[Frontiers in Bioscience 10, 3068-3077, September 1, 2005]

FINDING DOMINANT SETS IN MICROARRAY DATA

Xuping Fu [1], Li Teng[2], Yao Li [1], Wenbin Chen [3], Yumin Mao [1], I-Fan Shen [2] and Yi Xie [1]

[1] State Key Laboratory of Genetic Engineering, Institute of Genetics, School of Life Science, Fudan University, Shanghai 200433 PR, China, [2] Department of Computer Science and Engineering, Fudan University, Shanghai 200433 PR, China, and [3] Department of Mathematics, Fudan University, Shanghai 200433 PR, China

TABLE OF CONTENTS

1. ABSTRACT

Clustering allows us to extract groups of genes that are tightly coexpressed from Microarray data. In this paper, a new method DSF_Clust is developed to find dominant sets (clusters). We have preformed DSF_Clust on several gene expression datasets and given the evaluation with some criteria. The results showed that this approach could cluster dominant sets of good quality compared to kmeans method. DSF_Clust deals with three issues that have bedeviled clustering, some dominant sets being statistically determined in a significance level, predefining cluster structure being not required, and the quality of a dominant set being ensured. We have also applied this approach to analyze published data of yeast cell cycle gene expression and found some biologically meaningful gene groups to be dug out. Furthermore, DSF_Clust is a potentially good tool to search for putative regulatory signals.

2. INTRODUCTION

The advent of Microarray technology has enabled biologists to monitor the expression patterns of thousands genes in parallel during biological processes. How to make sense of the massive data sets produced by Microarray is now a great challenge. One of the aims is to identify, group, analyze genes that exhibit highly similar expression profiles. Such genes defined as coexpressed are probably coregulated. Clustering is most widely used for grouping analysis. There is a rich literature on cluster analysis and various techniques have been developed. For example, Eisen *et al*. popularized the use of hierarchical clustering and applied the average linkage hierarchical clustering algorithm to identify groups of co-regulated yeast genes (1). Tamayo *et al*. used self-organizing maps to identify clusters in the yeast cell cycle and human hematopoietic differentiation data sets (2). Other techniques such as k-means (3), principle component analysis (4), singular value decomposition (5), have been implemented and successfully been used to analyze high-dimensional Microarray data.

Clustering can be defined as the process of organizing objects into groups whose members are similar in some way. In Microarray data analysis, the algorithms intend to group genes whose expression profiles are sufficiently close to each other into small clusters. To find small clusters, in some common k-clustering approaches, such as k-means and self-organizing maps (SOM), the predefinition of the number of clusters (parameter of the algorithm) is required. However, the number of clusters present in the data is usually unknown in advance. Changing the parameter usually affects the final result considerably. In addition, k-clustering algorithms do not deal adequately with "noise". If genes are, despite a rather low correlation with other cluster members, forced to end up in one of the clusters. Hence the clusters contain a lot of "noise" and become less suitable for further analyses (6).

To address the issue of "noise", some graph theory based clustering methods are introduced. The intuitive idea of these methods is that clusters can be considered "densely populated areas" in data space. These areas ideally are well separated from each other (7). Based on the graph theoretic approach to clustering and segmentation, novel clustering algorithms, such as the corrupted clique model described by Ben-Dor *et al.* and iterative feature filtering using normalized cuts described by Xing *et al.* could be useful for Microarray data analysis (8,9).

In this paper, we present a new method, based on graph theoretic clustering, to find dominant sets in Microarray data. There is an analogy between the intuitive concept of a cluster and that of a dominant set (10). As a cluster, a dominant set is a set of entities that are similar under some criteria, and entities from different dominant sets are not similar. In a complete graph, a dominant set is considered as a subgraph only consisting of a cluster of nodes. When the concept is introduced to Microarray data, dominant sets turn out to be strictly equivalent to clusters of genes or samples.

In our approach, clusters are built and portrayed as unrelated entities. In contrast with k-means or SOM approach, it does not assume a given number of clusters and initial spatial structure of them, but determines cluster number and structure based on the data. In addition, some clusters are statistically determined. If appropriate parameters are selected, the members which do not show high correlations with other members in the dataset are not grouped into any dominant set, thus every member has high correlation with other ones in a dominant set, which results in very low noise in one dominant set. For convenience, in the following text, the algorithm is named as DSF_Clust.

3. ALGORITHM

3.1 Preliminaries

Dominant set is a novel concept that arises from the study of maximal subgraph problem in image clustering. An image can be represented as a similarity (edge weighted) graph, where the vertices represent individual pixels, and the weights on the edges reflect the similarity between pixel appearances. Graph-theoretic clustering reduces to a search for a complete subgraph which can be considered as a concept of the strictest definition of a cluster. A dominant set is a new formal definition of a cluster in the edge weighted graphs. In a complete graph, a dominant set is considered as a subgraph only consisting of a cluster of nodes.

We denote $P=(V,E,w)$ as an undirected edge-weighted graph without self loops, where $V=\{1,\cdots,N\}$ is the vertex set, $E\subseteq V\times V$ is the edge set, and w is the positive weight function that reflect similarity between pairs of linked vertices. Higher the w value is, closer the pairs of vertices are. There are two basic properties of the dominant set (cluster) definition: all elements within one dominant set should be similar to one another, and not similar to any element of other dominant sets. Based on the properties of internal homogeneity and external inhomogeneity, a dominant set can be defined as followed (10).

A non-empty subset of vertices $S\subseteq V$ such that $Q(T)>0$ for any non-empty $T\subseteq S$, is said to be dominant if:

1. $q_S(i)>0$, for all $i\in S$

2. $q_{S\cup\{i\}}(i)<0$, for all $i\notin S$. (equation 1)

Where $q_S(i)=\begin{cases}1, & if\,|S|=1\\ \sum\limits_{j\in S\setminus\{i\}} c_{S\setminus\{i\}}(j,i)q_{S\setminus\{i\}}(j), & otherwise\end{cases}$ (equation 2)

For the total weight of T, $Q(T)$ is defined to be:

$Q(T)=\sum\limits_{i\in T} q_T(i)$ (equation 3)

For any $i,j\in V$, if $j\notin S$, we define:

$c_S(i,j)=w_{ij}-\frac{1}{|S|}\sum\limits_{k\in S} w_{ik}$ (equation 4)

In the definition, $c_S(i,j)$ measures the similarity between nodes j and i, with respect to the average similarity between node i and its neighbors in S, and $q_S(i)$ is the function of the weights on the edges of the subgraph induced by S.

Let W be the weighted adjacency matrix of P, there is a correspondence between the problem of finding dominant sets in an edge-weighted graph and that of finding solutions of quadratic program as followed (11):
Maximize $f(X)=\frac{1}{2}XW^TX$, subject to $X\in Delta$ (equation 5)

Where $Delta=\{X\in IR^n:\ X\geq 0\ \ and\ \ e^TX=1\}$ is the standard simplex of IR^n. By virtue of this theoretical result,

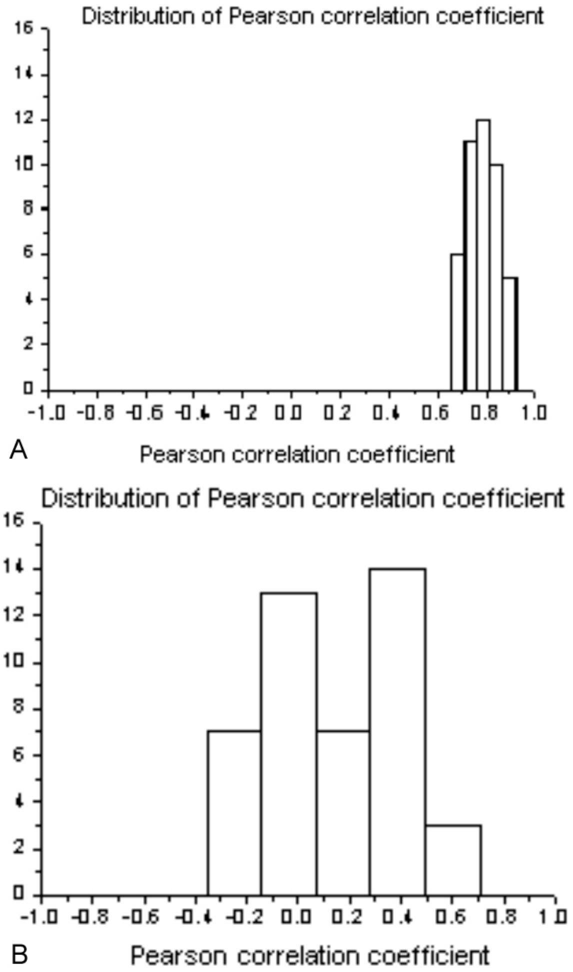


Figure 1. Distributions of Pearson correlation coefficients of elements set. (a) The set of 10 gene profiles with biggest $x_i(t)$ values after running equation 7 in 500 gene profiles. (b) The set of 10 randomly selected gene profiles from 500 gene profiles.

we can find a dominant set by first localizing a solution of program (equation 5) with the following replicator equation 6 and equation 7, and then picking up the support set X of the solution to form a dominant set (9). We consider the following dynamical system:

$$\dot{x}_i(t) = x_i(t)\left[(WX(t))_i - X(t)^T WX(t)\right], \text{(equation 6)}$$

where a dot signifies derivative with regard to time t, and its discrete time counterpart

$$x_i(t+1) = x_i(t)\frac{(WX(t))_i}{X^T(t)WX(t)}. \quad \text{(equation 7)}$$

The function $X^T(t)WX(t)$ is strictly increasing with increasing t along any non-stationary trajectory $X(t)$ under both continuous-time equation 6 and discrete-time equation 7. Furthermore, any such trajectory converges to a stationary point. Finally, a vector $X \in Delta$ is asymptotically stable under equation 6 and equation 7 if and only if X is a strict local maximizer of $X^T WX$ on $Delta$.

In the light of the dynamical properties, replicator equations naturally suggest themselves as a simple and useful heuristic for finding dominant sets. For we can find the strict local maximizer of $X^T WX$ in $Delta$ and also find a dominant set based on the vector X. In the following text, we will describe an approach to find dominant sets in Microarray data after running replicator equations.

3.2. Finding dominant sets in Microarray data

Microarray data can be represented by a real valued expression matrix I where I_{ij} is the expression level of gene i in the experiment (condition) j. Denote G as the set of genes and s_{ij} represents the similarity of the expression patterns for genes i and j. And each gene can be regard as a point in high dimensional space. Since the similarities between genes can also be presented by a weight matrix. The above graph theoretic based dominant set finding algorithm can be used to cluster genes in the similar way.

We find that after replicator equation 7 the value of $x_i(t)$ could be a criterion to judge if one point belongs to a dominant set. We performed equation 7 in randomly selected 500 gene expression patterns from Spellman *et al* (12). Here we chose to use only a single time course containing 18 time intervals. We can rearrange the 500 genes at the value of $x_i(t)$ after running equation 7. The distribution of Pearson correlation coefficients of the ten genes with biggest $x_i(t)$ values is shown in Figure 1a. In comparison, we randomly chose ten genes and draw their correlation coefficients distribution shown in Figure 1b.

Compared with random gene patterns (Figure 1b), the genes (Figure 1a) with bigger $x_i(t)$ value have tighter relations. The distribution of random datasets is sparse and the Pearson correlation coefficients are relatively low, while coefficients in Figure 1a are high. Assuming that the 10 genes with highest $x_i(t)$ value constitute a cluster, we found that out of the rest 490, only two genes' coefficients with the cluster center were higher than 0.8. And one of the two genes has 11th highest $x_i(t)$, and the other gene ranks 13 at the order of $x_i(t)$ value. Thus we can conclude that the genes with high $x_i(t)$ values would be grouped into a cluster. However, like many other partition algorithms, the cluster boundary is hard to define since $x_i(t)$ value is not obvious enough to distinguish the dominant set from non-dominant set. In the following section, we describe an approach to delimit the boundary between a dominant set and the rest of dataset when this algorithm is applied to Microarray data.

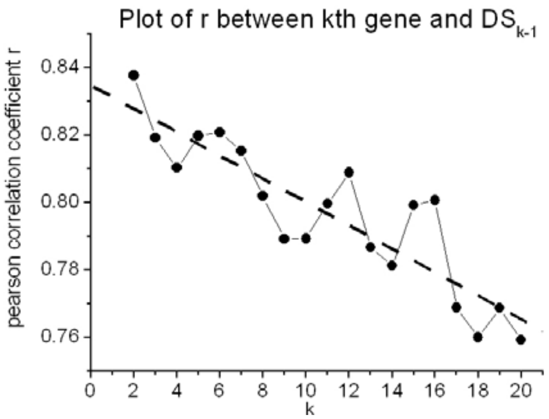


Figure 2. The Pearson correlation coefficients between kth gene and DS$_{k-1}$. Kth gene is the gene g_k and DS$_{k-1}$ is the dominant set made up by genes from g_1 to g_{k-1}

After running the above replicator equations, the genes ranked at $x_i(t)$ value. In the following text, the gene with k th biggest $x_i(t)$ value is named as g_k. Assuming that the genes from g_1 to g_{k-1} make up a dominant set and the center of the dominant set can be consider as the mean of the $k-1$ genes' expression profiles, then we calculated the Pearson correlation coefficient between g_k and the center. Figure 2 shows that the coefficients of $k=3,4,\cdots,19$. As the increment of k, the correlation coefficients tend to decrease. It shows that a gene has a stronger tendency to belonging to a dominant set if it has a bigger value of $x_i(t)$.

Therefore, definition of a dominant set is an iterative process. Genes enter iterations in $x_i(t)$ value order. That is, the bigger the value of $x_i(t)$ a gene is, the earlier the gene is used to determine whether it is within the dominant set. If a gene does not belong to a cluster, the affinities between the gene and the members in the cluster are lower than the affinities between the members within the cluster. Based on the assumption, a stop criterion can be created by two-sample t-test. The basic steps are described below:

(1) Initialization: This step initializes the members of a dominant set. Generally, the first three gene vectors are the initial members. However, if any Pearson correlation coefficient between a member and the mean of these three vectors is below a threshold, the first gene vector is removed from dataset and this step is repeated on the left dataset.

(2) Iterations: Only one gene is allowed in each iteration. When g_k is in this step, we denote Pearson correlation coefficients between g_k and $\{g_1, g_2, \cdots, g_{k-1}\}$ as $\{r_{1k}, r_{2k}, \cdots, r_{k-1k}\}$. In addition, the new center C_k and the Pearson correlation value r_k between g_k and C_k are calculated.

$$C_k = \frac{(k-1)*C_{k-1}+g_k}{k} \quad \text{(equation 8)}$$

(3) Termination: We use two sample t-test to create stop criterion. Assuming that $\{g_1, g_2, \cdots, g_{k-1}\}$ is a dominant set DS_{k-1}, we calculate the correlation coefficients $\{r_{1(k-1)}, r_{2(k-1)}, \cdots, r_{(k-2)(k-1)}\}$ between g_{k-1} and $\{g_1, g_2, \cdots, g_{k-2}\}$. Denote $\overline{r_k}$ and s_k as sample mean and sample variance of $\{r_{1k}, r_{2k}, \cdots, r_{k-1k}\}$.

$$\overline{r_k} = \frac{\sum_{i=1}^{k-1} r_{ik}}{k-1} \quad \text{(equation 9)}$$

$$s_k = \sqrt{\frac{\sum_{i=1}^{k-1}(r_{ik}-\overline{r_k})^2}{k-1}} \quad \text{(equation 10)}$$

The two-sample t-test is used to determine if the means of two samples $\{r_{1k}, r_{2k}, \cdots, r_{k-1k}\}$ and $\{r_{1k-1}, r_{2k-1}, \cdots, r_{k-2k-1}\}$ are equal. The two sample t-test is defined as:

H$_0$: $\overline{r_k} = \overline{r_{k-1}}$

H$_1$: $\overline{r_k} \neq \overline{r_{k-1}}$

t-statsitics: $$t\_value = \frac{\left|\overline{r_k} - \overline{r_{k-1}}\right|}{sqrt(\frac{s_{k-1}^2}{k-2} + \frac{s_k^2}{k-1})} \quad \text{(equation 11)}$$

If the t-statistics is higher than the t value of significance level alpha, the null hypothesis H$_0$ that the two means are equal is rejected. It means that g_k is significantly different from members of $\{g_1, g_2, \cdots, g_{k-1}\}$ in level of alpha and the iterations are terminated. Besides, to prevent low affinity members joining into a dominant set, threshold r_{ts} is defined. If r_k is below r_{ts}, the iteration is terminated. When the stop criterion is satisfied, genes of $\{g_1, g_2, \cdots, g_{k-1}\}$ constitute a dominant set.

In this approach, the mean of the corresponding expression profiles is calculated iteratively and subsequently the cluster center is moved to this mean profile. This approach moves the cluster in the direction where the "density" of profiles is higher.

3.3. Global Approach

The global approach (Table 1) is iterative. Each iteration find a dominant set using the above three steps. Before next iteration, the genes in the dominant set found in this step are removed. The rest of genes are used to find another dominant set in next iteration. There are two key parameters. Significant level alpha is the statistical character of

Table 1. The DSF_Clust global algorithm

```
Global Algorithm:
Initialize
  • Input Microarray data matrix I, significant level alpha, threshold $r_{ts}$
Iterate
  • Sort genes according to its corresponding $x_i$ value; let $g_k$ be the gene with kth biggest $x_i(t)$ value
  • Found one cluster using the gene list $\{g_1, g_2, \cdots, g_n\}$, until t value of t-test above the t value of alpha level or the Pearson correlation coefficient between $g_k$ and the dominant set $\{g_1, g_2, \cdots, g_{k-1}\}$ below $r_{ts}$
  • Take away the genes found in the cluster then continue with the genes left.
```

a dominant set and threshold r_{ts} controls the quality of a dominant set. In addition, this approach needn't predefinition of cluster number and structure. The Microarray data is grouped to small dominant sets based on internal structure of data. The source code is programmed in MATLAB and publicly available in website (http://www.chinagenenet.com/DSF_Clust). In the following text, a dominant set found by DSF_Clust are called a DS.

4. RESULTS AND DISCUSSION

4.1. Performance on three gene expression datasets

We used three gene expression datasets to evaluate the performance of our approach (13). The first dataset we used was a set of data about the response of human fibroblasts to serum (14). We chose a subset of 517 genes whose expression changed substantially in response to serum. According to Xu et al. work, the optimal number of clusters for this datasets is five (15). The second dataset was the budding yeast saccharomyces cerecisiae, with each gene having 18 time points. We selected four clusters (74 genes) in previous work (12,13). Genes in each of these four clusters shared similar pathway. Our third application was on the fluctuation of expression levels of approximately 6000 yeast genes over two cell cycles (17 time points) (16). We chose the subset consisting of 384 genes whose expression levels peaked at different time points corresponding to the five phases of the cell cycle.

According to the expected cluster numbers, we applied Kmeans approach in the three datasets. The resulted 4 clusters in first dataset, 5 clusters in second one and 5 clusters in third one are regarded as standards to evaluated our DSF_Clust results. We performed DSF_Clust in the cases of different r_{ts} s and alphas. For each trial we computed the closeness between the Kmeans results and the achieved DSs after using DSF_Clust. In convenience, we use the term 'cluster' to refer to a group of genes obtained by Kmeans approach. Two criteria introduced from other areas for closeness were used: sensitivity and specificity. Let N_{opk} be the number of entries of the k th cluster, and if this cluster is divided into several DSs, in which total entries' number is denoted as N_{Apk}. Then we define the specificity

$$SP = \frac{\sum_{i=1}^{n} N_{Apk}}{\sum_{i=1}^{n} N_{Opk}} \quad \text{(equation 12)}$$

Where n represents the respective expected cluster numbers in the three datasets. The specificity describes the effectiveness of DSF_Clust approach. Higher the specificity is, more entries belonging to one cluster are grouped into tight dominant sets. Let N_{Ank} be the number of entries of the k th DS in achieved structure and the genes in this dominant set can be fallen into several clusters, among which one cluster would have the maximum entries' number N_{Onk}. Then we define sensitivity

$$SE = \frac{\sum_{i=1}^{8} N_{Onk}}{\sum_{i=1}^{8} N_{Ank}} \quad \text{(equation 13)}$$

The sensitivity shows the rate of genes, which belong to one cluster, however, don't belong to one dominant set. Since the results of Kmeans partition are considered as standards, sensitivity means the accuracy of DSF_Clust approach. Higher the sensitivity is, fewer genes are grouped into 'wrong' dominant sets.

In addition, we calculated Pearson correlation coefficients between each member in a dominant set (cluster) and the center of the dominant set (cluster). The minimum coefficient MinR is regarded as an internal criterion to evaluate the quality of a dominant set (cluster). Intuitively, if MinR of a dominant set is close to 1, the members in the dominant set have tight correlations and very similar expression levels, whereas the MinR close to 0 means that some unrelated patterns or 'noise' are grouped into the dominant set. Thus MinR can show the goodness of a dominant set (cluster) quality.

Figure 3 shows sensitivities, specificities and MinRs of the DSF_Clust approach on the three dataset with various significant levels and correlation coefficient thresholds. In general, the sensitivities in three datasets are all above 0.85 even close to 1, except for in the alpha=0.99 condition. This ensures the correctness of the dominant sets finding results. When significant level increases from 0.80 to 0.99, the condition is more difficult to be satisfied on which dominant sets are separated in this level. Thus the sensitivity decreases with the increment of significant level alpha. If the correlation coefficient threshold r_{ts} is low and a significant level is not low enough to make a gene disjoint in a dominant set, some unrelated patterns join in a dominant set. Then the sensitivity is low, which is showed in the cases of alpha=0.99, $r_{ts} \le 0.5$ in dataset2 and dataset3. However, when significant level is below 0.99, although with the increment of r_{ts} the sensitivity increases slightly. Thus generally, the r_{ts} is independent on sensitivity.

Specificity indicates that the ratio of the genes grouped into dominant sets to the total genes. The others in total genes are considered as unrelated patterns. With the

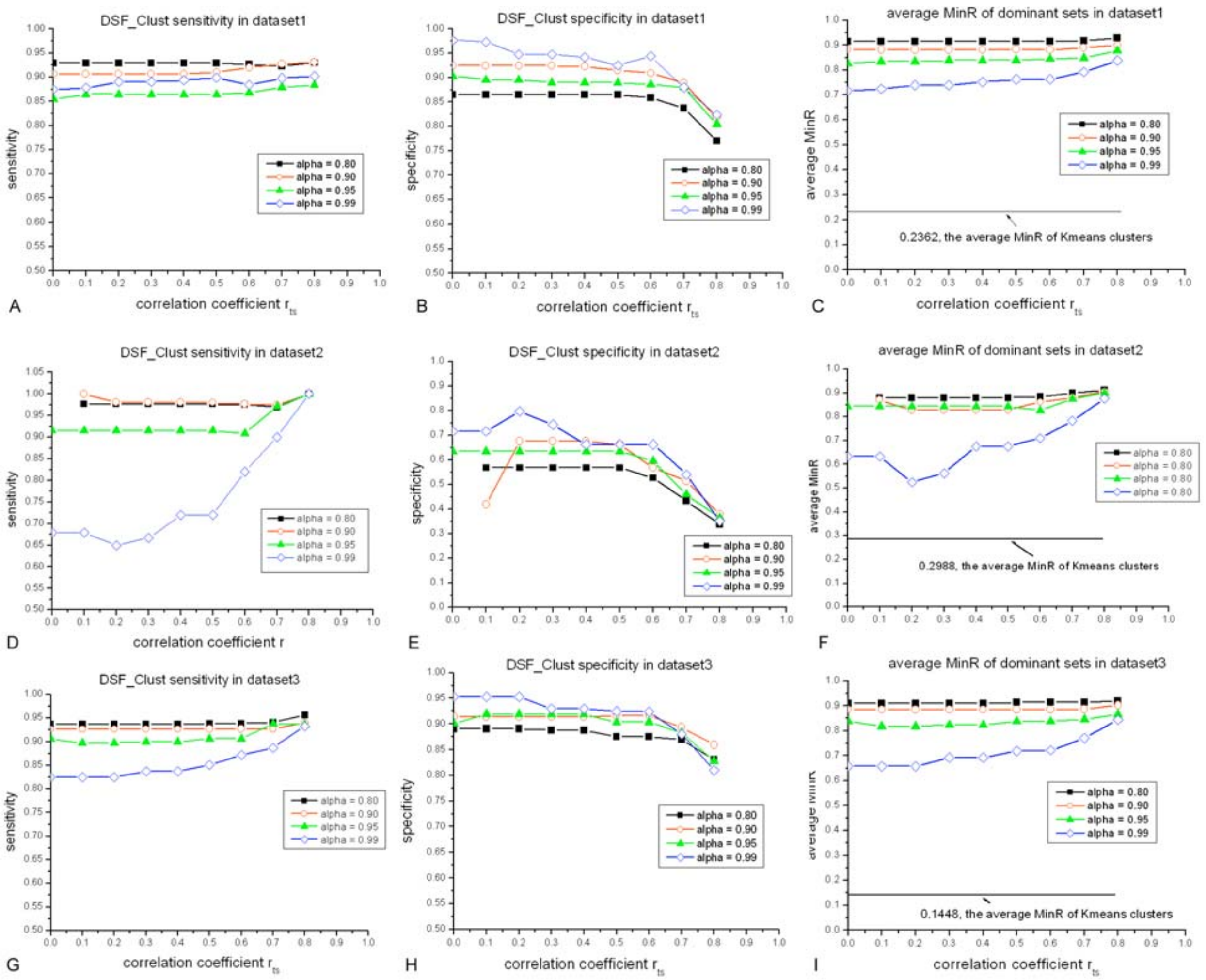


Figure 3. Sensitivities, specificities and MinRs calculated with different r_{ts} s and alphas on DSF_Clust approach. In each plot, the lines representing $alpha = 0.99$ (open diamonds), $alpha = 0.95$ (solid triangles), $alpha = 0.90$ (open circles) and $alpha = 0.80$ (solid rectangles) are curved. (a). sensitivity in dataset1 (b). specificity in dataset1 (c). MinR in dataset1. The straight line MinR=0.2362 represents the value of MinR obtained by Kmeans approach on dataset1 (d). sensitivity in dataset2 (e). specificity in dataset2 (f). MinR in dataset2. The straight line MinR=0.2988 represents the value of MinR obtained by Kmeans approach on dataset2 (g). sensitivity in dataset3 (h). specificity in dataset3 (i). MinR in dataset3. The straight line MinR=0.1448 represents the value of MinR obtained by Kmeans approach on dataset3.

increment of r_{ts} or the decreasing of significant level, the constraint is strengthened. Consequently, the number the genes belonging to one dominant set is decreasing, the specificity is also decreasing. Essentially, when r_{ts} increases or alpha decreases, a Kmeans cluster is splitting into several dominant sets that are satisfying the condition. But some genes originally belonging to a cluster cannot be grouped to any dominant set, then these genes are considered as "noise" for the dominant sets. The specificity is decreasing. However, in other aspect, the quality of a dominant set is improving and a dominant set appears to be "pure" after DSF_Clust approach.

The parameter MinR is used to evaluate the quality of the clusters obtained by one cluster approach. From the Figure 3, we can see that the MinRs of three datasets on Kmeans approach are only 0.2362, 0.2988, 0.1448, respectively. However, the MinRs on DSF_Clust approach are much higher than those on Kmeans approach. If the correlation coefficient threshold r_{ts} is chose to be 0.80, the MinRs are even higher than 0.90. Therefore, the patterns in a dominant set are very close to each other, and few "noise" patterns are contained in a dominant set. However, the improvement of dominant set quality is at the cost of the decreasing of specificity. That is, it is probable that some useful information would be lost after a Kmeans cluster is divided into few DSs. Yet highly tight correlated patterns are the purpose of most clustering approaches and they are also the base of subsequently further research, for example, regulatory elements finding.

In summary, with the increment of correlation coefficient threshold r_{ts}, the sensitivity and quality of dominant set is increasing, but specificity is decreasing. In addition, with the increment of significant level alpha, the sensitivity and quality of dominant set is decreasing, but specificity is increasing. Noted that in the condition of alpha=0.99, the magnitude of difference in sensitivity, specificity and MinR is a bit big. However, in other three conditions of alpha, they are only slightly changed. Based on the rules and practical instances, the parameters r_{ts} and alpha can be chosen in DSF_Clust approach.

4.2. Performance on yeast cell cycle

We applied our approach on a published data set in http://cellcycle-www.stanford.edu, which, from Spellman *et al*, consists of yeast Microarray experiments from varying conditions (12). Here we chose to use only one single time course, where gene expressions are measured in a single process of cell division cycles following alpha-factor block-and release. Moreover, we just focused on genes that have known functions characterized in data for easier interpretation and genes without missing values. After gene selection, only 1914 yeast gene expression profiles taken at 18 time intervals during two cell division cycles, synchronized by alpha arrest and release, were clustered.

It is common to normalize gene expression vectors before cluster analysis. Here we normalize the expression profiles so that, of each gene, the mean is 0 and the variance is 1. The parameters $alpha = 0.95$ and $r_{ts} = 0.8$ were used to find dominant sets from normalized data.

4.2.1. Finding co-regulated dominant sets

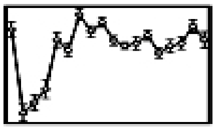
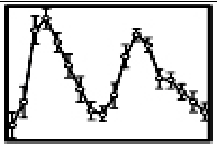
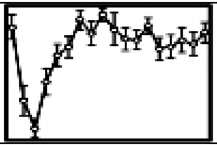
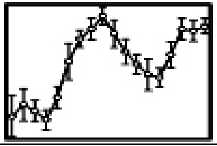
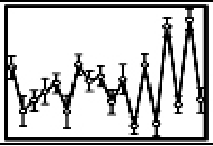
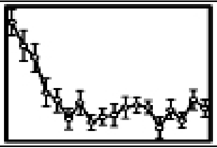
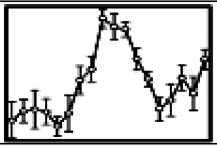
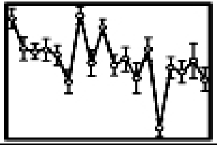
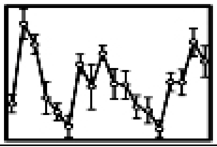
After running DSF_Clust, 1259 genes in total were grouped into 153 dominant sets. We mapped the genes in DSs to the functional categories in the MIPS database, http://mips.gsf.de/proj/yeast/catalogues/funcat/ version from 06.12.2001. Using the hypergeometric probability distribution, P values were calculated to associate DSs with each functional category (Table 2). Lower a P value (higher a –log10 p value) is, tighter the association is. It is likely that most of the ORFs belonging to these enriched functional categories are biologically significant members of the corresponding DSs. Here only 9 DSs were listed in Table 2. To see all DSs whose P values are above E-04, you can visit the website http://www.chinagenenet.com/DSF_Clust.

By using web-based tool FunSpec for rapid interpretation of yeast gene clusters (17), we found that many dominant sets are identified which have expression profiles close to previous work (12,18,19). Since the gene expression profiles cover about two full cell cycles, we firstly care about the cell cycle related dominant sets. Genes in DS3 are strongly cell cycle regulated and peak expression occurs in mid-G1 phase. This DS is corresponding to “CLN2” cluster described by Spellman *et al* (12). Most of the genes in this DS involve in DNA synthesis, replication, recombination and repair. The major cell cycle regulators: CLN1, CLN2, CLB5, CLB6 and SWI4 are contained in this DS. Many genes in this DS15 are included in “CLB2” cluster described by Spellman *et al*, where the genes are highly regulated with a peak in M phase. From the expression peak and genes’ function, it can be suggested that genes in this DS may play a role in budding and cell polarity. For the cell cycle was synchronized by alpha arrest and release, a group of genes directly involved in mating pheromone (alpha factor) would be found. DS21 is about pheromone response, some of which are known to be cell cycle regulated. It is suggested that some alpha specific genes are induced by transcription factor MAT alpha 1 binding to DNA cooperation with another factor MCM1. Most of genes in DS35 are involved in phosphate utilization. The cell-regulated genes are involved in transport of essential minerals and organic compounds across the cell membrane. They reach peak expression late in cell cycle during M and M/G1 phase. Most of genes in DS1 and DS9 participate in RNA processing. The genes include the RNA related enzymes or complexes, such as RNA polymerase, RNA processing complexes and DNA directed RNA polymerases. These genes are involved in tRNA or rRNA related processes, such as ribosome biogenesis and assembly, rRNA metabolism, rRNA modification, rRNA tRNA synthesis and transcription. In addition, we found many ribosome biogenesis dominant sets, in which the p-value of DS16 is smallest. These genes are mostly ribosome protein genes. A majority of the genes are functionally characterized by structural protein of ribosome, which means they are sub-components of ribosome machinery. However, some dominant sets are not shown in any reports and p values of mapping to functional categories reveal that these clusters may be significant, such as clusters of cytoplasmic and nuclear degradation (DS42) and purine ribonucleotide metabolism (DS124). These dominant sets can be done further research in their biological processes and activities.

4.2.3. Search for putative regulatory elements

It is assumed that genes with similar expression profiles, i.e. genes that are coexpressed may share something common in their regulatory mechanisms, i.e may be coregulated. Therefore, by clustering together genes with similar expression profiles, we find groups of potentially coregulated genes to search for putative regulatory signals (20). Searching the yeast promoter database SCPD (21), we found that most of the yeast cell cycle dominant sets described in our paper have their special regulatory elements (Table 3). DS3 is a dominant set involved in DNA synthesis, replication, recombination and repair. MBF and SBF are the two most direct cell cycle regulators. Their binding sites MCB (AGCGCGT) and SCB (CRCGAAA) can easily detected in these genes’ promoters. 25 genes of 29 in this DS contain MCB element, and SCB motif can be found in the upstream region of 10 genes. DS15 is supposed to be MCM1 regulated dominant sets. It is suggested that the majority of these genes contain either MCM1 site or the composite site MCM1+SFF and some genes are regulated through ECB element, a variant of the MCM1 site. We detected the motifs in the upstream region of these genes and found that 100 percent of these 15 genes contain CNNNWWRG element, which is very similar to MCM1 site. DS35 is characterized with

Table 2. Enrichment of DSs for ORFs with functional categories

DS		MinR	MIPS Functional category	N_{DS}	N_{FC}	D_P
1		0.90	rRNA processing	13	6	7.44E-08
3		0.80	DNA synthesis and replication DNA recombination and DNA repair	29 29	14 6	3.00E-16 3.25E-07
9		0.82	rRNA processing	29	10	5.50E-11
15		0.80	cytokinesis (cell division)	15	4	4.22E-06
16		0.80	ribosome biogenesis	20	18	8.13E-22
21		0.80	pheromone response, mating-type determination, sex-specific proteins	17	7	4.49E-07
35		0.78	phosphate utilization	11	4	2.86E-07
42		0.80	cytoplasmic and nuclear degradation	10	5	3.36E-06
124		0.78	purine ribonucleotide metabolism	7	4	6.51E-07

DS: the name of a dominant set. N_{DS}: Number of ORFs in a DS. MinR : the minimal correlation coefficient between the members of a DS and the DS center. N_{FC}: ORFs within Functional category in a DS. D_P: p-value, which shows the degree of enrichment of a cluster for ORFs within a particular functional category. If we noted respectively N_k and M_k the total number of ORFs in the clustering algorithm and the number of ORFs within a functional category, and the number of ORFs in the dominant set DS_k is noted n_k while the number of ORFs in DS_k within the functional category is noted m_k, then we can use the hypergeometric probability distribution to compute the probability(D_P, C_P) associated to each functional categoty, $p-value=\sum_{i=m_k}^{M_k}\left[\binom{M_k}{i}\binom{N_k-M_k}{n_k-i}\right]\Big/\binom{N_k}{n_k}$.

phosphate utilization then naturally the element of phosphate utilization, PHO4, are the leading motif of this dominant set. As expected, 8 genes of 11 contain this PHO4 motif. In addition, about 80 percent (14 genes in 18) of DS33 had the motif AGGNG. DS33 is involved mitochondrion metabolism, therefore we infer that the

Table 3. The elements of the DSs.

DS	Consensus	N_E	N_{DS}
DS1	GCGATGAG	6	13
DS3	WCGCGW (MCB)	25	29
	CNCGAAA (SCB)	10	29
DS15	CNNNWWRG	15	15
DS35	CACGT (PHO4)	8	11
DS33	AGGNG	14	18

N_E is the number of genes containing the element in this DS. N_{DS} is the number of the genes in the DS.

AGGNG motif may be functional important in the process of mitochondrion metabolism.

In searching for regulatory sites using Microarray data, the selection of appropriate clustering method is a challenge. Using common clustering methods, many unrelated profiles will be grouped into one cluster. Then the confusion or "noise" would be introduced. When using DSF_Clust approach with appropriate threshold, the resulted dominant sets will exhibit a tightly related expression profiles. It is effective to consider genes in a dominant sets as candidates for extraction of regulatory motifs. For example, in the above listed yeast cell cycle dominant sets, DS1 and DS9 are all tightly associated with rRNA processing and their profiles are very similar. In common clustering method, these two DSs will be merged into one cluster. However, the value of MinR of DS1 is much bigger than the correlation coefficient threshold r_{ts}. It means that if we assume all the expression data form a data space, the expressions in DS1 are close enough to each other that the DS1 can be separated out in significant level and the compulsory threshold r_{ts} is not needed. Moreover, when we search for their common regulatory elements, the motif GCGATGAG emerged in about fifty percent genes in DS1, however, only less than twenty percent in DS9. We supposed that this motif might be the key to result in essential separation of DS1 and DS9. It is anticipated that in future work DSF_Clust will be a useful tool to understand the regulatory mechanism.

5. SUMMARY

In this paper, we present a new approach named DSF_Clust to cluster Microarray data. The algorithm is designed to find clusters of significantly coexpressed genes in high-density areas of the data. It is an interesting new approach for some important issues in clustering are tried to deal with. Firstly, there is a significance level attached to a dominant set. Secondly, no pre specific cluster number is required. Number of dominant sets is automated determined based on the expression data used in clustering. Genes not exhibiting an expression profile significantly similar to the expression profile of other genes in the data set are not assigned to any one of the dominant sets. Thirdly, if we consider the closeness between the elements in one dominant sets as the criteria to evaluate the quality of a dominant set, then every dominant set is of good quality with high threshold r_{ts}. Thus this approach is a blend of unsupervised analysis and supervised analysis of Microarray expression data. We have shown that the unrelated patterns in a cluster produced by Kmeans approach will not grouped in any dominant sets produced by DSF_Clust. A "noisy" Kmeans cluster may be divided into few dominant set where elements are tightly close to each other in expression. When we used the DSF_Clust approach to cluster the published yeast cell-cycle Microarray data, a lot of dominant sets similar to clusters reported in other literatures are found. Some dominant sets of coregulated genes are easy to find putative regulatory elements. Our DSF_Clust approach is could be a helpful tool for finding motifs.

In our DSF_Clust approach, there are two key user-defined parameters, significance level alpha and Pearson correlation coefficient threshold r_{ts}. Definition of alpha has the advantage that a dominant set has a strict statistical meaning. The threshold r_{ts} ensures the qualities of the clustered dominant sets. When a dominant set is well-separated in data space, significance level alpha plays a key role to find the dominant set. However, in many cases, a dominant set is not separated well from other data in the level of significance alpha, then the threshold acts as " a guard" to prevent unrelated members joining in the cluster. When alpha increases, the constraint t-value t_{alpha} to separate a gene from a dominant set is getting small. Then a dominant set is easy to build. In addition, if correlation coefficient threshold r_{ts} is high, the members in a dominant set must have tight affinity. However, that one relative loose dominant set divides into few tight ones would happen.

Sometimes the results of our approach have a bit of disagreement with external biological knowledge. For the biological system is very complex, sometimes genes with different annotated functions or involved in different cellular process might have similar expression patterns. Then in these cases it is not enough to partition genes into disjoint clusters. The main purpose of our approach is to obtain dominant sets with tightly close patterns. In most cases, we can obtain detailed and meaningful groups. But at the same time, it is probable that two patterns with associated biological function are grouped into different dominant sets because their expression patterns have no enough affinity. Thus some information is lost. We should adjust the parameters of significant level and threshold according to the importance of the affinities between one dominant set and the lost information. In addition, our approach is not fast when the number of clustered patterns is very big. The process of iterations is time-consuming so that it may take much time to finding all dominant sets in real data. This problem will be addressed to optimize the algorithm in the future.

6. ACKNOWLEDGEMENT

This research is supported by a grant 2002AA2Z2002 from the National High Technology Research and Development Program of China (863 Program) and this research is also a part of project 60473104 and 30371422 supported by the National Natural Science Foundation of China.

7. REFERENCES

1. Eisen M. B, P. T. Spellman, P. O. Brown & D. Botstein: Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* 95, 14863-14868 (1998)
2. Tamayo P, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander & T. R. Golub: Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A* 96, 2907-12 (1999)
3. Tavazoie S, J. D. Hughes, M. J. Campbell, R. J. Cho & G. M. Church: Systematic determination of genetic network architecture. *Nat Genet* 22, 281-5 (1999)
4. Quackenbush J: Computational analysis of microarray data. *Nat Rev Genet* 2, 418-27 (2001)
5. Alter O, P. O. Brown & D. Botstein: Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A* 97, 10101-10106 (2000)
6. De Smet F, J. Mathys, K. Marchal, G. Thijs, B. De Moor & Y. Moreau: Adaptive quality-based clustering of gene expression profiles. *Bioinformatics* 18, 735-746 (2002)
7. Zahn C. T: Graph-theoretic methods for detecting and describing restalt clusters. *IEEE Trans.Comput* 68-86 (1971)
8. Ben-Dor A, R. Shamir & Z. Yakhini: Clustering gene expression patterns. *J Comput Biol* 6, 281-97 (1999)
9. Xing E. P & R. M. Karp: CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* 17 Suppl 1, S315 (2001)
10. Pavan, M & M. Pelillo: A new graph-theoretic approach to clustering and segmentation. *In CVPR'03 I* 145-152 (2003)
11. Motzkin T.S & E. G. Straus: Maxima for graphs and a new proof of a theorem of Turan. *Canad J Math* 17, 535-540 (1965)
12. Spellman P.T, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, D. Botstein & B. Futcher: Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* 9, 3297 (1998)
13. Ji X. L, J. Li_ling & Z. R. Sun.: Mining gene expression data using a novel approach based on hidden Markov medels. *FEBS Letters* 542, 125-131 (2003)
14. Iyer V. R, M. B. Eisen, D. T. Ross, G. Schuler, T. Moore, J. C. Lee, J. M. Trent, L. M. Staudt, J. Hudson Jr, 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, 83-7 (1999)
15. Xu Y, V. Olman, D. Xu: Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics* 18, 536-45 (2002)
16. Cho R. J, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart & R. W. Davis: A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2, 65-73 (1998)
17. Robinson M. D, J. Grigull, N. Mohammad & T. R. Hughes: FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics* 3, 35 (2002) [The software is available at http://funspec.med.utoronto.ca]
18. Getz G, E. Levine, E. Domany & M. Q. Zhang: Super-paramagnetic clustering of yeast gene expression profiles. *Physics A* 279, 457-464 (2000)
19. Tavazoie S, J. D. Hughes, M. J. Campbell, R. J. Cho & G. M. Church: Systematic determination of genetic network architecture. *Nat Genet* 22, 281-5 (1999)
20. Heyer L. J, S. Kruglyak & S. Yooseph: Exploring expression data: identification and analysis of coexpressed genes. *Genome Res* 9, 1106-1115 (1999)
21. Zhu J & M. Q. Zhang. SCPD: a promoter database of the yeast Saccharomyces cerevisiae. *Bioinformatics* 15, 607-611 (1999).

Abbreviations: CLN: Cyclin; CLB: B-type cyclin; SWI: Switching protein; MAT alpha: Mating type alpha specific genes; MCM: Mini Chromosome Maintenance; MBF: MCB binding factor; SBF: SCB binding factor; MCB: MluI cell cycle box; SCB: Swi4/Swi6 cell cycle box; SFF: Swi five factor;

Key words: Microarray data, dominant sets, cluster, correlation coefficient, significance level

Send correspondence to: Yi Xie, Institute of Genetics, School of Life Science, Fudan University, Shanghai, PR China, Tel: 86-021-65989936, Fax: 86-021-65985919, E-mail: yxie@fudan.edu.cn and I-fan Shen, Department of Computer Science and Engineering, Fudan University, Shanghai, PR China, Tel: 86-021-55664505, Fax: 86-021-55664506, E-mail: yfshen@fudan.ac.cn

http://www.bioscience.org/current/vol10.htm