Pacific Symposium on Biocomputing 5:476-487 (2000)

CLUSTER, FUNCTION AND PROMOTER: ANALYSIS OF YEAST EXPRESSION ARRAY

J. ZHU, M. Q. ZHANG

Cold Spring Harbor Lab, P. O. Box 100
Cold Spring Harbor, NY 11724

Gene clusters could be derived based on expression profiles, function categorization and promoter regions. To obtain thorough understanding of gene expression and regulation, the three aspects should be combined in an organic way. In this study, we explored the possible ways to analyze the large-scale gene expression data. Three approaches were used to analyze yeast temporal expression data: 1) start from clustering on the expression profiles followed by function categorization and promoter analysis, 2) start from function categorization followed by clustering on expression profiles and promoter analysis, and 3) start from clustering on the promoter region followed by clustering on expression profiles. For clustering analysis on the time-series data, we developed a largest-first algorithm, which provide a mechanism for quality control on clusters. For promoter analysis, we developed a core-extension algorithm.

Introduction

DNA chip technology enables the study of gene expression in a large scale. Yeast has been a popular model system for large-scale gene expression studies. Its advantages over other organisms include availability of the entire genomic sequence, relatively small size of genome and a large body of biomedical and genetic information. Up to now, half of its 6200 ORFs have been characterized (1). In addition, the promoter regions of more than 200 genes have been investigated in detail and more than 40 regulatory elements are well defined. With the available information, it is possible to combine function categorization and promoter analysis with the large-scale gene expression study to gain in-depth view on the genome structure and gene regulation.

Large-scale gene expression experiments are used to determine drug targets, identify co-regulated genes (2, 3, 4) and study the response to environmental conditions (4, 5) and the effect of a single gene on the entire genome (5, 7). Genes respond to experimental conditions in a dynamic way. Temporal expression experiments are used to capture the dynamic nature of gene expression. In such experiments, the expression data are collected over a time course. Examples of temporal experiments can be found in the studies of diauxic shift (5), sporulation (2) and cell cycles (3).

Co-regulated genes may share similar expression profiles, may be involved in related functions or regulated by common regulatory elements. There are different approaches to analyzing the large-scale gene expression data. The essence

is to identify gene clusters. For example, one can start from clustering on the expression profiles. For genes with similar expression patterns, identify their functions and putative regulatory elements in their promoter region. The study of the promoter region helps to understand gene co-regulation on the transcription level. Function categorization provides information on protein interaction and pathways. As an alternative, one can start from function categorization. For genes with related functions, study their expression patterns. Gene expression and regulation are complex biological processes. Genes involved in the same pathway or related functions unnecessarily have same expression patterns. It is important to understand what expression patterns are associated with a specific function. Finally, one can start from clustering genes based on their promoters. Genes carrying putative sites of a transcription factor may be related or unrelated in the transcription level. Few studies of regulatory elements on the genome level have been performed (8). Clustering study on genes sharing regulatory elements may provide clue on issues such as on what conditions those elements are active, their roles in activation and repression and their interactions with each other. Since each approach focuses on different aspect of the genome, they are equally important.

We tested the above three approaches using data from the sporulation experiment (2), which contains seven time steps. Before introduce the three approaches methods for clustering on expression profiles, function categorization and promoter analysis are described.

Methods

Largest-first clustering algorithm.

Clustering methods are essential to the discovery of expression patterns in the time series data. Two clustering methods are mainly used for this purpose. One is the classic hierarchical clustering method (9). The other is self-organizing map, SOM (10, 11). Both methods provide overviews on the entire data set. Researchers are also interested in clusters with special features, for example clusters with the similarity between each pair of genes being higher than a certain cutoff. To address such problem, we developed an algorithm based on the density search method. Due to the fact that the largest cluster always came out first, it was termed 'largest-first'. Using this approach, clusters could be derived based on their quality requirements.

For this algorithm, a data point corresponded to the measurement of temporal expression of a gene, which was shown as a multi-dimensional vector. The similarity between two data points was measured by the Pearson correlation, which has a range from -1 to 1. Prior to clustering, a similarity matrix was generated. At the beginning, all data points were placed in an initial data point pool, which was a collection of unclustered data points. The algorithm underwent iterations. One iteration generated one cluster. Clustered data points withdrew from the initial pool,

unclustered data points remained in the pool. In each round and for each data point in the data point pool, its density of neighbors was determined by calculating the number of data points with similarity to the selected data point above a certain cutoff, which was a parameter given by the user. The data point with the most number of neighbors was selected as the core for a new cluster. The next step was to grow the new cluster. For a data point to join a cluster, its similarity to the cluster should be greater than the cutoff. Once a cluster was updated, cluster members were screened against each other, the ones failed to maintain the same criterion withdrew from the cluster and returned to the pool. The algorithm stopped until no more cluster could be found.

The major advantage of this algorithm was to provide the mechanism for controlling the quality of clusters. The quality of an output cluster could be controlled by the similarity cutoff. The higher the cutoff the higher the quality. The algorithm was implemented in C.

We tested the largest-first algorithm using random sample which contained 1000 data points with 7 time steps. With the similarity cutoff = 0.9, the largest cluster found contained only three data points. With the similarity cutoff = 0.8, the size of largest cluster is 12. For a given data set, tests on the random data of the same dimensions are needed to determine the minimal size of a cluster to be considered significant. Besides providing a mechanism to control the quality of clusters, the largest-first algorithm provide a way to test the significance of a cluster, basically by controlling the size of a cluster. From the random test, the maximum expected size of clusters could be determined. For a cluster to be significant, its size should be several times larger than the maximum expected size.

Function categorization.

MIPS provides a function catalogue of all yeast genes. Among 6350 genes and ORFs documented in MIPS, 3529 are assigned to at least one function category. 12 major function categories are selected. Refer to table 1 and 2 for their name and number of consisting ORFs.

For an identified cluster, two parameters were needed to decide which function category was most related. The first parameter, designated as N, was the number of genes within a function category. The second one designated as -ln P, was the negative natural logarithm of the probability for finding N genes in a function category. The frequency of a function category was determined by dividing the size of the function category by total number of genes. The expected number of genes belonging to a function category (E) was equal to the cluster's size times the function category's frequency. The probability was computed assuming a Poisson distribution. The most related function category was the one with largest -ln P value and N much greater than E. As an alternative Z-scores could be used to determine the most significant function categories.

Promoter analysis.

The availability of genomic sequence of yeast enables the promoter analysis in the context of clusters. Table 1 shows transcription factors and regulatory elements provided by SCPD (12). Consensus sequences and their functions are also given. They are divided into two groups, uni-core and multiple-core according to the nature of recognition sites.

Table 1. Known transcription factors' binding sites and regulatory elements

Name	Consensus sequence	Related function
Uni-core		
ADR1	TCTTC	Carbohydrate utilization
GCN4	TGACTC	Amino-acid metabolism
GCR1	GMWTCCW	Carbohydrate utilization
GLN3	GATAAG	Nitrogen utilization
HAP2	ACCAATNA	CCAAT-binding factor
INO2	ATGTGAAWW	Lipid biosynthesis
MAC1	TTTGCTC	Reduction and utilization Fe, Cu
MATA1	TGATGTWR	Repress haploid genes in diploid cells
MATα1	WCAAYGNCAG	Activates alpha-specific genes
MATα2	TCNTGT	turn off a-specific genes
MCB	WCGCGW	Cell cycle control
MIG1	CCCCRSWWWWW	Glucose-repression
MSE	CACAAAA	Middle sporulation element
PDR1	TCCGYGGA	Detoxification
PHO4	CACGTK	Phosphate utilization
RAP1	RMACCCA	Transcription control
REB1	CCGGGTARNNR	Transcription control
RME1	GAACCTCAA	Meiosis and mitosis
ROX1	YYNATTGTTY	Repressor of hypoxic genes
SCB	CNCGAAA	Cell cycle control
STE12	ATGAAAC	Mating
SWI5	WACCAKY	Cell cycle control
UME6	WCGGCGGCWA	Nitrogen repression and induction of meiosis
YAP1	TTACTAA	Oxidative stress response
TBP	TATAWAW	TATA binding protein
SFF	GTMAACAA	Swi five factor, function with MCM1
ECB	GGAAAAD	Early cell cycle box
STRE	AGGGG	Stress response element
Multi-core		
ABF1	TCRNNNNNNACG	DNA-replication and transcriptional regulation
HAP1	CGGNNNTANNCGG	Heme-dependent activation
GAL4	CGGNNNNNNNNNNNCCG	Galactose-induction
LEU3	CCGGNNCCGG	Branched chain amino acid biosynthesis pathways
MCM1	DCCNNNWWRGG	Recruits coregulatory proteins for both gene activation and repression at a variety of loci
PUT3	CGGNNNNNNNNNNCCG	Proline utilization pathway
PPR1	CGGNNNNNNCCG	Regulating pyrimidine pathway

Uni-core motifs contain only one core region to be recognized by transcription factors.

Multi-core motifs contain several core regions.

The background frequency of a promoter element was determined using the promoter region of all yeast ORFs. The promoter region ranged from -500 to -1 regarding to the start of coding region. For an identified cluster, the promoter regions of all cluster genes were gathered as well. The putative sites of each promoter element were determined by searching through the promoter region for matches to the consensus sequences. The cluster frequency of a promoter element was determined by dividing the number of putative sites by the number of genes in the cluster. The expected number of putative sites could be estimated based on the background frequency. The probability (P) of finding certain number of putative sites could be calculated assuming a Poisson distribution. The significant putative motifs were those having a cluster frequency much higher than the background one and large -ln P. Using this method, non-specific signals such as TATA box and ploy(A) could be easily filtered.

There is also great interest in identifying unknown promoter elements. Multiple sequence alignments are commonly used to find common motifs in the promoter regions. Such examples including Gibbs sampler (13), Consensus (14) and MEME (15). Prior information on motif length and motif distribution is required for these approaches. Here, we present an algorithm called 'core-extension', which is based on k-tuple analysis of the promoter region, and requires no assumption on motif length and distribution. It is similar to approach described in (16).

Three types of k-tuples were employed. 5-tuple was a 5-mer with no mismatch. Degenerate 5-tuple was a 5-mer with one mismatch at position 2, 3 and 4, e.g. TNACT, TGNCT, and TGANT. N represented {A, T, C, G}. Degenerate 6-tuple was a 6 mer with one mismatch at position 2, 3, 4, 5.

The core-extension algorithm first selected significant k-tuples from the promoter region of a gene cluster. The selection procedure was similar to that for selecting known promoter elements. To demonstrate the procedure, 5-tuple was used as an example. The distributions of 5-tuples in the promoter region of all ORFs were used as a control. Over-represented 5-mers in a gene cluster were selected based on their Z-scores. The selected 5-mers were used as cores for sequence motifs. The initial motif was assumed to have the 5-tuple in the middle with extensions of 5 nucleotides on both sides. All sequences matching this motif were selected. A matrix was built upon. Each cell of the matrix contained the frequency count of a nucleotide at the corresponding position. For each position, a position score was computed as the standard deviation of nucleotide counts in all rows divided by 4. The score ranged from 0 to 1. It was a measurement on how conserved a position was. A score close to 1 indicated one nucleotide was dominant at that position. To validate a matrix, its ends were checked. If a position at either end of a matrix had a score lower than 0.3, it was dropped off from the matrix. The validation procedure stopped until both end positions had scores greater than 0.3. The matrix was used to select new putative sites using a cutoff. New putative sites were used to update the matrix, and the matrix was validated and applied to the next

round search. Updating and searching stopped until the matrix reached a stable stage. A consensus sequence was generated based on the final matrix output. The core-extension algorithm was suitable for finding motifs containing a single core. It could find conserved flanking regions around the core and incorporate mismatches in the core region. Other advantages include no requirement of knowledge on motif length and distribution. Degenerate 5-tuples and 6-tuples could be used to find motifs with less conserved cores.

Approach 1: start from expression profiles

For this approach, gene clusters were derived based on expression profiles. For each cluster, the most significant function category was determined and promoter elements were identified.

The sporulation experiment (2) contained seven time steps. The original data set included measurement on more than 6000 ORFs. Among them, 1870 represented characterized genes and contained no null time point data.

Using the largest-first clustering method, with similarity cutoff equal to 0.9 and cluster size larger than 30, 8 clusters were identified. The average expression profiles of each cluster are shown in figure 1. Cluster 1, 3, 4, 5 and 6 consisted of genes mainly repressed during sporulation. Other clusters contained induced genes. Most of them reached the highest expression level in the middle of sporulation. Table 2 summarizes the results from function categorization. The significance measurements of each function category were given in table 3. Table 4 shows the most significant function category for each cluster and identified putative regulatory elements. Genes in cluster 2 are induced in the middle of sporulation, most of them carried MSE, the middle sporulation element, in their promoter region. This is consistent with previous studies (17). Function categorization indicated that 30% of them were related to cell growth. In cluster 4, 72 genes were related to protein synthesis, 67 of them were ribosomal proteins. Putative RAP1 sites were identified in the promoter regions for most of them. RAP1 repressed the expression of ribosomal protein during sporulation (18). For clusters with similar profiles, their related functions and regulatory elements may be totally different. For example, cluster 2 and 8 all contained genes induced in the middle of sporulation. The dominant function for cluster 2 was cell growth. No major function was found for cluster 8. Their major promoter elements were also different, MSE for cluster 2 versus CCCCC for cluster 8. To understand the complexity of gene expression, one has to combine the analysis of expression pattern, function categorization and promoter analysis together.

Approach 2: start from function categorization

This approach was used to identify different expression profile associated with a function category. Expression data were first sorted by function catalogue prior to clustering. Clustering method was the largest-first described above. Promoter analysis was performed on identified clusters.

Using the same sporulation data described above, expression data were broken into 12 function categories. Here, we used genes related to cell growth as an example to demonstrate the procedure.

Among the 1870 genes selected from the sporulation data, 685 genes were related to cell growth. The result from the largest-first clustering on these genes' expression profiles is given in Figure 2. 11 clusters were identified with similarity cutoff = 0.8 and size greater than 10. Size of each cluster and identified putative regulatory elements are given in table 5.

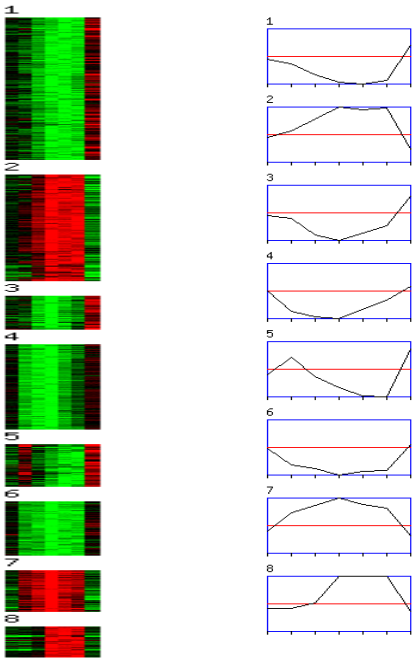

Figure 1. Clustering on the expression profiles using the largest-first algorithm. The similarity cutoff is 0.9. The minimal size of a cluster is 30.

Table 2. Function categorization of clusters in Figure 1

Cluster	1	2	3	4	5	6	7	8
Total number of genes	147	110	35	88	44	56	43	32
Metabolism(1047)*	50	33	18	6	14	22	11	11
Energy(246)	25	1	3	1	2	7	1	1
Cell growth(785)	27	38	4	3	6	7	16	6
Transcription(742)	18	23	5	5	9	11	9	6
Protein synthesis(346)	9	6	1	72	1	5	2	3
Protein destination(538)	17	17	7	9	8	9	8	5
Transportation facilitation(303)	13	11	2	0	4	6	4	4
Intracellular transport(450)	32	15	3	3	8	8	9	7
Cellular biogenesis(188)	7	8	2	1	2	4	3	3
Signal transduction(126)	3	1	0	1	0	0	0	0
Cell rescue, defense, death, aging(352)	18	8	3	1	1	7	5	2
Ionic homeostasis(121)	12	1	1	0	3	2	0	1

*The number in parentheses shows the number of genes related to that function.

Table 3. Significance measurement (-ln P) of function categories for clusters in Figure 1

Cluster	1	2	3	4	5	6	7	8
Metabolism(1047)	4.2	3.3	4.8	13	2.5	3.6	2.5	2.4
Energy(246)	10	5.7	1.6	4.4	1.6	2.9	1.9	1.4
Cell growth(785)	3.7	6.3	2.9	13	2.8	3.5	4.2	2.0
Transcription(742)	6.1	2.9	2.3	8.9	2.2	2.4	2.2	2.0
Protein synthesis(346)	3.4	3.3	2.2	90	2.9	1.8	2.1	1.5
Protein destination(538)	3.5	2.7	2.2	3.1	2.2	2.2	2.2	1.8
Transportation facilitation(303)	2.4	2.4	1.5		1.7	2.0	1.7	1.9
Intracellular transport(450)	6.8	2.6	1.8	5.8	2.5	2.1	3.0	2.8
Cellular biogenesis(188)	2.1	2.4	1.3	3.2	1.4	1.8	1.6	1.9
Signal transduction(126)	2.1	2.6		2.0				
Cell rescue, defense, death, aging(352)	3.0	2.6	1.6	6.7	2.9	2.1	1.9	1.6
Ionic homeostasis(121)	5.6	2.5	1.0		2.0	1.3		1.0

P is the probability of find N genes (as shown table2) belonging to a function category in a cluster.

Cluster 1, 4, 6 had different expression profiles but they seemed to be controlled by the same transcription factors MCB, with consensus ACGCGT. Cluster 1 and 6 were mainly repressed during sporulation, while cluster 4 was induced in the later stage. It indicated that MCB might function differently on different gene cluster. Cluster 2 also consisted of repressed genes, however, most of them carried sequences similar to REB1's binding sites. Compared to cluster 6, cluster 2 had a lagged repression. This might be due to the difference between functions of MCB and REB1. Cluster 7 genes were induced in the middle of sporulation. Instead of MSE, STRE was dominant in their promoter region. STRE

was the dominant in the promoter region of cluster 1 in figure 1, which was repressed during sporulation. The same element was involved in repression as well as induction under the same condition. The results indicted a transcription factor may play different roles on different gene clusters.

Approach 3: start from promoter

It is important to find the relationship between promoter elements and expression profiles. Approach 1 presents a way to identify putative promoter elements for a given expression pattern. It is only one aspect of the problem. Another aspect is to identify expression patterns for a promoter element. For this approach, genes containing certain motifs were selected first. Then clustering was performed.

MSE was identified as the major regulatory element for cluster 2 in Figure 1, which was induced in the middle of sporulation. Among 1870 genes selected from the sporulation data, 528 genes contained MSE (CACAAAA) in their promoter region.

Figure 3 shows the clustering result on MSE containing genes from the largest-first algorithm. The similarity cutoff was 0.8. Besides clusters induced in the middle of sporulation, there were cluster corresponding to induction in the early and late stage, and those repressed at various stages. It indicated MSE alone was not enough to determine the expression pattern. It might function through interactions with other elements. For example, cluster 5 was induced in the early middle of sporulation. Promoter analysis showed cluster 5 was also rich in STRE (AGGGG) besides MSE.

Conclusion

Clustering analysis, function categorization and promoter analysis help to gain thorough overviews of the expression data. There is no logic order to define which analysis should come first. Currently, the most commonly used approach is the approach 1 described in the text. However, other approaches also provide useful information. No single method is good enough. It is important to combine different approaches together.

Acknowledgement

This work was supported by NIH grant HG01696 to M. Q. Zhang.

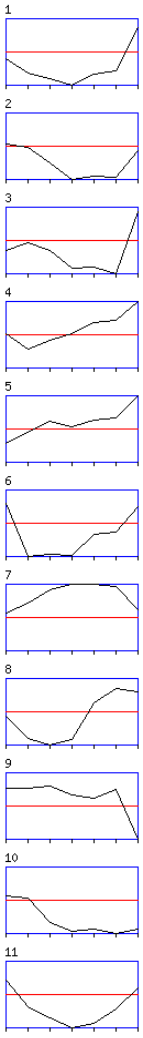

Figure 2. Clustering on expression profiles of 685 cell growth related genes using largest-first algorithm. Similarity cutoff is 0.8.

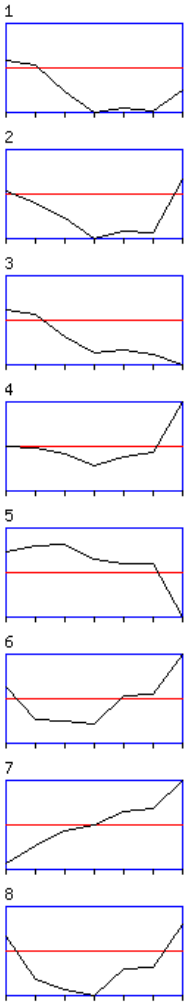

Figure 3. Clustering on expression profiles of 528 MSE containing genes using largest-first algorithm. Similarity cutoff is 0.8.

Table 4. Functions and regulatory elements for clusters in Figure 1

Cluster	Function*	Regulatory elements
1	Energy	AGGGG (STRE)
2	Cell growth	CACAAAA (MSE)
3	Metabolism	TNCCACAC
4	Protein synthesis	ACCCATACAT (RAP1)
5		
6	Metabolism	GCGCAAAA
7	Cell growth	AGGCGCCT
8		CCCCC

*The most significant function was determined by combining information in Table 1 and 2.

Table 5. Size and promoter elements for clusters in Figure 2

Cluster	Size	Promoter elements
1	100	ACGCGT (MCB)
2	89	TTACCCG (REB1)
3	30	
4	38	ACGCGW (MCB)
5	21	
6	24	ACGCGT (MCB)
7	23	AGGGG (STRE)
8	18	
9	16	
10	14	
11	10	

Reference:

1. J.M. Cherry, C Adler, C Ball, S.A. Chervitz, S.S. Dwight, E.T. Hester, Y. Jia, G. Juvik, T. Roe, M. Schroeder, S. Weng, D. Botstein. Nucleic Acids Res 1998, 26:73-79
2. S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P.O. Brown, I. Herskowitz. Science 1998, 282:699-705
3. P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, B. Futcher. Mol Biol Cell 1998, 9:3273-3297
4. F.P. Roth, J.D. Hughes, P.W. Estep, G.M. Church, Nat Biotechnol 1998, 16:939-945
5. J.L. DeRisi, V.R. Iyer, P.O. Brown. Science 1997, 278:680-686
6. E.G. Jennings, R.A. Young. Trends Genet 1999, 15:202-204
7. D.A. Pearce, T. Ferea, S.A. Nosel, B. Das, F. Sherman. Nat Genet 1999, 22:55-58
8. M.Q. Zhang. 1999, Genome Res, in press.
9. M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein. Proc Natl Acad Sci U S A 1998, 95:14863-14868
10. P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E.S. Lander, T. Golub. Proc Natl Acad Sci U S A 1999, 96:2907-2912
11. P. Toronen, M. Kolehmainen, G. Wong, E. Castren. FEBS Lett. 1999, 451:142-146
12. J. Zhu, M.Q. Zhang. Bioinformatics. In press
13. C.E. Lawrence, S.F. Altschul, M.S. Boguski, J.S. Liu, A.F. Neuwald, J.C. Wootton. Science 1993, 262:208-214
14. G.Z. Hertz, G.W. Hartzell, G.D. Stormo. Comput Appl Biosci. 1990, 6:81-92
15. M.Q. Zhang. 1999, Comp & Chem. 23:233-250.
16. T.L. Bailey, C. Elkan. Ismb 1995, 3:21-29.
17. N. Ozsarac, M.J. Straffon, H.E. Dalton, I.W. Dawes. Mol Cell Biol 1997, 17:1152-1159
18. K. Mizuta, R. Tsujii, J.R. Warner, M. Nishiyama. Nucleic Acids Res 1998, 26:1063-1069