응용통계연구 제 18권 2호, 2005년, pp. 355-369

행렬도를 이용한 유전자발현자료의 탐색적 분석*

박미라[1)]

요 약

마이크로어레이 실험에서는 유전자의 기능과 상호작용의 이해를 돕기 위한 방안으로 유전자발현자료의 시각화방법이 많이 사용되고 있다. 행렬도는 유전자와 샘플들을 동시에 그려볼 수 있어서, 유전자 또는 샘플의 군집이나 유전자-샘플간 연관작용을 알아보는데 더욱 유용하게 쓰일 수 있다. 본고에서는 마이크로어레이실험에서 행렬도를 이용하여 유전자의 군집 및 연관성을 알아보는 방법을 소개하고, 추가점기법을 이용하여 새로운 샘플을 분류하는 방법을 제안하였다. Golub et al.(1999)의 백혈병데이터와 Alizadeh et al.(2000)의 림프구데이터, Ross et al.(2000)의 NCI60 종양조직데이터를 이용하여 유용성을 살펴보았으며, 계층적 군집분석 및 k-평균 군집분석 등 다른 기법을 이용한 결과와 비교하고 이러한 기법을 행렬도와 연계하는 방안을 살펴보았다.

주요용어: 유전자발현, 마이크로어레이, 행렬도, 추가점기법, 군집분석, 분류

1. 서론

유전체연구 분야에서는 서열분석이나 단백질 구조분석에서부터 마이크로어레이 및 2D-PAGE 등의 기능유전체학 분야에 이르기까지 다양한 주제에 대해 통계학적 문제가 생물학적 문제와 교차되고 있다. 특히 최근 cDNA와 oligonucleotide 마이크로어레이칩(microarray chip) 기술의 발달로 수천 개의 유전자에 대한 발현양상을 동시에 관찰할 수 있게 됨으로써 방대한 자료가 생성되게 되었으며, 복잡한 생물학적 연관을 쉽게 파악할 수 있는 시각화방법의 적용이 필수적이 되었다. 표준화(normalization) 등과 같은 전처리를 거친 유전자발현프로필의 군집방법으로 현재 계층적 군집분석(hierarchical clustering)이나 k-평균 군집분석, SOM (self-organizing map)과 같은 분리기법이 많이 사용되고 있으며, 시각화방법으로는 주로 덴드로그램(dendrogram)을 이용하여 개체간 관계를 표현하는 방법이 사용되고 있다(Tibshirani et al., 1999; Tamayo et. al., 1999). 또한 주성분분석이나 대응분석을 이용하여 군집을 표현하는 시각화도 시도되었다(Raychadhuri et al., 2000; Alter et al., 2000; Kishino and Waddle, 2000; Fellenberg et al., 2001). 계층적 군집분석의 덴드로그램의 경우 같은 덴드로그램이라도 노드의 배열 순서를 달리하면 노드간의 거리가 달라져 보이는 약점이 있으며, 군집하고자 하는 유전자의 수가 많을 때에는 덴드로그램도 복잡해지므로 결과를 한눈에 파악하기가 더 이상 쉽지 않다. k-평균 군집분석이나 SOM 등 자동적인 분리방법에 따른 오류도 발생하며, 서로 배타적인 그룹을 형성하게 되어 여러 세포활동에 동시

* 본 연구는 한국과학재단 목적기초연구(R05-2003-000-11954-0)지원으로 수행되었음.

1) (301-832) 대전시 중구 용두동 143-5, 을지의과대학교 의예과, 조교수
E-mail: mira@eulji.ac.kr

관여하는 유전자의 분석으로는 적절하지 않기도 한다. 또한 이 경우에는 단순히 분류된 군집별로 유전자의 프로파일을 꺾은선 그래프나 색상표로 그리는 이외에는 마땅히 시각화 방법이 없다. 이들은 모두 유전자 또는 샘플 둘 중 하나의 분석결과를 주고, 유전자와 샘플간의 관련을 동시에 보여주지는 않는다. 한편, 대응분석(correspondence analysis)을 적용하면 카이제곱거리로 표현되는 유전자간, 샘플간의 거리를 파악할 수 있고 유전자와 샘플을 대응하여 관계를 살펴볼 수도 있다. 그러나 이는 기본적으로 분할표를 분석하기 위한 방법으로서, cDNA 실험데이터와 같이 음수를 갖게 되는 경우에는 데이터를 모두 양수로 바꾸어서 적용해야 하며, 이 때 어떤 값을 더해주느냐에 따라 결과가 다르게 나타나게 된다.

행렬도(biplot)는 데이터의 행(개체)과 열(변수)을 공간상의 점으로 표현하여 개체간, 변수간의 관계를 파악하게 되며, 또한 행과 열을 중첩하여 개체와 변수간의 관계를 탐색하고 원자료값을 재생성할 수 있는 다변량 분석기법이다(Gabriel, 1971). 본고에서는 마이크로어레이실험에서 행렬도를 이용하여 유전자 및 샘플의 군집과 그들간의 연관성을 알아보는 방법을 소개하고, 추가점기법을 이용하여 새로운 유전자 또는 샘플을 알려진 범주로 분류하는 방법을 제안할 것이다. 마이크로어레이 실험의 유형별로 cDNA 마이크로어레이 실험인 Alizadeh et al.(2000)의 림프구데이터와 Ross et al.(2000)의 NCI60데이터, 그리고 oligonucleotide 마이크로어레이실험인 Golub et al.(1999)의 백혈병 데이터를 이용하여 유용성을 살펴보았으며, 계층적 군집분석 및 k-평균 군집분석 등 다른 기법을 이용한 결과와 비교하고 이러한 기법을 행렬도와 연계하는 방법을 살펴보았다.

2. 마이크로어레이 실험과 데이터 형태

마이크로어레이실험은 1cm가량의 작은 칩위에 수천, 수만 개의 유전자를 붙여놓고 검체에서 추출된 mRNA의 발현 정도를 한 번의 실험으로 조사하는 방법이다. 마이크로어레이는 사용되는 뉴클레오타이드에 따라 cDNA 칩과 oligonucleotide 칩으로 분류될 수 있다.

많이 사용되는 2-channel cDNA 마이크로어레이방식은 EST(Expression Sequence Tages)의 모든 염기서열을 슬라이드에 붙여 상보서열을 가진 유전자를 식별하는 방식으로, 두 개의 다른 환경에서 얻어진 세포로부터 mRNA를 추출하여 이를 역전사시킬 때 각각 다른 색깔의 형광물질을 띤 염기를 집어넣는다. 흔히 실험검체에 적색의 형광물질(Cy5)을, 대조검체에 녹색의 형광물질(Cy3)을 넣어 합성된 두 개의 cDNA를 같은 양으로 섞어서 하나의 cDNA 마이크로어레이칩에 결합시킨다(Brown and Bostein, 1999). 결합이 안된 유전자들을 씻어낸 칩을 레이저로 읽어 각 유전자의 발현강도를 측정하게 된다. R과 G를 각각 배경강도가 조정된 적색과 녹색의 발현강도라고 했을 때 $\log_2(R/G)$를 각 점에 대한 발현강도 데이터로 놓는 경우가 많다.

Affymetrix사가 개발한 oligonucleotide 마이크로어레이방식은 각 oligonucleotide와 25개 염기서열의 중앙점인 13번 염기서열을 변형시킨 oligonucleotide를 나란히 배열하여 결합량을 비교한다(Lipshultz et al., 1999). 각 유전자마다 11 ~ 20쌍의 PM(Perfect match)와 MM(Mis-match)의 차이를 측정하게 된다. 이 때는 레이저판독기로 직접 발현강도를 측정하게 되어, cDNA 마이크로어레이실험에서처럼 두 샘플의 결합의 상대적 비율로 측정되

는 것이 아니고 단일샘플의 절대적 발현량을 측정하게 된다. 각 유전자별로 11 ~ 20쌍의 PM−MM에 대한 평균이 발현데이터로 사용된다.

3. 행렬도와 추가점기법의 응용

n개 mRNA 샘플에 대한 p개의 유전자의 발현정도를 측정했을 때 전처리후 얻어지는 데이터는 다음과 같은 $n \times p$ 행렬

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

로 표현할 수 있다. 이때 x_{ij}는 i번째 mRNA 샘플에서의 j번째 유전자의 발현수준이 된다. 실험에 따라서는 Spellman et al.(1998)이나 Chu et al.(1998)과 같이 p개 유전자를 시간대별로 측정하여 시간에 따른 변화를 측정하기도 한다. 대부분 원 데이터값을 사용하기보다는 중심화 또는 표준화한 데이터로 변환후 사용한다. 행렬 X의 비정칙치분해(SVD;singular value decomposition)는

$$X = UDV^t$$

와 같다. 여기서 U와 V는 각각 $n \times r$, $p \times r$행렬이고 $U^tU = V^tV = I_r$로 직교정규열(orthonormal column)을 갖는다. D는 비정칙치 $\lambda_1, \cdots, \lambda_r$을 원소로 갖는 $r \times r$ 대각행렬이며, r은 X의 계수이다. $0 \le \alpha \le 1$일 때 $D^\alpha = diag(\lambda_1^\alpha, \cdots, \lambda_r^\alpha)$로 정의하고 $G = UD^\alpha$, $H = VD^{1-\alpha}$로 정의하면 $X = UDV^t = GH^t$로 표현할 수 있게 된다. 또한 G와 H의 처음 s개의 열로 구성된 $n \times s$, $p \times s$ 행렬을 각각 $G_{(s)}$와 $H_{(s)}$라고 했을 때 X는

$$\underset{n \times p}{X_{(s)}} \quad = \quad \underset{n \times s}{G_{(s)}} \ \underset{s \times p}{H_{(s)}^t}$$

에 의해 잘 근사된다(Householder and Young, 1938). 이를 이용하여 $G_{(s)}$와 $H_{(s)}$의 각 행을 s차원의 샘플좌표점과 유전자 좌표점으로 표현한다. 흔히 2차원($s = 2$) 또는 3차원($s = 3$) 그래프를 그린다. α값의 선택에 따라 다른 성질의 행렬도가 생성되는데 $\alpha = 1$로 행렬도를 그렸을 때에는

$$XX^t \approx G_{(s)}G_{(s)}^t$$

가 되므로 이 때 그려진 행렬도 중 행그림(샘플그림)에서 찍힌 점들간의 거리는 샘플들사이의 유클리디안 거리에 근사하는 성질을 갖게 된다. 열그림(유전자그림)은 주성분축이 유전자들의 어떤 선형결합으로 형성되는지(주성분계수)를 보여준다. 한편 $\alpha = 0$을 선택한 경우에는

$$X^tX \approx H_{(s)}H_{(s)}^t$$

가 되므로 행렬도의 열그림(유전자그림)에서 유전자들이 이루는 사잇각이 유전자들간의 공분산(또는 상관)을 나타내게 된다. 이때 행그림(샘플그림)에서 점들간의 거리는 샘플들사이의 마할라노비스(Mahalanobis)거리로 근사된다(cf.Gabriel,1971;최용석,1999;허명회,1999).

α값을 어떻게 선택하든지에 관계없이 구해진 행그림과 열그림을 중첩하면 항상 i번째 샘플과 j번째 유전자가 만드는 내적이 유전자발현자료 x_{ij}를 나타내게 된다. 이 때 s차원 그래프에 대한 적합도는

$$GOF_{(s)} = 1 - \frac{|X - X_{(s)}|^2}{|X|^2} = \frac{\sum_{j=1}^{s} \lambda_j^2}{\sum_{j=1}^{r} \lambda_j^2}$$

로 정의된다. 여기서 $|X|^2 = trace(X^t X)$이다.

데이터를 모두 얻게 된 후에 새로운 정보가 추가되는 경우에 이미 그려진 그래프위에 추가점을 덧입힐 수 있다(Lebart et al., 1984). 이러한 추가점기법을 이용하면 알려지지 않은 유전자나 샘플의 속성을 이해할 수 있다. n_s개의 추가 샘플이 있는 경우 추가된 데이터를 $n_s \times p$행렬 $Z_+ = (z_{+ij})$라고 하자. $\alpha = 1$일때 원 그래프가 중심화된 데이터를 사용했을 경우에는 추가점도 마찬가지로

$$x_{+ij} = (z_{+ij} - \bar{x}_j)$$

로 변환하고 표준화한 데이터의 경우에는

$$x_{+ij} = (z_{+ij} - \bar{x}_j)/s_j$$

로 변환하여 변환된 추가행렬 $X_+ = (x_{+ij})$을 만든다. 여기서 $\bar{x}_j$와 s_j는 각각 원 데이터의 행렬도 그림에서 사용된 j번째 유전자의 평균 및 표준편차이다. 관계식

$$XV = (UDV^t)V = UD$$

을 이용하여 X_+V의 처음 s열을 새로운 개체점으로 사용한다. 마찬가지 방식으로 p_s개의 추가된 유전자가 있을 때 $\alpha = 0$인 경우에는 변환된 추가행렬 $(X^+)^tU$의 첫 s개의 열을 좌표로 추가 유전자를 그래프상에 표현하게 된다. 이러한 방법으로 새로운 환자의 샘플이 생겼을 때 기존의 그래프상에 덧찍힌 추가점이 어떤 샘플과 유사한 위치에 찍히는지 파악함으로써 환자의 질환을 분류할 수 있고, 마찬가지로 새로운 유전자를 잘 알려진 유전자들 중 어떤 유형과 유사한지 파악할 수 있을 것이다.

이러한 추가점 기법을 이용하여 성별이나 다른 기전을 표시하는 추가행렬을 생성하여 기존 그래프 위에 덧찍음으로써 유전자나 샘플의 성질을 표시할 수도 있다. 특이치(outlier)가 있는 경우에도 이를 제외하고 행렬도를 그리고 난 후에 특이치를 추가점으로 하여 이의 위치를 표시할 수 있으며, 반복실험을 한 경우 2차 실험결과를 추가자료로 하여 시간에 따른 이들의 변화를 한 그래프에서 확인할 수 있을 것이다.

4. 발현자료 분석 결과

행렬도 방법을 cDNA 칩 및 oligonucleotide 칩을 이용하여 얻은 세 개의 마이크로어레이 데이터에 적용하여 보았다. 유전자발현은 oligonucleotide 칩에 의한 결과일 때는 절대적인 값이지만 cDNA 칩의 경우에는 참조샘플과 비교하여 구해진 상대적인 값이 된다. 이 데이터들은 각 유전자별로 평균을 0으로 중심화시켰으나 각 변수값이 모두 유전자발현강도로 변수들의 측정수준이 모두 같다고 판단하여 분산을 1로 표준화시키지는 않았다. 발생된 결측치에 대해서 k-nearest neighbor algorithm을 사용하여 결측치를 추정하였고 k는 5로 지정하였다(Troyanskaya et al., 2001). 이때 neighbor는 유전자이고 neighbor간의 거리는 유전자간 상관에 근거한 것이다. 행렬도를 위한 분석과 그래프는 SAS/IML 및 SAS/GRAPH, Sigma Plot을 이용하였다. 계층적 군집분석과 k-평균 군집분석 결과는 미국 University of California at Berkeley의 Eisen 연구실에서 만들어진 두 프로그램, Cluster와 Treeview 프로그램을 이용하여 구하였다(cf. http://rana.lbl.gov/index.htm). 사용된 세 개의 데이터는 다음과 같다.

림프구 데이터:

이 데이터는 성인 림프성 질병의 유전자발현연구에서 나온 것으로 cDNA 마이크로어레이 실험에서 얻어진 것이다(Alizadeh et al., 1999). 원 데이터는 96개의 샘플에서 구해진 것이나 여기서는 이 중 세가지 림프성 질병, B-cell chronic lymphocytic leukemia(B-CLL), follicular lymphoma(FL), diffuse large B-cell lymphoma(DLCL)만을 취하여 모두 62개의 샘플(11개의 B-CLL, 9개의 FL, 42개의 DLCL)에 대한 4026개의 유전자 발현값이다. 데이터는 형광강도비에 밑이 2인 로그를 취한 값이다.
(cf. http://genome-www.stanford.edu/lymphoma)

백혈병 데이터:

이 데이터는 총 3571개의 유전자의 발현값으로 구성되어 있으며 세 종류의 샘플로 구분되어 있다. 38개의 B-cell acute lymphomablastic leukemia(ALL)과 9개의 T-cell ALL, 25개의 acute myeloid leukemia(AML)로 나뉘어진다(Golub et al., 1999). 유전자발현수준은 Affymetrix사의 고밀도 oligonucleotide array로 측정된 것이다. 데이터는 Dudoit et al.(2002)에서와 같은 전처리 과정을 거쳐 얻어진 것이다.
(cf. http://www.genome.wi.mit.edu/MPR)

NCI60 데이터:

이 데이터는 미 국립암센터의 항암물질 프로젝트로부터 생성된 것으로 cDNA 마이크로어레이 실험에서 얻어진 것이다(Ross et al., 2000). 유방암(7개), 중추신경계암(5개), 결장암(7개), 백혈병(6개), 흑색종(8개), 폐종양(9개), 난소암(6개), 전립선암(2개), 신장암(9개), 그리고 1개의 알려지지 않은 조직 등 다양한 종양 조직으로부터 나온 세포주(cell line)들의 데이터이다. 데이터는 60개의 샘플과 1375개의 유전자로 구성되어 있으며, 형광강도비에 밑이 2인 로그를 취한 값이다.
(cf. http://genome-www.stanford.edu/nci60)

분석결과 림프구 데이터와 백혈병 데이터의 경우 샘플들의 군집이 2차원 그래프로 잘 분류되고 유전자간, 유전자-샘플간의 해석이 가능하였다. NCI60 데이터의 경우에도 같은 유형의 질병들은 대부분 그래프 상에서 가까운 위치에 놓이는 경향을 보였으며, 유전자간, 유전자-샘플간의 해석이 가능하였다. 그림4.1은 림프구 데이터에서 $\alpha = 1$로 놓았을 때의 샘플그림이다. 가까이 놓인 두 점들은 비슷한 유전자 프로파일을 가짐을 나타낸다. 알려진 세포(cell)의 유형을 색상과 기호로 표현하였는데 DLCL은 적색의 원형, FL은 흑색의 삼각형, B-cell CLL은 청색의 사각형으로 그렸다. 그림에서 보는 바와 같이 대체로 같은 종류의 세포는 비슷한 위치에 다른 형태의 세포는 멀리 찍혔으며, 제1축은 DLCL인지 아닌지를 설명하고 FL과 CLL은 제2축에 의해서 분리, 설명됨을 알 수 있다. 그림4.2는 림프구 데이터에서 $\alpha = 1$로 놓았을 때의 유전자그래프이다. 분석을 위해서는 관심이 있는 유전자들을 별도로 표시하여 위치를 보거나 또는 일정 지역의 유전자를 선택하여 ID를 확인하는 방법으로 탐색해 나갈 수 있을 것이다. 여기서는 발현패턴이 뚜렷한 유전자들 가운데에서 10개를 선택하여 그래프상에 숫자로 표현하였다. 선택된 점들의 유전자명과 clone ID는 1=(Fibronectin 1;clone=139009), 2=(Unknown UG Hs.106127 ESTs;clone=358163), 3=(Fibronectin 1;clone=139009), 4=(FCERI=Fc epsilon receptor gamma chain;clone=235155), 5=(CD10;clone=200814), 6=(CD1C;Clone=428103), 7=(TCL-1;clone=1241524), 8=(TCL-1;clone=200018), 9=(CD23A;clone=1352822), 10=(Unknown;clone=1352493)이다. 비슷한 위치에 놓인 유전자끼리 서로 연관이 있게 된다. 예컨대 유전자 (1,2,3,4)와 (5,6), 그리고 (7,8)과 (9,10)이 각각 비슷한 발현패턴을 가진다는 것을 알 수 있다.

그림4.1과 그림4.2를 중첩하여 보면 세포와 유전자간의 관계를 알 수 있다. 예를 들어 유전자 1-4는 DLCL0002와 DLCL0026방향으로 길게 누워 있으므로 두 세포를 비롯한 DLCL에서 큰 값을 가진다는 것을 알 수 있다. 유전자 1-4를 비롯한 이 부근의 유전자들은 DLCL에서는 많이 발현하고 FL 및 CLL에서는 발현하지 않는 모습을 보인다. Alizadeh et al.(2000)에서의 결과와 비교하여 해석해보면 이들은 "림프노드(Lymph node)"신호로 정의되는 것으로 CD14, CSF-1 receptor와 같이 단핵구(monocytes) 및 대식세포(macrophage)의 표시자로 알려진 유전자들과 Fibronectin과 같이 extracellular matrix의 생성과 재구축에 관여하는 유전자들이다. 유전자 5,6은 FL들의 방향으로는 내적이 크지만 CLL들과의 내적은 작으므로 FL에서 많이 발현하고 CLL에서는 발현이 적어 두 세포를 구분짓는 역할을 한다는 것을 알 수 있다. Alizadeh et al.(2000)의 결과와 비교해 보면 이들을 비롯하여 이와 유사한 패턴의 유전자들이 "배중심 B세포(Germinal Center B cell)"의 특성을 갖는 유전자들로 분류된다. 이들은 휴면 Blood B cell과 활성 Blood B cell 양쪽으로부터 모두 구별되는 유전자그룹으로 알려져 있다. 한편 유전자 7,8의 경우에는 FL10:CD19+와 FL11:CD19+을 비롯한 FL 방향으로 뻗어있어 FL세포에서 많이 발현되며, 유전자 9,10은 CLL68과 CLL71같은 CLL세포에서 더 큰 발현값을 갖는 유전자들임을 알 수 있다. 림프구 데이터에서 2차원 행렬도의 적합도는 37.6%이다. 절대적인 값으로 평가하면 적합도가 낮은 편이나 유전자발현 데이터의 경우에 차원축소 정도를 감안하면 무시할 수 없는 값이라 생각된다.

그림4.3은 중심화된 발현수치를 나타내는 색상표중에서 이 예에서 선택된 10개 유전자

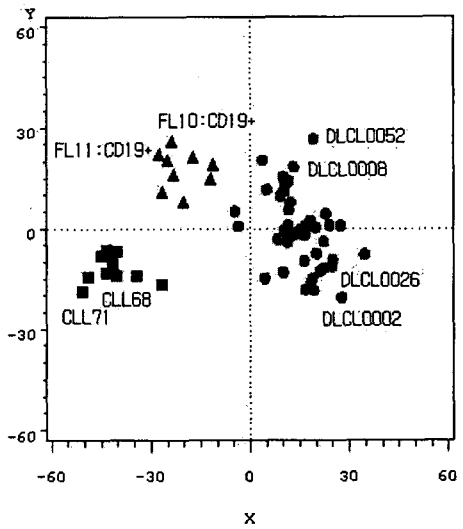

그림 4.1: 림프구 데이터의 샘플그림

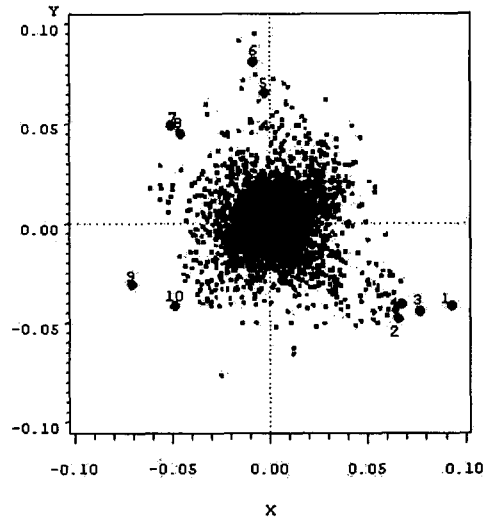

그림 4.2: 림프구 데이터의 유전자그림

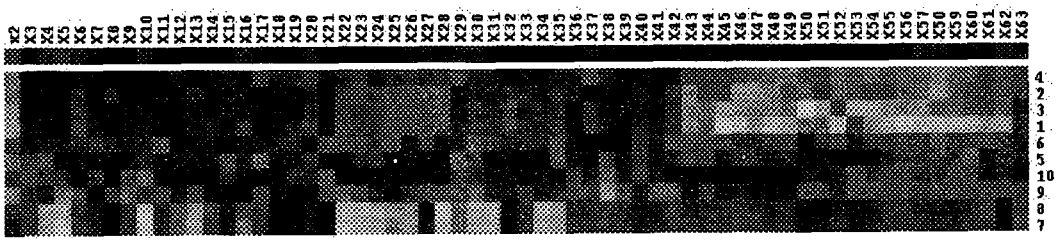

그림 4.3: 림프구 데이터 중 선택된 10개 유전자의 자료
(x2-x42, x63:DLCL, x43-X51:FL x52-x62:CLL)

의 색상만 추린 것이다. 색상이 붉을수록 양으로 큰 값을 가지며, 색상이 푸를수록 음으로 큰 값이고, 검을 수록 발현비가 0에 가까운 값임을 의미한다(흑백인쇄시 밝은 색으로 보여지는 것이 녹색계통의 색상이고 약간 어두운 색이 붉은색, 아주 어두운 것이 검은색이다). 그림에서 x2-x42와 x63은 DLCL 세포이며, x43-x51은 FL 세포이고, x52-x62는 CLL 세포를 가리킨다. 이때 DLCL0002=x29, DLCL0026=x23, FL10:CD19+=x46, FL11:CD19+=x49, CLL68=x53, CLL71=x58과 같다. 오른쪽의 숫자는 그림4.2에서 사용된 유전자번호를 표시한 것이다. 우리가 2차원 행렬도로부터 해석한 내용을 이 패턴에서 확인할 수 있다.

백혈병 데이터의 경우 $\alpha = 1$로 놓았을 때의 샘플그림을 보면 제1축의 양의 방향에 AML이, 음의 방향으로 ALL이 자리잡고 있으며(그림4.4), ALL은 크기에 따라 다시 B cell-ALL 부분과 T cell-ALL부분으로 따로 군집되어 있어 제1축의 해석에 의한 샘플의 구분이 가능하다. 여기서 AML은 흑색의 삼각형, B cell-ALL은 청색의 사각형, 그리고 T cell-ALL은 적색

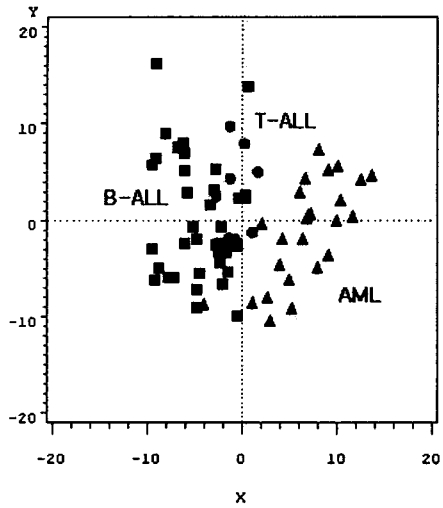

그림 4.4: 백혈병 데이터의 샘플그림

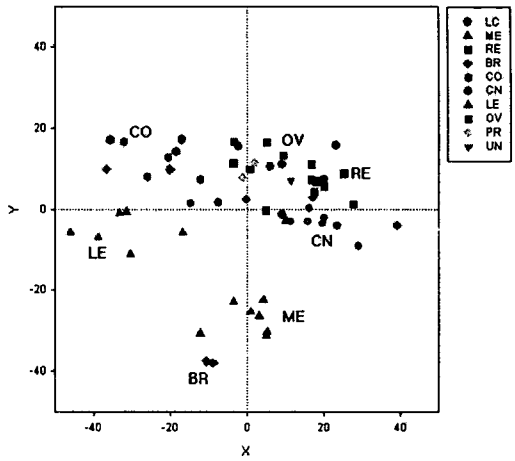

그림 4.5: NCI 60 데이터의 샘플 그림

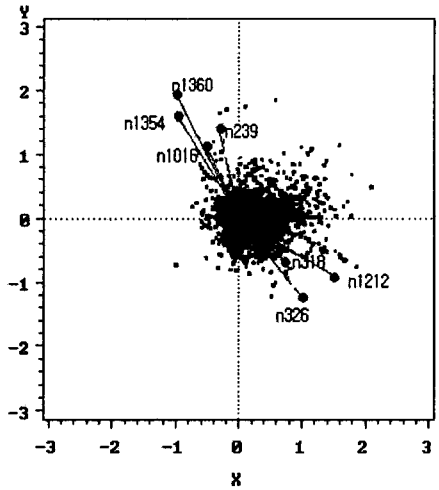

그림 4.6: NCI 60 데이터의 유전자 그림

의 원형으로 표시되었다. 이 그래프의 적합도는 69.4%로 높은 편이다. 그림4.5는 NCI 60데이터에서 $\alpha = 1$ 로 했을 때의 유전자그림으로 흑색종=ME, 폐암=LC, 중추신경계암=CN, 신장암=RE, 유방암=BR, 난소암=OV, 백혈병=LE, 결장암=CO, 전립선암=PR, 알려지지 않은 조직=UN으로 표시되었다. 결장암과 백혈병의 모든 세포주는 제1축의 음의 방향으로 비슷한 위치에 각각 군집되어 있으며, 흑색종은 멜라닌생성부족으로 보고된 사례(LOX-

IMVI)를 제외한 나머지 7개 세포주가 제2축의 음의 방향에 군집되어 있다. 또한 신장암과 중추신경계암의 경우에도 비슷한 위치에 군집되어 있다. 반면 유방암은 이질적인 경향을 보이며 폐암의 경우에도 흩어져 있는 것을 확인할 수 있다.

한편 그림4.6은 NCI 60데이터에서 $\alpha = 0$으로 했을 때의 유전자그림이다. 원점에서 각 유전자를 연결한 선분간의 각도가 작을수록 유전자간의 관련이 있다. 예를 들면 n1016은 n1360, n1354, n239등과 양의 방향으로 관련이 있다. 실제 이들의 상관계수를 구해보면 각각 0.79, 0.81, 0.79로 이러한 해석을 옳다는 것을 알 수 있다. 반면, n1016은 반대방향으로 찍힌 n318, n1212, n326등과는 음의 관련이 있음을 나타낸다. 이들의 상관계수는 각각 -0.61, -0.60, -0.57이다. 한편 각 선분의 길이는 각 유전자의 표준편차의 크기를 나타낸다. 각 유전자들의 실제 표준편차는 n1360=2.86, n1354=2.48, n239=2.00, n1212=2.20, n326=2.09, n1016=1.74, n318=1.40이다. 이 경우에는 2차원 그림의 적합도가 28.9%로 그다지 높지는 않은 편이다.

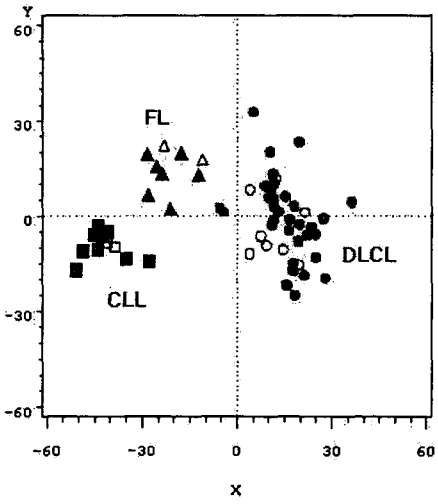

그림 4.7: 림프구 데이터의 샘플추가그림

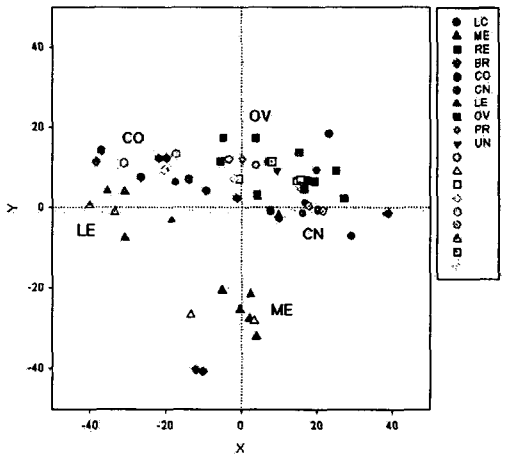

그림 4.8: NCI 60 데이터의 샘플추가그림

제안된 추가점기법의 유용성을 보기 위해서 랜덤하게 원래 데이터의 80%를 골라 행렬도를 작성하고 나머지 20%의 데이터를 새로운 데이터로 간주하여 추가점기법을 이용하여 좌표를 구하였다(그림4.7, 그림4.8). 그림에서 추가된 점들은 속이 빈 도형으로 표시되었으며 색상 및 도형의 모양은 원래 속한 그룹을 의미한다. 림프구 데이터와 NCI 60데이터 모두에서 이 추가점들이 원래의 그룹에 가까이 찍혀져 있음을 알 수 있다. 따라서 추가점들이 원래의 그룹으로 잘 분류되며 샘플에 대한 정보가 없는 새로운 개체를 기존에 알려진 카테고리에 분류하는데 도움을 줄 수 있다는 것을 알 수 있다.

다른 군집분석방법과의 비교를 위해 같은 데이터를 k-평균 군집분석 및 계층적 군집분석방법으로 분석하였다. 그림4.9는 림프구 데이터를 계층적 군집분석의 평균연결(average

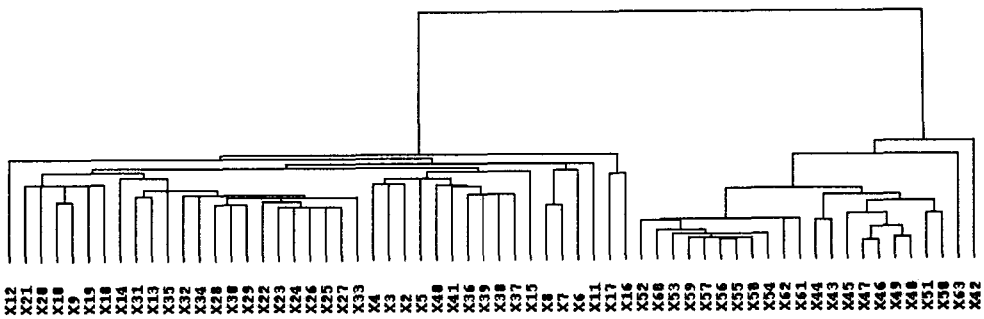

그림 4.9: 림프구 데이터의 샘플 덴드로그램 (average linkage결과)

그림 4.10: 림프구 데이터의 샘플 덴드로그램 (single linkage 결과)

linkage)방법으로 분석한 결과이다. 행렬도와 유사하게 x2-x41이 가까이에 묶여져 있고, x43-x51, x52-x62이 가까이에 묶여져 있음을 확인할 수 있다. 오른쪽 끝의 x42와 x63은 행렬도(그림4.1)에서 원점에 가깝게 찍힌 두 샘플로서 DLCL샘플이지만 FL이나 CLL과 가까이 있었음을 확인할 수 있다. 그림4.10은 같은 데이터를 단순연결(single linkage)방법을 이용하여 분석한 것으로 결과가 조금 다르다.

그림4.11과 그림4.12는 $k = 3$인 k-평균 군집분석을 적용하여 나온 두 개의 결과를 행렬도에 표시한 것이다. k-평균 군집분석에 의해 다른 군집으로 분류된 것은 서로 다른 도형으로 표시하였다. 모두 1000회의 반복을 시행한 것이지만 초기치가 다르게 설정되었으며(랜덤하게 주어짐) 결과가 약간 다르다. 그림4.11의 경우에는 DLCL의 세 샘플(x63, x42, x2)을 제외하고는 원래의 샘플분류에 맞게 분류되어 있지만, 그림4.12의 경우에는 FL과 CLL을 한 그룹으로 묶고 DLCL이 두 그룹으로 나뉘어졌다. k-평균 군집분석이나 SOM등의 분리기법은 이와 같이 항상 같은 결과를 주지는 않고, 또한 결과가 서로 배타적인 그룹을 생성하게 된다. 반면, 행렬도의 경우에는 그래프의 형태를 결정하면 항상 같은 결과를 보여준다는 장점이 있으나 연구자에게 어떤 결론을 내려주지는 않는다. 그림4.11, 그림4.12와 같이 하나의 그래프에 두 방법의 결과를 동시에 표현하면 연구자에게 보다 많은 정보를 줄 수 있어 군집을 판단하는데 도움을 줄 것이다. 유전자의 경우도 마찬가지이다. 유전자들이 그룹별로 뚜렷이 분리되지 않고 공간상에 고루 흩어져 있는 경우에는 k-평균군집분석이나 SOM등의 방법과 같이 각 유전자들을 특정 군집에 할당하는 것은 오히려 혼동을 줄 수 있

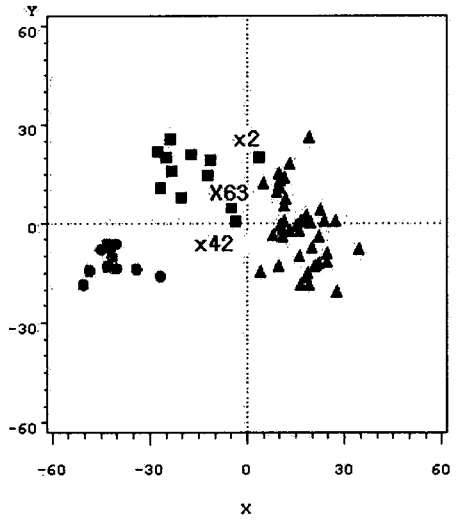

그림 4.11: 림프구 데이터의 샘플그림: 행렬도와 k-평균군집분석결과 ($k = 3$) 합성1

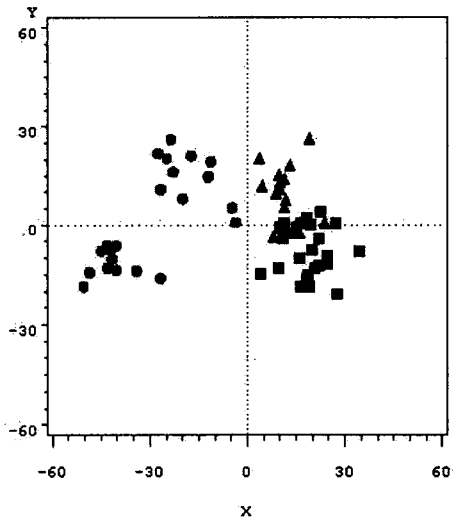

그림 4.12: 림프구 데이터의 유전자그림: 행렬도와 k-평균군집분석 결과 ($k = 3$) 합성2

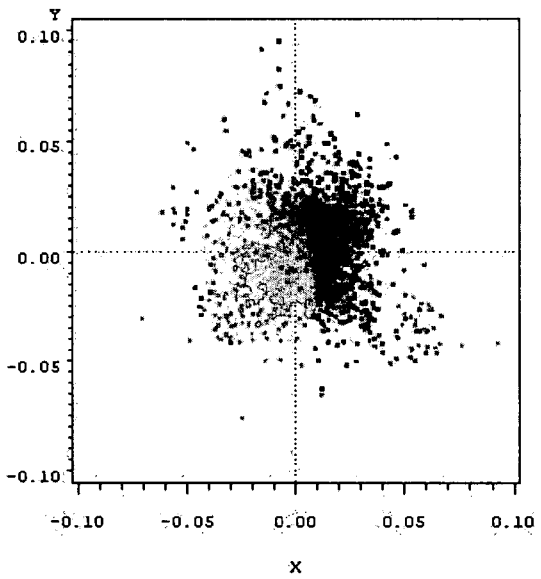

그림 4.13: 림프구 데이터의 유전자그림: 행렬도와 k-평균군집분석 결과 ($k = 10$) 합성

다. 이 때에는 특정 유전자가 어떤 군집에 속하느냐보다는 이웃하는 유전자들이 무엇인지 어떤 샘플들과 관련이 있는지를 알아보는 것이 더 유용하며, 행렬도는 이러한 탐색적 분석에 효과적이다. 그러나 역시 연구자가 유전자를 몇 개 군집으로 할당하고 싶을 때에는 직관적인 정보외에 유전자가 속하는 군집을 결정해주지 않는다는 단점이 있으므로, 두 결과를 같이 표현하는 것이 도움이 될 것이다(그림4.13 참조). k-평균 군집분석이나 SOM의 단점 중 하나는 군집의 개수 또는 그리드의 수를 사전에 정해야 하는 것으로 군집수의 결정은 쉽지 않으며, 군집수 결정을 위한 여러 방법이 있기는 하지만(Dudoit et al., 2002), 만족할 만한 척도가 없는 것이 현실이다. 따라서 많은 경우 군집수를 달리하여 반복적으로 분석할 필요가 있는데, 행렬도의 그래프에 분리방법에 의한 군집을 표시하면 두 개의 결과를 모두 보면서 분류된 군집의 성질을 파악할 수 있어 적정한 군집수를 판단하는 데에도 도움이 될 것이다.

5. 결론 및 토의

DNA 마이크로어레이 실험의 데이터 분석에서 발현프로필의 군집분석은 큰 부분을 차지하고 있다. 방대한 실험자료를 보다 쉽게 이해하기 위해서는 시각적 접근방식이 필수적이다. 분석결과에서 보는 바와 같이 행렬도의 사용은 cDNA 마이크로어레이 실험 및 oligonucleotide 마이크로어레이 실험 모두에서 관련 있는 유전자들을 조사하고, 샘플들을 분류하며, 유전자와 샘플간의 연관관계를 탐색적으로 알아보는데 매우 유용하였다. 또한 추가점 기법을 응용하면 새로운 유전자 또는 샘플의 분류에 활용될 수 있음을 보였다.

데이터에 대한 사전지식이 별로 없을 때 각 분석방법은 나름의 장단점을 가지고 있으므로 여러 방법을 적용해보고 비교를 통해서 의미있는 결과를 도출해내는 것이 필요하다. 하나의 결과가 다른 방법에서의 결과와 유사하면 자연스런 분할로 볼 수 있을 것이다. 또한 여러 방법들을 적절한 순서로 섞어서 사용할 수 있다. 예컨대 계층적 군집분석을 적용한 후 이의 결과를 이용하여 k-평균 군집분석의 초기조건을 개선한다거나(Jain et al., 1999), 먼저 주성분분석을 하여 주요 두 축의 값을 초기치로 하여 SOM을 적용한다거나(Kohonen, 1995) 하는 등의 작업을 생각할 수 있다. 본 고에서 제안한 바와 같이 SOM이나 k-평균 군집분석 같은 분리기법과 행렬도를 하나의 그래프 안에 표현하는 것도 각 방법의 단점을 상호보완하는 길이 되겠다.

행렬도에서 α를 어떤 값으로 놓느냐에 따라 결과는 조금 달라지지만 전반적인 그래프 모양이나 해석은 크게 달라지지 않았다. 하지만 근본적으로 $\alpha = 1$인 경우에는 행공간의 차원축소기법이며, $\alpha = 0$인 경우에는 열공간의 차원축소기법이라고 할 수 있으므로, 샘플들의 분류에 더 관심이 있을 때에는 $\alpha = 1$로, 유전자간의 관계에 보다 관심이 있을 때에는 $\alpha = 0$으로 하는 것이 좋을 것이다.

그 밖에 실험데이터 분석을 위해 선결해야 할 문제가 있었는데 하나는 데이터를 평균이 같도록 중심화만 할 것이냐, 분산까지 같도록 표준화할 것이냐의 문제로 이를 처리하는 방법은 중심화만 하는 경우(cf. Raychaudhuri et al., 2000)와 표준화를 하는 경우(cf. Alon et al., 1999; Getz et al., 2000)로 엇갈린다. 변수간 단위가 모두 같으므로 굳이 표준화를 할 필

요가 없다고 판단되어 본고에서는 중심화만 시킨 데이터를 사용하였다. 또 다른 문제는 무엇을 변수로, 무엇을 개체로 놓느냐의 문제이다. 여러 선행연구에서 유전자를 개체로 간주하고 분석을 하였으나(cf. Alter et al., 2000; Getz et al., 2000; Holter et al.,2000; Landgrebe et al., 2002), 유전자를 변수로 보는 경우도 있었다(cf. Dudoit et al., 2002). 마이크로어레이 실험에서는 샘플수는 작고 유전자수는 크기 마련이므로 유전자를 개체로 하는 것이 데이터 다루기가 한결 수월할 것이다. 그러나 각 샘플이 독립적이고 한 샘플에 대해 여러 유전자를 분석하는 실험의 경우에는 샘플을 개체로 하는 것이 더 타당할 것으로 생각된다.

참고문헌

최용석 (1999). <행렬도의 이해와 응용>, 부산대학교 출판부, 부산.

허명회 (1999). <다변량 수량화>, 자유아카데미, 서울.

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I.., Yang, L., Marti, G.E., Moore, T., Hudson, Jr J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Bostein, D., Brown, P.O., and Staudt, L.M.(2000). Different type of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503-511.

Alon, U., BarKai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci.*, **96** , 6745-6750.

Alter, O., Brown, P. O., Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling, *Proc. Natl. Acad. Sci.*, **97**, 10101-10106.

Brown, P. O. and Bostein, , D. (1999). Exploring the new world of genome with DNA microarrays, *Nature Genetics*, **21**(Suppl.1):33-37.

Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast, *Science*, **282**, 699-705.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77-87.

Fellenberg, K., Hauser, N. C,, Brors, B., Neutzner, A., Hoheisel, J. (2001). Correspondence analysis applied to microarray data, *Proc. Natl. Acad. Sci.*, **98**, 10781-10786.

Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58**, 453-466.

Getz, G., Levine, E. and Domany, E. (2000). Coupled two-way clustering analysis of gene microarray, *Proc. Natl. Acad. Sci.*, **97**, 12079-12084.

Golub, TR., Slonim, DK., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, JP., Coller, H., Loh, ML., Downing, JR., Caligiuri, MA., Bloomfiled, CD., and Lander, ES.(1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531-537.

Householder, A.S., Young, G. (1938). Matrix approximation and latent roots, *American mathematical Monthly*, **45**, 165-171.

Holter, N. S., Mitra, M., Maritan, A., Cieplak, M., Banavar, J. R., Fedoroff . (2000). Fundamental patterns underlying gene expression profiles: Simplicity from complexity, *Proc. Natl. Acad. Sci.*, **97**, 8409-8414.

Jain, A. K., Murty, M. N. and Flymn, P. J. (1999). Data clustering: a review, *ACM Computing Survey*, **31**(3).

Kohonen, T. (1995), *Self-Organizing Map*, Springer-verlag, Berlin.

Kishino, H. and Waddle, P. (2000). Correspondence Analysis of Genes and Tissue Types and Finding Genetic Links from Microarray Data, *Genome Informatics*, **11**, 83-95.

Landgrebe, J., Wurst, W. and Welzl, G. (2002). Permutation-validated principal components analyisis of microarray data, *Genome Biology*, **3**, research0019.1-0019.11.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis*, New York, John & Wiley.

Lipshultz, R.J., Fodor, S., Gengers, T., and Lockhart, D. (1999). High-density synthetic oligonucleotide arrays, *Nature Genetics*, supplement **21**, 20-24.

Raychadhuri, S., Stuart, J.M., Altman, R. B. (2000). Principal components analysis to summarize microarray experiments: Application to sporulation time series, *Pac Symp Biocomput*, 455-466.

Ross, DT., Scherf, U., Eisen, MB., Perou, CM., Rees, C., Spellman, P., Iyer, V., Jeffrey, SS., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, JC., Lashkari, D., Shalon, D., Myers, TG., Weinstein, JN., Botstein, D., and Brown, PO. (2000). Systematic variation in gene expression patterns in human cancer cell lines, *Nat Genet.*, Mar24-3; 227-235.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B. (1998). Comprehensive identification of cell cycle-regulated ganes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell*, **9**, 3273-3297.

Tamayo, P., Slonim, D., mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R. (1999). Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci.*, **96**, 2907-2912.

Tibshirani, R., Hastie, T., Eisen, M., Ross, D., Botstein, D., Brown, P. (1999). Clustering methods for the analysis of DNA microarray data, *Tech Report, Dept. of Health Research and Policy, Stanford Univ.*, www.stat.stanford.edu/~tibs/lab/publications.html.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P.O., Hastie, T., Tibshiranni, R., Bostein, D. and Altman, R. B. (2001). Missing value estimation methods for DNA microarrays, *Bioinformatics*, **17**, 520-525.

[2004년 6월 접수, 2005년 1월 채택]

Exploratory Analysis of Gene Expression Data Using Biplot*

Mira Park [1)]

ABSTRACT

Genome sequencing and microarray technology produce ever-increasing amounts of complex data that needs statistical analysis. Visualization is an effective analytic technique that exploits the ability of the human brain to process large amounts of data. In this study, biplot approach applied to microarray data to see the relationship between genes and samples. The supplementary data method to classify new sample to known category is suggested. The methods are validated by applying it to well known microarray data such as Golub et al.(1999), Alizadeh et al.(2000), Ross et al.(2000). The results are compared to the results of several clustering methods. Modified graph which combine partitioning method and biplot is also suggested.

Keywords: Gene expression data, Microarray, biplot, Supplementary data analysis, Clustering, Classification

* This work is supported by grant No. R05-2003-000-11954-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

1) Assistant Professor, Dept. of Premedicine, Eulji University, 143-5 Yongdu-dong, Chung-gu, Daejon, 301-832, Korea.
E-mail : mira@eulji.ac.kr