주성분 자기조직도를 이용한 마이크로어레이 자료의 분석

박미라1) 장유진2) 허명회3)

요 약

마이크로어레이자료의 분석에 있어서 주성분 자기조직도(principal component SOM)의 유용성을 알아보고, 흔히 사용되는 다른 군집분석방법과 비교하였다. 또한 MST(minimal spanning tree)를 이용하여 주성분자기조직도 결과의 적합성을 알아보았다.

주요용어 : 자기조직도(SOM), 마이크로어레이실험, 유전자발현

1. 서론

최근 cDNA와 올리고뉴클레오타이드 마이크로어레이칩 기술의 발달로 수천 개의 유전자 발현양상을 동시에 관찰할 수 있게 되었다. 마이크로어레이 자료분석의 주요 목적 중 하나는 유사한 패턴을 가진 유전자(또는 샘플)의 군집을 파악하는 것으로, 이를 통해 같은 기능을 가진 집단의 탐색이나 알려지지 않은 유전자들의 성질을 파악할 수 있게 된다. 유전자발현프로필의 군집방법으로 계층적 군집분석이나 k-평균 군집분석, 주성분분석 등이 적용되고 있다. Kohonen(1995)에 의해 개발된 자기조직도(Self-organizing Map;SOM)도 여러 연구에서 유전자의 패턴인식에 유용한 것으로 알려져 있다(Tamayo et al., 1999; Toronen et al., 1999). 그러나 이의 적용에는 노드의 수와 형태, 비 연속적인 지도제공 등의 여러 문제가 있다. 이러한 문제에 대한 대안으로서 Huh(2003)는 주성분 자기조직도(principal component SOM; PC-SOM)를 제안한 바 있다. 이 연구에서는 잘 알려진 두 종류의 마이크로어레이 데이터를 이용하여 유전체자료의 분석에 있어서 주성분 자기조직도의 유용성을 연구하였다. 또한 이를 기존의 군집분석방법과 비교하고, MST(minimal spanning tree)를 이용하여 주성분자기조직도 결과의 적합성을 알아보았다.

2. 주성분 자기조직도의 기본 알고리즘

주성분 자기조직도(PC-SOM)의 기본 아이디어는 1차원 자기조직도(SOM) 산출이후 입력개

* 이 연구는 한국학술진흥재단 지원으로 수행되었음 (R14-2003-002-01001-0)

1) 대전시 중구 용두동 143-5, 을지의과대학교 의예과, 조교수.
2) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 석사과정.
3) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수.

체들을 각각 대표점과 잔차로 분리하고 여기서 발생한 잔차들로 별개의 1차원 SOM을 다시 구하여 기존의 결과에 교차적으로 붙여 2차원 SOM을 만드는 것이다(Huh,2003). 기본형 PC-SOM을 구하기 위한 구체적인 알고리즘은 다음과 같다.

1) 1축의 노드수(c_1)를 정한 뒤 주성분분석을 실시하여 제 1 고유값 λ_1을 구한다. 1축의 사영점들의 실질적인 범위를 $(-2.5\sqrt{\lambda_1}, 2.5\sqrt{\lambda_1})$이라고 하면, 노드의 간격을 일정하게 하기 위한 구간의 초기치는 $interval = 5\sqrt{\lambda_1}/(c_1-1)$이 된다.
2) 1차원 SOM을 구하여 승자 노드 $k_1(i)$와 해당하는 중량 $w_{k1}(i)$를 구한다. 입력개체 x_i를 중량 $w_{k1}(i)$와 잔차 $x_i - w_{k1}(i)$로 나누고 입력개체를 잔차로 대체한다.
3) 다시 주성분분석을 실시하여 제 1 고유값 λ_2를 구한다. 이 때 1축과 2축의 그리드간격을 같게 하기 위한 제 2축의 노드수는
$$c_2 = round(5\sqrt{\lambda_2}/interval) + 1$$
이 된다. 여기서 $round(\cdot)$는 반올림 함수이다.
4) c_2개의 노드를 갖는 1차원 SOM을 만들어 승자노드와 중량들의 리스트를 구한다. 이에 따라 입력개체 x_i는 2차원 SOM에서 노드 $(k_1(i), k_2(i))$로 배치된다.

이러한 절차를 반복적용하면 3차원 이상의 지도를 작성할 수도 있으며, 이 때 2축의 노드수는 자동적으로 정해지게 된다. 한편 보간형 PC-SOM은 비연속적인 지도를 주는 SOM의 문제점을 보완하기 위한 방법으로, 입력개체를 해당노드의 중량으로 할당하는 대신 인접중량들의 가중치로 대체하여 지도를 보다 연속적으로 만드는 방법이다. 입력개체 x에 대한 승자노드의 중량을 w_k, 이의 왼쪽과 오른쪽 인접 노드 중량을 각각 w_l, w_r라고 했을 때, 승자노드 w_k에 단순할당하는 대신 이 두 인접노드를 연결하는 선분상에서 입력개체에 가장 가까운 점을 찾아 할당하는 방법이다. 예컨대 왼쪽노드와 오른쪽 노드의 연결선을 각각 7등분한 점을 부노드(subnode)라고 하고 그 중의 한 점으로 할당하는 방법을 쓸 수 있다. PC-SOM의 장점 중 하나는 각 중량의 성분을 관찰함으로써 SOM의 각 축이 어떤 변수적 특성을 갖는지 알 수 있다는 점이다.

3. 마이크로어레이 자료와 분석결과

cDNA 칩 및 올리고뉴클레오타이드 칩을 이용하여 얻은 마이크로어레이 데이터에 자기조직도를 적용하여 보았다. 유전자발현은 올리고뉴클레오타이드 칩에 의한 결과일 때는 절대적인 값이지만 cDNA 칩의 경우에는 참조샘플과 비교하여 구해진 상대적인 값이 된다. 이 데이터들은 각 유전자별로 평균을 0으로 중심화시켰으며, 분산을 1로 표준화시킨 경우와 표준화하지 않은 두 가지 경우에 대해 모두 분석을 실시하였다. 발생된 결측치에 대해서 k-nearest neighbor algorithm을 사용하여 결측치를 추정하였고 k는 5로 지정하였다. 이때 이웃(neighbor)은 유전자이고 이웃간의 거리는 유전자간 상관에 근거한 것이다. 분석과 그래프는 SAS/IML 및 SAS/GRAPH를 이용하였다. 사용된 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(DLBCL)만을 취하여 모두 62개의 샘플(11개의 B-CLL, 9개의 FL, 42개의 DLBCL)에 대한 4026개의 유전자 발현값이다. 데이터는 강도비에 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).

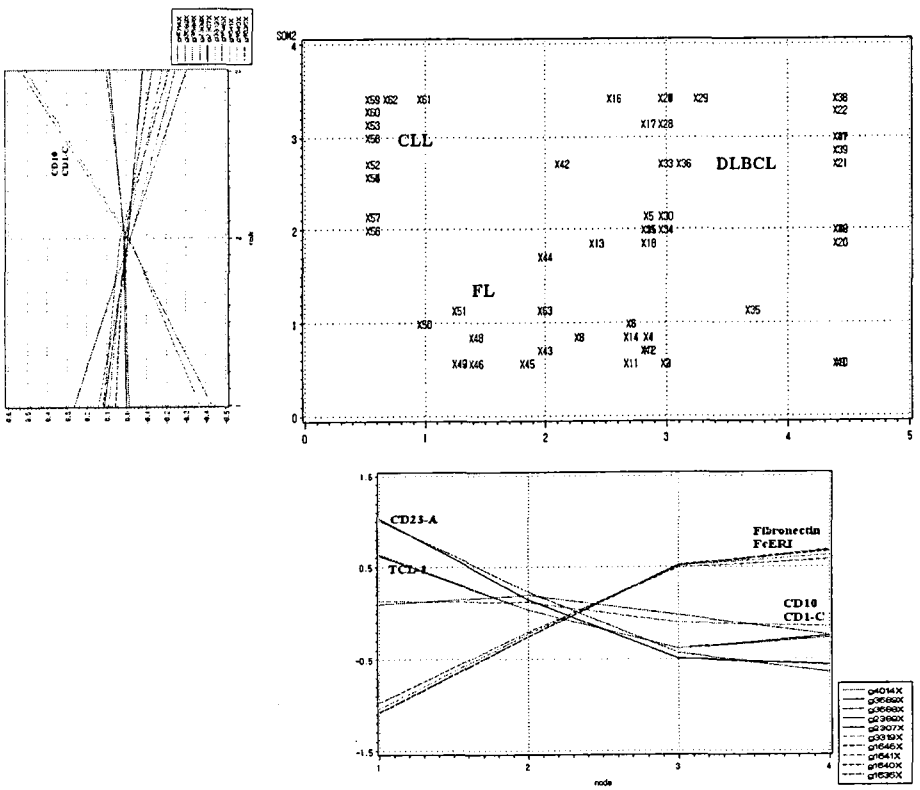
데이터의 일부 분석결과를 보면 다음과 같다. <그림 1>은 림프구데이터에 4×3 보간형 PC-SOM을 적용한 결과이다. 분석자가 1축의 노드수로 4를 정하였고, 2축의 노드수는 3으로 계산되었다. 윗줄 오른쪽 그림에서 x2-x42와 x63은 DLBCL 세포이며, x43-x51은 FL 세포이고, x52-x62는 CLL 세포를 가리킨다. 세 그룹이 잘 분리되어 있음을 알 수 있다. 1축의 우측에는 주로 DLBCL셀들이, 좌측에는 주로 CLL 및 FL 셀들이 포진하고 있으며, CLL과 FL셀은 2축에 의해 분리되고 있다. 이의 아래쪽 그림은 몇 개의 유전자에 대한 변수그림으로서 예컨대 Fibronectin 과 FcERI같은 유전자들은 1축의 노드가 증가함에 따라 증가하는 경향을 보여, 이들이 주로 FL 및 CLL보다는 주로 DLBCL에서 발현하고 있음을 시사한다. 실제로 이들은 "림프노드(Lymph node)" 신호로 정의되는 유전자들로 알려져 있으며, DLBCL셀과 연관이 있는 것으로 알려져 있다. 한편 CD23-A나 TCL-1과 같은 유전자들은 반대의 패턴을 보여 이들이 DLBCL보다는 주로 FL 및 CLL보다는 주로 DLBCL에서 발현하고 있음을 시사한다. 왼쪽 위의 그림은 2축에 대한 변수그림이다. 여기서 CD10이나 CD1-C등의 유전자들은 FL셀과 CLL 셀을 구분짓는 역할을 한다는 것을 알 수 있다. 이들은 "배중심 B세포(Germinal Center B cell)"의 특성을 갖는 유전자들로 알려져 있다.

<표 1>은 백혈병데이터의 기본 PC-SOM 결과이다. 각 열은 1축의 노드를, 각 행은 2축의 노드를 의미하며, 표안에는 할당된 셀이름이 있다. A로 시작되는 것은 ALL셀을 의미하고, B로 시작되는 것은 B-cell ALL, T로 시작하는 것은 T-cell ALL을 의미한다. 1축에 의해 ALL과 AML이 대부분 잘 나누어지고 있음을 알 수 있다. 한편 <그림 2>는 백혈병데이터의 보간형 PC-SOM 결과의 노드를 MST(minimal spanning tree)를 이용하여 연결한 것이다. 따로 표기가 없는 왼쪽의 점들은 B-cell ALL이며, 오른쪽의 따로 표기가 없는 점들은 AML 셀들이다. 여기서는 5개의 변(edge)을 제거하여 6개의 자연적인 군집을 생성하였으며, 군집결과 기대한대로 각 그룹을 잘 분할해주고 있음을 알 수 있다.

4. 결론

두 종류의 마이크로어레이 데이터에 PC-SOM을 적용한 결과 같은 그룹에 속하는 샘플들이 같은 군집으로 잘 분리되었으며, 변수그림을 통해 쉽게 각 유전자와 샘플들간의 연관성을 파악할 수 있었다. 또한 보간형 PC-SOM을 사용함으로써 비교적 손쉬운 방법으로 특정 군집으로의 단순할당보다 정밀하게 샘플간의 거리를 표현할 수 있었다. 이 방법은 Kohonen의 SOM을

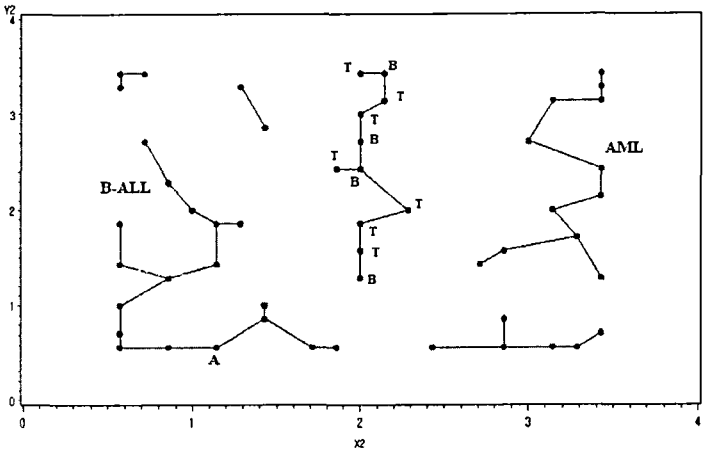
단순 적용한 경우와 비교할 때 2축의 노드수를 사전에 정하지 않아도 된다는 이점이 있으며, 보다 연속적인 출력을 가능하게 함으로써 향상된 시각화를 제공한다. 또한 SOM에서는 제공하지 않는 변수그림을 같이 그릴 수 있다는 장점이 있다. 한편 주성분분석의 경우에는 연속적인 그래프를 제공하지만 특정군집으로의 할당을 정해주지는 않는다. 따라서 마이크로어레이자료분석에 있어서 PC-SOM의 활용은 이 두 방법의 장점을 잘 절충하는 방안이라 할 수 있겠다.



<그림 1> 림프구 데이터의 4x3 보간형 PC-SOM 결과와 변수그림

Node	1	2	3
1	B3_1, B9_1, B11_1, B27_1, B31_1, B33_1, B41_1, B49_1, B11_2, B31_2, B33_2	T5_1, T7_1, B15_1, B17_1, T19_1, T29_1, B35_1	A67_1, A71, A75, A77, A61_1, A45_2, A65_2, A67_2
2	B51_1, B53_1, B3_2, B7_2, B27_2, B29_2	T13_1, T21_1, T23_1, T47_1, B13_2, T35_2	A73, A57_1, A63_1, A65_1, A49_2, A47_2, A59_2, A69_2
3	B37_1, B39_1, B43_1, B45_1, B15_2, B17_2, B19_2, B21_2, B23_2, B9_2, B39_2, B41_2, A63_2, B25_2	B25_1, B55_1, B5_2, B37_2, A57_2	A69_1, A59_1, A43_2, A51_2, A53_2, A55_2, A61_2

<표 1> 백혈병데이터의 기본 PC-SOM 결과



<그림 2> 백혈병데이터의 보간형 PC-SOM에 대한 MST 결과

참고문헌

Alizadeh, A.A., Eisen, M.B., Davis, R.E., et al. (2000) Different type of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503-511.

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.

Golub, T.R., Slonim, D.K., Tamayo, P., et al. (1999)Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537.

Huh, M. H. (2003) Principal Components Self-Organizing Map PC-SOM, *Korean Journal of Applied Statistics.* 16(2), 321-333.

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

Tamayo, P., Slonim, D., mesirov, J., et al. (1999) Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation, *Proc. Natl. Acad. Sci.* 96:2907-2912.

Toronen, P., Kolehmainen, M., Wong, G., and Castren, E. (1999) Analysis of gene expression data using self-organizing maps, *Federation of European Biochemical Societies.* 451., 142-146.