응용통계연구 (2008)
21(6), 911–921

평행좌표 플롯을 활용한 유전자발현 자료의 시각화

박미라[1] · 곽일엽[2] · 허명회[3]

[1]을지대학교 의과대학 예방의학교실, [2]고려대학교 통계학과, [3]고려대학교 통계학과

(2008년 9월 접수, 2008년 9월 채택)

요 약

유전자발현 자료로부터 유용한 생물학적 정보를 얻기 위해 여러 시각화 방법이 개발되어 왔다. 본 논문에서는 평행좌표 플롯(parallel coordinate plot: PCP)을 이용하여 유전자발현 패턴을 찾아내어 표현하고자 하였다. 평행좌표 플롯의 두 변형인 ePCP(enhanced parallel coordinate plot)와 APCP(Andrews' type parallel coordinate plot)를 림포마(lymphoma) 자료에 적용하여 62개 샘플을 의미있게 배열하고 300개 유전자를 평활 곡선으로 표현하였다.

주요용어: 유전자발현, 림포마 자료, 평행좌표 플롯, EPCP, APCP.

1. 서론

최근의 정보생물학분야에서는 자료시각화 방법이 중요시되고 있다. 큰 규모의 자료가 생성되는 유전자칩 실험 분석의 경우 heatmap이나 dendrogram 등의 그래프가 빈번하게 사용되고 있다. 이 외에 보다 효과적인 자료시각화를 위해 다수의 방안이 시도된 바 있다 (Park 등, 2005; Prasad와 Ahson, 2006; Santamaria 등, 2008).

본 연구에서는 평행좌표 플롯(parallel coordinate plot: PCP)을 이용하여 유전자발현(gene expression) 자료를 시각적으로 표현하고자 한다. 평행좌표 플롯은 평면에 평행하게 놓인 좌표축을 통과하는 연결선으로 고차원의 다변량 관측(multivariate observation)을 표현하는 방법이다 (Inselberg, 1985). 다시 말하여, p개 샘플(또는 조건)에서 n개의 유전자 발현값으로 구성된 유전자발현 자료에 대한 평행좌표 플롯은 가로로 평행하게 p개의 변수 축을 등간격으로 설정하고 유전자별로 샘플 발현값을 각 축에서 최소 0 · 최대 1이 되도록 변환하여 변수 축에 하나씩 타점한 뒤 이들을 곧은 선으로 연결한 그래프이다. 즉, p개 샘플(또는 조건)의 n개 유전자발현 자료는 n개의 $p-1$번 꺾인 연결 선으로 표현된다. 최근 Cheng 등 (2008)과 Santamaria 등 (2008)의 연구에서 평행좌표 플롯이 마이크로어레이 자료의 분석에 활용되었다.

평행좌표 플롯에서는, 변수 축을 연결하는 선분들이 대부분 교차하지 않는 패턴을 나타내면 두 변수가 양의 상관이 있고 선분들이 대부분 교차하는 패턴을 나타내면 음의 상관이 있는 것으로 해석된다. 따라서 평행좌표 플롯으로 인접 변수간 연관성을 유추할 수 있다. 그러나 각 변수가 임의의 순서로 놓임으로써 플롯이 등락을 수없이 거듭하는 복잡한 형태가 될 수 있으며, 변수들이 놓이는 순서에 따라 그

정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행되었음(R14-2003-002-01001-0).
[1](301-832) 대전시 중구 용두동 143-5, 을지대학교 의과대학 예방의학교실, 부교수. E-mail: mira@eulji.ac.kr
[2](136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 대학원 석사과정 졸업.
[3]교신저자: (136-701) 서울시 성북구 안암동 5-1, 고려대학교 통계학과, 교수. E-mail: stat420@korea.ac.kr

래프의 모양이 완전히 달라지게 된다. 또한 변수 축 사이의 선이 관측 간 단순 연결에 불과하다는 등의 단점이 있다. 본 연구에서는 이러한 문제에 대한 대안으로 개발된 ePCP(enhanced Parallel Coordinate Plot)와 APCP(Andrews' type Parallel Coordinate Plot)의 두 변형을 활용할 것이다. 2절에서 ePCP와 APCP에 대하여 소개하고 특성을 비교하였으며, 3절에서 실제 유전자발현자료에 두 방법론을 적용하여 유용성을 살펴보았다. 마지막으로 4절에서 본 연구를 맺는다.

2. 평행좌표 플롯의 변형

2.1. ePCP의 알고리즘과 특성

Huh와 Park (2008)에서 제안된 ePCP(enhanced parallel coordinate plot)는 기존의 평행좌표 플롯을 개선한 버전이라고 할 수 있다. 개선사항은 (i) 특이점에 매우 민감한 최소값과 최대값을 일정하게 되도록 각 변수를 표준화하는 대신 평균 0, 표준편차 1이 되도록 변수를 표준화하고, (ii) 상호 연관도를 고려하여 변수들의 배열 순서를 정하고, (iii) 변수 축 간 간격에 변수 간 인접도(proximity)를 반영하며, (iv) 꺾은 선 대신 평활 곡선으로 관측을 표현하는 것 등이다. 이를 실현하기 위한 알고리즘은 다음과 같다.

n개 관측에 대한 p개 변수 자료를 $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$로 표기하되 $\boldsymbol{x}_j' = (x_{1j}, x_{2j}, \ldots, x_{nj})$가 j번째 변수·n개 관측값의 표준화 변환(평균 0 · 표준편차 1)이라고 하자($j = 1, \ldots, p$). 그러면, 각 변수는 반지름이 $\sqrt{n-1}$ 인 초구(超球; hyper-sphere)상에 있게 된다. 변수 $\boldsymbol{x}_j$와 $\boldsymbol{x}_k$간 초구상 거리는

$$d(\boldsymbol{x}_j, \boldsymbol{x}_k) = \sqrt{n-1}\cos^{-1}(\mathrm{corr}(\boldsymbol{x}_j, \boldsymbol{x}_k)), \quad j, k = 1, \ldots, p$$

이고 p개 변수를 $\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2}, \ldots, \boldsymbol{x}_{s_p}$로 배열하는 경우 초구 상에서 총 거리는

$$D = d(\boldsymbol{x}_{s_1}, \boldsymbol{x}_{s_2}) + \cdots + d(\boldsymbol{x}_{s_{p-1}}, \boldsymbol{x}_{s_p})$$

가 된다. 총 거리 D를 최소화하는 방법으로 Traveling Salesman's Problem 등 여러가지가 있겠으나 현실적인 방편은 Hurley (2004)의 endlink 알고리즘을 활용하는 것이다.

Endlink 알고리즘 (Hurley, 2004)은 다음과 같이 구성된다. (i) 변수 간 거리가 가장 짧은 두 변수를 묶어 한 그룹의 양 끝으로 한다. (ii) 변수 그룹과 변수 (그룹)간 거리를 그룹의 양 끝 변수와 변수(그룹의 양 끝) 간 거리로 정의한다. (iii) 변수 그룹 간 가장 짧은 거리를 찾아내 끝과 끝을 연결한다. (iv) 앞의 두 단계를 전체가 하나로 연결된 그룹이 만들어질 때까지 반복한다. 이렇게 해서 모든 변수들이 일렬로 정렬되면, 상관 정도가 큰 변수들은 가깝게, 상관 정도가 작은 변수들은 멀리 놓인다. 이 때 평행좌표 플롯에 두 인접변수 $\boldsymbol{x}_j$와 $\boldsymbol{x}_k$ 사이의 간격을 $d(\boldsymbol{x}_j, \boldsymbol{x}_k)$에 비례하게 두면 변수들의 초구상 위치가 사실과 가깝게 된다. 한편으로, 평행좌표 플롯이 초구상의 최단 연결선을 나타내게 됨으로써 변수 축과 변수 축 사이가 거의 평활한 곡선으로 이어진 그래프를 얻을 수 있다. 이러한 평행좌표 플롯(ePCP)은 초구상에서 관측별로 첫 번째 변수값에서 출발하여 마지막 변수값에 종착하는 연결된 궤적을 보여준다.

2.2. APCP의 알고리즘과 특성

Kwak과 Huh (2008)에서 제안된 ACPC(Andrews' type parallel coordinate plot)은 앤드류스 플롯(Andrews' plot)을 평행좌표 플롯의 형태로 바꿔서 보여준다. 그 방법은 다음과 같다.

퓨리에 급수의 서로 독립인 p개 기저함수를 써서 앤드류스 궤도(Andrews' path)를

$$\boldsymbol{a}_p(t) = \left(\frac{1}{\sqrt{2}}, \sin t, \cos t, \sin 2t, \cos 2t, \ldots\right)', \quad 0 \le t < 2\pi$$

로 정의하자. 그러면 앤드류스 (Andrews, 1972) 플롯은 n개의 관측별로 각각 정의되는

$$f_{\boldsymbol{x}_i}(t) =< \boldsymbol{a}_p(t),\ \boldsymbol{x}_i >, \quad i = 1, \ldots, n$$

을 2차원 평면에 한 그래프로 표현한 것으로 볼 수 있다. 여기서 $< \boldsymbol{a}, \boldsymbol{x} >$은 두 벡터 $\boldsymbol{a}$와 $\boldsymbol{x}$간 내적이다.

p가 홀수일 때, 앤드류스 궤도 $\boldsymbol{a}_p(t)$를 회전하여 같은 반경의 초구 상에서 또 하나의 궤도인 $\boldsymbol{b}_p(t)$를 다음과 같이 만든다:

$$\boldsymbol{b}_p(t) = U_p \boldsymbol{a}_p(t),$$

여기서 U_p는

$$U_p = \begin{pmatrix} \boldsymbol{a}_p'(t_1) \\ \vdots \\ \boldsymbol{a}_p'(t_p) \end{pmatrix}$$

로 정의되는 $p \times p$인 정규직교행렬이다.

이제 회전된 앤드류스 궤도 $\boldsymbol{b}_p(t)$와 관측벡터 $\boldsymbol{x}_i$를 내적하여, 새로운 앤드류스 함수

$$g_{\boldsymbol{x}_i}(t) =< \boldsymbol{b}_p(t),\ \boldsymbol{x}_i >, \quad i = 1, \ldots, n$$

를 만들어내 그래프로 나타내면 그것은 새로운 앤드류스 플롯이며 동시에 평행좌표 플롯이 된다. 왜냐하면 $U_p' U_p = I_p$가 되기 때문이다(I_p는 $p \times p$ 항등행렬).

p가 짝수인 경우에는 항상 0의 값을 갖는 변수가 추가된 확장 관측

$$\boldsymbol{x}_i^+ = (0, x_{i1}, \ldots, x_{ip})', \quad i = 1, \ldots, n$$

으로 자료셋을 치환한 다음 위의 알고리즘을 적용하면 된다.

앤드류스 플롯은 (i) 평균 관측의 앤드류스 함수가 개별 앤드류스 함수들의 평균과 같고, (ii) 관측 간 유클리드 제곱 거리가 앤드류스 함수 간의 제곱 거리에 비례한다는 장점이 있는데, APCP는 이러한 앤드류스 플롯의 장점을 계승한다. 또한 APCP는 sine과 cosine의 결합으로 표현되므로 모든 점들에서 미분가능한 평활 곡선으로 구성된다.

3. 유전자발현 자료의 시각화

3.1. 림포마 자료와 분석과정

2절에서 소개한 두 종류의 평행좌표플롯을 p개 샘플에서의 n개 유전자 발현값 자료에 적용해보기로 한다. 이 절에서 예로 사용할 자료인 림포마 자료는 성인 림프성 질병의 유전자발현연구 중 cDNA 마이크로어레이 실험에서 얻어진 것이다 (Alizadeh 등, 2000). 분석자료는 96개의 샘플에서 구해진 4,026개의 유전자 형광강도비의 밑이 2인 로그 변환값이다 (URL: http://genome-www.stanford.edu/lymphoma). 이 연구에서는 세 가지 림프성 질병, diffuse large B-cell lymphoma(DLCL: $X2-X42$, $X63$), B-cell chronic lymphocytic leukemia(B-CLL: $X52-X62$). follicular lymphoma(FL: $X43-X51$) 만을 취하여 모두 62개의 샘플을 사용하였다(즉, 11개의 B-CLL셀, 9개의 FL셀, 42개의 DLCL셀).

또한 4,026개의 유전자 중에서 Kruskal-Wallis 검정을 적용하여 세 질병을 가장 잘 분류하는 300개의 유전자를 선별하였다. 그리고 이 유전자들을 k-평균 군집분석을 통해 8개의 군집으로 나누어 시각화할

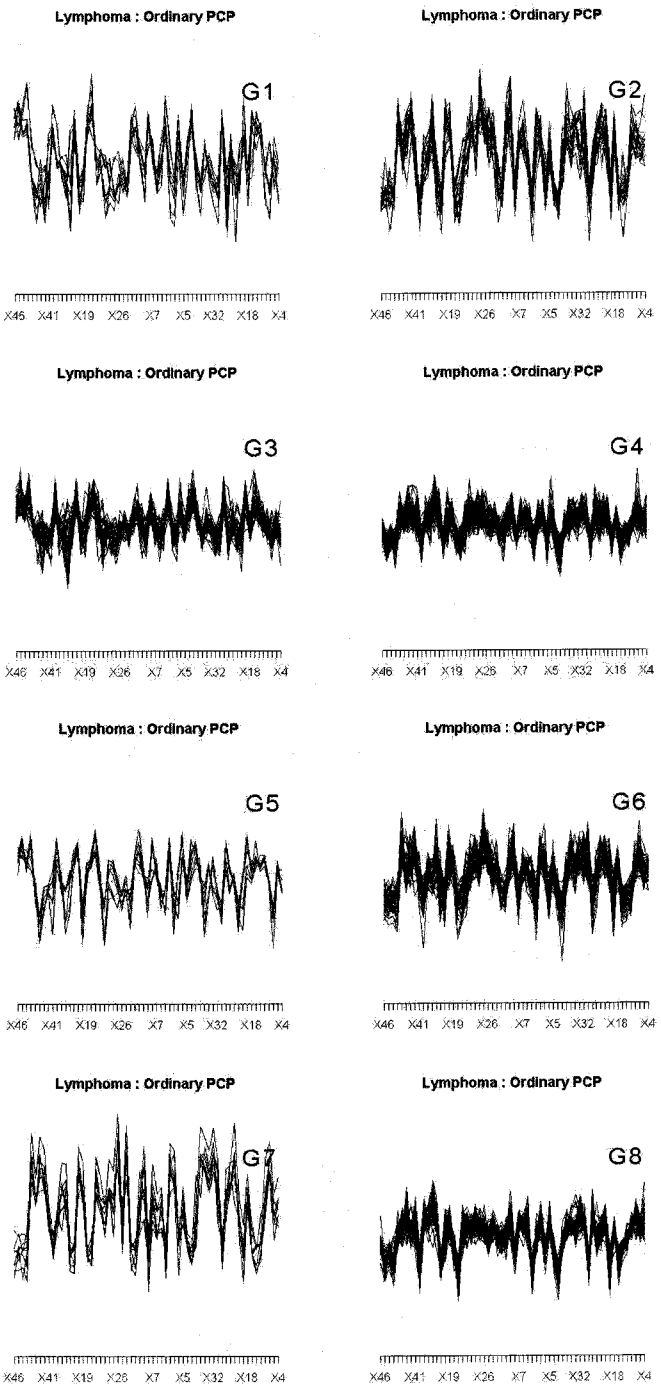

그림 **3.1.** 림포마 자료에 대한 보통 PCP(모든 유전자 군집)

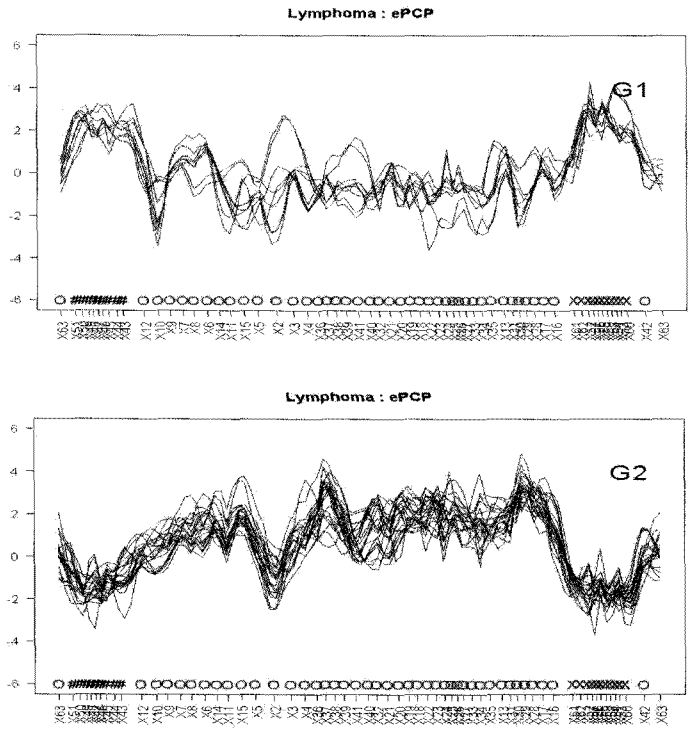

그림 3.2. 림포마 자료에 대한 ePCP(유전자 군집 1과 군집 2)

것이다. 1개의 그래프 안에 수 많은 유전자를 표현해 넣는 것은 실제로 무리이기 때문이다 (Santamaria 등, 2008). 평행좌표 플롯에서 첫 번째 샘플과 마지막 샘플 간 거리를 보기위해 마지막 샘플 다음에 첫 샘플을 넣었다.

그림 3.1은 보통 평행좌표플롯(PCP)이다. 변수순서를 임의화하여 분석자료를 초기 원자료에 준하는 상태로 만든 다음 곧바로 플롯이 만들어졌으므로, 군집으로 나누어 그렸음에도 불구하고 군집간 패턴 차이를 쉽게 파악할 수 없다. 이와 비교하여 상대적으로 ePCP와 APCP에 어떤 장점이 있는가를 다음 소절에서 보기로 한다.

3.2. 분석결과

그림 3.2가 림포마 자료의 8개 군집($G1$–$G8$) 중 군집 1($G1$)과 군집 2($G2$)의 ePCP를 보여준다. Endlink 알고리즘에 의해서 결정된 샘플의 순서는 $X63$-$X51$-$X50$-$X49$-$X48$-$X47$-$X46$-$X45$-$X44$-$X43$-$X12$-$X10$-$X9$-$X7$-$X8$-$X6$-$X14$-$X11$-$X15$-$X5$-$X2$-$X3$-$X4$-$X36$-$X37$-$X38$-$X39$-$X41$-$X40$-$X32$-$X21$-$X20$-$X19$-$X18$-$X22$-$X23$-$X24$-$X25$-$X26$-$X27$-$X33$-$X34$-$X35$-$X13$-$X31$-$X30$-$X29$-$X28$-$X17$-$X16$-$X61$-$X62$-$X52$-$X57$-$X56$-$X55$-$X59$-$X58$-$X54$-$X53$-$X60$-$X42$-($X63$)이다. 그래프 하단에 B-CLL셀은 녹색의 X, FL셀은 청색의 #, DLCL셀은 적색의 O로 나타내었다. 순서를 보면 대체로 같은 종류의 셀들이 한 곳에 모여 있다. 다만, $X63$과 $X42$는 DLCL셀임에도 불구하고 다른 DLCL셀들과는 떨어져 있어 특이한 셀임을 짐작할 수 있다.

샘플 축 간의 간격은 샘플 간의 인접도(proximity)를 나타낸다. 좁은 간격은 인접도가 강함을 의미한다. 즉 두 변수 간 상관이 높음을 나타낸다. 그리고 넓은 간격은 인접도가 약함을 말한다. 즉 두 변수 간 상관이 작음을 의미한다. 샘플 간 간격이 $X43$과 $X12$ 사이(FL과 DLCL의 경계), $X16$과 $X61$ 사이(DLCL과 B-CLL의 경계) 그리고 $X42$의 좌우, $X2$의 좌우, $X63$의 좌우에서 넓다. 이런 식으로 샘플들의 군집을 시각적으로 볼 수 있다. 또한 B-CLL셀과 FL셀은 같은 셀들 내에서 간격이 좁아 서로 응집성이 강하고 반면 DLCL셀들은 상대적으로 간격이 넓어서 B-CLL셀이나 FL셀들에 비해 응집성이 약음을 알 수 있다.

그림 3.3에서 8개의 군집별로 각 군집의 그래프적 특성을 살펴보면, $G1$은 FL셀과 B-CLL셀에서 높은 발현값을 갖는 특성을 갖고 있으며, $G2$의 경우에는 반대로 모든 B-CLL셀과 $X2$셀에서는 아주 낮은 값을 갖고, FL값도 약간 낮은 편인 유전자들의 집합이다 (그림 3.2 참조). 반면 $G3$의 경우 DLCL에 비해 B-CLL에서 약간 많이 발현하는 유전자들이며, $G4$는 반대의 경우이다. 나머지도 마찬가지 방식으로 해석이 가능하다. $G1$부터 $G8$까지 해당하는 유전자 수는 각각 9, 22, 40, 98, 8, 53, 8, 62개이다. 이제까지 ePCP로 몇 가지 사항을 쉽게 알 수 있었으나 보통 PCP인 그림 3.1에서는 그것이 쉽지 않다.

림포마 자료의 APCP에서도 앞과 유사한 해석을 얻을 수 있다. 그림 3.4는 270개의 분석유전자 자료를 군집평균으로 대치하여 표현한 APCP인데 각 군집에 할당된 유전자 수에 비례하게 선의 굵기를 지정하여 군집의 크기를 나타내었다. APCP의 특성상 그래프상의 두 곡선 간 평균제곱거리는 두 유전자 간 유클리드 제곱거리에 근사하게 된다. 가장 굵은 선으로 그려진 $G4$와 $G8$은 다른 군집에 비해 많은 유전자가 속한 큰 군집이며 그래프에서 아주 가까이 그려져서 두 군집이 유사하다는 것을 알 수 있다. 오른쪽의 B-CLL셀을 기준으로 봤을 때 세 군집($G1$, $G3$, $G5$)은 B-CLL에서 높은 값을, 나머지 다섯 군집($G2$, $G4$, $G6$, $G7$, $G8$)은 낮은 값을 갖는 경향이 있다.

같은 샘플에 대해 추가한 유전자가 있는 경우 평행좌표 플롯의 기존 패턴과 비교함으로써 어느 것에 가장 가까운지 알아볼 필요가 있다. 이를 위해서 유전자 300개 중 임의로 10%를 떼어내어 유보하고 나머지 270개의 유전자 자료에 대한 APCP를 그려보고 나서, 별도 그래프에 유보된 30개 유전자의 APCP 곡선을 그렸다.

그림 3.5는 유보된 30개 유전자 자료에 대한 APCP이다. 대조를 쉽게 하기 위해서 편의상 30개의 유보 유전자에 대한 APCP 그래프를 전체 유전자 자료의 k-평균 군집화에서 생긴 군집번호로 나누어 표출하였다. 분석유전자의 APCP와 유보유전자의 APCP와 유사함을 확인할 수 있다. 유보유전자 중 $G1$에 속한 것은 없어 그림 3.5의 첫 그림은 빈칸이다. 그림 3.5가 그림 3.4에 시각적으로 대응하고 있음을 볼 수 있다.

4. 맺음말

유전자 발현자료의 시각화는 생물학적 정보의 탐색에 매우 중요한 도구이다. 본 연구에서는 고차원의 유전자발현패턴을 표현하기 위한 방법으로서 평행좌표플롯(PCP)기법의 적용을 고려하였다. 보통의 평행좌표플롯은 다변량자료의 시각화방법으로서 많이 사용되어 온 방법이나, 이를 샘플수와 유전자수가 모두 큰 대규모 발현자료에 단순 적용하게 되면 수많은 곡선이 엉키는 결과를 낳게 되어 의미 있는 발현패턴을 파악하기 어렵다 (Cheng 등, 2008; Santamaria 등, 2008). 그림 3.1과 같이 사전 군집분석을 통해 유전자를 군집화하고 각 군당 유전자수를 적게 만든 경우에도 샘플의 순서가 임의로 배열되어 있는 경우 또는 샘플 순서에 아무 의미가 없는 경우에는 샘플별로 등락이 거듭되는 지그재그 형태의 그래프가 생성되어 해석이 쉽지 않다.

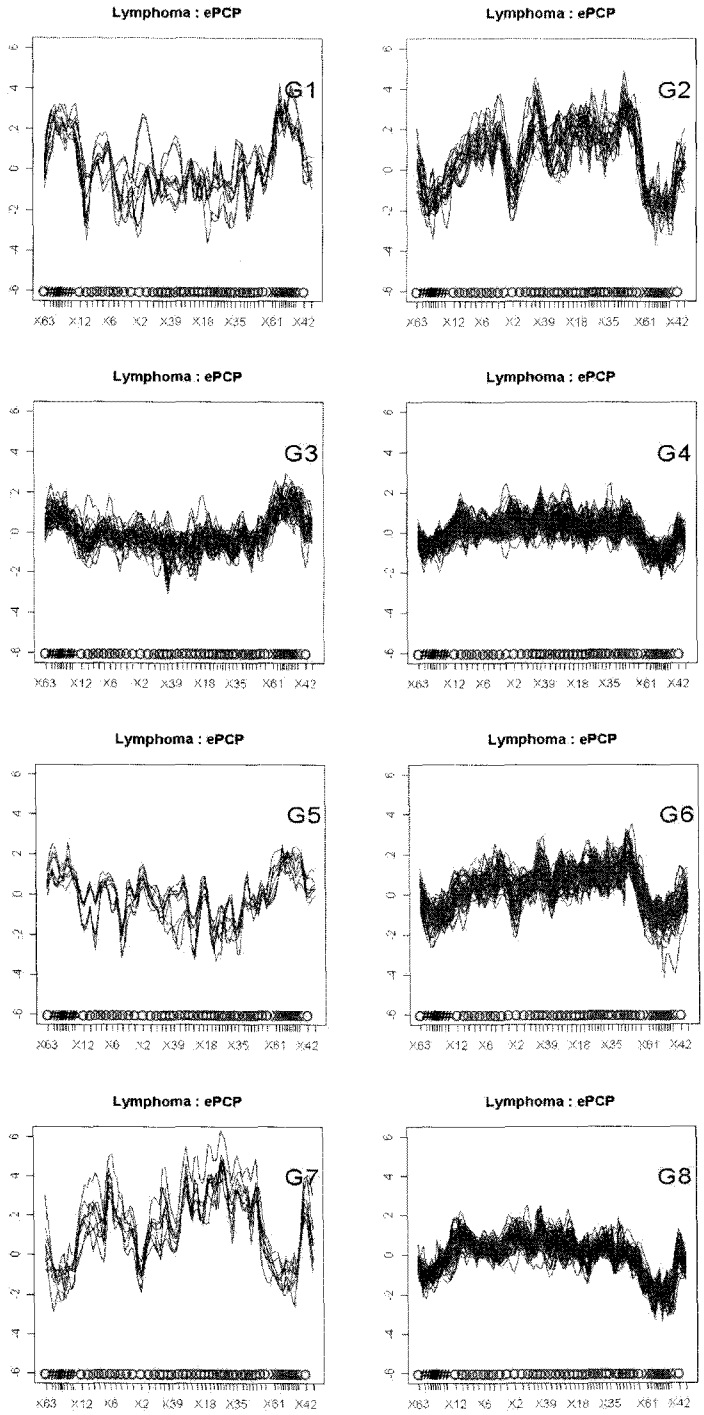

그림 3.3. 림포마 자료에 대한 EPCP(모든 유전자 군집)

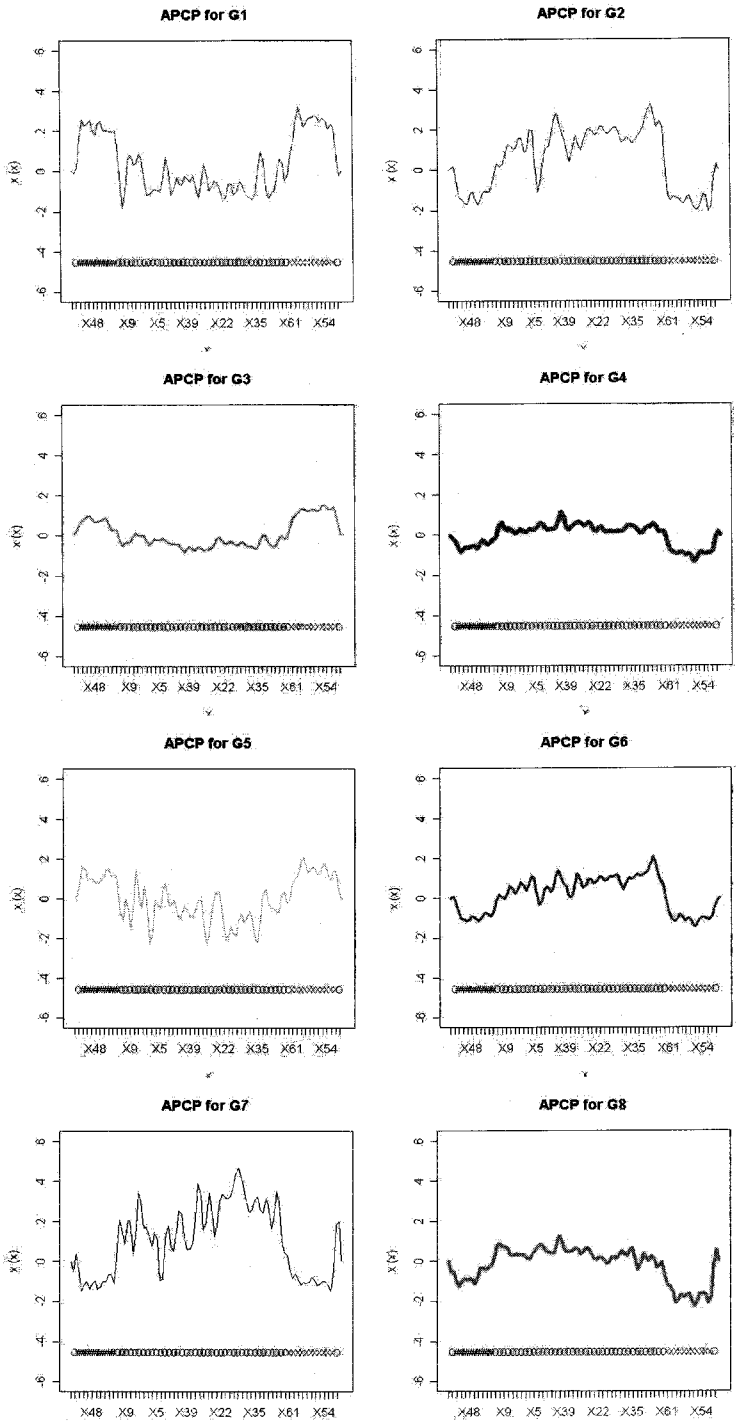

그림 **3.4.** 림포마 자료에 대한 APCP(분석유전자 270개)

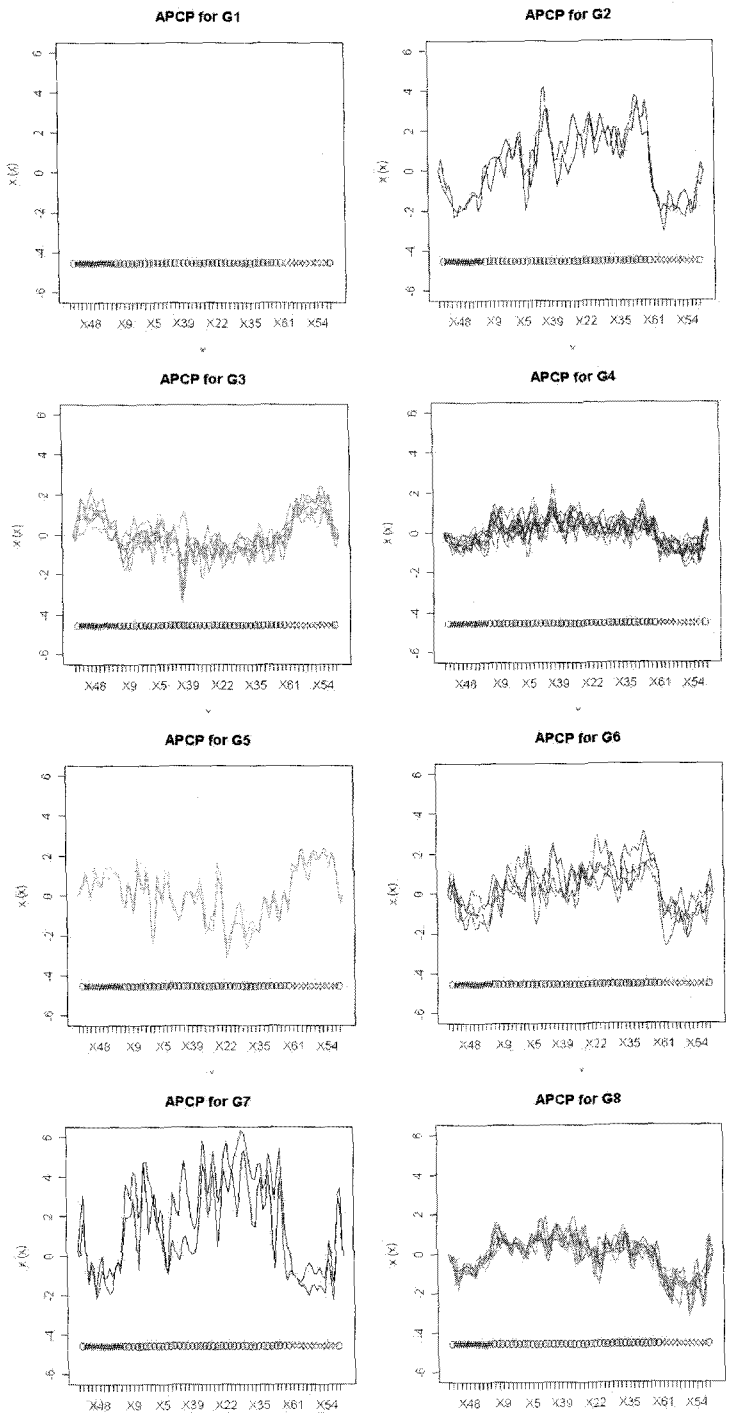

그림 3.5. 림포마 자료에 대한 APCP(유보유전자 30개)

본 연구에서는 이러한 단점을 보완하기 위한 방법으로 기존 평행좌표플롯의 변형인 ePCP와 APCP를 적용하였다. 림포마 자료에 대해 유전자를 사전 군집화한 후 ePCP를 적용한 결과, 거리가 가까운 샘플들은 가까이 놓이도록 순서를 재배정함으로써 샘플간의 관계를 파악할 수 있었으며, 거리에 따라 샘플간의 간격이 달라지게 하여 샘플의 군집에 관한 정보를 추가로 나타낼 수 있었다. 실제로 기존 연구에서 B-CLL, FL, DLCL 셀들로 구분된 바 있는 셀의 군집이 그래프에서 잘 드러났으며, 같은 DLCL 셀이라도 응집력이 약한 셀들은 간격이 넓게 나타나서 이에 대한 정보를 추가로 얻을 수 있었다. 또한 군집간 패턴을 비교하여 어떠한 유전자 군이 어떠한 셀들에서 주로 발현하는지 유전자와 샘플의 결합패턴을 파악할 수 있었다. 한편 APCP 결과에서는 그래프상의 두 곡선간 거리(평균제곱거리)를 유전자간의 거리(유클리드 제곱거리)로 해석할 수 있어, 각 군집별로 평균발현 그래프를 비교해봄으로써 유전자 군집간 거리와 특성을 파악할 수 있었다. 평균그래프를 그린 경우에는 그래프의 굵기를 해당유전자수에 비례하게 그림으로써 각 군집의 크기를 표현하였다. 이와 같이 유전자발현자료에 대해 ePCP와 APCP를 적용함으로써 기존 PCP의 단점을 보완하여, 많은 자료를 표현하면서도 한결 보기가 수월하고, 샘플에 대해서도 보다 많은 정보를 주는 그래프를 얻을 수 있다. 최근 유전자 자료의 시각화를 위한 다양한 시도가 있어 왔으나 ePCP와 APCP의 활용이 이에 대한 또 하나의 방안이 될 수 있을 것이다.

참고문헌

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, J. 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). Distinc type of diffuse large b-cell lymphoma identified by gene expression profiling, *Nature*, **403**, 503–511.

Andrews, D. F. (1972). Plots of high-demensional data, *Biometrics*, **28**, 125–136.

Cheng, K. O., Law, N. F., Siu, W. C. and Liew, A. W. (2008). Identification of coherent patterns in gene expression data using and efficient biclustering algorithm and parallele coordinate visualization, *BMC Bioinformatics*, **9**, 1–28.

Huh, M. H. and Park, D. Y. (2008). Enhancing parallel coordinate plots, *Journal of the Korean Statistical Society*, **37**, 129–133.

Hurley, C. B. (2004). Clustering visualizations of multidimensional data, *Journal of Computational & Graphical Statistics*, **13**, 788–806.

Inselberg, A. (1985). The plane with parallel coordinates, *The Visual Computer*, **1**, 69–91.

Kwak, I. Y. and Huh, M. H. (2008). Andrews' plot for extended uses, *Communications of the Korean Statistical Society*, **15**, 87–94.

Park, M., Jang, Y. J. and Huh, M. H. (2005). Analysis of gene expression data using PC-SOM, In *Proceedings of the 55th Session of International Statistical Institute*, 313.

Prasad, T. V. and Ahson, S. I. (2006). Visualization of microarray gene expression data, *Bioinformation*, **1**, 141–145.

Santamaria, R., Therón, R. and Quintales, L. (2008). A visual analytics approach for understanding biclustering results from microarray data, *BMC Bioinformatics*, **9**, 1–19.

Applications of Parallel Coordinate Plots for Visualizing Gene Expression Data

Mira Park[1] · Il-Youp Kwak[2] · Myung-Hoe Huh[3]

[1]Dept. of Preventive Medicine, Eulji University;
[2]Dept. of Statistics, Korea University; [3]Dept. of Statistics, Korea University

(Received September 2008; accepted September 2008)

Abstract

Visualization of the gene expression data on a low-dimensional graph is helpful in uncovering biological information contained in the data. In this study, we focus on two modified versions of the parallel coordinate plot. First one is the ePCP(enhanced parallel coordinate plot) which shows "near smooth" connecting curves between axes spaced proportionately to the proximity of re-ordered variables. Second one is APCP(Andrews' type parallel coordinate plot) which is obtained by rotating Andrews'plot that has a form of the parallel coordinate plot. Visualization procdures using ePCP and APCP are given for the lymphoma data case.

Keywords: Gene expression, lymphoma data, parallel coordinate plot, ePCP, APCP.

This work was supported by the Korea Research Foundation Grant founded by Korean Goverment (MOEHRD, R14-2003-002-01001-0).

[1]Associate Professor, Dept. of Preventive Medicine, Eulji University, Daejeon 301-832, Korea.
E-mail: mira@eulji.ac.kr

[2]Graduate Student, Dept. of Statistics, Korea University, Seoul 136-701, Korea.

[3]Corresponding author: Professor, Dept. of Statistics, Korea University, Seoul 136-701, Korea.
E-mail: stat420@korea.ac.kr