two-way Clustering

3.1 Preliminary Steps

We represent the gene expression data as a n_g by n_s matrix: $\widehat{\mathbf{M}} = \{m_{i,j} | i = 1, 2, \dots, n_g, j = 1, 2, \dots, n_s\}$, where there are n_s columns, one for each sample, and n_g rows, one for each gene. One row of genes is also called a *gene vector*, denoted as $\vec{g}_i = \langle m_{i,1}, m_{i,2}, \dots, m_{i,n_s} \rangle$. Thus, a gene vector contains the values of a particular attribute for all samples.

3.1.1 Data Normalization

Data sometimes need to be transformed before being used [21]. For example, attributes may be measured using different scales, such as centimeters and kilograms. In instances where the range of values differs widely from attribute to attribute, these differing attribute scales can dominate the results of the cluster analysis. It is therefore common to normalize the data so that all attributes are on the same scale.

The following are two common approaches to data normalization for each gene vector:

$$m'_{i,j} = \frac{m_{i,j} - \bar{m}_i}{\bar{m}_i}, \quad (1)$$

or

$$m'_{i,j} = \frac{m_{i,j} - \bar{m}_i}{\sigma_i}, \quad (2)$$

where

$$\bar{m}_i = \frac{\sum_{j=1}^{n_s} m_{i,j}}{n_s}, \quad \sigma_i = \frac{\sqrt{\sum_{j=1}^{n_s} (m_{i,j} - \bar{m}_i)^2}}{n_s - 1}$$

and $m'_{i,j}$ denotes the normalized value for gene vector i of sample j , $m_{i,j}$ represents the original value for gene i of sample j , n_s is the number of samples, \bar{m}_i is the mean of the values for gene vector i over all samples, and σ_i is the standard deviation of the i^{th} gene vector.

3.1.2 Similarity Measure

Many methods of cluster analysis depend on some measure of similarity (or distance) between the vectors to be clustered. Although *Euclidean distance* is a popular distance measure for spatial data, the *correlation coefficient* [14] is widely believed to be more suitable for pattern-discovery approaches because it measures the strength of the linear relationship between two vectors. This measure has the advantage of calculating similarity on the basis only of the pattern and not the absolute magnitude of the spatial vector. The formula of the correlation coefficient between two vectors $X = (x_1, x_2, \dots, x_k)$ and $Y = (y_1, y_2, \dots, y_k)$ is:

$$\rho_{X,Y} = \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^k (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^k (y_i - \bar{y})^2}}, \quad (3)$$

where k is the length of vectors X and Y , and

$$\bar{x} = \frac{1}{k} \sum_{i=1}^k x_i, \quad \bar{y} = \frac{1}{k} \sum_{i=1}^k y_i.$$

We will use the *correlation coefficient* as the similarity measure for the proposed clustering approach so that pattern similarities between genes or samples in each group will be revealed regardless of their spatial proximity.

3.2 Main Algorithm

In interrelated two-way clustering, both genes and samples are simultaneously clustered. The algorithm, illustrated in Figure 2, is an iterative procedure based on \bar{M} with n_g genes and n_s samples. The idea is to delineate the relationships between gene clusters and sample groups while iteratively clustering through both genes and samples to extract important genes and classify samples simultaneously. Within each iteration there are five main steps:

Step 1: clustering on genes.

The task of this step is to cluster n_g genes into k groups, denoted as G_i ($1 \leq i \leq k$), each of which is an exclusive subset of the entire gene set. The clustering method can be any method which takes the number of clusters as an input parameter, such as K-means or SOM [23, 22].

Step 2: clustering on samples.

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 [10]), represented by $S_{i,a}$ and $S_{i,b}$.

Step 3: clustering results combination.

This step combines the clustering results of the previous steps. By *Step 2*, we get k pairs of samples clusters $S_{i,a}, S_{i,b}$ ($i = 1, 2, \dots, k$). Then we choose one cluster from each pair and find all possible intersection of these k sample clusters, denoted as sample groups C_j ($1 \leq j \leq 2^k$). Without loss of the generality, let $k = 2$. Then the samples can be divided into four groups:

- C_1 (intersection of $S_{1,a}$ and $S_{2,a}$);
- C_2 (intersection of $S_{1,a}$ and $S_{2,b}$);
- C_3 (intersection of $S_{1,b}$ and $S_{2,a}$);
- C_4 (intersection of $S_{1,b}$ and $S_{2,b}$).

Figure 3 illustrates the results of this combination. In the figure, 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”. The green color (second line) represents cluster results based on G_1 and blue color (third line) indicates the 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 ; and C_4 includes samples marked as “b” based on G_1 and marked as “b” based on G_2 .

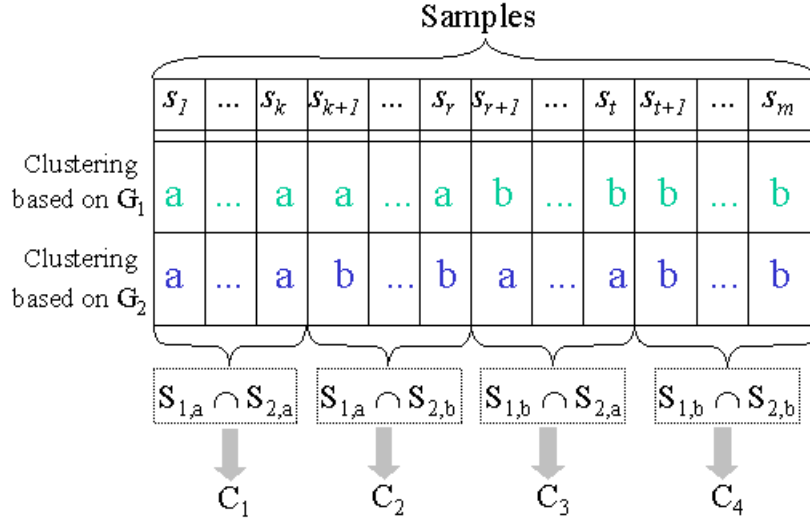


Figure 3: Clustering results combination when $k = 2$. s_1, s_2, \dots, s_m in the first line represent samples.

If $k = 3$, there will be eight possible sample groups. In general, the number of possible sample groups will reach 2^k . Usually k is set to be 2 to reduce the computational complexity.

Step 4: heterogeneous groups detection.

Among the sample groups C_1 to C_{2^k} , we choose two distinct groups C_s and C_t such that each sample in C_s is in the different cluster with each sample in C_t during clustering in Step 2. (C_s, C_t) is called a *heterogeneous group*. For example, if $k = 2$, 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 ($1 \leq i \leq k$). (C_s, C_t) is then a 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. We use these heterogeneous groups as the representation of the original sample partition.

Step 5: gene dimension deduction.

In this step, we reduce genes based on the sample patterns in the heterogeneous groups. To find genes whose expression patterns are strongly correlated with the class distinction within the heterogeneous group, we build on-off patterns according to the class distribution of each heterogeneous group and sort genes by their degree of correlation with the patterns. For example, for the heterogeneous group (C_1, C_4) , two patterns $\vec{p}_1 = \langle 0, 0, \dots, 0, 1, 1, \dots, 1 \rangle$ and $\vec{p}_2 = \langle 1, 1, \dots, 1, 0, 0, \dots, 0 \rangle$ are introduced. The pattern \vec{p}_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, \vec{p}_2 includes $|C_1|$ one's followed by $|C_4|$ zeros. For each pattern, we calculate correlation coefficient defined in Equation (3) between the on-off patterns and each gene vector \vec{g}_i :

$$\text{corr}_1(\vec{g}_i) = \rho_{\vec{g}_i, \vec{p}_1} \quad \text{and} \quad \text{corr}_2(\vec{g}_i) = \rho_{\vec{g}_i, \vec{p}_2}. \quad (4)$$

Genes have high correlation coefficient with either \vec{p}_1 or \vec{p}_2 are considered to manifest the heterogeneous group pattern. So the final *correlation value* is defined as:

$$\text{corr}(\vec{g}_i) = \max(\text{corr}_1(\vec{g}_i), \text{corr}_2(\vec{g}_i)). \quad (5)$$

We then sort all genes according to the correlation values in descending order. Gene dimension deduction is performed by eliminating some genes from the end of this ranked list. We examine the genes in the second half of the ranked list, choose a “shaving point” between two genes with the largest difference in *correlation values*, and remove the genes below the shaving point. The remaining gene set is denoted as G' . Figure 4 provides an example in which the line shows the ranked *correlation values* of all genes. As we can see, in the second half of the list (genes 7 ~ 13), the largest difference occurs between gene 8 and 9. Thus, the shaving point will be set between these genes and genes 9 to 13 will be filtered out. The semantic meaning of this shaving criterion is that, while each gene between 2 to 8 shows slightly less relevance to the sample pattern than the previous one, group 9 and following are much less relevant and they can be shaved.

It is appropriate to select the shaving point from the second half of the ranked lists so that too many genes will not be removed in a single step, particularly when the largest difference appears between the first few genes. Since the heterogeneous group pattern determined earlier in the process may not exactly match the actual partition.

Similarly, for the other heterogeneous group (C_2, C_3) , another reduced gene sequence G'' is generated.

Now the question is which gene subset should be chosen for the next iteration, G' or G'' ? The semantic meaning behind this is to select a heterogeneous group which is a better candidate to represent the empirical

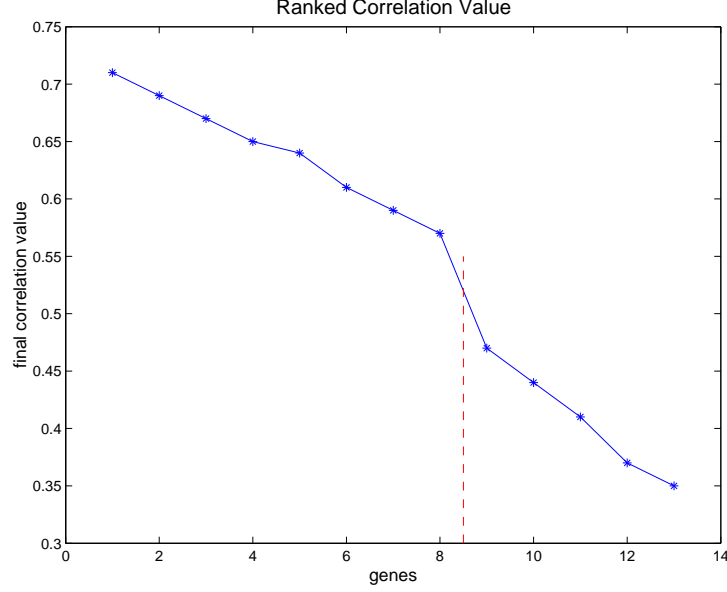


Figure 4: Distribution of ranked correlation coefficient value for a set of genes.

distribution of samples because G' and G'' are generated based on the corresponding heterogeneous groups. In the previous work, cross-validation method [51, 19] was applied to identify heterogeneous group and select genes. In general, we hope that the samples in each part of a given heterogeneous group will have high correlation to one another and have large difference from the samples in the other part when measured over the reduced gene vectors. We therefore measure the likelihood of a given heterogeneous group to be representative pattern for samples by combining its correlation within each part and dissimilarity between two part. For example, the likelihood of representation of heterogeneous group (C_1, C_4) can be defined as:

$$\delta_{C_1, C_4} = \exp\left(-\left(\frac{1}{\log|C_1| \cdot COR(C_1)} + \frac{1}{\log|C_4| \cdot COR(C_4)}\right)\right) \times DIS(C_1, C_4). \quad (6)$$

$COR(C_i)$ (Equation 7) measures the accumulation of variance of each gene vector over samples in C_i . If the variance value of each gene vector is low, then this gene vector is stable for this set of samples. Thus, if $COR(C_i)$ is low, then the genes involved generally correlate to manifest a single function over samples in C_i . This set of highly-correlated samples may represent a single condition for the entire sample pattern of empirical interest.

$$COR(C_i) = \frac{|C_i| \cdot |G'|}{\sum_{i \in G'} \sum_{j \in C_i} (m_{i,j} - \bar{m}_{i,C_i})^2}, \quad \text{where } \bar{m}_{i,C_i} = \frac{1}{|C_i|} \sum_{j \in C_i} m_{i,j}. \quad (7)$$

The dissimilarity $DIS(C_i, C_{i'})$ of two parts within the heterogeneous group is defined as the average of the pairwise dissimilarity between all pairs of objects in the different parts. This is expressed by the following equation:

$$DIS(C_i, C_{i'}) = \frac{1}{|C_i| \times |C_{i'}|} \sum_{p_1 \in C_i, p_2 \in C_{i'}} \frac{1}{\rho_{\bar{m}_{p_1}, \bar{m}_{p_2}}} \quad (8)$$

where

$$\begin{aligned}\vec{m}_{p_1} &= \langle m_{1,p_1}, m_{i_2,p_1}, \dots, m_{|G'|,p_1} \rangle, \\ \vec{m}_{p_2} &= \langle m_{1,p_2}, m_{i_2,p_2}, \dots, m_{|G'|,p_2} \rangle.\end{aligned}$$

\vec{m}_{p_1} is a sample in sub-matrix C_i , \vec{m}_{p_2} is a sample in sub-matrix $C_{i'}$ and $\rho_{\vec{m}_{p_1}, \vec{m}_{p_2}}$ is the correlation coefficient between \vec{m}_{p_1} and \vec{m}_{p_2} defined in Equation 3.

The likelihood is calculated for each heterogeneous group. When the heterogeneous group which has higher likelihood value is found, its corresponding reduced genes is selected for the next iteration.

3.3 Class Validation and Termination

After one iteration involving detection of sample pattern and selection of genes, a certain number of genes will be shaved. The remaining genes and the entire samples then form a new gene expression matrix from which a new iteration starts.

We will now discuss the issue of determining when sufficient iterations have been performed. Ideally, iterations will be terminated when a stable and significant pattern of samples has emerged. Thus, the iteration termination criterion involves determining the measurement and threshold which identifies a “stable and significant” pattern.

The propose of clustering samples is based on identifying groups of empirical interesting patterns in the underlying samples. In general, we hope that the samples in a given group will be similar (or related) to one another and different from (or unrelated to) the samples in other groups. The greater the similarity (or homogeneity) within a group, and the greater the difference between groups, the better or more distinct the partition.

As described above, after each iteration, we use the remaining genes to classify samples and then use the *coefficient of variation* (CV) to measure how “internally-similar and well-separated” this partition is:

$$CV = \frac{1}{N} \sum_{k=1}^N \frac{\sigma_k}{\|\vec{\mu}_k\|}, \quad (9)$$

where N represents the cluster number, μ_k indicates the center of group k , and σ_k represents the standard deviation of group k . Assuming there are t objects $\{\vec{v}_1, \vec{v}_2, \dots, \vec{v}_t\}$ in group k , each object is a n -dimensional vector $\vec{v}_j = \langle m_{1,j}, m_{2,j}, \dots, m_{n,j} \rangle$. The center of group k is defined as:

$$\vec{\mu}_k = \langle \bar{m}_{1,k}, \bar{m}_{2,k}, \dots, \bar{m}_{n,k} \rangle,$$

where

$$\bar{m}_{i,k} = \frac{1}{|t|} \sum_{j=1}^t m_{i,j}, \quad (i = 1, 2, \dots, n).$$

And the standard deviation of group k is defined as:

$$\sigma_k = \frac{\sqrt{\sum_{i=1}^t \|\vec{v}_i - \vec{\mu}_k\|^2}}{t - 1}.$$

It is clear that, if the dataset contains an “internally-similar and well-separated” partition, the standard deviation of each group will be low, and the CV value is expected to be small. Thus, based on the coefficient of variation, we may conclude that small values of the index indicate the presence of a “good” pattern.

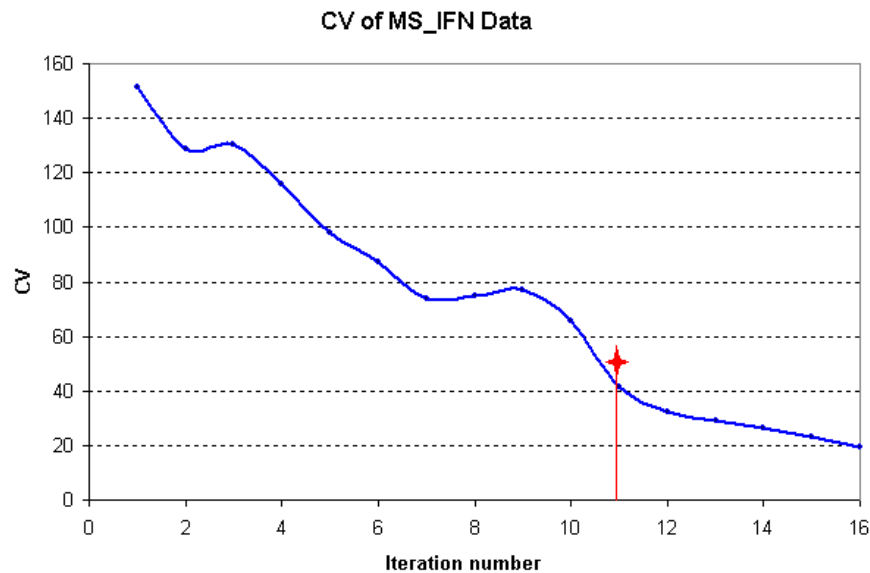


Figure 5: The CV value after each iteration.

We examine the *coefficient of variation* values after each iteration and terminate the algorithm after an iteration with a *CV* value much smaller than the previous. In the example in Figure 5, we have applied the interrelated two-way clustering approach to a data set by monitoring the *CV* until the gene number is very small. This illustrates the change in *CV* value in relation to gene numbers. In this figure, the *CV* value drops abruptly after a certain iteration (here, when the gene number is less than 100 after the eleventh iteration), and the iteration can then stop. This termination point is indicated by a red star.

Another applicable termination condition involves checking whether the number of genes is small enough to guide sample class prediction. This number is highly dependent on the type of data. For example, in a typical biological system, the number of genes needed to fully characterize a macroscopic phenotype and the factors determining this number are often unclear. Experiments also show that, for certain data, gene numbers varying from 10 ~ 200 can all serve as good predictors [19]. For our microarray data experiments, we have chosen 100 as a compromise termination number; e.g. when the number of genes falls below 100, the iteration stops. This termination condition is used only as a supplementary criterion.

Genes that remain will be regarded as the selected genes resulting from this interrelated two-way clustering approach. They are then used to cluster the samples for a final result. Since the number of genes is relatively small, the traditional clustering methods can be applied to the selected genes. The remaining genes can also be treated as “predictors” to establish cluster labels such as disease symptoms and control condition for the next batch of samples.

4 Performance Evaluation

In this section, we will analyze the effectiveness of the proposed approach through experiments on various data sets.

4.1 External Evaluation Criteria

4.1.1 Rand Index

A measurement of “agreement” (called the “Rand Index” [43, 60]) between the ground-truth of the sample partition and the clustering result was used to evaluate the performance of the algorithm. Given a set of samples $S = \{s_1, s_2, \dots, s_{n_s}\}$, suppose $P = \{p_1, p_2, \dots, p_k\}$ is the actual partition according to the conditions of empirical interest, and $Q = \{q_1, q_2, \dots, q_m\}$ is a partition of samples resulting from the clustering algorithm which satisfies $S = \bigcup_{i=1}^k p_i = \bigcup_{j=1}^m q_j$, $p_i \cap p_{i'} = \emptyset$ ($1 \leq i, i' \leq k$) and $q_j \cap q_{j'} = \emptyset$ ($1 \leq j, j' \leq m$). Let a represent the number of pairs of samples that are in the same class in P and in the same cluster in Q , b represent the number of pairs of samples that are in the same class in P but not in the same cluster in Q , c be the number of pairs of samples that are in the same class in Q but not in the same cluster in P , and d be the number of pairs of samples that are in different classes in P and in different clusters in Q . Thus, a and d measure the agreement of two partitions, while b and c indicate disagreement. The formula of the *Rand Index* [43] is:

$$RI = \frac{a + d}{a + b + c + d}. \quad (10)$$

The *Rand Index* lies between 0 and 1. When the two partitions match perfectly, the *Rand Index* is 1. In our experiments, we calculate a *Rand Index* value between the ground-truth and the result of each potential method to evaluate the quality of the clustering algorithms. In these tests, a higher the *Rand Index* value indicates better algorithm performs.

4.1.2 Interactive Visualization

A linear mapping tool [8, 64] which maps the n -dimensional dataset onto two-dimensional space is used to view the changes in sample distribution during the iterative process.

Let vector $\vec{P}_g^* = (x_{g1}, x_{g2}, \dots, x_{gn})$ represent a data element in the n -dimensional space. Equation (11) describes the mapping of \vec{P}_g^* onto a two-dimensional point \vec{Q}_g^* :

$$\vec{Q}_g^* = \sum_{i=1}^n (\lambda_i * x_{gi}) \vec{S}_i \quad \lambda_i \in [-1, 1] \quad (11)$$

where λ_i is an adjustable weight for each dimension (coordinate) with a default value is 0.5, n is the number of dimensions of the input space, and \vec{S}_i ($i = 1, 2, \dots, n$) are unit vectors which divide the center circle of the display into n equal directions, i.e., $\vec{S}_i = 2\pi/n * i$. The mapping Formula (11) replicates the correlation relationship of the input space onto the two-dimensional images. Note that point $(0, 0, \dots, 0)$ in the input space will be mapped onto the two-dimensional center $(0, 0)$ (assuming all dimension weights are equal). Additionally, all points in the format (a, a, \dots, a) will also be mapped to the center. If \vec{X} and \vec{Y} have the same pattern; i.e., ratios of each mapped pair, these vectors will be mapped onto a straight line across the center of the 2D display space. All vectors with same pattern as \vec{X} and \vec{Y} will be mapped onto that line. This mapping method takes the advantage of graphical visualization techniques to reveal the underlying data patterns.

Table 1: Rand Index value reached by applying two traditional clustering methods.

Data Set (Sample #)		MS_IFN	MS_CON	Leukemia
		28	30	72
K-means	Equation (1)	0.4841	0.4920	0.5027
	Equation (2)	0.4815	0.4851	0.5070
SOM	Equation (1)	0.5238	0.4920	0.4945
	Equation (2)	0.4815	0.4920	0.5027

4.2 Experimental Results

We will now present experimental results using three microarray data sets. The first two data sets are from a study of multiple-sclerosis patients collected by the Neurology and Pharmaceutical Sciences Departments of the State University of New York at Buffalo [39]. Multiple sclerosis (MS) is a chronic, relapsing, inflammatory disease, and interferon- β (IFN- β) has offered the main treatment for MS over the last decade [62]. The MS dataset includes two groups: the MS_IFN group, containing 28 samples (14 MS, 14 IFN), and the MS_CON group, containing 30 samples (15 MS, 15 Control). Each sample is measured over 4132 genes. The third data set is based on a collection of leukemia patient samples reported in (Golub et al., 1999) [19]. The matrix includes 72 samples (47 ALL vs. 25 AML). Each sample is measured over 7129 genes. The ground-truth of the partition, which includes such information as how many samples belong to each cluster and the cluster label for each sample, is used only to evaluate the experimental results.

To evaluate the performance of the proposed algorithm, we compared its performance in classifying the samples with two popularly-used traditional clustering methods: *K-means* ($K=2$) and *self-organizing maps* (use 2×2 SOM). Table 1 provides result obtained by directly applying the clustering algorithms to high gene-dimension matrices without an interrelated two-way analysis. Both clustering algorithms were applied to the matrix after data normalization according to Equations (1) and (2). This table indicates that performance of SOM is slightly superior to that of k-means for the two MS datasets. The k-means algorithm performed better with the leukemia data set. However, neither of these two methods resulted in a very good matching rate.

The proposed interrelated two-way clustering approach was also applied to the same gene expression matrices. The results obtained were dependent on the following parameters:

- Basic clustering algorithm: K-means or SOM
- Choice of *data normalization* method: Equation 1 or Equation 2

Figure 6 provides clustering results of the multiple sclerosis and leukemia datasets with all possible combinations of the above parameters. The horizontal axis indicates the different datasets. The four different colors are used to represent the various combination of the basic clustering algorithms and data normalization methods. These results indicate that, while the Rand Index value varies from 0.6 to 0.9 for different parameter combinations and different datasets, the index is consistently higher than the results obtained by directly applying cluster methods. The figure also shows that the optional measurement or combination of parameters is highly dependent upon the application, the environment, and the distribution of the data.

In Figure 7, the interactive visualization tool is used to show the distribution of samples before and after the interrelated two-way clustering procedure. This result is based upon using 2×2 SOM as the basic

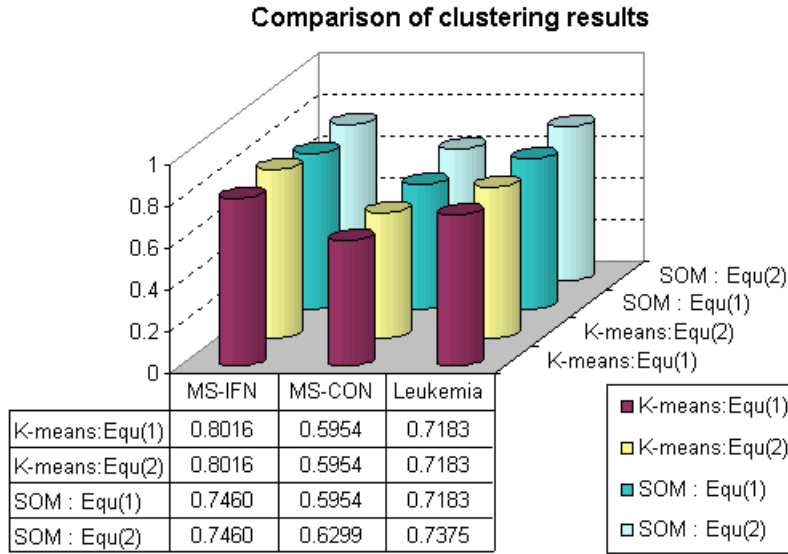


Figure 6: Rand Index values for the three datasets by using the interrelated two-way clustering method.

clustering method and Equation 1 as the data normalization function. As indicated by this figure, prior to the application of the iterative approach, the samples are uniformly scattered, with no obvious clusters. As the iterations proceed, sample clusters progressively emerge until in Figure 7(B), the samples are clearly separated into two groups. The green and red dots indicate the actual partition of the samples, while the two dashed circles show the clusters resulting from the interrelated two-way clustering approach, with arrows pointing out the incorrectly-classified samples. This visualization provides a clear illustration of the iterative process. Here, it selected 96 genes and classified 28 samples into two groups. 11 samples are in group one, matching the MS disease samples. Another 17 samples are in group two, of these, 14 are from the IFN treatment group and 3 are incorrectly matched.

Figures 6 and 7 therefore illustrate the effectiveness of the interrelated two-way clustering method for such high-dimensional gene data.

5 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. This approach can detect significant patterns within samples while dynamically selecting significant genes which manifest the conditions of actual empirical interest. We have shown that, during the iterative clustering, reducing genes can improve the accuracy of sample class discovery, which in turn will guide further genes reduction. We have demonstrated the effectiveness of this approach through experiments conducted with two multiple-sclerosis data sets and a leukemia data set. These experiments indicate that this appears to be a promising approach for unsupervised sample clustering on gene array data sets.

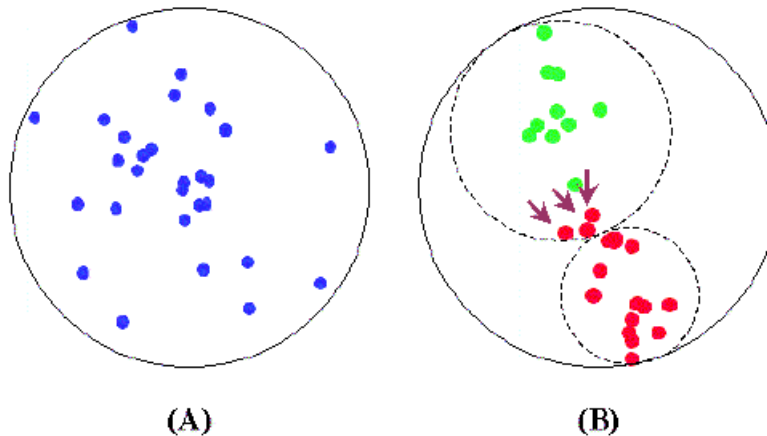


Figure 7: The interrelated two-way clustering approach as applied to the MS_IFN group. (A) shows the distribution of the original 28 samples. Each point represents a sample mapped from the intensity vectors of 4132 genes. (B) shows the distribution of the same 28 samples after the interrelated two-way clustering approach. The 4132 genes have been reduced to 96 genes, therefore each sample is a 96-dimension vector.

References

- [1] Alizadeh A.A., Eisen M.B., Davis R.E., Ma C., Lossos I.S., Rosenwald A., Boldrick J.C., Sabet H., Tran T., Yu X., Powell J.I., Yang L., Marti G.E. et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, Vol.403:503–511, February 2000.
- [2] Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D. and Levine A.J. 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.
- [3] Alter O., Brown P.O. and Bostein D. Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, Vol. 97(18):10101–10106, August 2000.
- [4] Azuaje, Francisco. Making genome expression data meaningful: Prediction and discovery of classes of cancer through a connectionist learning approach, 2000.
- [5] Barash Y. and Friedman N. Context-specific bayesian clustering for gene expression data. *Bioinformatics, RECOM01*, 2001.
- [6] Ben-Dor A., Friedman N. and Yakhini Z. Class discovery in gene expression data. In *Proc. Fifth Annual Inter. Conf. on Computational Molecular Biology (RECOMB 2001)*, 2001.
- [7] Ben-Dor A., Shamir R. and Yakhini Z. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.
- [8] Bhadra D. and Garg A. 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.

- [9] Biedl T., Brejova B., Demaine E.D., Hamel A.M. and Vinar T. Optimal Arrangement of Leaves in the Tree Representing Hierarchical Clustering of Gene Expression Data. Technical report 2001-14, University of Waterloo, Canada, 2001.
- [10] Brazma, Alvis and Vilo, Jaak. Minireview: Gene expression data analysis. *Federation of European Biochemical societies*, 480:17–24, June 2000.
- [11] Brown M.P.S., Grundy W.N., Lin D., Cristianini N., Sugnet C.W., Furey T.S., Ares M.Jr. and Haussler D. Knowledge-based analysis of microarray gene expression data using support vector machines. *Proc. Natl. Acad. Sci.*, 97(1):262–267, January 2000.
- [12] Chen J.J., Wu R., Yang P.C., Huang J.Y., Sher Y.P., Han M.H., Kao W.C., Lee P.J., Chiu T.F., Chang F., Chu Y.W., Wu C.W. and Peck K. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. *Genomics*, 51:313–324, 1998.
- [13] DeRisi J., Penland L., Brown P.O., Bittner M.L., Meltzer P.S., Ray M., Chen Y., Su Y.A. and Trent J.M. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genetics*, 14:457–460, 1996.
- [14] Devore, Jay L. *Probability and Statistics for Engineering and Sciences*. Brook/Cole Publishing Company, 1991.
- [15] Eisen M.B., Spellman P.T., Brown P.O. and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, Vol. 95:14863–14868, 1998.
- [16] Ermolaeva O., Rastogi M., Pruitt K.D., Schuler G.D., Bittner M.L., Chen Y., Simon R., Meltzer P., Trent J.M. and Boguski M.S. Data management and analysis for gene expression arrays. *Nature Genetics*, 20:19–23, 1998.
- [17] Furey T.S., Cristianini N., Duffy N., Bednarski D.W., Schummer M., and Haussler D. Support Vector Machine Classification and Validation of Cancer Tissue Samples Using Microarray Expression Data. *Bioinformatics*, Vol.16(10):909–914, 2000.
- [18] Getz G., Levine E. and Domany E. Coupled two-way clustering analysis of gene microarray data. *Proc. Natl. Acad. Sci. USA*, Vol. 97(22):12079–12084, October 2000.
- [19] Golub T.R., Slonim D.K., Tamayo P., Huard C., Gassenbeek M., Mesirov J.P., Coller H., Loh M.L., Downing J.R., Caligiuri M.A., Bloomfield D.D. and Lander E.S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, Vol. 286(15):531–537, October 1999.
- [20] Hakak Y., Walker J.R., Li C., Wong W.H., Davis K.L., Buxbaum J.D., Haroutunian V. and Fienberg A.A. . Genome-Wide Expression Analysis Reveals Dysregulation of Myelination-Related Genes in Chronic Schizophrenia. *Proc. Natl. Acad. Sci. USA*, Vol. 98(8):4746–4751, April 2001.
- [21] Han, Jiawei and Kamber, Micheline. *Data Mining: Concept and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, August 2000.
- [22] Hartigan, J.A. *Clustering Algorithm*. John Wiley and Sons, New York., 1975.

- [23] Hartigan, J.A. and Wong, M.A. Algorithm AS136: a k-means clustering algorithms. *Applied Statistics*, 28:100–108, 1979.
- [24] Hastie T., Tibshirani R., Boststein D. and Brown P. Supervised harvesting of expression trees. *Genome Biology*, Vol. 2(1):0003.1–0003.12, January 2001.
- [25] Heller R.A., Schena M., Chai A., Shalon D., Bedilion T., Gilmore J., Woolley D.E. and Davis R.W. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA*, 94:2150–2155, 1997.
- [26] Herrero J., Valencia A. and Dopazo J. A hierarchical unsupervised growing neural network for clustering gene expression patterns. *Bioinformatics*, 17:126–136, 2001.
- [27] Holter NS., Mitra M., Maritan A., Cieplak M., Banavar JR., Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci*, 97(15):8409–8414, 2000.
- [28] Iyer V.R., Eisen M.B., Ross D.T., Schuler G., Moore T., Lee J.C.F., Trent J.M., Staudt L.M., Hudson Jr. J., Boguski M.S., Lashkari D., Shalon D., Botstein D. and Brown P.O. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.
- [29] Jiang M., Ryu J., Kiraly M., Duke K., Reinke V. and Kim S.K. Genome-Wide Analysis of Developmental and Sex-Regulated Gene Expression Profiles in *Caenorhabditis Elegans*. *Proc. Natl. Acad. Sci. USA*, Vol. 98(1):218–223, January 2001.
- [30] Jiang S., Tang C., Zhang L., Zhang A. and Ramanathan M. 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.
- [31] Jorgensen, Anna. Clustering excipient near infrared spectra using different chemometric methods. Technical report, Dept. of Pharmacy, University of Helsinki, 2000.
- [32] Kohonen T. *Self-Organization and Associative Memory*. Springer-Verlag, Berlin, 1984.
- [33] Li, Wentian. Zipf’s Law in Importance of Genes for Cancer Classification Using Microarray Data. Lab of Statistical Genetics, Rockefeller University, April 2001.
- [34] Luscombe N. M., Greenbaum D. and Gerstein M. Review: What is bioinformatics? An introduction and overview. *Yearbook of Medical Informatics*, pages 83–99, 2001.
- [35] Manduchi E., Grant G.R., McKenzie S.E., Overton G.C., Surrey S. and Stoeckert C.J.Jr. Generation of patterns from gene expression data by assigning confidence to differentially expressed genes. *Bioinformatics*, Vol. 16(8):685–698, 2000.
- [36] Martin K.J., Graner E., Li Y., Price L.M., Kritzman B.M., Fournier M.V., Rhei E. and Pardee A.B. High-Sensitivity Array Analysis of Gene Expression for the Early Detection of Disseminated Breast Tumor Cells in Peripheral Blood. *Proc. Natl. Acad. Sci. USA*, Vol. 98(5):2646–2651, February 2001.
- [37] Mody M., Cao Y., Cui Z., Tay K.Y., Shyong A., Shimizu E., Pham K., Schultz P., Welsh D. and Tsien J.Z. Genome-Wide Gene Expression Profiles of the Developing Mouse Hippocampus. *Proc. Natl. Acad. Sci. USA*, Vol. 98(15):8862–8867, July 2001.

- [38] Moler E.J., Chow M.L. and Mian I.S. Analysis of Molecular Profile Data Using Generative and Discriminative Methods. *Physiological Genomics*, Vol. 4(2):109–126, 2000.
- [39] Nguyen LT., Ramanathan M., Munschauer F., Brownschidle C., Krantz S., Umhauer M., et al. Flow cytometric analysis of in vitro proinflammatory cytokine secretion in peripheral blood from multiple sclerosis patients. *J Clin Immunol*, 19(3):179–185, 1999.
- [40] Park P.J., Pagano M., and Bonetti M. A Nonparametric Scoring Algorithm for Identifying Informative Genes from Microarray Data. In *Pacific Symposium on Biocomputing*, pages 52–63, 2001.
- [41] Pavlidis P., Weston J., Cai J. and Grundy W.N. Gene Functional Classification from Heterogeneous Data. In *RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology*, pages 249–255. ACM Press, 2001.
- [42] Perou C.M., Jeffrey S.S., Rijn, M.V.D., Rees C.A., Eisen M.B., DRoss O.T., Pergamenschikov A., Williams C.F., Zhu S.X., Lee, J.C.F., Lashkari D., Shalon D., Brown P.O., and Bostein D. Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl. Acad. Sci. USA*, Vol. 96(16):9212–9217, August 1999.
- [43] Rand, W.M. Objective criteria for evaluation of clustering methods. *Journal of the American Statistical Association*, 1971.
- [44] Schena M., Shalon D., Davis R.W. and Brown P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270:467–470, 1995.
- [45] Schena M., Shalon D., Heller R., Chai A., Brown P.O., and Davis R.W. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA*, Vol. 93(20):10614–10619, October 1996.
- [46] Shalon D., Smith S.J. and Brown P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*, 6:639–645, 1996.
- [47] Slonim D.K., Tamayo P., Mesirov J.P., Golub T.R. and Lander E.S. Class Prediction and Discovery Using Gene Expression Data. In *RECOMB 2000: Proceedings of the Fifth Annual International Conference on Computational Biology*. ACM Press, 2000.
- [48] Speed, Terry. cDNA Microarrays on Glass Slides. Institute for Pure and Applied Mathematics (IPAM) Functional Genomics/Expression Arrays Fall 2000 tutorial. <http://www.ipam.ucla.edu/programs/fg2000/tutorials.html>.
- [49] Spellman PT., Sherlock G., Zhang MQ., Iyer VR., Anders K., Eisen MB., et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, 1998.
- [50] Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. 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.
- [51] Tang C., Zhang L., Zhang A. and Ramanathan M. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *Proceeding of BIBE2001: 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, Bethesda, Maryland, November 4-5 2001.

- [52] Tavazoie, S., Hughes, D., Campbell, M.J., Cho, R.J. and Church, G.M. Systematic determination of genetic network architecture. *Nature Genet*, pages 281–285, 1999.
- [53] Thomas J.G., Olson J.M., Tapscott S.J. and Zhao L.P. An Efficient and Robust Statistical Modeling Approach to Discover Differentially Expressed Genes Using Genomic Expression Profiles. *Genome Research*, Vol. 11(7):1227–1236, 2001.
- [54] Tusher V.G., Tibshirani R. and Chu G. Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proc. Natl. Acad. Sci. USA*, Vol. 98(9):5116–5121, April 2001.
- [55] Vingron, M. and Hoheisel, J. Computational Aspects of Expression Data. *J. Mol. Med.*, 77:3–7, 1999.
- [56] Virtaneva K., Wright F.A., Tanner S.M., Yuan B., Lemon W.J., Caligiuri M.A., Bloomfield C.D., Chapelle A. de la and Krahe R. Expression Profiling Reveals Fundamental Biological Differences in Acute Myeloid Leukemia with Isolated Trisomy 8 and Normal Cytogenetic. *Proc. Natl. Acad. Sci. USA*, Vol. 98(3):1124–1129, January 2001.
- [57] Welford S.M., Gregg J., Chen E., Garrison D., Sorensen P.H., Denny C.T. and Nelson S.F. Detection of differentially expressed genes in primary tumor tissues using representational differences analysis coupled to microarray hybridization. *Nucleic Acids Research*, 26:3059–3065, 1998.
- [58] Welsh J.B., Zarrinkar P.P., Sapinoso L.M., Kern S.G., Behling C.A., Monk B.J., Lockhart D.J., Burger R.A. and Hampton G.M. Analysis of Gene Expression Profiles in Normal and Neoplastic Ovarian Tissue Samples Identifies Candidate Molecular Markers of Epithelial Ovarian Cancer. *Proc. Natl. Acad. Sci. USA*, Vol. 98(3):1176–1181, January 2001.
- [59] Xing E.P. and Karp R.M. Cliff: Clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics*, Vol. 17(1):306–315, 2001.
- [60] Yeung, Ka Yee and Ruzzo, Walter L. An empirical study on principal component analysis for clustering gene expression data. Technical Report UW-CSE-2000-11-03, Department of Computer Science & Engineering, University of Washington, 2000.
- [61] Yeung K.Y., Haynor D.R. and Ruzzo W.L. Validating Clustering for Gene Expression Data. *Bioinformatics*, Vol.17(4):309–318, 2001.
- [62] Yong V., Chabot S., Stuve Q. and Williams G. Interferon beta in the treatment of multiple sclerosis: mechanisms of action. *Neurology*, 51:682–689, 1998.
- [63] Zhang H., Yu C.Y., Singer B., and Xiong M. Recursive Partitioning for Tumor Classification with Gene Expression Microarray Data. *Proc. Natl. Acad. Sci. USA*, Vol. 98(12):6730–6735, June 2001.
- [64] Zhang L., Tang C., Shi Y., Song Y., Zhang A. and Ramanathan M. VizCluster: An Interactive Visualization Approach to Cluster Analysis and Its Application on Microarray Data. In *Second SIAM International Conference on Data Mining*. Hyatt Regency Crystal City, Arlington, VA, April 2002.
- [65] Zou S., Meadows S., Sharp L., Jan L.Y. and Jan Y. N. Genome-Wide Study of Aging and Oxidative Stress Response in *Drosophila Melanogaster*. *Proc. Natl. Acad. Sci. USA*, Vol. 97(25):13726–13731, December 2000.