한국생물공학회지 제16권 제4호
Korean J. Biotechnol. Bioeng.
Vol. 16, No. 4, 381-388(2001)

DNA칩 데이터 분석을 위한 유전자발현 통합분석 프로그램의 개발

양 영 렬 · †허 철 구
한국생명공학연구원 유전체 연구센터 바이오인포메틱스 Lab.
(접수 : 2001. 7. 16., 게재승인 : 2001. 8. 23.)

Program Development of Integrated Expression Profile Analysis System for DNA Chip Data Analysis

Young Lyeol Yang and Cheol Goo Hur†
Bioinformatics Laboratory, Genome Research Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoun-dong, Yusong-gu, Daejon 305-333, Korea
(Received : 2001. 7. 16., Accepted : 2001. 8. 23.)

A program for integrated gene expression profile analysis such as hierarchical clustering, K-means, fuzzy c-means, self-organizing map(SOM), principal component analysis(PCA), and singular value decomposition(SVD) was made for DNA chip data anlysis by using Matlab. It also contained the normalization method of gene expression input data. The integrated data anlysis program could be effectively used in DNA chip data analysis and help researchers to get more comprehensive analysis view on gene expression data of their own.

Key Words : DNA chip, expression profile analysis, hierarchical clustering, K-means, fuzzy c-means, SOM, PCA, SVD, matlab

서 론

1990년에 시작된 인간유전체연구(Human Genome Project)의 급속한 진전에 따라 현재의 생물학 연구는 유전체를 기반으로 한, 생물을 통합적으로 분석, 이해하는 방향으로 연구의 패러다임이 바뀌고 있다. 또한, 이러한 연구를 가능하게 하는 새로운 생물학적 기술들-예를 들면 DNA 칩 기술(1), 프로테오믹스(proteomics)(2) 등-의 개발들이 뒤따르고 있으며, 이러한 방법들은 기존의 생물학적 연구방법과는 달리 대규모 데이터의 생산, 분석을 요구하고 있다. 이런 관점에서 대규모의 생물학적 데이터를 관리, 저장하고, 이들을 분석함으로써 유용한 생물학적 정보를 찾아내는 바이오인포매틱스(bioinformatics)의 기능이 향후 다양한 생물학 연구에 더욱 중요해질 전망이다(3). 각 생물의 유전체 연구를 통해 궁극적으로 알고자 하는 것은 유전체의 서열정보뿐만 아니라 각 유전자가 갖는 생물학적 기능, 그리고 각 유전자간의 상호작용 및 유전자와 환경과의 상호작용에 의해 나타나는 다양한 생명현상을 이해하고, 이것을 이용하는 데 있을 것이다(4-6). 현재까지 각 단계별로 다양한 연구가 진행되고 있으며, 이러한 연구들은 생물을 좀 더 근본적으로 이해하고 활용하는 데 도움을 줄 것으로 생각된다.

본 논문에서는 많은 기능 유전자들이 환경적인 변화에 따라 발현되는 양상을 통합적으로 볼 수 있는 DNA 칩의 데이터 분석 시스템의 개발에 대해 다루고자 한다. DNA 칩은 DNA 염기의 상보적인 결합원리를 이용하여 수 많은 oligo nucleotide나 cDNA와 같은 유전자 탐침(probe)을 고체 표면(예, glass) 위에 심은 것으로 실험 대상이 되는 세포나 조직 내에서 발현되는 mRNA의 발현양상을 볼 수 있게 한다. 이미 Affymetrix, Clontech, NEN 등과 같은 회사는 연구용으로 다양한 DNA칩을 시판하고 있으며, 연구소 및 실험실에서도 생물체의 유전자 발현양상을 보기 위해 cDNA 칩을 만들어 실험에 응용해 오고 있다(7,8). DNA칩의 데이터분석과 관련해서 중요한 부분으로 칩의 이미지(image) 데이터의 분석과 유전자발현 분석이 있다. DNA 칩의 경우 실험의 재현성에 영향을 줄 수 있는 요소들이 많이 있기 때문에 DNA 칩 제작상의 품질관리, 실험조건, 분석조건의 관리가 매우 중요하다고 볼 수 있다(9). DNA 칩의 유전자 발현 양상을 분석하기 위해서 hierarchical clustering(HC), K-means, self-organizing map(SOM), principal component analysis(PCA) 등과 같은 방법들이 응용이 되어왔으며(10-12), 최근에 들어와서 gene shaving(13), singular value decomposition(SVD) (14,15), support vector machine(SVM)(16) 등과 같은 새로운 분석 방법이 개발되고 있다. 많은 상업적인 분석 소프트웨어들은 HC,K-means, SOM, PCA 등의 분석방법을 지원하고 있으며, 인터

†Corresponding Author : Bioinformatics Laboratory, Genome Research Center, Korea Research Institute of Bioscience and Biotechnology, 52 Eoun-dong, Yusong-gu,Daejon 305-333, Korea
TEL : +82-42-860-4478, FAX : +82-42-860-4308
E-mail : hurlee@mail.kribb.re.kr

Table 1. Web sites for getting information about Matlab-based clustering

Name	Web site for downloading	Comments
Genlab	http://genlab.tudelft.nl	Gene expression &Network Laboratory
DCPR	http://neural.cs.nthu.edu.tw/matlab/DCPR	Data clustering and pattern Recognition
SOMtoolbox	http://www.cis.hut.fi/projects/somtoolbox	SOM toolbox for Matlab

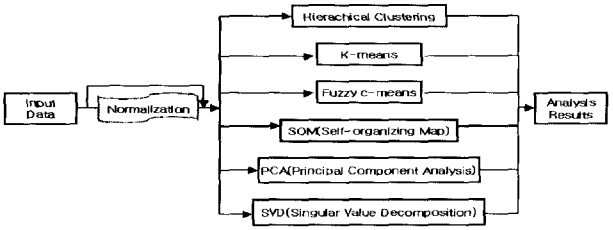


Figure 1. Major components in integrated gene expression profile analysis program.

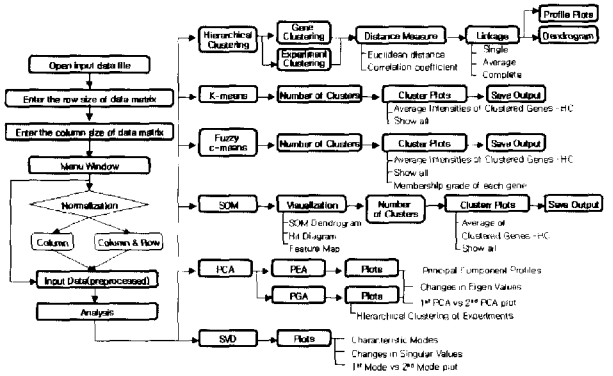


Figure 2. Executive flow diagram of integrated gene expression profile analysis program. (PEA:Principal Experiment Analysis, PGA: Principal Gene Analysis)

넷을 통하여 freeware로 사용할 수 있는 Xcluster, J-express 등도 이 정도 범위의 분석을 지원하고 있다(17). 클러스터링 방법은 유전자 발현 패턴의 유사성을 토대로 분석하는 방법으로 DNA 칩 데이터의 1차적인 분석방법이 되며, 이것을 바탕으로 미지의 유전자의 기능예측 및 클러스터내 유전자들의 전사조절 인자의 모티프 서열탐색, 유전자 네트워 구성 등에 활용된다.

본 논문에서는 매트랩(Matlab)(18)을 기반으로 인터넷에서 사용할 수 있는 자원들을 모아 주어진 DNA 칩 데이터의 HC, K-means, Fuzzy c-means, SOM, PCA, SVD 등의 통합적인 유전자발현 분석 프로그램을 개발한 과정과, 기존의 논문에 보고된 효모의 sporulation 데이터를 가지고 통합분석 프로그램를 수행한 결과를 제시하고, 결론적으로 이 프로그램이 실제적인 DNA 칩의 데이터 분석에 효율적으로 응용될 수 있음을 보이고자 한다.

재료 및 방법

유전자 발현 통합 분석 프로그램 개발을 위해 매트랩(Matlab ver. 6.0, MathWorks)을 사용하였으며, 프로그램 개발을 위하여 인터넷을 통하여 다운로드받아 참고 및 응용이 된 매트랩 M파일들은 Table 1과 같다. 매트랩은 행렬(Matrix) 연산을 기본으로 다양한 분석 도구 상자(toolbox)을 제공하고 있어, 수치해석, 제어시스템의 설계 등 공학분야에서 많이 활용되고 있으며, 행렬 형태의 유전자 발현 데이터의 해석에 효과적으로 활용될 수 있다.

DNA 칩 통합분석 프로그램은 기존에 많이 사용되고 있는 HC, K-means, SOM, PCA와 이외에 Fuzzy c-means 방법, 최근에 발표된 SVD방법이 포함되도록 하였으며, 매트랩에서 그래픽 유저 인터페이스(GUI) 형태의 프로그램으로 개발하였다.

분석 프로그램의 구성도 및 실행도를 Figure 1과 Figure 2에 나타내었다.

통합 분석 프로그램은 하나의 데이터 파일을 다양하게 분석할 수 있도록 하였으며, 기존에 발표된 각종 분석 방법들의 결과들을 쉽게 볼 수 있도록 개발하였다. 클러스터링(clustering)을 위해 입력 데이터의 정규화가 필요할 경우 열 기준과 행/열 기준 두 가지가 가능하도록 하였다. 특히, HC의 경우 유전자(Gene)와 실험에 대해 둘다 가능하도록 하였으며, 실험의 HC은 PCA의 Principal Gene Analysis(PGA)의 결과를 가지고도 할 수 있도록 디자인하였다. PCA도 Principal Gene Analysis(PGA)와 Principal Experiment Analysis(PEA)가 둘 다 가능하도록 하였으며, SOM의 경우 기존의 상업적인 소프트웨어 및 freeware 소프트웨어들이 지원하지 않는 feature map, Best Matching Unit(BMU), Hit Diagram 등도 지원하도록 하였다. 그리고, K-means, Fuzzy c-means(19), SOM에 의한 클러스터링의 경우 각 클러스터내 유전자들의 평균 발현 패턴의 HC이 가능하도록 하였으며, 이것은 클러스터간의 유연관계에 대한 정보를 제공하게 된다. Fuzzy c-means 클러스터링의 경우에는 각 유전자가 어떤 클러스터내에 속하는 정도를 나타내는 소속 정도(membership grade)를 볼 수 있도록 하였으며, 이것은 하나의 유전자가 어떤 클러스터와 어느 정도 연관되어있는 지에 대한 정보를 줄 수 있다. 유전자 발현 통합 분석 프로그램의 최종 결과들은 그래프 형태로 볼 수 있도록 하였으며, 각 유전자별 해당 클러스터에 대한 최종 결과는 텍스트 파일로 저장하여 사용할 수 있도록 하였다. 통합 분석 시스템의 분석능을 알아보기 위해 스탠포드 대학의 Pat Brown 랩에서 1998년 사이언스지에 발표한 효모의 포자형성관련 DNA 칩 발현 데이터(http://cmgm.stanford.edu/pbrown/sporulation/) 를 사용하였다. DNA 칩 데이터는 6118개의 효모 유전자에 대해 질소원 제한후 포자가 형성되어가는 과정을 7개의 시간(0/0.5/2/5 /7/9/11.5 시간)에 대해 유전자 발현을 나타낸 데이터이다(20).

결과 및 고찰

Figure 2의 순서에 따라 매트랩 기반의 유전자 발현 통합 분석시스템의 그래픽 유저 인터페이스 중요 화면을 Figure 3

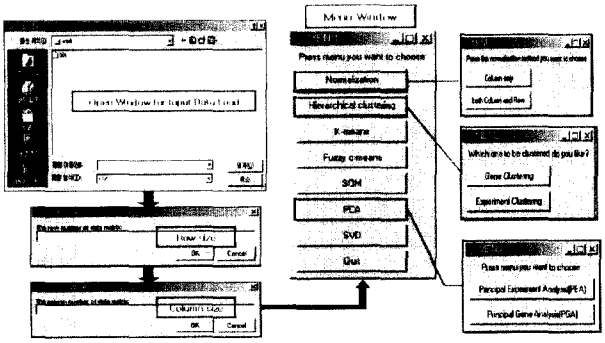


Figure 3. Graphics user interface of integrated gene expression profile analysis program.

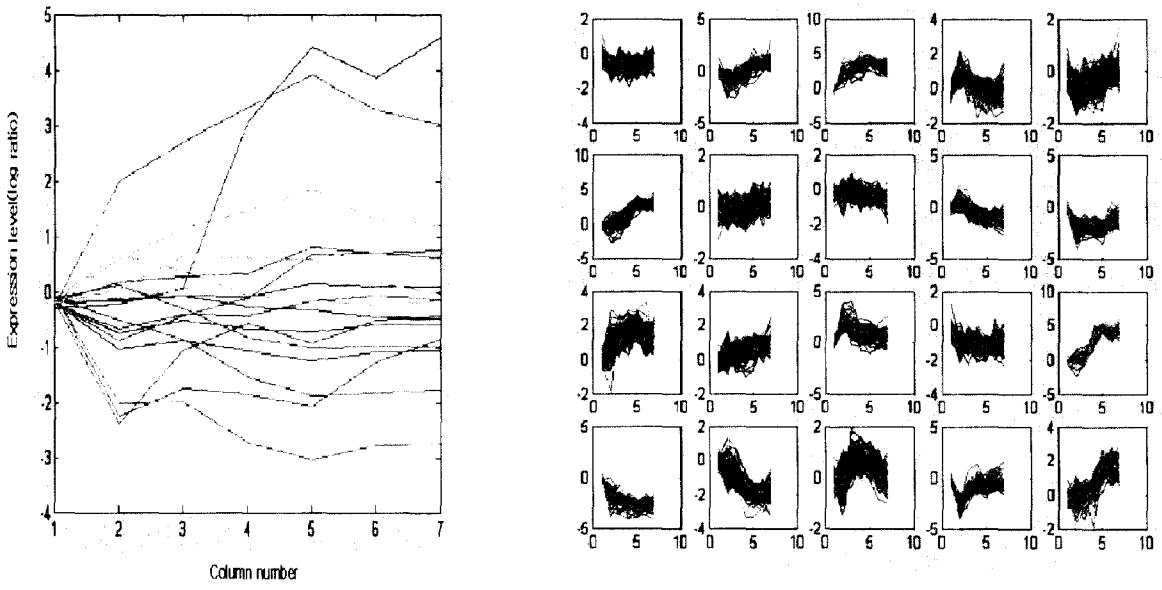
에 나타내었다. 효모의 포자형성 DNA 칩 데이터 파일을 입력한 후 행과 열의 개수를 입력하게 되면, Normalization/HC/K-means/Fuzzy c-means/SOM/PCA/SVD/Quit 등의 메뉴화면이 뜨게 되고, 여기서 필요한 분석 방법을 선택하면 주어진 데이터에 대한 분석이 가능하다. K-means, SOM, Fuzzy c-means, PCA, SVD 등은 6118개 전체 유전자에 대한 데이터를 가지고 분석하였으며, HC의 경우 clustering dendrogram을 전체 유전자에 대해 보기가 어렵기 때문에 효모 포자형성 데이터의 일부분인 116개의 유전자 데이터를 가지고 실행하였다.

전체 데이터에 대한 K-means(12)의 실행결과를 Figure 4에 나타내었다. K-means 클러스터의 개수는 20개로 하였으며, 데이터 정규화는 실시하지 않았다. Figure 4(a)는 클러스터 20개의 그룹 유전자들의 평균 발현 데이터의 양상을 보여준 것이고, Figure 4(b)는 각 클러스터에 대한 전체 유전자들의 발현 양상을 보여주는 것이다.

SOM(21)의 경우도 K-means와 동일한 조건에서 분석하였으며, 이때도 클러스터의 개수는 20개로 하였다. SOM의 결과는 Figure 5, 6에 나타내었다. SOM은 인공지능을 이용한 클러스터링 방법으로 데이터 map에 대하여 뉴런(neuron)으로 구성된 feature map이 있다. Feature map의 각 뉴런이 SOM 분석 후 각 클러스터의 중심이 되며, 데이터 map의 실제 data와 매핑이 되는데 이 때 거리상으로 가장 가까운 뉴런이 그 data의 BMU가 된다. Hit Diagram은 각 뉴런에 연결된 데이터의 수를 나타내며, feature map profile은 각 뉴런에 연결된 실제 data의 각 좌표별 데이터값들을 시각화한 것으로 profile의 패턴이 유사하면 DNA칩 데이터 행렬의 열(column)의 데이터 패턴이 유사한 것으로 해석할 수 있으며, 이러한 결과는 칩 데이터의 열에 대해 HC를 한 결과와 비교할 경우 유사한 결과를 보여주게 된다. Figure 5(b)를 보면 열 2/3이 그리고 4/5, 6/7이 유사한 패턴을 주는 데 Figure 10 (a)의 HC의 결과와 같은 경향임을 알 수 있다. SOM 툴박스의 경우 다양한 결과 분석 툴을 제공하며, SOM dendrogram, 그리고 각 뉴런간의 거리를 계산해서 distance profile을 보여주는 등 다양한 기능을 사용할 수 있다. 이러한 시각적인 기능들은 클러스터의 개수가 주어진 입력 데이터에 어느 정도 있는 지에 대한 정보도 제공하게 된다.

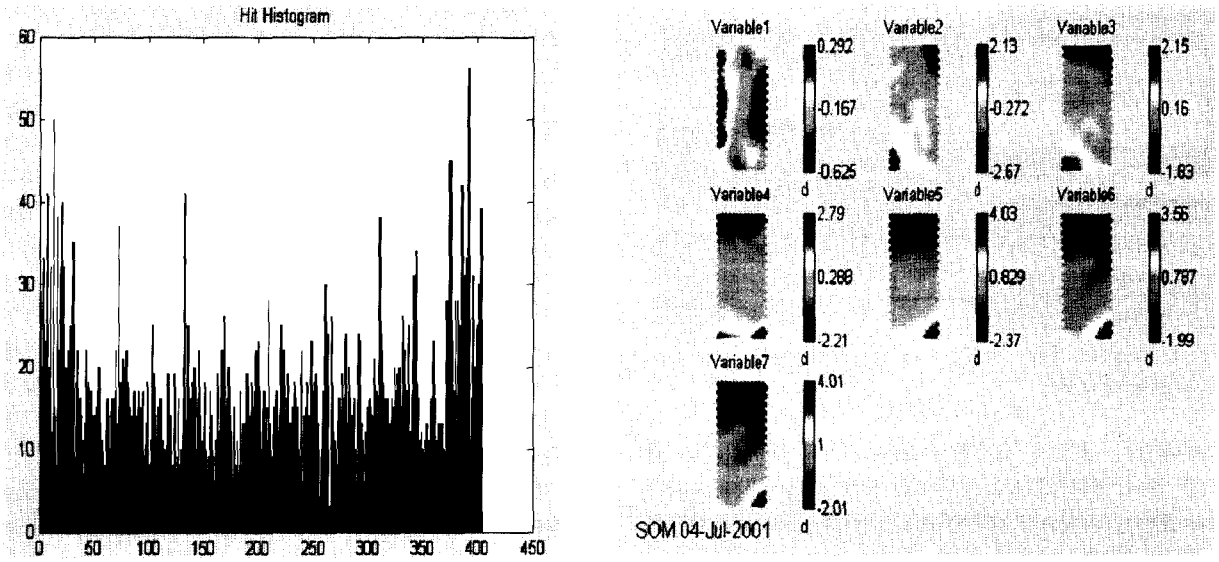
Fuzzy c-means 방법에 의한 클러스터링 결과는 Figure 7,8에 나타내었다. 이 방법은 클러스터링 방법에 퍼지이론을 도입한 것으로 최종 결과는 각 유전자들이 각 클러스터에 속해 있는 정도를 나타내는 소속 정도(membership grade)에 대한 정보와 이것을 토대로 한 클러스터링 결과이다. 본 알고리즘에서는 해당 유전자의 소속 정도가 가장 큰 클러스터를 그 유전자가 속하는 클러스터로 결정하지만, 소속 정도에 대한 정보는 클러스터간의 유사성 및 한 유전자가 여러 클러스터에 속할 수 있는 여지를 제공함으로써 한 유전자가 가지는 다양한 생물학적 기능에 대한 정보를 얻을 수도 있게 된다.

같은 입력 데이터를 가지고, 클러스터의 개수도 동일한 조건(20개)에서 클러스터링을 했지만, Figure 4(a), Figure 6(a),



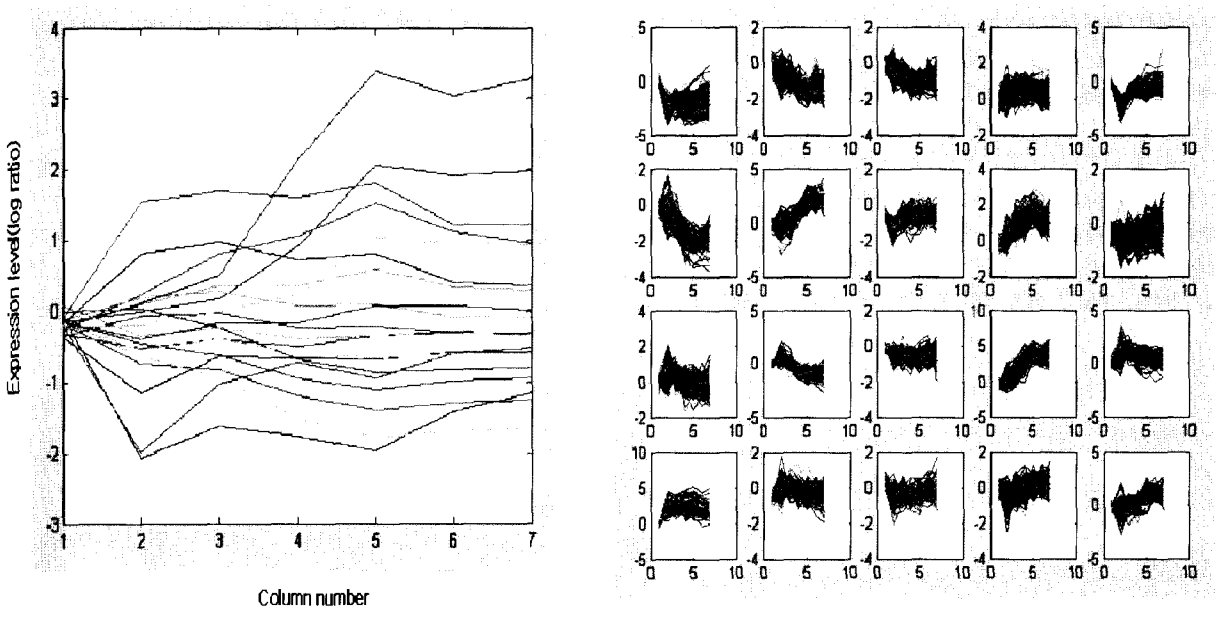
(a) (b)

Figure 4. K-means clustering (a) average gene expression pattern of each cluster, (b) expression patterns of all genes in each cluster. [No nomalization/20 clusters]



(a) (b)

Figure 5. SOM clustering (a) Hit histogram of each neuron in feature map (b) Profiles of each components in each map unit. [No nomalization/feature map size=31*13=403 neurons/20-clusters]

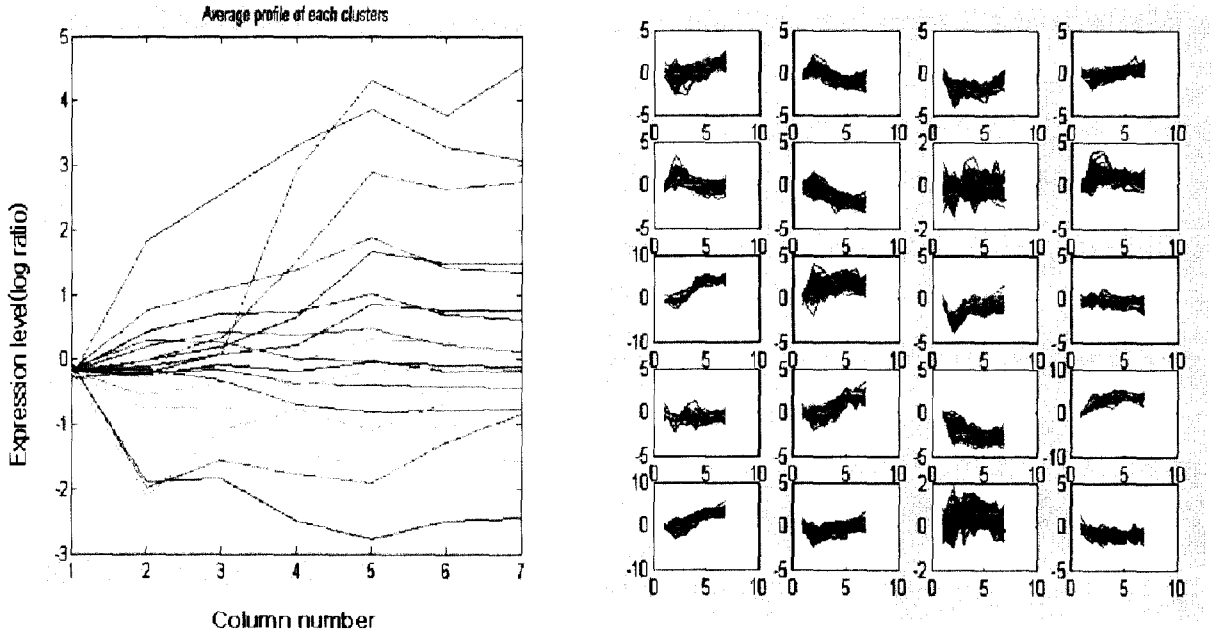


(a) (b)

Figure 6. SOM clustering (a) average gene expression pattern of each cluster, (b) Expression patterns of all genes in each cluster. [No nomalization/20-clusters]

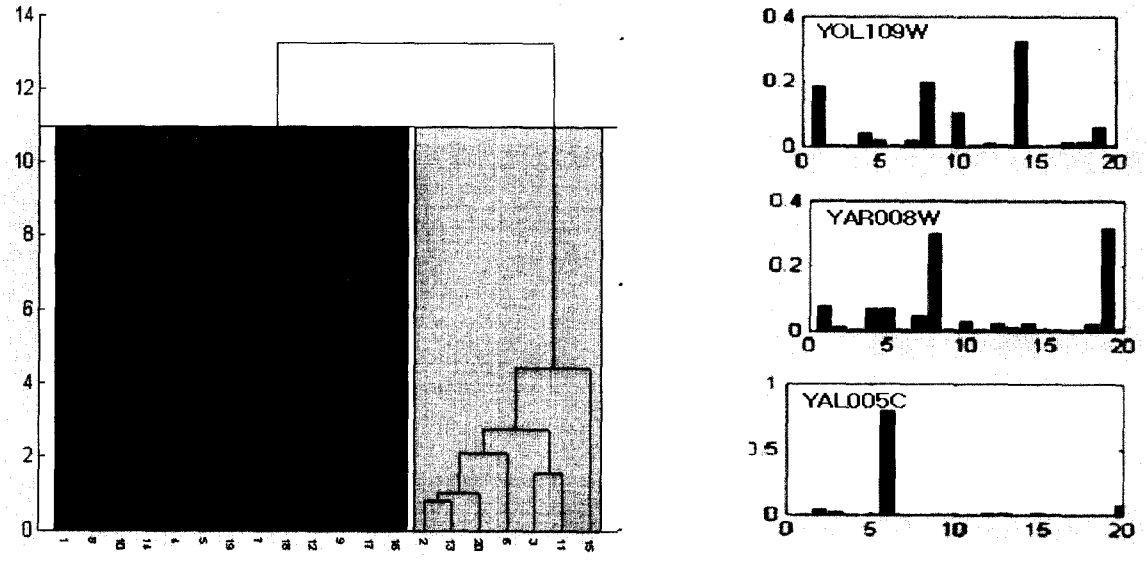
Figure 7(a)에서 보듯이 K-means, SOM, Fuzzy c-means의 결과가 차이가 있는 이유는 이들 알고리듬의 초기화가 랜덤(random)하게 이루어지기 때문이다. 따라서, 어느 경우가 클러스터링의 정확한 답이다라고 생각하는 것은 옳지 않으며 주어진 조건에서 나온 하나의 결과로 해석하는 것이 바람직하다.

Figure 8 (a)는 Fuzzy c-means 분석 결과중에서 각 클러스터의 평균 유전자 발현 패턴을 가지고 HC를 한 결과를 보여준다. K-means나 SOM의 경우 모두 이 방법을 적용시킬 수 있으며 이 방법으로 각 클러스터간의 유연관계를 볼 수 있다. Figure 8(b)는 Fuzzy c-means 분석 결과로 얻을 수 있는 것으로 각 유전자들이 각 클러스터내에 어느 정도 속해 있는지의 정도를 알 수 있다. 한 유전자가 전체 클러스터에 속하는 정도를 0-1의 범위로 표시한 것으로 이 값이 클수록 해당 클러스터에 속하는 정도가 크다고 말할 수 있다. 한 유전자가 각 클러스터에 속하는 소속 정도를 모두 합하면 1이 된다. 이것을 근거로 해서 예를 들면 유전자 YAL005C의 경우 클러스터 6번에 아주 강하게 속한 반면, YOL109W는 클러스터 14번에 속한 정도가 가장 크지만 클러스터 1, 8번에도 어느 정도 속함을 알 수 있다.



(a) (b)

Figure 7. Fuzzy c-means clustering (a) average gene expression pattern of each cluster, (b) expression patterns of all genes in each cluster. [No nomalization/20-clusters]



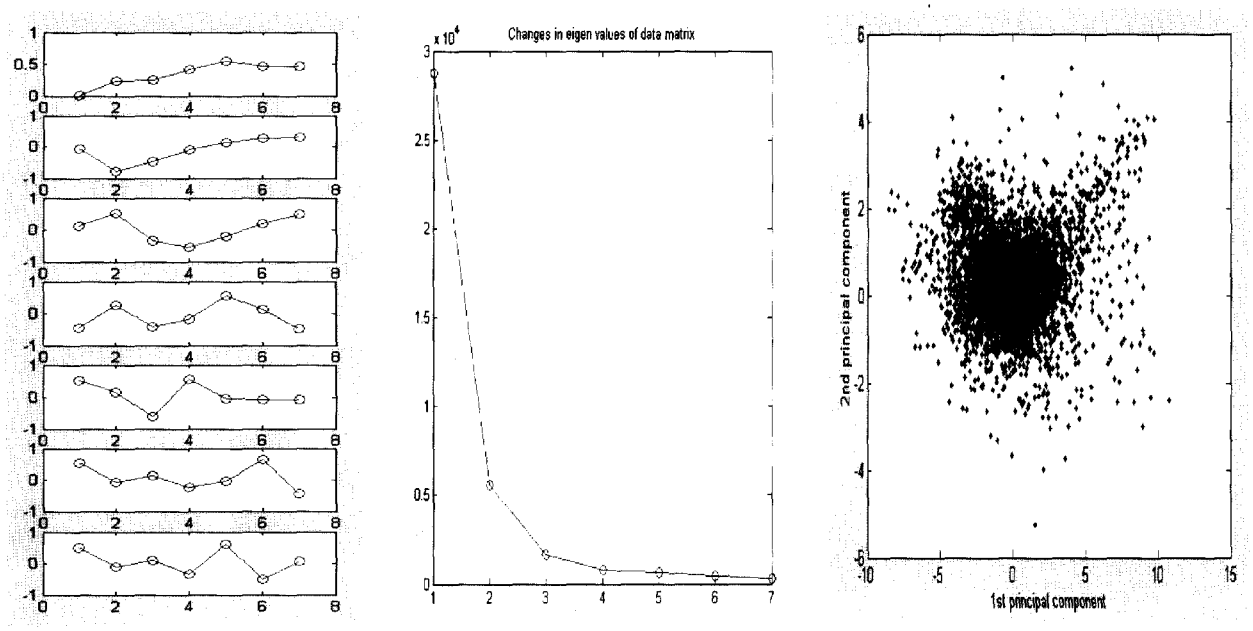
(a) (b)

Figure 8. Fuzzy c-means clustering (a) hierarchical clustering of average patterns of each cluster[Euclidean distance/complete linkage] (b) membership grades of three genes [YOL109W,YAR008W,YAL005C].

유전자 YAR008W의 경우는 클러스터 8, 19에 속하는 정도가 각각 0.2986, 0.3190으로 거의 비슷한 것을 알 수 있다. 이 경우 YAR008W는 클러스터 19번에 해당하는 것으로 결정하게 된다. 하지만 YAR008W가 클러스터 8번의 특성도 비슷하게 가질 것으로 생각해도 되는 것이다. 이러한 결과는 Figure 8(a)와 같이 클러스터의 평균 패턴을 HC를 한 결과와 많이 유사함을 볼 수 있다. Fuzzy c-means 클러스터링을 할 경우 이와 같이 각 유전자들이 속할 수 있는 클러스터에 대한 정보를 얻을 수 있게 된다.

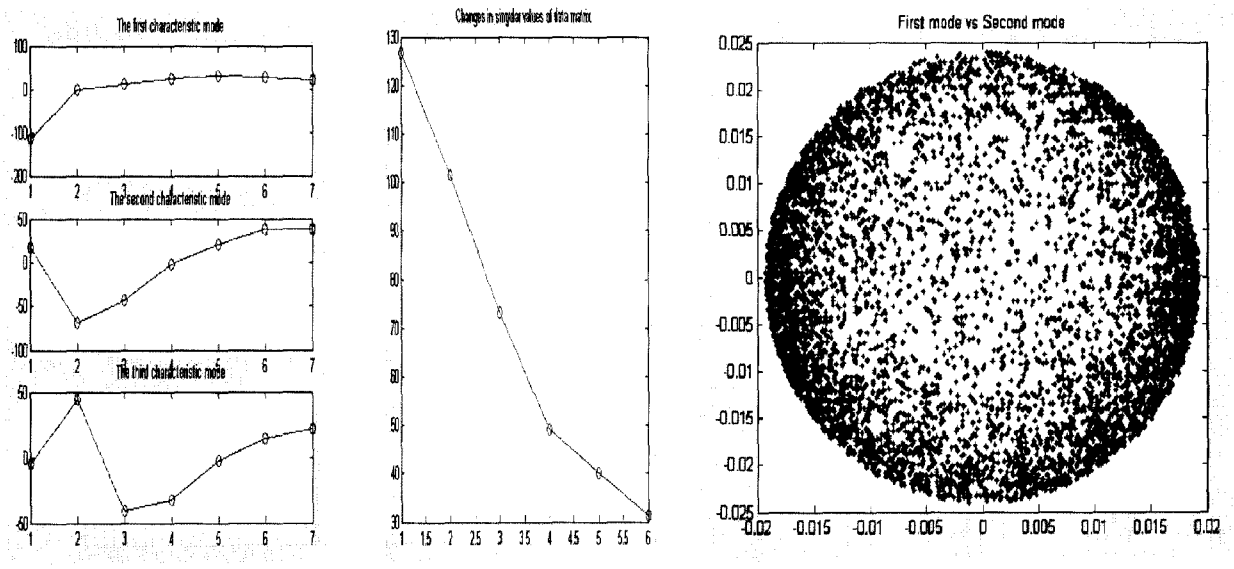
PCA(22-24)는 행렬의 고유값 분석(Eigen Analysis)을 도입하여 변수의 차원을 줄이는 방법으로, 이 방법에 의해 새로운 변수인 principal component (PC)가 생성이 되며, principal component는 기존 변수의 선형조합(linear combination)에 의해 표현되어진다. 이 방법은 다차원변수의 데이터를 저차원(예를 들면, 2,3차원)의 principal component에 대해 표현함으로써 데이터의 분포 형태를 시각화시킬 수 있다(Figure 9(c)). Figure 9에 PCA 분석 결과를 나타내었다.

Figure 9은 PCA 중 PEA에 대해 나타내었으며, PGA를 할



(a) (b) (c)

Figure 9. Pricipal component analysis (a) coefficients of principal components (b) changes in eigen value (c) data plot on between 1st and 2nd principal component. [No data normalization /Principal Expriment Analysis(PEA)]

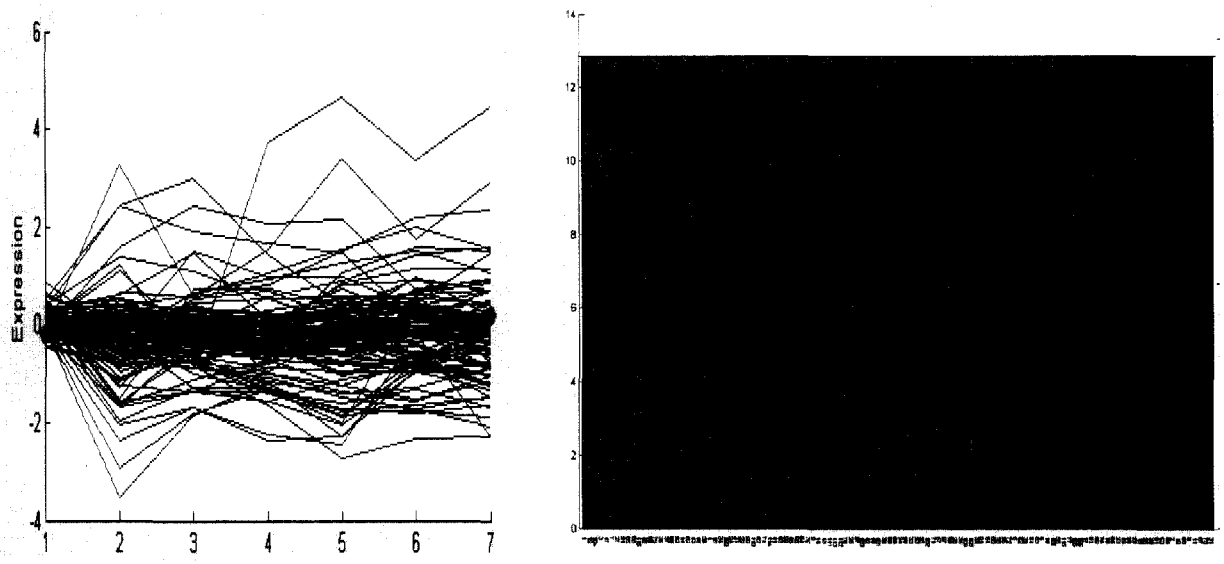


(a) (b) (c)

Figure 10. SVD analysis (a) characteristic modes(1st,2nd,3rd) (b) change in singular values (c) plot of the coefficients for characteristic mode 1 against the coefficients for characteristic mode 2. [Row/column normalization,]

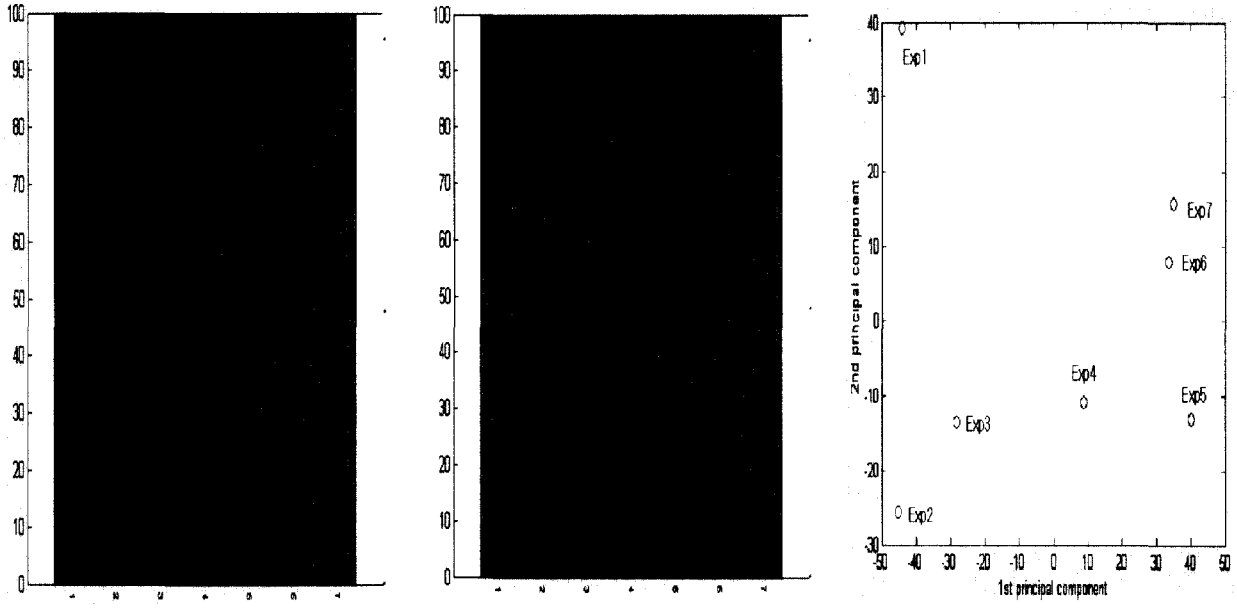
경우 DNA 칩의 각 실험조건의 유전자 발현 패턴의 유사성에 대해 알 수 있다. 각 실험조건의 발현 패턴의 유사성은 HC에 의해서도 알 수 있고(Figure 10(a)), 앞에서 언급했듯이 SOM의 feature map을 가지고도 알 수 있으며(Figure 5(b)), Figure 10(b)와 같이 PGA에 의해서도 알 수 있게 된다.

K-means/SOM과 같은 기존의 클러스터링(clustering) 방법들은 클러스터(cluster)간의 상호관계에 대해 아무런 정보를 주지 못하는 반면에 SVD는 클러스터 간의 상호관계에 대한 정보를 제공할 수 있다. Neal 등(14)은 이 방법을 이용하여 효모의 세포주기(cell cycle) 및 포자형성(sporulation) 관련 클러스터간의 관계를 파악하였으며, 유전자발현 양상이 복잡하고 다양하지만 환경에 대한 세포의 반응으로 나타나는 유전자발현 양상은 크게 단순한 2-3개의 특성 모드(characteristic mode)에 의해 결정되어지며, 이러한 특성모드 몇 개에 의해 전체 유전자 발현 양상을 모사할 수 있음을 밝혔다. 그리고, 주요 특성 모드의 계수(coefficient)를 그래프로 나타내면 클러스터의 특성에 따라 타원형의 형태, 계수의 분산형태가 달라지는 것을 관찰하였다. 이러한 특성 때문에 SVD방법이 다양하게 응용될 수 있다. Figure 10은 효모의 포자형성 데이터의 SVD 분석 결과를 나타낸다. Figure 10(a)는 효모의 포자형성



(a) (b)

Figure 11. Hierarchical clustering (a) expression profiles of 116 genes (b) clustering dendrogram). [distance measure: Euclidean method, complete linkage]



(a) (b) (c)

Figure 12. Clustering results of experimental conditions (a) hierarchical clustering (b) hierarchical clustering after PGA (c) plot of each experiments on between 1st and 2nd principal component in PGA(Exp1(0 hr, Exp2(0.5 hr), Exp3(2 hr), Exp4(5hr), Exp5(7hr), Exp6(9hr),Exp7(11.5hr). [distance measure: Euclidean method, complete linkage]

관련 주요 발현 패턴의 특성 모드 3개를 나타내며, Figure 10(b)는 SVD 결과에서 singular value의 변화를 나타낸다. 이 값들의 편차가 클수록 특성 모드가 두드러지게 나타나게 되며, 소수의 특성 모드에 의해 전체 유전자 발현 양상이 결정되게 된다. Figure 10(c)는 각 유전자 발현 패턴을 결정하는 1st, 2nd 특성 모드 계수(coefficient)를 표시한 것으로 논문에 의하면 클러스터링된 유전자에 대해 이 그림을 그리게 되면 각 클러스터간의 관계및 클러스터의 특성에 대한 정보를 얻을 수 있게 된다(14,15).

마지막으로 HC(25) 분석 결과를 보면, 이 경우에는 6118개 유전자중 116개를 택하여 실시하였다. 너무 많은 유전자를 가지고 할 경우 HC에 의한 dendrogram을 보기가 어렵기 때문에 일부분만을 가지고 시행하였다. HC은 유전자와 실험에 대해 각각 시행할 수 있으며, 패턴의 유사성을 측정하기 위해서 유클리디언 방법(Euclidean method)과 상관계수 방법(correlation coefficient method) 두개가 지원되며, linkage method로 single/ average/ complete 방법이 지원된다. 결과는 Figure 11에 나타내었다.

이상과 같이 하나의 입력 데이터를 매트랩 기반 유전자 발현 통합분석 시스템내의 기능을 사용하여 다양하게 분석할 수 있으며, 각 분석 방법을 통하여 원하는 정보들을 얻을 수 있었다. 특기할 사항은 기존의 HC방법외에 SOM의 feature map을 통하여서도 실험방법들의 유사성을 볼 수 있으며, PGA를 통한 principal component들에 대한 데이터 플롯을 할 경우 유사성의 관계를 시각화시킬 수 있다는 것이다(Figure 12(c))

Figure 12(a), 12(b)의 종축의 번호는 각각 실험조건을 나타내는 것으로 입력 데이타의 실험시간에 관한 것이다. Figure 12(a), 12(b)를 보면 2/3이 그리고 4/5, 6/7이 서로 유연관계가 큰 것으로 나타나는 데, Figure 12(c)에서 이 관계를 확인할 수 있다. 그리고, Fuzzy c-means 클러스터링 방법을 이용하여 각 유전자가 각 클러스터에 어느 정도 속하는지에 대한 정보를 얻도록 하였으며, K-means, SOM, Fuzzy c-means 분석 결과로 얻어지는 각 클러스터의 평균 유전자 발현패턴의 HC을 통하여 클러스터간의 유연관계도 볼 수 있도록 구성하였다.

Figure 4(b), Figure 6(b), Figure 7(b)와 같이 K-means, SOM , Fuzzy c-means 방법에 의해 클러스터링을 할 경우 최종적으로 각 유전자별 입력 발현 데이터와 그 유전자가 속한 클러스터링 번호가 한 행으로 저장될 수 있도록 하였다. 따라서, 클러스터링 번호별로 그룹지어진 유전자들을 쉽게 확인할 수 있도록 하였다.

요 약

DNA칩의 유전자 발현 데이터의 통합적 분석을 위하여 매트랩을 기반으로 한 통합분석 프로그램을 구축하였다. 이 프로그램은 유전자 발현 분석을 위해 일반적으로 많이 쓰는 방법인 Hierarchical clustering(HC), K-means, Self-organizing map(SOM), Principal component analysis(PCA)를 지원하며, 이외에 Fuzzy c-means방법과 최근에 발표된 Singular value decomposition(SVD) 분석 방법도 지원하고 있다. 통합분석프로그램의 성능을 알아보기 위하여 효모의 포자형성(sporulation)과정의 유전자발현 데이터를 사용하였으며, 각 분석 방법에 따른 분석 결과를 제시하였으며, 이 프로그램이 유전자 발현데이타의 통합적인 분석을 위해 효과적으로 사용될 수 있음을 제시하였다.

REFERENCES

1. The chipping forecast(1999), Special supplement to *Nature Genetics*, **21**
2. Naaby-Hansen S, Waterfield MD, and Cramer R.(2001), Proteomics - post-genomic cartography to understand gene function, *Trends Pharmacol Sci.*, **22**(7), 376-84
3. Luscombe N M, Greenbaum D, and Gerstein M.(2001), What is bioinformatics? An introduction and overview, *International Medical Informatics Association Yearbook.*, 83-100
4. Somogyi R.(1997), Level-by level inference from large-scale gene expression data, *Functional Genomics*, **2**(5), 1-16
5. Voit E. O. and Radivoyevitch T.(2000), Biochemical systems anlaysis of genome-wide expression data, *Bioinformatics*, **16**(11), 1023-1037
6. Paul S., Douglas A.B., and John H.B.(2000), Modeling transcriptional control in gene networks-methods, recent results and future directions, *Bulletin of Mathematical Biology*, **62**, 247-292
7. Papers on microarray data analysis: http://linkage.rockefeller.edu/wli/microarray/
8. Stanford Microarray Database: http://genome-www4.stanford.edu/MicroArray/SMD/
9. Mark Schena(2000), Microarray Biochip Technology, A BioTechniques Books Publication, Eaton Publishing
10. Herzel H, Beule D., Kielbas S., and Korbel J.(2001), Extracting information from cDNA arrays, *CHAOS*, **11**(1) ,98-107
11. Quackenbush J.(2001), Computational analysis of microarray data, *Nature Reviews Genetics*, **2**, 418-427
12. Jain A.K., Murty M<.N. and Flynn P.J.(1999), Data clustering:A review, *ACM Computing Surveys*, **31**(3), 264-323
13. Hastie T. and Thbshirani R., et.al (2000), 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns,*Genome biology*, **1**(2), 1-21
14. Neal S. H. Madhusmita M, et.al(2000), Fundamental patterns underlying gene expression profiles:Simplicity from complexity, *PNAS*, **97**(5), 8409-8414
15. Neal S. H Amos M, et.al(2001), Dynamic modeling of gene expression data, *PNAS*, **98**(4), 1693-1698
16. Terrence S. F., Nello C., et.al(2000), Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics*, **16**(10), 906-914
17. Databases and other software tools for Gene Expression, http://www.microarray.nl/others.html
18. Matlab: http://www.mathworks.com/products/
19. Guthke R and Hahn D., et.al(2000), Gene expression data mining for functional genomics, *ESIT* 2000, **14**(15), 170-177
20. Chu S, DeRisi J, Eisen M, and et.alMulholland J, Botstein D, Brown PO,and Herskowitz I.(1998), The transcriptional program of sporulation in budding yeast. *Science*, **282**(5389), 699-705
21. Kohonen, T.(1997), Self-organizing maps, Springer Verlag, 2nd Edition
22. A. van Ooyen(2001), Theoretical Aspects of Pattern Analysis In: *New Approaches for the Generation and Analysis of Microbial Fingerprints* (eds. L. Dijkshoorn, K. J. Towner & M. Struelens), Elsevier, Amsterdam, in press.
23. Johnson R. A. and Wichern D. W.(1998), Applied Mutivariate Statistical Analysis, Prentice-Hall International, Inc., 4th edition
24. Raychaudhuri S., Stuart J. M., and Altman R. B.(2000), Principal component analysis to summarize microarray experiments:Application to sporulation time series, *Pacific Symposium on Biocomputing* 2000, 452-463
25. Claverie J. M.(1999), Computational methods for the identification of differential and coordinated gene expression, *Human Molecular Genetics*, **8**(10), 1821-1832