다차원 척도 법을 이용한
DNA microarray 자료의 시각화

권 성우, 한 종훈

(sw74@postech.ac.kr)

포항 공과 대학교 화학 공학과

1. 서론

많은 양의 유전체 데이터를 분석하기 위해서는 새로운 방식의 분석 기술이 필요하다. 이러한 기술들 중에서 대표적인 것이 바로 Microarray 기술이다. Microarray 기술에는 DNA chip, protein chip, lab-on a chip 등이 있다. DNA chip은 고형체에 고정 시킨 DNA와 mRNA나 다른 DNA를 잡종 형성 시켜 만들게 되는데 특정 상태의 유전자 발현 양상을 연구 할 때 사용하고 있다.

DNA microarray 데이터의 경우 변수로 사용되는 유전자의 개수가 5000개에서 2만 4000개 사이로 변수 개수가 많은 특징이 있다. 따라서 이것을 적절하게 시각화하기 위해서는 자료를 함축적으로 표현하는 다차원 척도 법 (multidimensional scaling)을 많이 사용한다. 본 글에서는 다차원 척도 법에 대해서 알아보고자 한다.

2. 다차원 척도 법의 정의

다차원 척도법은 군집 분석에서와 마찬가지로 자료에 내재된 구조를 찾아내어 자료를 함축적으로 표현하고자 하는 자료 축약 형 다변량 분석 기법 이다. 다차원 척도법에서는 개체들 사이의 유사성 또는 비유사성을 평가하는데 사용될 수 있는 기준을 찾아내어 각 기준에 대하여 각 개체를 다차원 공간상에 시각적으로 표현하게 된다. 따라서 분석 목표는 각 개체들의 유사성 혹은 비유사성 (거리)를 설명할 수 있는 내재된 의미 있는 차원을 찾아내는 것이다.

3. 비 유사성과 거리와의 관계

n개 개체가 있을 때 개체 i와 개체 j사이에 관측된 비유사성을 δ_{ij} 라고 했을 때 비유사성 행렬은 Δ은 아래와 같이 표시된다.

$$\Delta = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1n} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & \delta_{nn} \end{pmatrix}$$

비유사성 행렬에서 대각 원소는 0이 되고, 대칭 행렬이 된다.

한편 개체들 사이의 비유사성 자료로부터 각 개체들을 공간상에 좌표화하기 위해서는 개체들 사이의 상대적 거리를 계산하여야 한다. 두 개체 i와 j사이의 거리를 d_{ij} 라고 했을 때 거리 행렬 D는 아래와 같다.

$$D = \begin{pmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{n1} & \cdots & d_{nn} \end{pmatrix}$$

거리 행렬 또한 대각 원소는 0이 되고, 대칭 행렬이 된다.

다차원 척도의 기본 개념은 두 개체 사이의 거리 d_{ij} 가 두 개체 사이의 비유사성 δ_{ij} 와 대응 관계에 있어야 된다는 것이다. 예를 들면 작은 비유사성은 작은 거리와 큰 유사성은 큰 거리와 대응 관계가 성립되어야 하므로 d_{ij} 와 δ_{ij} 사이의 함수 관계는 어떤 증가 함수의 형태를 갖게 될 것이다. 여기서 d_{ij} 와 δ_{ij} 함수 관계는 사전에 알려져 있지 않기 때문에 입력 자료의 측정 척도에 따라 몇 개의 모형으로 구분하여 가정한다. 보통 DNA microarray 자료의 경우 비율 척도 이기 때문에 기울기가 양수이며 원점을 통과하는 직선적인 함수 형태를 갖는 함수로 가정하고 최소 자승 회귀 방법을 통해 거리의 추정 값인 $\hat{d}_{ij}$ 을 구한다.

4. 적합도 측정

d_{ij} 와 δ_{ij} 사이의 관계를 최적화 시키는 함수 형태를 추정하기 위하여 개체들의 거

리 d_{ij}와 가정된 모형에 의하여 추정된 거리 $\hat{d}_{ij}$ 사이의 적합도를 측정하는데 보통 stress 척도를 사용한다.

$$\text{stress} = \left| \frac{\sum_{i \neq j}^{n} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i \neq j}^{n} d_{ij}^{\,2}} \right|$$

스트레스 척도에서 분자는 d_{ij}와 $\hat{d}_{ij}$ 사이의 차이 자승 합을 나타내고, 분모는 일반적으로 차원이 증가하면 거리 d_{ij}는 증가하기 때문에 서로 다른 차원의 적합도를 비교하기 위하여 사용된 정규화 값을 나타낸다. 만일 두 거리가 동일 하면 ($d_{ij} = \hat{d}_{ij}$) 스트레스 값은 0이 되고 d_{ij}는 δ_{ij}의 함수에 의해 완벽하게 추정된다는 것을 의미하게 된다. 또한 스트레스 값이 크다면 실제 거리 d_{ij}와 추정된 거리 $\hat{d}_{ij}$의 차이가 크다는 의미가 된다. 다차원 척도의 적합도를 통계적으로 검정할 수 있는 방법은 없으나 일반적으로 다음과 같은 스트레스 값에 대한 평가 기준이 사용되고 있다.

스트레스 값	적합도 평가
0.2이상	아주 나쁘다
0.2	나쁘다
0.1	보통이다
0.05	좋은 편이다
0.025	매우 좋은 편이다
0	완벽하다

표 1. 다차원 척도 모델의 평가 기준

5. 최적화 과정

다차원 척도에서 기본적인 전제 조건은 공간에서 개체들 사이의 거리가 개체들 사이의 비유사성 정도와 일치해야 한다는 것이다. 따라서 일치도를 높이기 위해서는 d_{ij}와 δ_{ij} 사이의 관계를 최적화 시키는 함수를 반복적으로 추정하는 과정을 거치게 된다. 반복 과정은 출발 좌표로부터 구한 스트레스를 작게 할 수 있는 새로운 좌표

로 개체 점들을 이동시키는 과정을 의미한다. 예를 들면 d_{ij}가 $\hat{d}_{ij}$ 보다 크다면 i번째 점은 차이의 특정 비례량만큼 j번째 점으로 이동하게 되고, 역으로 d_{ij}가 $\hat{d}_{ij}$ 보다 작다면 j번째 점은 차이의 특정 비례량 만큼 i번째 점으로 이동하여 새로운 좌표를 구하게 된다. 매 과정마다 구한 스트레스 값이 설정한 값보다 작거나, 새로운 과정을 반복하여도 스트레스 값 차이의 감소가 매우 미미하다면 반복 과정을 중지하고 마지막 단계에서 구한 좌표로 공간상에 표시하게 된다. 이러한 스트레스를 최소화 할 수 있는 최적화 함수를 구하는 방법으로는 수치 최적화 방법인 최대 급경사 (steepest descent) 방법이나, 가우스-뉴우튼 방법이 사용된다.

6. 다차원 척도 방법을 이용한 DNA microarray 자료의 시각화

암 환자의 DNA microarray 자료를 이용하여 암 환자의 암을 분류하고 각 암에 대해 특이적으로 과다 발현 하거나 억제 되는 유전자 (marker gene)를 찾는 것이 최근 연구 동향이다. 즉, 다 변량 통계 기법을 통해 암 환자를 분류하고 암 환자 분류에 주요한 영향을 주는 유전자를 결정한다. 그러나 보통 특이적으로 발현하는 유전자의 개수가 70개에서 500개 사이가 되므로 이들 차원에서 각 환자의 표본을 시각화 하지 못한다. 따라서 이때 많이 사용하는 기법이 다차원 척도 방법이다 (그림 1). 분석한 자료를 시각화 시키는 것은 최종 결과를 보여 주는 것이며, 대부분의 생명 공학자나 의학자들은 수치 보다는 시각화된 그림을 선호하기에 것이기에 매우 중요한 분석 방법이다.

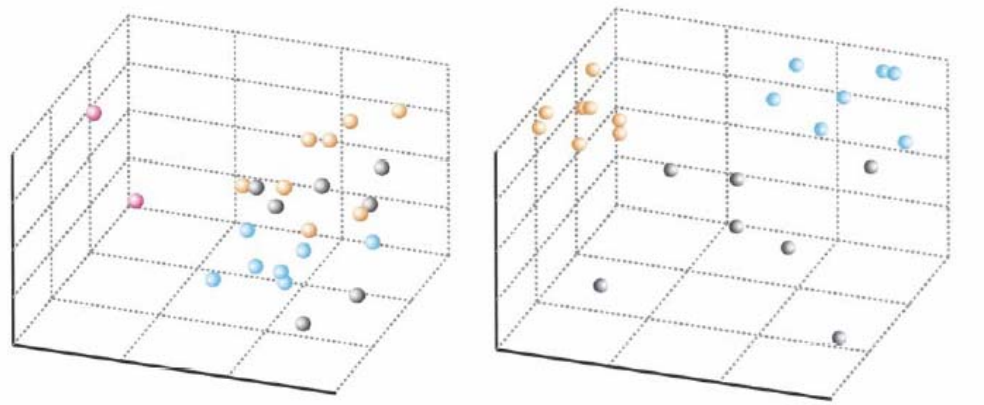

그림 1. 유방암 환자의 DNA microarray 자료를 다차원 척도 방법으로 시각화 한 것이다. 각 색은 암의 종류를 의미 한다. 왼쪽은 5600여 개의 유전자를 다차원 척도 방법으로 시각화 한 것이고 오른쪽은 5600여 개에서 53개의 주요한 유전자를 선택한 후에 다차원 척도 방법으로 시각화 한 것이다. 암 종류의 분리가 오른쪽이

왼쪽 보다 더 잘 된 것을 알 수가 있으며, 이를 통해 53개의 선택한 유전자들이 각 암의 종류를 나누는데 관여 하는 것을 알 수가 있다.

참고 문헌

Borg, I., and J. Lingoes. (1987). *Multidimensional Similarity Structure Analysis*. Springer-Verlag.

Ingrid, H et al., (2001). Gene-expression profiles in hereditary breast cancer. *New England Journal of Medicine.* 344. 539-548.

Null, C.H. and Sarle, W.S. (1982). Multidimensional Scaling by Least Squares. *Proceedings of the Seventh Annual SAS Users Group International Conference*.

Schiffman, S.S., Reynolds, M.L., and Young, F.W. (1981). *Introduction to Multidimensional Scaling*. New York: Academic Press.