한국생물공학회지 제16권 제6호
*Korean J. Biotechnol. Bioeng.*
Vol. 16, No. 6, 538-546(2001)

# DNA chip 통합분석 프로그램을 이용한 효모의 세포주기 유전자 발현 통합 데이터의 분석

양 영 렬 · †허 철 구
한국생명공학연구원 국가유전체 정보센타
(접수 : 2001. 10. 20., 게재승인 : 2001. 12. 14.)

# Analysis of Combined Yeast Cell Cycle Data by Using the Integrated Analysis Program for DNA chip

Young Lyeol Yang and Cheol Goo Hur†
National Center for Genome Information, Korea Research Institute of Bioscience and Biotechnology, 52 Eoun-dong, Yusong-gu, Daejon 305-333, Korea
(Received : 2001. 10. 20., Accepted : 2001. 12. 14.)

An integrated data analysis program for DNA chip containing normalization, FOM analysis, various kinds of clustering methods, PCA, and SVD was applied to analyze combined yeast cell cycle data. This paper includes both comparisons of some clustering algorithms such as K-means, SOM and fuzzy c-means and their results. For further analysis, clustering results from the integrated analysis program was used for function assignments to each cluster and for motif analysis. These results show an integrated analysis view on DNA chip data.

**Key Words** : DNA chip, expression profile analysis, integrated data analysis program, function assignment, motif analysis, Matlab

## 서 론

DNA 칩의 유전자 발현 양상을 분석하기 위해서 종래에는 hierarchical clustering (HC), K-means, self-organizing map (SOM), principal component analysis(PCA) 등과 같은 방법들이 많이 사용되어 왔지만(1-4), 최근에 들어와서 gene shaving (5), singular value decomposition(SVD) (6,7), support vector machine(SVM)(8) 등과 같은 새로운 분석 방법이 개발되어 응용되고 있다. 현재 많은 통계적 분석 방법들이 DNA chip 데이터 분석등에 응용이 되고 있으며, 새로운 분석 도구들이 지속적으로 개발되고 있다. DNA chip 데이터를 분석함에 있어 고려되어야 할 몇가지 사항들을 정리하면 다음과 같다.

첫째, DNA chip 이미지 데이터의 quality에 대한 문제로 여기에는 칩을 만드는 단계에서의 품질관리 및 실험단계에서의 정확한 프로토콜 확립이 중요하다고 할 수 있다(9). 그리고, 실험 목적에 맞는 실험 디자인도 중요한 요소가 될 수 있다. 실험자들 입장에서는 적은 비용으로 짧은 몇번의 실험으로 원하는 결과를 얻고자 하지만, 현재 내부 품질관리 수준 및 DNA chip이 갖고 있는 pin to pin variation, spot to spot variation, chip to chip variation, dye effects 등을 고려하여 조심스러운 실험 디자인이 필요하다고 본다. 최근 보고에 의하면(10) 칩간의 영향이 dye effect나 spot간의 영향보다 크기 때문에, 할 수 있다면 통계적 분석이 가능하도록 한 칩에 하나의 유전자에 대해 여러 개의 spot을 찍어 실험하는 것이 바람직하다고 할 수 있다. 물론 칩 제조 단계에서 높은 품질관리가 보장된다면 적은 수의 spot으로 실험을 해도 되지만 그렇지 않다면 칩내 spot replicates, 그리고 chip replicates에 대한 부분을 이후 통계적 분석이 가능하도록 사전에 잘 계획하는 것이 중요하다.

둘째, 이미지 데이터의 통계적인 분석이다(11). 상기에서 언급되었지만 칩의 이미지 데이터에 존재하는 여러 가지 변이(variation)을 정량화하고 보정할 수 있는 통계적인 분석이 중요하다. 이러한 통계적 분석을 통하여 칩 데이터에 대한 신뢰도를 제시할 수 있다. 여기에서는 control spot의 선정문제, 그리고 통계적 보정 방법들이 중요한 요소가 될 수 있다.

셋째, 이미지 데이터 분석 후 유전자 발현 데이터의 분석 방법으로 실험 목적에 맞는 분석 방법의 선정이 중요하다(12,13). 기존의 클러스터링 분석방법은 유사한 발현 패턴을

†Corresponding Author : National Center for Genome Information, Korea Research Institute of Bioscience and Biotechnology, 52 Eoun-dong, Yusong-gu, Daejon 305-333, Korea
TEL : +82-42-879-8520, FAX : +82-42-879-8559
E-mail : hurlee@mail.kribb.re.kr

가지는 유전자들을 묶어내는 방법으로 미지의 유전자에 대한 기능 예측에 유용하게 사용될 수 있다. 하지만 같은 클러스터내에 들어가는 유전자들이 반드시 동일한 기능을 갖는다고 할 수 없기 때문에 클러스터의 기능할당을 통한 통계적 유의성을 따져보는 것이 중요하다. 그리고, 최근에 들어와서는 발현패턴을 근거로 한 클러스터링 방법외에 같은 기능을 갖는 유전자들을 분류한 후 발현패턴에 의한 클러스터링 분석을 하거나 동일한 모티프를 갖는 유전자들을 분류한 후 클러스터링하는 다양한 분석방법들이 시도되고 있다(14).

클러스터링 분석에 있어서 중요한 문제는 실제 데이터내에 존재하는 클러스터의 개수에 대한 사전지식이 부족하다는 문제와 여러가지 클러스터링 방법중 어느 것을 사용하는 것이 좋은 지에 대한 기준이 부족하다는 것이다. 그러나 최근에 발표된 FOM 분석 방법(15)이나 Gap statistic(16)은 이 문제에 있어 좋은 기준을 제시해 주고 있다. 그리고, 클러스터링을 할 때는 사전에 통계적으로 신뢰성이 낮은 데이터에 대한 필터링이 또한 중요하다. 만약 실험 목적이 실험한 sample간에 발현차이가 큰 유전자들을 찾는 것이라면 기존의 클러스터링 방법은 적합하지 않고, Gene shaving 방법이 적합하며, 유전자 네트워크 상의 유전자간의 상호연관성을 보고 싶다면 SVD 방법을 사용하는 것이 바람직하다. 그리고, 유전자들에 대한 sample의 분류가 목적이라면 SVM, discriminant analysis 방법들이 사용될 수 있다. 이와 같이 실험목적에 따라 그에 부합되는 좋은 분석방법들이 있기 때문에 실험자는 분석에 앞서 분석 목적에 맞는 방법의 선정 및 그 방법에 대한 어느 정도의 이해가 필요하다.

본 논문에서는 본 실험실에서 개발된 DNA chip 통합분석 프로그램(17)을 이용한 효모의 세포주기관련 통합 발현 데이터의 분석에 대하여 보고하고자 한다. 이 프로그램은 매트랩 기반으로 작성이 되었으며 이전에 보고된 HC, K-means, Fuzzy c-means, SOM, PCA, SVD 통합분석 프로그램외에 최근에 발표된 FOM(Figure of Merit) 분석을 추가하여 분석 데이터 내에 존재하는 클러스터 개수의 추정 및 클러스터링 알고리즘의 비교가 가능하도록 하였으며, 분석 프로그램의 클러스터링 결과는 각 클러스터의 기능할당(function assignment), 모티프 분석(motif analysis) 등에 응용을 하였다. 본 논문은 DNA chip을 이용한 새로운 유전자의 기능분석을 위한 클러스터링 방법, 통계적 분석방법에 의한 각 클러스터의 기능할당, 모티프 분석에 대한 방법을 제시하였다.

## 재료 및 방법

본 논문에 사용된 유전자 발현 통합 분석 프로그램(17)은 매트랩(Matlab ver. 6.0, MathWorks)을 기반으로 작성되었으며, 이 프로그램에서 지원되는 기능들을 정리하면 Table 1과 같다.

본 프로그램을 이용한 분석 순서는 다음과 같다. 첫째, 주어진 입력 데이터에 대한 정규화에 대한 여부가 결정된 후, 둘째 FOM 분석을 통해 어떤 클러스터링 방법을 사용하는 것이 가장 좋은 결과를 줄 수 있는지 여부와 입력 데이터에 존재하는 클러스터의 개수를 추정하게 된다. 이후 HC/K-means/SOM/Fuzzy c-means와 같은 클러스터링 분석을 행하게 된다. 이와는 별도로 PCA는 주성분 분석을 통해 실제 입력 데이터들이 어떻게 공간상에 분포되어 있는가에 대한 정보를 줄 수 있으며, 주어진 입력데이타의 개수가 적고 공간상에 데이터의 분포가 어느 정도 잘 분리되어 분산되어있는 경우 데이터에 존재하는 클러스터의 개수에 대한 정보를 얻을 수도 있다. SVD 분석 방법은 유전자 네트워크상에 상호연관성이 있는 유전자들에 대한 정보를 제공함으로써 기존의 클러스터링 방법이 제공하지 못하는 정보를 제공하게 된다.

본 논문에서는 기존에 발표된 효모의 세포주기관련 DNA 칩 발현 통합 데이터를 활용하여(http://genome-www.stanford.edu/cellcycle/data/rawdata/combined.txt), FOM 분석을 통한 클러스터링 알고리즘의 성능비교 및 클러스터 개수의 추정, 클러스터링 결과의 비교 및 클러스터의 기능할당 및 모티프 분석 결과를 제시하였다. 효모의 세포주기 관련 통합 발현 데이터는 4개의 다른 방법(centrifugal elutriation, alpha-factor, Cdc15, Cdc28)에 대한 시간에 따른 세포배양 데이터와 G1 cyclin Cln30, B-type cyclin Clb2p를 사용한 G1말기 유도실험(induction experiments) 데이터가 포함되어 있으며, 총 6178개 유전자에 대해 77개 실험조건에서의 발현데이터로 구성되어 있다. 이 데이터 중 77개 조건에 대해 모든 데이터를 다 갖고 있는 유전자는 509개로 본 논문에서는 77개 조건, 509개 유전자에 대한 발현 데이터를 결국 분석하였다. 효모의 유전자에 대한 기능 분류는 MIPS의 Yeast ORF에 대한 Function catalogue (http://mips.gsf.de /proj/CYGD/db/index.html)를 따랐

**Table 1.** Main functions supported by integrated data analysis program

| Functions | Purposes | Comments |
|---|---|---|
| Normalization | Normalization of expression data | Row/column/(row&column) |
| Hierarchical clustering(HC) | Clustering data into dendrogram | Hard clustering |
| K-means clustering | Clustering data into K number of clusters | |
| Self organizing map(SOM) | Clustering based on neural network | |
| Fuzzy c-mens clustering | Clustering based on fuzzy set theory | Fuzzy clustering |
| Principal component analysis(PCA) | Multidimensional scaling method & visualization | |
| Singular value decomposition(SVD) | Finding characteristic modes and some interacting genes in genetic networki | Supports genetic network analysis |
| FOM analysis | Comparison of clustering algorithms and estimation of the number of clusters in dataset | Cluster validation method |

**Table 2.** Function catalogue of yeast ORFs in MIPS

| Function classification | Number of ORFs | Abbreviation(in this paper) |
|---|---|---|
| Metabolism | 1062 | MET |
| Energy | 252 | ENG |
| Cell growth, cell division and DNA synthesis | 833 | CGW |
| Transcription | 787 | TSP |
| Protein synthesis | 357 | PST |
| Protein destination | 586 | PDT |
| Transport facilitation | 310 | TPF |
| Cellular transport and transport mechanisms | 495 | CTP |
| Cellular biogenesis | 205 | BGN |
| Cellular communications/signal transduction | 133 | CCM |
| Cell rescue, defense, cell death and aging | 363 | RSC |
| Ioinc homeostasis | 123 | HSS |
| Celluar organization | 2254 | - |
| Transport elements, viral and plasmid proteins | 116 | - |
| Classification not yet clear-cut | 150 | - |
| Unclassified proteins | 2437 | - |

(This data was based on the results in MIPS on Jun 22, 2001)

으며(Table 2), 클러스터내 각 유전자의 모티프 분석은 SCPD (The Promoter Database of *Saccharomyces cerevisiae*) 데이터베이스를 활용하여. Yeast ORF의 upstream region 500 bp를 뽑아 각 유전자에 존재하는 49개의 알려진 Yeast ORF에 대한 분석 결과를 활용하였다(http://cgsigma.cshl.org/jian/).

## 결과 및 고찰

### FOM 분석에 의한 K-means, SOM, Fuzzy c-means의 클러스터링 알고리즘 성능비교

본 논문에서는 입력 발현 데이터의 정규화를 위해 행에 대한 정규화를 시도하였으며, 이 경우 Euclidean distance를 발현 패턴의 유사성 척도로 사용하더라도 상관계수(correlation coefficient)를 사용했을 때와 마찬가지로 발현세기는 차이가 있더라도 패턴의 유사성을 갖는 유전자들을 동일한 클러스터로 묶어낼 수가 있게 된다(18). 이와 같이 행에 대한 정규화가 된 입력 데이터를 FOM분석을 이용하여 K-means, SOM, Fuzzy c-means의 클러스터링 알고리즘 성능을 비교하여 보았다. FOM 분석은 Yeung 등(15)에 의해 제안된 방법으로 클러스터링 결과의 quality를 비교할 수 있는 파라메터로 FOM (figure of merit)를 사용하였는 데, 이 값은 전체 입력 데이터 중 각각의 열에 해당하는 실험조건을 하나씩 빼어가면서 클러스터링을 한 후, 앞서 뺀 실험조건을 클러스터링 결과를 검증하는 데 사용하는 방법으로 각 클러스터에 속한 유전자들의 발현 평균값과 해당 클러스터에 속한 유전자들간의 발현정도의 차이를 모두 합한 값이 된다. 따라서, 이 값이 작을수록 좋은 클러스터링 quality를 주는 알고리즘이 된다. 본 논문에서는 K-means, SOM, Fuzzy c-means 클러스터링 방법을 서로 비교하기 위해 클러스터 개수 5, 10, 15, 20, 30. 50, 100개의 클러스터 개수에 대해 FOM값을 서로 비교하여 보았다. Figure 1에 그 결과가 나와있는 데, 주어진 결과를 보면 Fuzzy c-means 클러스터링 결과가 다른 두개에 비해 전체 클러스터링 개수 범위에서 더 나은 결과를 보여주고 있다.

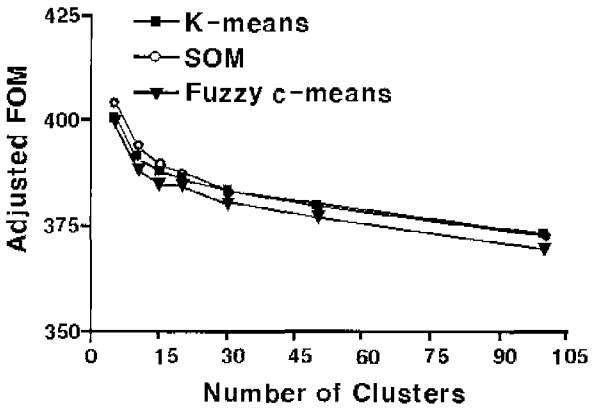


**Figure 1.** Results of FOM analysis with regard to three clustering algorithms such as K-means, SOM and Fuzzy c-means.

K-menas와 SOM의 경우는 서로 비슷한 결과를 보여주고 있다. 이러한 경향은 일반적이기 보다는 주어진 입력데이터에 따라 그리고 클러스터링 조건에 따라 조금씩 차이가 있게 된다.

Figure 1에서 보면 클러스터의 개수가 20개 정도에서 그래프의 기울기가 상당히 줄어들게 되는 데, 이와 같이 FOM 그래프에서 elbow같이 꺽어지는 부분이 주어진 data에 존재하는 클러스터의 개수로 추정하게 된다(15). 따라서, 본 논문에서는 이후 클러스터링을 위한 클러스터 개수는 20개로 정하였다.

### K-means, SOM, Fuzzy c-means 클러스터링 결과

FOM 분석 결과에 따라 20개의 클러스터 개수에 대해 상기 클러스터링 분석 방법을 이용하여 주어진 효모의 세포주기 관련 통합발현 데이터의 클러스터링을 실시하였다. 매트랩 기반의 통합분석 프로그램은 각 클러스터링 방법에 대해 클러스터의 평균 발현패턴(Figure 2), 평균 발현패턴을 근간으로 한 각 클러스터간의 유연관계를 보여주는 HC결과 (Figure 3), 각 클러스터내의 유전자들의 개수에 대한 결과

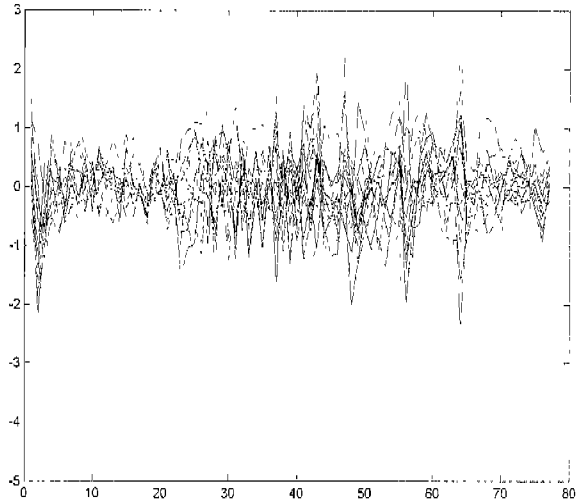

**Figure 2.** Average patterns of each cluster in K-means.

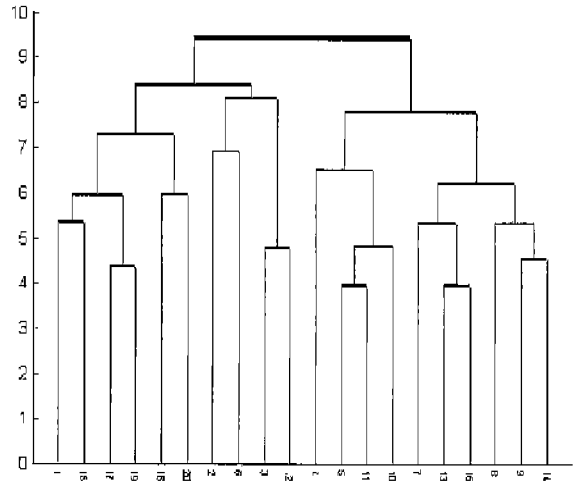

**Figure 3.** Hierarchical clustering of each cluster in K-means clustering.

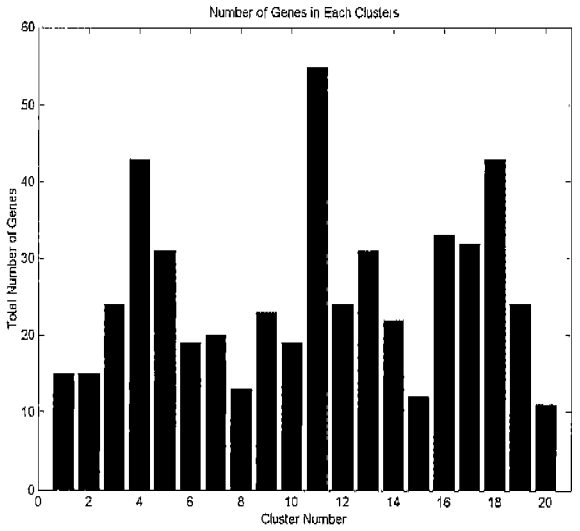


**Figure 4.** The number of genes in each cluster(K-means clustering).

(Figure 4), 각 클러스터별 유전자들의 전체 발현패턴에 대한 결과(Figure 5)를 그래프로 제시하며, 분석과정에서 나온 각

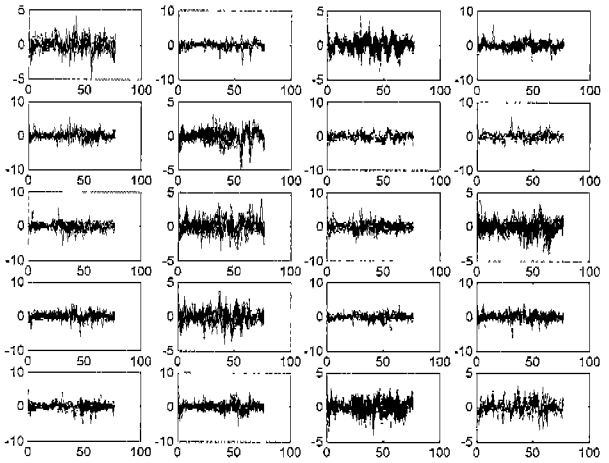

**Figure 5.** Expression patterns of overall genes in each cluster (K-means clustering).

**Table 3.** Statistics fr the number fo genes in each cluster regarding clustering methods

| | K-means | SOM | Fuzzy c-means |
|---|---|---|---|
| Standard deviation | 11.6 | 8.6 | 8.0 |
| Mode | 24 | 24 | 24 |
| Range | 44 | 33 | 31 |

유전자의 클러스터 정보는 텍스트 형태의 파일로 저장할 수 있다. 이 외에 SOM의 경우는 각 뉴런의 feature map에 대한 그래프, SOM dendrogram등이 추가되며, Fuzzy c-means 클러스터링의 경우 각 유전자가 20개의 클러스터에 속하는 정도를 나타내는 귀속정도(membership grade)에 대한 정보등이 더 추가된다(data not shown)(17). Figure 2-5까지는 K-means 클러스터링 결과만을 분석예로 보여주고 있으며, 다른 두개의 클러스터링 결과는 생략하였다. 데이터로 제시하지는 않았지만 각 클러스터링 방법에 따라 평균 발현패턴이나 클러스터내 유전자의 개수등에 있어서 서로 차이가 나타나게 된다.

Table 2에는 각 클러스터링 방법에 따른 클러스터내 유전자들의 개수에 대한 통계적인 수치가 나와있는데, Fuzzy c-means나 SOM의 결과가 K-means 방법에 비해 각 클러스터내 유전자 개수의 편차가 적으며, 최대/최소 유전자 개수의 차이(range)도 적게 나타남을 알 수 있다. 이 논문에서는 K-means 클러스터링을 위한 초기화 방법으로 랜덤초기화(random initialization) 방법을 사용하고 있는 데, 이 경우 클러스터링을 할 때마다 초기 랜덤화에 따른 로컬 최적화에 따른 결과를 나타내기 대문에 SOM이나 Fuzzy c-means보다는 각 클러스터내 유전자 개수의 편차 및 range가 크게 나타날 수 있다. 이러한 결과는 K-means 클러스터링 방법에서 초기화 조건으로 어떤 방법을 사용하느냐에 따라 결과는 달라지게 된다. 이에 반해 SOM이나 Fuzzy c-means의 경우 K-means보다는 좀더 균일한 분석 결과를 보여주게 된다.

### 각 클러스터의 기능 할당(Function assignment)및 모티프 분석(Motif analysis)

클러스터링에 의해 발현 패턴이 유사한 유전자들을 각각의 클러스터로 묶어낼 수 있는데, 이러한 방법은 기능이 알려

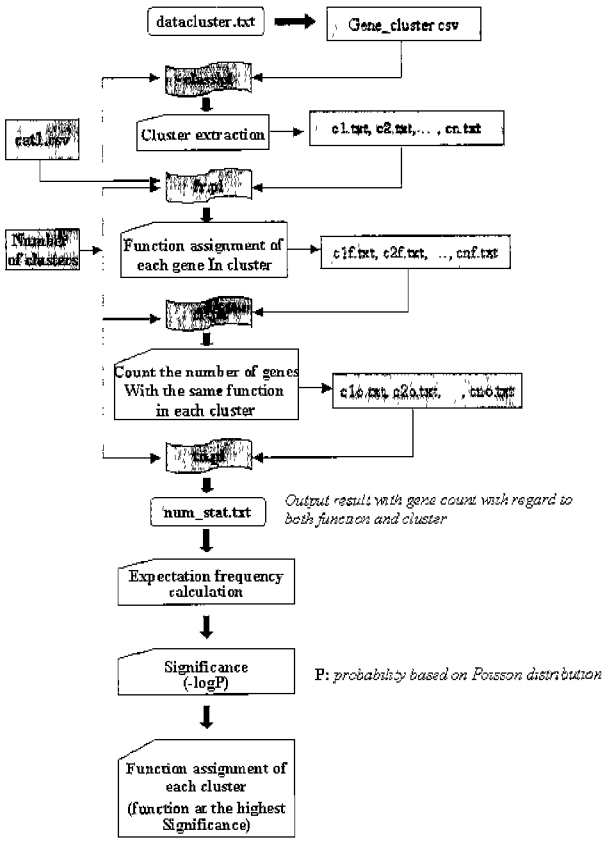


**Figure 6.** Schematic workflow for the function assignment to each cluster after clustering.

져 있지 않은 유전자들과 기능을 이미 알고 있는 유전자들을 발현패턴의 유사성을 근거로 클러스터링함으로써 미지 유전자의 기능에 대한 정보를 줄 수 있다. 하지만 발현 패턴이 유사한 유전자들이 반드시 유사한 기능을 갖는 것은 아니고 역으로 기능이 유사한 유전자들이라도 반드시 동일한 발현 패턴을 보이지는 않기 때문에 각 클러스터가 어떤 기능에 속하는 지를 통계적 유의성을 가지고 분석하는 것이 필요하다. Zhu 등(14)은 효모의 ORF를 기능별로 분류해놓은 MIPS 데이터베이스의 Function catalog 데이터를 토대로 각 클러스터의 기능을 할당하였는 데, 구체적인 방법은 Figure 6에 나타내었다. 각 클러스터링 분석후 얻어지게 되는 데이터(datacluster.txt) 중 각 유전자의 해당 클러스터 번호를 정리한 데이터(gene_cluster.csv)를 따로 정리하여, 이 정보를 토대로 각 클러스터에 할당된 유전자들을 따로 분류하였다. 이후 각 클러스터안에 있는 유전자들을 MIPS 데이터베이스의 yeast ORF에 대한 function catalogue 데이터(cat1.csv)와 비교하여 각 클러스터별 같은 기능을 갖고 있는 유전자의 개수를 계수하게 되며(num_stat.txt), 이후 이 값들을 토대로 Poisson 확률을 구하여 각 클러스터별 기능의 중요도(-logP)를 계산하게 된다. 이 때 각 클러스터별 동일 기능 유전자의 개수가 기대값(클러스터내 유전자 총 개수 각 기능별 유전자의 frequency) 보다 높게 나타나며, -logP의 값이 가장 큰 값을 갖는 기능을 해당 클러스터의 기능으로 할당하게 되며, 이 값이 3.5이상일 때 유의성이 높은 것으로 간주하였다(14).

각 클러스터의 모티프 분석은 먼저 509개의 유전자에 대하여 SCPD에 있는 49개의 알려진 모티프에 대하 분석한 정보(motif_info.csv)를 토대로 각 클러스터안에 있는 유전자들의 모티프의 계수를 각 클러스터별 모티프별로 정리하게 된다. 이후 분석은 클러스터의 기능할당에 사용한 방법을 그대로 사용하여 모티프의 중요도를 확률적으로 계산하여 각 클러스터별 중요 모티프를 찾게 된다(Figure 7). 분석을 위해 필요한 프로그램은 Perl로 작성하였으며, 각 단계별 프로그램들은 배치파일(batch file)로 묶어, 클러스터링한 후 얻은 결과 데이터(gene_cluster.csv)와 클러스터 개수에 대한 정보가 주어지게 되면 자동으로 단계별 분석이 이루어지도록 하였으며, 확률 계산은 매트랩 기반 프로그램을 작성하여 계산하였다. Table 4,5,6은 각 클러스터링 방법에 따른 클러스터의 기능할당과 관련하여 각 클러스터별 기능의 중요도에 대한 분석결과 데이터를 보여주고 있다. 분석결과에 따르면 3개의 클러스터링 방법 모두가 단백질 합성(protein synthesis, PST)과 관련되어 유의성이 높은 클러스터를 찾아주었다. K-means 클러스터링의 경우 클러스터 2,6번이 중요도 7.93, 13.64를 나타내었으며, SOM의 경우 클러스터 3번이 14.08의 중요도를 나타내었으며, Fuzzy c-means의 경우 클러스터 3,4번이 각각 9.00, 13.21의 높은 값으로 단백질 합성과 관련된 기능으로 할당되었다. 이와 같이 통계적 유의성을 따질 때 전술한 클러스터 외에 나머지 클러스터들은 특정 기능을 갖는다고 말하기가 어렵다는 것을 알 수 있다.

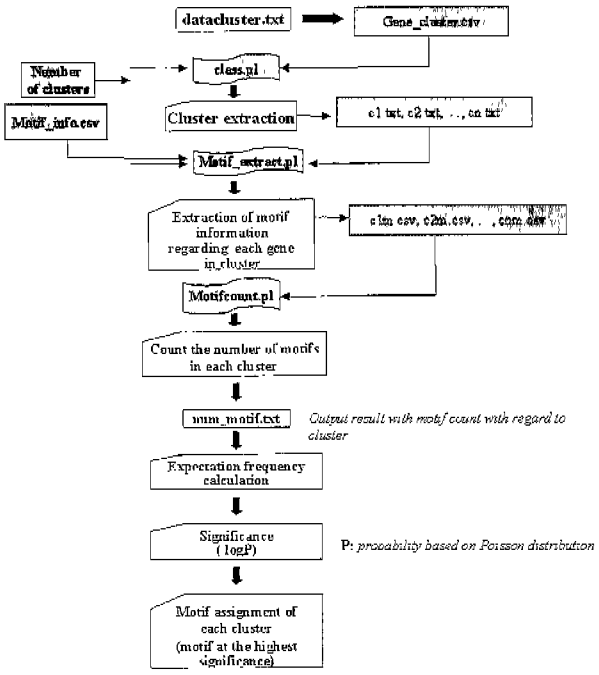


**Figure 7.** Schematic workflow for the motif assignment to each cluster after clustering.

통계적 유의성이 높은 단백질 합성 관련 클러스터를 살펴보면, 각 클러스터내 기지의 유전자와 미지의 유전자가 함께 있는 클러스터가 존재함을 알 수 있으며(Table 6), 각 클러스터링 방법별로 얻어진 단백질 합성 관련 유전자들간에는 조금씩 차이가 있음을 알 수 있다. 미지의 유전자중 3가지 클러스터링 방법 모두에 의해 공통적으로 단백질 합성 관련 유전자로 판별된 것은 YCR029C-a와 YNR053C였다. K-means

**Table 3.** Significance(-logP) of each function within each cluster(K-means clustering)

| | TPF | TSP | BGN | CTP | RSC | PST | PDT | MET | HSS | ENG | CGW | CCM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 0.02 | 0.91 | 0.02 | 0.04 | 0.03 | 0.03 | 0.05 | 0.08 | 0.01 | 0.02 | 0.07 | 0.01 |
| C2 | 0.39 | 1.00 | 0.26 | 0.63 | 0.46 | **7.93** | 0.83 | 1.34 | 0.16 | 0.31 | 0.58 | 0.17 |
| C3 | 0.43 | 1.12 | 0.94 | 0.85 | 0.44 | 0.50 | 0.83 | 0.66 | 0.18 | 0.35 | **1.84** | 0.55 |
| C4 | 0.53 | 1.99 | 0.44 | 0.63 | 1.37 | 0.89 | 0.66 | 1.33 | 0.46 | **3.22** | 1.04 | 0.34 |
| C5 | 0.97 | 0.96 | **1.96** | 0.60 | 0.49 | 0.67 | 0.70 | 0.87 | 0.23 | 0.70 | 0.68 | 1.02 |
| C6 | 0.47 | 1.18 | 0.31 | 0.74 | 0.55 | **13.64** | 0.57 | 1.03 | 0.19 | 0.37 | 1.25 | 0.20 |
| C7 | **2.30** | 1.25 | 0.32 | 0.53 | 1.47 | 0.56 | 0.72 | 0.81 | 0.54 | 0.39 | 0.77 | 0.21 |
| C8 | 0.05 | 0.12 | 0.03 | 0.08 | 0.06 | 0.06 | 0.09 | 0.58 | 0.02 | 0.04 | **0.65** | 0.02 |
| C9 | 0.46 | 0.51 | 0.19 | 0.44 | 0.45 | 0.33 | 0.65 | **0.93** | 0.12 | 0.23 | 0.57 | 0.13 |
| C10 | **1.79** | 0.71 | 0.29 | 0.85 | 0.44 | 0.50 | 0.77 | 0.73 | 0.57 | 0.44 | 1.19 | 0.19 |
| C11 | **2.18** | 2.31 | 0.57 | 0.88 | 0.90 | 1.42 | 0.79 | 1.70 | 0.50 | 1.69 | 1.47 | 0.54 |
| C12 | **1.58** | 0.75 | 0.54 | 1.14 | 0.34 | 0.33 | 0.45 | 0.58 | 0.12 | 0.23 | 0.57 | 0.13 |
| C13 | **1.00** | 0.70 | 0.81 | 0.57 | 0.89 | 0.47 | 0.67 | 0.77 | 0.51 | 0.43 | 0.72 | 0.24 |
| C14 | **1.58** | 0.75 | 0.19 | 0.44 | 0.45 | 0.33 | 0.55 | 0.58 | 0.69 | 1.07 | 0.57 | 0.13 |
| C15 | 0.62 | 0.44 | 0.10 | 0.24 | 0.17 | 0.17 | **0.97** | 0.44 | 0.06 | 0.12 | 0.44 | 0.06 |
| C16 | 0.79 | 0.76 | 0.44 | 1.25 | 0.72 | 0.89 | 0.73 | 0.80 | 0.31 | 0.47 | **1.01** | 0.86 |
| C17 | 0.32 | **1.54** | 0.21 | 0.51 | 0.44 | 0.36 | 0.60 | 0.59 | 0.13 | 0.25 | 1.06 | 0.64 |
| C18 | 1.14 | 0.79 | 0.67 | 0.63 | 0.72 | 0.57 | 0.66 | 0.80 | **1.52** | 0.47 | 0.83 | 0.34 |
| C19 | 0.44 | 0.71 | 0.29 | 0.59 | 0.67 | 0.50 | 0.57 | 0.73 | 0.18 | 0.84 | **1.08** | 0.19 |
| C20 | 0.10 | 0.49 | **0.89** | 0.16 | 0.11 | 0.11 | 0.56 | 0.45 | 0.04 | 0.08 | 0.26 | 0.04 |

(The shaded blocks have lower number of genes than expectation frequency)

**Table 4.** Significance(-logP) of each function within each cluster(SOM clustering)

| | TPF | TSP | BGN | CTP | RSC | PST | PDT | MET | HSS | ENG | CGW | CCM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 0.45 | 0.87 | 1.09 | 0.55 | 0.77 | 0.39 | 0.47 | 0.66 | 0.14 | 0.27 | 0.72 | 1.39 |
| C2 | 0.69 | 1.25 | 0.88 | 0.79 | 0.63 | 0.56 | 0.57 | 0.70 | 0.54 | 0.39 | 0.99 | 0.21 |
| C3 | 0.52 | 1.31 | 0.34 | 0.82 | 0.60 | **14.08** | 0.57 | 1.76 | 0.20 | 0.41 | 0.88 | 0.22 |
| C4 | 0.49 | 0.71 | 0.72 | 1.10 | **1.11** | 0.78 | 0.78 | 0.76 | 0.27 | 0.65 | 0.78 | 0.46 |
| C5 | 0.39 | **1.69** | 0.26 | 0.63 | 0.44 | 0.44 | 0.73 | 0.65 | 0.16 | 0.31 | 1.21 | 0.58 |
| C6 | 2.52 | 1.94 | 0.58 | 0.71 | 0.64 | **3.29** | 0.73 | 0.97 | 0.74 | 0.86 | 0.86 | 0.46 |
| C7 | **2.62** | 1.38 | 0.65 | 0.65 | 0.71 | 0.92 | 0.73 | 1.79 | 0.32 | 0.91 | 1.47 | 0.35 |
| C8 | 0.78 | 0.64 | 0.48 | **1.91** | 0.44 | 0.44 | 0.51 | 0.65 | 0.16 | 0.46 | 0.67 | 0.17 |
| C9 | **0.84** | 0.19 | 0.05 | 0.69 | 0.09 | 0.08 | 0.14 | 0.49 | 0.03 | 0.06 | 0.20 | 0.03 |
| C10 | 0.20 | 0.50 | 0.13 | 0.31 | 0.23 | 0.22 | 0.44 | 0.88 | 0.08 | 0.16 | **1.05** | 0.80 |
| C11 | 0.10 | 0.49 | 0.06 | 0.16 | 0.11 | 0.11 | 0.18 | 0.45 | 0.04 | 0.08 | **1.00** | 0.04 |
| C12 | 1.07 | 0.83 | 0.45 | 0.77 | 0.60 | 0.58 | 0.62 | 0.97 | 0.20 | **2.61** | 0.88 | 0.22 |
| C13 | 0.45 | 0.57 | 0.23 | 0.45 | 0.44 | 0.79 | 0.60 | **1.09** | 0.14 | 0.47 | 0.92 | 0.15 |
| C14 | 0.45 | 1.37 | 0.44 | 0.57 | 0.92 | 0.61 | **1.67** | 0.71 | 0.21 | 0.43 | 0.74 | 0.23 |
| C15 | 0.44 | 0.93 | 0.24 | 0.46 | 0.43 | 0.42 | 0.49 | 1.34 | 0.62 | **3.22** | 0.99 | 0.16 |
| C16 | 0.89 | 0.57 | 0.21 | 0.44 | 1.35 | 0.36 | 0.60 | 0.69 | **1.50** | 0.25 | 0.57 | 0.14 |
| C17 | **4.13** | 0.77 | 0.50 | 0.62 | 1.41 | 0.86 | 0.65 | 1.35 | 0.92 | 0.61 | 0.76 | 0.33 |
| C18 | 0.49 | 1.25 | 0.45 | 0.78 | 0.99 | 0.56 | 0.59 | 1.59 | 0.19 | **2.70** | 0.65 | 0.21 |
| C19 | 0.45 | 0.87 | 1.09 | 0.65 | 0.44 | 0.44 | 0.47 | 1.18 | 0.14 | 0.27 | **1.37** | 0.62 |
| C20 | 0.20 | 0.68 | 0.13 | 0.31 | 0.23 | 0.22 | **0.81** | 0.59 | 0.08 | 0.16 | 0.66 | 0.08 |

(The shaded blocks have lower number of genes than expectation frequency)

의 경우 이외에 YDL041W, YFR056C가 Fuzzy c-means의 경우는 YAR009C, YBR219C가 MIPS 데이터베이스에서는 기능이 분류되어있지 않지만, 본 분석에 의하면 통계적으로 유의성높게 단백질 합성과 관련된 유전자로 분석되었다. 그리고, 각 클러스터내 유전자들의 발현패턴을 가지고 HC을 하게되면 클러스터내 기지의 유전자들과 미지의 유전자들간의 유연관계를 dendrogram형태로 더 자세히 볼 수 있게 된다(data not shown).

Table 7은 K-means 클러스터링후 얻어진 결과에 대해 각 클러스터별 모티프 분석결과를 나타낸 것이다. 전술한 바와 같이 SOM, Fuzzy c-means 클러스터링 방법에 대해서도 동일한 방법으로 분석할 수 있다(data not shown). 3가지 경우 TBP(TATA binding protein) 모티프가 logP의 값이 K-means의 경우 5.35(C11), SOM의 경우 4.59(C6), Fuzzy c-means의 경우 5.48(C16)로 통계적으로 유의성이 높게 나타났다. 이 때 각각의 클러스터 기능은 TPF(transport facilitation))로 할당되었지만 통계적 유의성은 3보다 적은 값을 나타내었다. 그리고, 3가지 클러스터링 방법에 따른 각 클러스터의 모티프 분

**Table 5.** Significance(-logP) of each function within each cluster(Fuzzy c-meanc clustering)

| | TPF | TSP | BGN | CTP | RSC | PST | PDT | MET | HSS | ENG | CGW | CCM |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 0.69 | 0.79 | 0.32 | 1.14 | 0.57 | 0.56 | 0.59 | 1.59 | 0.19 | 0.78 | 0.65 | 0.21 |
| C2 | 0.81 | 0.57 | 0.49 | 0.59 | 0.43 | 0.42 | 0.59 | 0.79 | 0.15 | 0.29 | 0.70 | 0.16 |
| C3 | 0.39 | 1.00 | 0.26 | 0.63 | 0.46 | 9.00 | 0.83 | 1.34 | 0.16 | 0.31 | 0.67 | 0.17 |
| C4 | 0.42 | 1.06 | 0.28 | 0.67 | 0.49 | 13.21 | 0.53 | 1.43 | 0.17 | 0.33 | 1.12 | 0.18 |
| C5 | 0.44 | 1.25 | 0.45 | 0.79 | 0.63 | 0.56 | 0.72 | 0.71 | 0.54 | 1.95 | 0.84 | 0.21 |
| C6 | 1.60 | 0.83 | 0.34 | 0.77 | 0.62 | 0.46 | 0.96 | 0.80 | 0.53 | 0.43 | 0.65 | 0.22 |
| C7 | 0.02 | 0.06 | 0.02 | 0.04 | 0.03 | 0.03 | 0.05 | 0.08 | 0.01 | 0.02 | 0.88 | 0.01 |
| C8 | 0.22 | 0.64 | 0.15 | 0.35 | 0.26 | 0.25 | 0.76 | 0.81 | 0.09 | 0.18 | 0.62 | 0.09 |
| C9 | 0.43 | 1.12 | 0.94 | 0.85 | 0.44 | 0.50 | 0.83 | 0.66 | 0.18 | 0.35 | 1.84 | 0.55 |
| C10 | 0.54 | 0.75 | 0.83 | 0.86 | 0.92 | 0.61 | 0.58 | 0.93 | 0.21 | 1.22 | 0.93 | 0.50 |
| C11 | 2.58 | 0.67 | 0.28 | 1.30 | 0.44 | 0.47 | 0.57 | 0.70 | 1.30 | 0.33 | 1.12 | 0.18 |
| C12 | 0.43 | 0.59 | 0.28 | 0.67 | 0.69 | 0.47 | 0.53 | 0.92 | 0.17 | 0.87 | 0.85 | 0.18 |
| C13 | 0.34 | 1.44 | 0.23 | 0.55 | 0.40 | 0.39 | 0.60 | 0.66 | 0.14 | 0.27 | 0.72 | 0.62 |
| C14 | 0.45 | 0.87 | 0.51 | 0.45 | 0.44 | 0.39 | 0.91 | 0.74 | 0.14 | 0.27 | 1.82 | 0.15 |
| C15 | 0.61 | 1.56 | 1.90 | 0.98 | 1.22 | 0.69 | 0.60 | 0.85 | 0.24 | 1.11 | 0.69 | 0.48 |
| C16 | 2.63 | 1.36 | 0.49 | 0.79 | 1.40 | 1.17 | 0.77 | 0.90 | 0.41 | 0.57 | 0.80 | 0.43 |
| C17 | 1.39 | 1.00 | 0.44 | 0.57 | 0.84 | 0.49 | 0.60 | 0.76 | 1.78 | 0.44 | 0.79 | 0.26 |
| C18 | 0.44 | 0.64 | 0.48 | 0.63 | 0.71 | 0.44 | 0.73 | 1.26 | 0.16 | 2.26 | 0.67 | 0.17 |
| C19 | 3.17 | 1.05 | 0.43 | 0.92 | 0.82 | 0.72 | 0.62 | 1.08 | 0.49 | 0.44 | 0.71 | 0.27 |
| C20 | 0.49 | 0.62 | 0.59 | 0.44 | 0.29 | 0.47 | 0.43 | 0.57 | 0.10 | 0.20 | 0.60 | 0.72 |

(The shaded blocks have lower number of genes than expectation frequency)

**Table 6.** Genes assigned to protein synthesis within each cluster after function assignment of each cluster

| K-means clustering | | | | SOM clustering | | | | Fuzzy c-means clustering | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cluster 2 | Function | | | Cluster 3 | Function | | | Cluster 3 | Function | | |
| YCR029C-a | | Unknown | | YBL027W | PST | | | YAR009C | | Unknown | |
| YDL041W | | Unknown | | YBR084C-a | PST | | | YBR219C | | Unknown | |
| YER074W | PST | | | YBR181C | PST | | | YCR029C-a | | Unknown | |
| YER075C | CGW | | | YCR029C-a | | Unknown | | YER074W | PST | | |
| YGL147C | PST | | | YDR064W | PST | | | YGL147C | PST | | |
| YGR118W | PST | | | YER074W | PST | | | YGR118W | PST | | |
| YGR214W | PST | PDT | CGW | YGL147C | PST | | | YGR214W | PST | PDT | CGW |
| YIL148W | PST | PDT | | YGR118W | PST | | | YIL148W | PST | PDT | |
| YKL081W | PST | | | YGR214W | PST | PDT | CGW | YJR027W | | Unknown | |
| YLR048W | PST | PDT | | YJL189W | PST | | | YKL081W | PST | | |
| YLR333C | PST | | | YJR145C | PST | | | YLR048W | PST | PDT | |
| YLR344W | PST | | | YKL081W | PST | | | YLR333C | PST | | |
| YLR388W | PST | | | YKR094C | PST | PDT | | YLR344C | PST | | |
| YNL067W | PST | | | YLR185W | PST | | | YLR388W | PST | | |
| YNR053C | | Unknown | | YLR333C | PST | | | YLR406C | PST | | |
| Cluster 6 | Function | | | YLR344W | PST | | | YNL067C | PST | | |
| YBL027W | PST | | | YNL067W | PST | | | YNR053C | | Unknown | |
| YBR084C-a | PST | | | YNR053C | | Unknown | | Cluster 4 | Function | | |
| YBR181C | PST | | | YOR234C | PST | | | YBL027W | PST | | |
| YDR012W | PST | | | YPL198W | PST | | | YBR084C-a | PST | | |
| YDR025W | PST | | | | | | | YBR181C | PST | | |
| YDR064W | PST | | | | | | | YDR025W | PST | | |
| YFR056C | | Unknown | | | | | | YDR064W | PST | | |
| YHL001W | PST | | | | | | | YHL001W | PST | | |
| YJL189W | PST | | | | | | | YJL189W | PST | | |
| YJR145C | PST | | | | | | | YJR145C | PST | | |
| YKL180W | PST | | | | | | | YKL180W | PST | | |
| YKR094C | PST | PDT | | | | | | YKR094C | PST | PDT | |
| YLR185W | PST | | | | | | | YLR185W | PST | | |
| YLR367W | PST | | | | | | | YLR367W | PST | | |
| YMR306W | MET | | | | | | | YOL120C | PST | | |
| YOL120C | PST | | | | | | | YOR234C | PST | | |
| YOR234C | PST | | | | | | | YPL143W | PST | | |
| YPL143W | PST | | | | | | | YPL198W | PST | | |
| YPL198W | PST | | | | | | | | | | |

**Table 7.** Significance(-logP) of each motif within each cluster(K-means clustering)

| | C1 | C2 | C3 | C4 | C5 | C6 | C7 | C8 | C9 | C10 | C11 | C12 | C13 | C14 | C15 | C16 | C17 | C18 | C19 | C20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ABF1 | 0.77 | 0.71 | 0.83 | 0.99 | 0.88 | 0.77 | 1.06 | 0.69 | 1.99 | 0.91 | 1.66 | 2.28 | 0.86 | 1.08 | 0.68 | 0.96 | 0.88 | 0.95 | 1.05 | 0.71 |
| ACE2 | 0.66 | 0.66 | 0.91 | 0.89 | 0.84 | 0.76 | 0.72 | 0.63 | 1.24 | 1.22 | 1.04 | 0.81 | 0.76 | 1.08 | 0.71 | 0.88 | 0.76 | 0.84 | 0.70 | 0.57 |
| ADR1 | 0.98 | 1.04 | 1.11 | 1.24 | 1.11 | 1.08 | 1.02 | 0.99 | 1.06 | 1.20 | 1.26 | 1.03 | 1.17 | 1.13 | 1.13 | 1.41 | 1.10 | 1.17 | 1.77 | 1.06 |
| AP-1 | 0.95 | 0.73 | 1.06 | 1.05 | 1.17 | 0.80 | 1.06 | 0.89 | 0.89 | 085 | 1.10 | 0.86 | 0.90 | 0.87 | 0.75 | 0.93 | 0.91 | 1.02 | 1.00 | 0.70 |
| ATF | 0.59 | 1.51 | 0.88 | 0.91 | 1.24 | 0.67 | 0.79 | 0.79 | 0.65 | 0.59 | 1.35 | 0.75 | 0.71 | 1.32 | 0.76 | 0.79 | 0.71 | 0.79 | 0.75 | 0.49 |
| BAS2 | 1.18 | 1.38 | 1.40 | 1.39 | 1.34 | 1.24 | 1.71 | 1.11 | 1.23 | 1.18 | 1.89 | 1.41 | 1.28 | 1.30 | 1.64 | 1.47 | 1.90 | 2.62 | 1.50 | 1.89 |
| CPF1 | 0.26 | 0.23 | 0.43 | 0.52 | 0.69 | 0.30 | 0.35 | 1.10 | 0.39 | 0.46 | 1.58 | 0.78 | 0.68 | 0.44 | 0.21 | 0.45 | 0.52 | 051 | 0.39 | 0.54 |
| CSRE | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| CuRE | 0.45 | 1.11 | 0.57 | 1.06 | 0.58 | 0.47 | 0.58 | 0.44 | 0.81 | 0.65 | 1.12 | 0.54 | 0.68 | 0.79 | 1.76 | 0.66 | 0.67 | 0.86 | 0.81 | 0.41 |
| ECB | 1.07 | 100 | 1.24 | 1.47 | 1.07 | 1.35 | 1.45 | 1.11 | 1.00 | 0.96 | 2.09 | 1.01 | 1.05 | 1.05 | 0.87 | 2.48 | 1.11 | 2.16 | 1.00 | 0.86 |
| GAL4 | 0.24 | 0.22 | 0.44 | 0.58 | 0.46 | 0.28 | 0.33 | 0.21 | 0.37 | 1.58 | 0.69 | 2.79 | 1.14 | 0.35 | 0.20 | 0.53 | 0.50 | 0.49 | 0.44 | 0.19 |
| GC/FAR | 0.40 | 0.44 | 0.60 | 0.75 | 0.58 | 0.71 | 1.03 | 0.35 | 0.60 | 0.48 | 1.03 | 0.62 | 0.77 | 0.98 | 0.86 | 0.57 | 0.57 | 0.84 | 0.60 | 0.30 |
| GCN4 | 1.72 | 1.30 | 1.49 | 1.62 | 1.64 | 1.37 | 1.39 | 1.30 | 2.49 | 1.43 | 1.73 | 1.68 | 1.47 | 1.52 | 1.30 | 1.64 | 1.66 | 1.79 | 1.71 | 1.28 |
| GCR1 | 1.10 | 2.16 | 2.45 | 1.35 | 1.29 | 1.66 | 1.17 | 1.56 | 1.20 | 1.34 | 1.77 | 1.70 | 1.28 | 1.22 | 1.21 | 1.72 | 1.26 | 1.36 | 1.20 | 1.05 |
| HAP1 | 0.03 | 0.02 | 0.04 | 0.83 | 0.05 | 0.03 | 0.04 | 0.02 | 0.04 | 0.03 | 0.11 | 0.04 | 0.99 | 0.04 | 0.02 | 0.06 | 0.05 | 0.07 | 0.04 | 0.02 |
| HSTF | 1.08 | 1.26 | 1.06 | 1.67 | 1.69 | 1.35 | 1.05 | 1.01 | 1.14 | 1.06 | 1.64 | 1.20 | 1.73 | 1.06 | 1.01 | 1.35 | 1.35 | 1.56 | 1.48 | 1.27 |
| LEU3 | 0.01 | 0.01 | 0.02 | 0.04 | 0.02 | 0.01 | 0.02 | 0.01 | 0.02 | 0.02 | 0.05 | 0.02 | 1.27 | 0.02 | 0.01 | 0.03 | 0.03 | 0.04 | 0.02 | 0.01 |
| MATalpha2 | 0.62 | 1.08 | 1.07 | 1.06 | 1.54 | 0.94 | 0.71 | 1.62 | 0.73 | 0.66 | 0.95 | 0.79 | 1.11 | 0.81 | 0.64 | 0.79 | 0.77 | 0.91 | 0.79 | 0.59 |
| MCB | 0.59 | 0.57 | 6.61 | 0.87 | 0.98 | 1.04 | 0.71 | 0.93 | 1.04 | 0.65 | 1.42 | 0.71 | 2.07 | 0.75 | 0.89 | 1.59 | 1.48 | 0.90 | 0.77 | 0.54 |
| MCM1 | 0.71 | 0.74 | 0.83 | 0.95 | 1.38 | 0.88 | 4.50 | 0.66 | 1.03 | 0.86 | 1.04 | 0.87 | 0.94 | 1.24 | 0.69 | 0.86 | 1.53 | 1.38 | 1.02 | 0.64 |
| MIG1 | 0.32 | 0.29 | 0.44 | 0.61 | 1.41 | 0.37 | 0.73 | 0.28 | 0.48 | 0.44 | 0.76 | 0.49 | 0.47 | 0.47 | 0.27 | 0.85 | 0.90 | 0.93 | 2.34 | 0.24 |
| MSE | 0.94 | 0.76 | 1.00 | 2.13 | 0.90 | 0.97 | 1.52 | 1.31 | 1.01 | 0.80 | 1.19 | 1.04 | 1.01 | 0.93 | 0.76 | 1.27 | 0.92 | 1.03 | 0.91 | 0.98 |
| NBF | 0.58 | 0.68 | 1.05 | 2.04 | 0.69 | 0.57 | 0.65 | 0.47 | 0.66 | 0.73 | 0.83 | 0.72 | 0.66 | 1.08 | 0.47 | 0.68 | 0.67 | 1.76 | 1.12 | 0.45 |
| ORC | 0.15 | 0.63 | 0.26 | 0.46 | 0.29 | 0.18 | 0.21 | 0.64 | 0.50 | 0.19 | 0.64 | 0.23 | 0.46 | 1.10 | 0.13 | 0.86 | 0.45 | 0.72 | 0.50 | 0.12 |
| PDR3 | 0.13 | 0.12 | 0.21 | 0.80 | 1.78 | 0.61 | 0.57 | 0.11 | 0.19 | 0.16 | 0.66 | 0.20 | 0.25 | 0.19 | 0.11 | 0.28 | 0.26 | 0.44 | 0.19 | 0.10 |
| PHO2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| PHO4 | 0.64 | 0.84 | 0.76 | 0.86 | 1.08 | 1.47 | 1.72 | 0.59 | 0.91 | 0.72 | 1.26 | 0.72 | 0.82 | 0.71 | 0.58 | 1.09 | 0.79 | 0.88 | 0.75 | 0.97 |
| PPR1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| PUT3 | 0.12 | 0.10 | 0.19 | 1.43 | 0.52 | 0.13 | 0.16 | 0.10 | 0.17 | 0.63 | 0.44 | 0.18 | 0.22 | 0.58 | 0.75 | 0.25 | 0.24 | 0.45 | 0.17 | 0.09 |
| RAP1 | 0.64 | 2.90 | 0.76 | 0.94 | 1.66 | 1.45 | 1.22 | 1.11 | 0.76 | 0.67 | 1.03 | 0.77 | 0.80 | 0.74 | 0.67 | 0.84 | 1.04 | 0.86 | 0.76 | 0.57 |
| REB1 | 0.45 | 0.44 | 0.71 | 0.82 | 0.59 | 0.48 | 0.57 | 0.44 | 0.57 | 0.86 | 0.77 | 0.57 | 1.51 | 0.84 | 0.69 | 0.65 | 0.82 | 0.71 | 0.57 | 0.44 |
| repressor_of_CAR1 | 0.09 | 0.08 | 0.61 | 0.48 | 0.58 | 0.10 | 0.12 | 0.08 | 0.14 | 0.11 | 0.37 | 0.64 | 0.57 | 0.65 | 0.08 | 0.20 | 0.18 | 0.26 | 0.64 | 0.07 |
| RLM1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| RME1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| ROX1 | 0.44 | 0.43 | 1.58 | 0.69 | 0.57 | 0.55 | 0.47 | 0.43 | 0.50 | 0.63 | 0.96 | 0.58 | 0.57 | 1.76 | 0.44 | 0.58 | 0.70 | 0.87 | 0.85 | 0.44 |
| SCB | 0.91 | 0.83 | 0.83 | 0.94 | 0.92 | 1.57 | 0.73 | 1.26 | 0.86 | 1.15 | 1.16 | 1.41 | 0.93 | 1.01 | 0.77 | 0.93 | 0.87 | 0.89 | 0.86 | 0.86 |
| SFF | 0.40 | 0.44 | 0.60 | 0.66 | 1.18 | 0.46 | 0.45 | 0.44 | 0.60 | 0.48 | 0.83 | 0.62 | 0.77 | 2.04 | 0.45 | 0.57 | 0.57 | 0.73 | 0.46 | 0.30 |
| SMP1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| STE12 | 1.77 | 2.02 | 1.04 | 1.17 | 2.67 | 1.11 | 1.29 | 1.04 | 1.12 | 1.49 | 1.34 | 1.36 | 1.12 | 1.01 | 0.92 | 1.66 | 1.12 | 1.25 | 1.61 | 1.16 |
| STRE | 0.86 | 1.07 | 2.47 | 3.72 | 1.46 | 1.50 | 0.88 | 0.80 | 0.90 | 1.06 | 1.35 | 1.11 | 0.99 | 0.90 | 1.25 | 1.43 | 1.69 | 1.15 | 0.95 | 0.87 |
| SW15 | 1.14 | 1.13 | 1.20 | 2.17 | 1.18 | 1.68 | 1.02 | 0.87 | 2.07 | 0.94 | 2.74 | 2.25 | 1.06 | 1.24 | 1.32 | 1.10 | 1.07 | 1.13 | 1.31 | 0.86 |
| T4C | 0.55 | 0.17 | 0.32 | 0.45 | 0.44 | 0.22 | 0.26 | 0.58 | 0.47 | 0.23 | 0.53 | 0.29 | 0.81 | 0.28 | 0.16 | 0.42 | 0.44 | 1.01 | 0.47 | 1.39 |
| TBP | 1.12 | 3.15 | 1.26 | 0.37 | 1.32 | 4.26 | 1.65 | 1.22 | 2.65 | 1.73 | #NUM! | 1.56 | 1.57 | 1.33 | 4.37 | 1.30 | 1.44 | 1.70 | 1.66 | 2.38 |
| TEA1 | 0.12 | 0.10 | 0.19 | 1.43 | 0.52 | 0.13 | 0.16 | 0.10 | 0.17 | 0.63 | 0.44 | 0.18 | 0.22 | 0.58 | 0.75 | 0.25 | 0.24 | 0.45 | 0.17 | 0.09 |
| UASINO | 0.46 | 0.27 | 0.44 | 1.91 | 0.56 | 0.45 | 0.40 | 0.26 | 0.43 | 0.83 | 0.78 | 0.43 | 0.45 | 0.43 | 0.49 | 0.60 | 0.46 | 1.06 | 0.44 | 0.22 |
| UASPHR | 0.83 | 0.95 | 1.58 | 102 | 1.03 | 0.83 | 1.36 | 1.19 | 0.90 | 0.83 | 1.09 | 0.90 | 1.06 | 0.86 | 0.81 | 2.35 | 1.18 | 1.12 | 0.98 | 1.25 |
| UASRAD | 0.14 | 0.13 | 0.23 | 0.43 | 0.48 | 0.16 | 0.54 | 0.67 | 0.21 | 0.17 | 0.46 | 0.22 | 0.27 | 0.53 | 0.68 | 0.46 | 0.47 | 0.44 | 0.52 | 0.11 |
| URSRHR | 0.01 | 0.01 | 0.02 | 0.04 | 0.02 | 0.01 | 0.02 | 0.01 | 1.37 | 0.02 | 0.05 | 0.02 | 0.02 | 0.02 | 0.01 | 0.03 | 0.03 | 0.04 | 0.02 | 0.01 |
| XBP1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

(The shaded blocks have lower number of genes than expectation frequency)

석 결과를 앞서 설명한 클러스터의 기능할당 분석결과에서 얻어진 단백질 합성 관련 클러스터에 대해서 적용시켜 보았다. 이 경우 3가지 클러스터링 방법 모두에 대해 RAP1 (Transcription control)와 GCR1(Carbohydrate utilization) 모티프가 상대적으로 유의성이 높게 나타났다. 이러한 결과는 이 두가지 모티프가 서로 연관성이 높다는 기존의 데이터와 잘 일치하고 있다(http://cgsigma.cshl.org/cgi-bin/jz/getfactorlist). 하지만 클러스터의 기능할당 정도의 유의성이 높다고 하여 반드시 모티프 분석에서도 높은 유의성을 갖는 모티프가 나타나는 것은 아님을 알 수 있다. 이러한 이유 때문에 동일한 모티프를 갖고 있는 유전자들 중심으로 유전자 발현 데이터를 분석하는 시도가 이루어지고 있으며, 이러한 분석결과는

기존의 클러스터링 방법과는 다른 정보를 제공하게 된다 (19,20). 본 논문에서는 기존에 알려진 모티프에 대하여 분석하였지만, 각 클러스터내 새로운 모티프를 찾기 위해서는 MEME나 HMMER등을 사용하여 분석할 수 있다(21).

## 요 약

효모의 세포주기 관련 유전자 발현 통합 데이터를 사용하여 본 연구실에서 개발한 유전자 발현 통합 분석프로그램을 사용하여, 클러스터링 알고리즘의 성능을 비교하고 데이터내에 존재하는 클러스터 개수를 추정하기 위해 FOM 분석을 적용하였으며, 이 분석방법을 통하여 K-means, SOM, Fuzzy c-means 클러스터링 방법의 성능을 서로 비교하였다. 클러스터 개수를 추정한 다음 3가지 클러스터링 방법에 대한 클러스터링 결과 비교, 클러스터의 기능할당 및 모티프 분석을 시도하였다. 본 논문에서 제시하는 분석 방법은 DNA chip 발현 데이터의 일반적인 분석방법으로 유전자 발현 패턴의 유사성을 토대로 한 클러스터링 방법에 근간을 두고 있다. 본 논문에서는 클러스터링한 후 각 클러스터의 기능할당 및 모티프 분석에 대한 일반적인 분석방법을 제시하였으며, 본 연구실에서 개발한 유전자 발현분석 통합 프로그램이 효율적으로 사용될 수 있음을 보여주고 있다.

## REFERENCES

1. Herzel, H., D. Beule, S. Kielbas, and J. Korbel (2001), Extracting information from cDNA arrays, *CHAOS*, **11**(1), 98-107
2. Quackenbush, J. (2001), Computational analysis of microarray data, *Nature Reviews Genetics*, **2**, 418-427
3. Jain, A. K., M. N. Murty, and P. J. Flynn (1999), Data clustering:A review, *ACM Computing Surveys*, **31**(3), 264-323
4. Sherlock, G.(2000), Analysis of large-scale gene expression data, *Current opinion in immunology*, **12**, 201-205
5. Hastie, T., R. Thbshirani, et al. (2000), 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns, *Genome biology*, **1**(2), 1-21
6. Holter, N. S., M. Mitra, et al. (2000), Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *PNAS*, **97**(5), 8409-8414
7. Holter, N. S., A. Maritan, et al. (2001), Dynamic modeling of gene expression data, *PNAS*, **98**(4), 1693-1698
8. Terrence, S. F., N. Cristianini., et al. (2000), Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**(10), 906-914
9. Zubritsky, E. (2000), Spotting a microarray system, *Americal Chemical Society*, December 1, 761-767
10. Tseng, G. C., M. K. Oh, L. Rohlin, J. C. Liao and W. H. Wong (2001), Issues in cDNA microarray analysis: quality filtering, channel normalization, modes of variations and assessment of gene effects, *Nucleic Acids Research*, **29**(12), 2549-2557
11. Kuklin, A., S. Shah, B. Hoff, and S. Shams(2000), Information processing issues and solutions associated with microarray technology, *LRA*, **12**, 317-327
12. Claverie, J. M. (1999), Computational methods for the identification of differential and coordinated gene expression, *Human Molecular Genetics*, **8**(10), 1821-1832
13. Raychaudhuri, S., P. D. Sutphin, J. T. Chang and R. B. Altman (2001), Basic microarray analysis: grouping and feature reduction, *Trends in Biotechnology*, **19**(5), 189-193
14. Zhu, J. and M.Q. Zhang (1999), Cluster, function and promoter: Analysis of yeast expression array, *PSB*, 476-487
15. Yeung, K. Y., D. R. Haynor, and W. L. Ruzzo (2001), Validating clustering for gene expression data, *Bioinformatics*, **17**(4), 309-318
16. Tibshirani, R., G. Walther, and T. Hastie (2001), Estimating the number of clusters in a dataset via the Gap statistic, *JRSS-B*, 1-21
17. Yang, Y. L. and C. G Hur (2001), Program development of integrated expression profile analysis system for DNA chip data analysis, *Korean J. Biotechnol. Bioeng.*, **16**(4), 381-388
18. Lee, J. K. (2001), Statistical bioinformatics for gene expression data, *The 2th International Symposium on Bioinformatics*, 103-124
19. Jakt, L. M., L. Cao, K. S. E. Cheah, and D. K. Smith (2001), Assessing clusters and motifs from gene expression data, *Genome Research*, **11**, 112-123
20. Pilpel, Y., P. Sudarsanam, and G.M. Church (2001), Identifying regulatory networks by combinatorial analysis of promoter elements, *Nature Genetics*, **29**, 153-159
21. Gibas, C and P. Jambeck (2001), Developing Bioinformatics Computer Skills, 205-214, O'Reilly & Associates, Inc.