응용통계연구 제 17권 3호, 2004년, pp. 419-430

되돌림설계를 이용한 마이크로어레이 실험 자료의 분석 *

이선호[1)]

요 약

마이크로어레이 기술은 한번에 수만개의 유전자를 동시에 분석할 수 있는 고효율, 고가의 새로운 연구 도구로 자리잡았으며 마이크로어레이 실험 자료의 올바른 분석을 위해서는 실험 목적에 맞는 실험계획법의 확립과 통계분석법의 적용이 중요하다. 본 논문에서는 마이크로어레이 자료에서 여러 군 사이에서 발현의 차이를 보이는 유전자를 찾을 수 있는 되돌림 설계를 소개하고 ANOVA 모형을 이용하여 분석하는 방법을 제시한다. 연세대학교 암전이연구센터의 되돌림설계를 이용한 백혈병 자료를 MA-ANOVA(Wu et al.(2003))를 이용하여 분석하였다.

주요용어: 마이크로어레이 실험, ANOVA 모형, 되돌림설계, 분산분석,집락분석

1. 서론

인류의 건강과 생명의 현상을 해석하기 위한 게놈 차원의 유전자 기능과 변화에 대한 연구가 중요시 되면서 1995년 미국 Stanford 대학에서 개발된 마이크로어레이 기술은 하나의 어레이 위에서 수 천-수만개 유전자의 발현 양상을 동시에 밝히면서 유전자간의 상호작용을 포괄적으로 관찰할 수 있어 생명과학 분야의 고효율 분석 연구 도구로 자리잡았다.

서로 다른 처리를 하였거나 종류가 다른 두 개의 표본(specimen)에서 추출한 mRNA에 각각 Cy3-dUTP(Cy3)와 Cy5-dUTP(Cy5)의 형광염기를 집어넣고 역전사 (reverse transcription) 시킨다. 이때 합성된 두 가지 cDNA를 같은 양으로 섞어서 유전자와 보합반응을 시키면 이들의 발현정도는 mRNA에 표지된 형광염기에 따라 다른 색으로 나타난다. 즉, Cy3는 녹색 형광을 띠고 Cy5는 적색 형광을 띠는데 스캐너를 이용하여 측정한 두 형광강도는 유전자와 각 표본의 발현강도를 수치화한 것이다. 마이크로어레이 실험은 동시에 수만개 유전자의 보합작용을 가능하게 만들었고 여기서 얻은 정보를 분석하여 생명현상의 기전을 이해하고 발현의 형태가 다른 유전자(특이유전자)들을 찾아낼 수 있게 된 것이다. 또한 특이유전자들을 바탕으로 한 암 발병과 전이에 관한 연구는 암의 본질 규명과 이를 이용한 조기진단 및 치료 방안을 제시할 수 있다(Golub et al.(1999), Alizadeh et al.(2000), Huang et al.(2003), Nutt et al.(2003)).

마이크로어레이 실험은 연구의 목적, 비교하는 mRNA sample의 종류와 양, 가능한 슬라이드의 수 등에 따라 실험계획이 변화한다. 또한 실험의 특성상 여러 종류의 변이가 발생

* 본 연구는 한국과학재단 목적기초연구(R04-2003-000-10145-0)지원으로 수행되었음.

1) (143-747) 서울시 광진구 군자동 98, 세종대학교 응용수학과, 부교수
E-mail: leesh@sejong.ac.kr

하기 때문에 반복실험이 중요하지만 경제적인 이유로 충분한 반복실험이 이루어지지 못하는 경우가 대부분이다. 그러므로 안정적인 실험 결과와 효율적인 자료분석을 위해서는 실험 목적과 주어진 제약조건 뿐만 아니라 분석 방법까지 염두에 둔 실험계획법 확립이 중요하다.

본 논문에서는 3개 이상 처리군 사이에서 발현 형태의 차이를 비교할 수 있는 되돌림 설계를 이용한 마이크로어레이 실험 자료의 분석 방법을 다루었다. 2장에서는 되돌림 설계를 설명하고, 3장에서는 마이크로어레이 자료분석을 위한 ANOVA 모형을 설명하였다. 4장에서는 되돌림설계를 하여 얻은 9개 어레이의 백혈병 자료를 바탕으로 급성백혈병, 재생불량성빈혈과 정상 상태에서 발현 형태가 다른 유전자들을 검색하였다.

2. 되돌림설계

cDNA 마이크로어레이 실험은 두 가지 형광 표지를 이용한 크기가 2인 블럭을 구성하므로 세 가지 이상의 처리군, $V_1, \cdots, V_v$ 을 비교하기 위해서는 불완전 블럭설계(incomplete block design) 중에서 분석 목적을 충족할 수 있는 실험설계를 구상하는 것이 필수적이다.

'$A \rightarrow B$'는 Cy3를 표지한 처리 A의 표본과 Cy5를 표지한 처리 B의 표본을 같은 어레이에서 보합반응시킨 것을 의미할 때 $V_1 \rightarrow V_2$, $V_2 \rightarrow V_3$, $\cdots$, $V_v \rightarrow V_1$ 으로 비교하는 방법을 되돌림설계(loop design)(그림 2.1))라 한다.

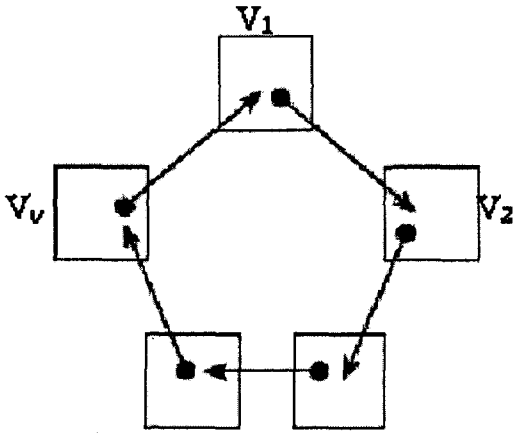

그림 2.1: 마이크로어레이 실험의 되돌림설계

이 설계방법에는 뚜렷하게 상반되는 장단점이 있다. 실험자 입장에서는 비교대상의 모든 표본에 Cy3와 Cy5의 두 가지 형광물질을 표지하는 과정이 번거롭기 때문에 지금까지 발표된 마이크로어레이 관련 논문들에서 되돌림 설계를 이용한 실험은 극히 드물다. 만약 실험이 끝난 후 새로운 표본을 추가하는 경우 이와 함께 짝을 이룰 표본들의 RNA가 남아 있을 때는 가능하지만 한 환자에서 채취가능한 RNA는 소량이므로 실제로는 불가능하다.

또한 실험결과 중 상태가 좋지 않은 어레이가 끼어 있을 때는 그를 제외한 나머지를 대상으로 분석하는 것도 불가능하다. 그러나 통계학자의 입장에서 보면 각 표본에 대해 두 가지 형광물질을 모두 표지하여 반응값을 관찰함으로써 분석의 목적이 되는 처리 효과가 염료효과와 서로 균형(balance)을 이루고 이것은 ANOVA 모형을 이용한 분산분석을 가능하게 하는 큰 장점이 되는 것이다. 또한 일반적으로 사용되는 준거설계(reference design)에서는 분석 대상인 각 표본에 대해 단 하나의 관찰값만을 얻어 이상값이 있을 경우 결과에 많은 영향을 미치지만 되돌림 설계에서는 두개의 관찰값을 얻게 되어 자료 분석이 훨씬 효율적이라는 큰 장점이 있다. 그러나 되돌림 설계에서 서로 멀리 떨어져 위치한 두 표본을 비교하는 경우 이 두 표본들 을 연결하는 사이에 끼인 다른 표본들에 의한 간접적 효과가 부수적인 변이를 발생하는 문제도 있다.

환경이 물고기들의 유전자 발현에 미치는 영향을 알아내기 위한 연구에서 Oleksiak et al.(2002)은 서식지가 다른 세 종류의 어류에서 각 5마리씩의 물고기로부터 8번 반복하여 RNA를 추출하고 이것을 되돌림 설계를 이용하여 보합반응시켜 60장의 어레이 자료를 얻었다. 두 번의 독립적인 분산분석을 하여 세 종류 어류 사이에 발현의 차이를 보이는 유전자와 15마리 물고기 사이에 발현의 차이를 보이는 유전자를 검색하였다.

3. 되돌림 설계의 마이크로어레이 자료분석을 위한 모형 설정

3.1. 변이의 종류

마이크로어레이 실험의 가장 간단한 형태는 한 가지 요인에 대한 유전자 발현의 변화를 관찰하는 것으로, 세포조직이나 종양 종류, 약물의 처리에 따른 차이들을 알 수 있고 시간에 따른 생물학적 진행과정을 관찰할 수 있다. 그런데 마이크로어레이 실험 과정에는 다른 실험보다 관찰값에 변이를 발생하는 요소가 더 많이 존재한다. 우선 여러 개의 어레이를 이용한 실험에서 사용된 어레이가 동일하지 않기 때문에 변이가 일어날 수 있는데 이를 어레이 효과(array effect, A)라 한다. 그리고 실험에 참여한 표본들의 성질이 서로 동일하지 않거나 다른 처리를 통해 얻어진 것이고 또한 이들과 반응하는 어레이에 점적된 수많은 유전자들도 그 특성이 다르므로 보합반응을 하였을 때 발현의 정도가 다르다. 이렇게 발생한 차이를 처리효과(variety effect, V)와 유전자 효과(gene effect, G)라 한다. 또한 각 표본에 형광염기를 표지하였을 때 색의 화학 특성상 녹색 형광이 적색 형광보다 발현강도가 약간 더 높게 나타나는 형광염료효과(dye effect, D)가 있다. 흔한 경우는 아니지만 유전자에 따라 특별히 발현 정도에 차이를 보이는 형광염기가 있기도 하다. 이런 여러가지 영향으로 마이크로어레이 실험 후 얻어진 발현값들은 큰 차이를 보이는 것이다.

3.2. 마이크로어레이 자료분석을 위한 ANOVA 모형 설정

n개의 유전자가 점적되어 있고 v가지 서로 다른 처리를 한 a개의 어레이 실험 자료에서 y_{ijkg}는 i번째 어레이($i = 1, \cdots, a$)에서 j번째 형광표지(j=1: Cy5 표지, j=2 : Cy3 표지)와 k번째 처리($k = 1, \cdots, v$)을 한 표본과 반응한 g번째 유전자($g = 1, \cdots, n$)의 발현강도를 로그 변환한 값이라고 하자. 이 때 y 에 영향을 주는 기본적인 요인은 A, D, V, G의 주효과와

이들의 교호작용(interaction)으로 나타나는 효과들이다. 마이크로어레이 실험의 특성상 여러 교호작용중 A와 D의 교호작용은 V와 교락(confound)되어 있고, A와 V의 교호작용은 D와, D와 V는 A와 교락되어 있다. 또한 3요인 이상의 교호작용도 주요인,또는 2요인의 교호작용 효과와 교락되어 있어 이들의 분석에 아래의 ANOVA 모형이 제시되었다(Kerr and Churchill(2001a), Kerr, Martin and Churchill(2000)).

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + (DG)_{jg} + \epsilon_{ijkg} \tag{3.1}$$

여기서 μ는 전체 관찰값들의 평균이고, ϵ은 기대값이 0인 서로 독립인 확률오차를 의미하며 $\sum_i A_i = \sum_j D_j = \sum_k V_k = \sum_g G_g = \sum_k (VG)_{kg} = \sum_g (VG)_{kg} = \sum_i (AG)_{ig} = \sum_g (AG)_{ig} = \sum_j (DG)_{jg} = \sum_g (DG)_{jg} = 0$ 을 만족한다. 여기서 제시된 고정모수효과모형(fixed effect model)은 계산상의 편리에 의해 사용되었고 Wolfinger et al.(2001)은 변량 모형(random effect model)이 더 좋은 경우도 있음을 보였다.

분석 모형에서 발현강도를 그대로 사용하지 않고 로그 변환을 하는 이유는 승법적으로 연결되어 발현강도의 차이에 영향을 주는 효과들을 변수변환을 함으로써 가법적인 관계로 바꾸어 모형에 대한 분석이 쉬워지기 때문이다. 또한 변환하지 않은 자료와 제곱근이나 역수 등의 다른 변환을 사용한 자료들과 비교한 결과 로그 변환이 발현강도 분석에 제일 적합하였다(Sapir and Churchill(2000), Cui et al.(2003)).

실제로 한 마이크로어레이 실험에서는 동일한 유전자들이 점적된 어레이를 사용한다. 그러므로 실험에서 어레이, 염료와 처리의 모든 가능한 조합에서 유전자들은 같은 수로 반복되는 양상을 나타내므로 G는 다른 효과들 A, D, V와 직교한다. 그리고 마이크로어레이 자료분석에 ANOVA 모형을 적용할 때의 장점은 일반적으로 염료나 어레이의 차이를 보정해 주기 위한 표준화 과정이 따로 필요하지 않고 모형에 A와 V의 효과항을 추가함으로써 대신할 수 있다는 것이다.

우리 관심의 대상인 주어진 표본의 성질에 따라 발현의 차이를 보이는 유전자를 검색하기 위해서는 V와 G의 상호작용인 (VG) 의 추정값인 0이 아닌 유전자를 찾으면 된다.

3.3. 되돌림 설계 실험 자료의 분석

v가지 처리를 비교하기 위해 v장의 어레이가 필요한 되돌림 설계에서는 각 처리마다 염료 효과가 균형을 이루므로 효과 (VG)를 추정하는데 (DG)가 영향을 미치지 않는다. 그러므로 (3.1)의 ANOVA 모형에서 모수 (DG)를 생략할 수 있고 이것은 자유도에 여유가 생겨 식 (3.2)를 이용한 분산분석을 가능하게 한다.

$$y_{ijkg} = \mu + A_i + D_j + V_k + G_g + (VG)_{kg} + (AG)_{ig} + \epsilon_{ijkg} \tag{3.2}$$

분석에 사용되는 각 모수들의 최소제곱추정량을 구하기 위해서는 잔차 제곱합 $RSS = \sum_{ijkg} (y_{ijkg} - \mu - A_i - D_j - V_k - G_g - (VG)_{kg} - (AG)_{ig})^2$ 를 최소화하는 모수들을 찾으면 된다. 즉 RSS를 각 모수들에 대해 편미분한 식을 연립하여 구할 수 있다. '.'는 각 변수의 해당 index의 모든 가능한 경우를 평균한 것을 뜻할 때 $\alpha_{ig} = y_{i..g} - y_{i...} - y_{...g} + y_{....}$,

$\beta_{kg} = y_{..kg} - y_{..k.} - y_{...g} + y_{....}$, $\gamma_{kg} = \beta_{kg} - (\alpha_{k-1,g} + \alpha_{kg})/2$ 라 하자. 또한 $m = [v/2]$ 일 때 $(VG)_{kg}$는 아래와 같이 추정할 수 있다(Kerr and Churchill(2001a)).

$$\frac{v}{2}(V\hat{G})_{kg} = \begin{cases} m(m+1)\gamma_{kg} + \sum_{i=1}^{m-1}(m-i)(m-i+1)(\gamma_{k-i,g} + \gamma_{k+i,g}) & v: \text{홀수} \\ m^2\gamma_{kg} + \sum_{i=1}^{m-1}(m-i)^2(\gamma_{k-i,g} + \gamma_{k+i,g}) & v: \text{짝수} \end{cases}$$

g번째 유전자가 서로 다른 처리 사이에서 발현의 차이를 보이는지 검정하지 위한 가설은 모형 (3.2)에서 아래와 같이 표현된다.

$$H_0 : VG_{1g} = VG_{2g} = \cdots = VG_{vg} = 0 \quad \text{대} \quad H_1 : H_0\text{은 사실이 아님.} \tag{3.3}$$

H_0과 H_1 에서의 g번째 유전자의 잔차제곱합을 RSS_{0g}과 RSS_{1g}이라 하고 자유도를 df_0, df_1이라 하면 분산분석에서 가설 (3.3)에 대하여 아래의 검정통계량이 사용된다.

$$F = \frac{(RSS_{0g} - RSS_{1g})/(df_0 - df_1)}{RSS_{1g}/df_1} = \frac{\sum_{k=1}^{v}(V\hat{G})_{kg}^2/(v-1)}{RSS_{1g}/df_1} \tag{3.4}$$

ϵ이 정규분포를 따르면 검정통계량 F는 자유도가 $(v-1, n-1)$인 F분포를 따른다.

마이크로어레이 실험에서 자료가 방대하다는 것은 유전자 수가 많기 때문이고 표본의 수는 극히 적으므로 주어진 한 유전자가 특이유전자인지 판단하기 위한 통계량 (3.4)은 검정력이 낮다. 이런 단점을 보완하기 위해 모든 유전자들이 동일 분산을 갖는다는 가정아래 합동분산추정량을 사용하여 검정력을 높일 수도 있지만 가정에 위배되는 경우 허위양(false positives)을 발생시키는 위험이 있다.

4. 자료분석

4.1. 실험 자료

연세대학교 의과대학에서는 백혈병의 진단에 쓰이는 유전자를 찾기 위해 세브란스병원에 내원한 급성백혈병(Acute leukemia, L) 및 재생불량성빈혈 (Aplastic Anemia, A)로 진단된 환자와 정상인(Normal, N)의 유전자 발현을 비교하였다. 각 군의 대상자 골수를 채취하자마자 골수내 기질세포를 분리하여 total RNA를 추출하였는데 1인의 검체에서 추출되는 RNA양이 적어서 각 군당 2명의 RNA를 합쳐서 사용하였다. Total RNA 4μg을 증폭(amplification)시킨 후 17,000개의 인간유전자로 구성된 cDNA 마이크로어레이에 보합반응을 시켰다. 이 때 실험은 되돌림설계를 시행하였고 각 실험을 3번 반복하여 총 9개의 마이크로어레이 자료를 얻었다. 식 (3.2)에서 $i = 1, \cdots, 9$, $k = 1(L), 2(N), 3(A)$ 임을 알 수 있다.

9장의 어레이를 이용한 실험에서 모든 유전자는 18개의 관찰값을 얻을 수 있다. 되돌림설계를 이용한 우리의 백혈병 자료는 세가지 처리, L,N와 A 각각에 대해 Cy3 형광표지와 Cy5 형광표지에 대한 보합반응 결과를 균형있게 3개씩 모두 6개의 관찰값을 얻었다. 만약 준거설계를 시행하였다면 비교 대상인 세 처리 모두 Cy5 형광 표지에 대한 3개씩의 결과만 얻을 수 있고, 공통 준거의 역활을 할 뿐 우리의 관심 대상이 아닌 Cy3 형광표지를 한 표본에서 9개의 관찰값을 얻었을 것이다.

4.2. 자료 정리와 표준화 작업

암전이센터에서 사용한 어레이는 스팟이 17664개 있고 이를 32개의 프린트 팁을 사용하여 유전자를 점적하였다. 백혈병자료는 16835개 스팟에 유전자를 점적하였지만 9장 어레이 모두에서 발현강도를 측정할 수 있는 유전자는 15593개였다. 이중 2-4회 중복된 유전자는 각 발현의 평균값을 구하여 대표값으로 삼으니 분석 대상의 서로 다른 유전자는 14429개($g = 1, \cdots, 14429$)로 줄어들었다.

실험이 안정적이지 못할 경우 결측치가 많이 생기므로 결측치를 대치하는 방법이 중요하다. Troyanskaya et al.(2001)가 여러가지 대치방법을 비교하고 경우마다 효율적인 방법이 다름을 보였다.

마이크로어레이 실험은 다른 실험보다 변이를 유발하는 과정이 많은데 Cy3 형광 염료가 Cy5 형광 염료보다 강도가 높은 경향과 유전자를 점적하는 프린트 팁과 점적하는 위치에 따라 발현강도에 차이가 나타나는 체계적 변동원을 제거하여야 실험의 목적인 처리 효과를 올바로 비교할 수 있다. 비교 대상이 되는 처리와는 무관하게 발생되는 이러한 변동을 제거하는 과정이 표준화(normalization)이고 본격적인 자료 분석 전에 반드시 선행되어야 할 작업이다. 이에 대한 연구는 마이크로어레이 실험의 다른 분야와는 달리 이미 많은 연구 결과가 나와 있다(Callow et al.(2000), Dudoit et al. (2002), Yang et al.(2000,2002), Kim et al.(2003))

어레이의 한 스팟에서 얻은 두 표본의 로그변환 발현강도 (y_{i1kg}, y_{i2kg})로부터 두 값의 차이를 $M = y_{i1kg} - y_{i2kg}$, 평균을 $A = (y_{i1kg} + y_{i2kg})/2$라 할 때, 분석 대상 유전자의 M-A 산포도를 그리면 특이 유전자 비율이 낮다는 가정 아래에서는 각 점들이 $M = 0$을 중심으로 일정한 띠 안에 분포되어 있을 것이기 때문에 전체 유전자의 M값의 중위수를 0으로 보정하는 전체 중위수 보정(global normalization), 강도가 낮은 지역에서 M값이 낮게 나타나는 경우 비모수적 추세선을 $M = 0$와 일치하도록 하는 강도 의존적 보정(intensity dependent normalization), 프린트 팁에 의한 공간 효과와 강도 의존적 추세를 동시에 보정하여 주는 프린트 팁별 강도 의존적 보정(intensity dependent print tip normalization) 등 다양한 표준화 방법이 있다.

그림 4.1은 백혈병 자료 중 한 개 어레이의 관찰값에 대한 M-A 산포도와 32개 팁별 상자그림이고 나머지 8개의 어레이도 모두 비슷한 형태를 보였다. M-A 산포도에서 보면 전체적으로 강도 의존적 추세를 보이는 동시에 32개 팁별로 추정한 lowess 적합선에도 차이가 있음을 볼 수 있다. 또한 1.8K 어레이는 32개의 프린트 팁에 의한 평면이 구성되어 있는데 상자그림을 보면 이중 프린트 팁이 4, 8, $\cdots$, 32에 해당하는 오른쪽 끝에 위치한 8개의 블록에서 다른 프린트 팁들과 구별되는 양상을 보였기 때문에 프린트 팁별 강도 의존적 방법을 사용하여 자료를 표준화하였다. 이 때 강도 의존적 보정은 보정되어야 할 값 c_{ikg}를 이용하여 관찰값 (y_{i1kg}, y_{i2kg})을 ($y_{i1kg} - c_{ikg}/2$, $y_{i2kg} + c_{ikg}/2$) 로 변환시킨다(Kerr et al.(2002)).

프린트팁별 중위수 보정 외에 퍼짐 보정(scale normalization)과 슬라이드간 보정(multi-slide normalization) 등을 고려할 수 있으나 백혈병 자료의 각 어레이 관찰값에 대해 상자그림을 그려본 결과 이에 대한 표준화는 필요없다고 판단되었다.

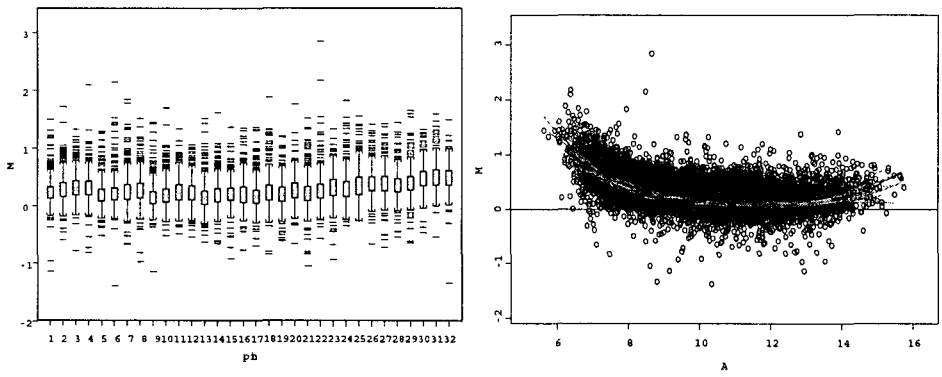

그림 4.1: 한 개 슬라이드 자료에 대한 팁별 상자그림, 전체 MA 산포도와 팁별 lowess 적합선

4.3. 특이 유전자 검색

g번째 유전자가 A, L과 N 사이에서 발현의 차이를 보이는지 검색하기 위한 가설은 (3.2)에서 아래와 같이 설정할 수 있다.

$$H_0 : VG_{1g} = VG_{2g} = VG_{3g} = 0 \quad 대 \quad H_1 : H_0 \text{ 은 사실이 아님} \tag{4.1}$$

위의 가설을 검정하는 분산분석으로 미국 Jackson 연구소의 Churchill group에서 개발한 마이크로어레이 자료분석을 위한 software package인 MA-ANOVA(MicroArray Analysis Of VAriance)(Wu et al.(2003))를 이용하였다.

분산분석에서 F-검정을 하기 위한 전제조건은 자료들이 정규성과 등분산을 만족하는 것인데 마이크로어레이 자료는 이런 가정을 만족시키지 못하므로 정규성에 기인한 신뢰구간의 계산이나 등분산성에 바탕을 둔 F-검정이 불가능하다. MA-ANOVA는 잔차를 permutation하는 방법을 사용하여 검정을 하였고 비모수 F-검정의 단점을 보완하기 위해 (3.4)의 F-통계량(F_1)외에 자료의 다른 측면을 보여줄 수 있는 수정된 두가지 통계량(F_2, F_3)을 곁들여 제공하였다.

F_1은 고전적인 F-통계량으로 분모는 $Var(\epsilon_{ijkg}) = \sigma_g^2$ 의 가정에서 추정한 $\hat{\sigma_g^2}$가 사용된다. 분모가 한개의 유전자에 국한하여 구해진 통계량 F_1은 실험에 참여한 표본의 수가 적으므로 검정력이 낮다는 단점이 있고(Cui et al.(2004)) 이를 개선하기 위해 통계량 F_3는 $Var(\epsilon_{ijkg}) = \sigma^2$ 의 가정 아래 모든 유전자로부터 구한 합동 분산 추정량 $\hat{\sigma_{pool}^2}$을 분모로 사용하였다. 또한 F_1과 F_3의 절충안으로 제시된 F_2는 분모에 $\hat{\sigma_g^2}$와 $\hat{\sigma_{pool}^2}$의 평균을 사용하였다.

그림 4.2는 14429개의 유전자를 각 유전자의 F_1 통계량의 유의수준과 F_3의 통계량의 값을 수직축과 수평축으로 하여 산포도로 나타낸 것이다. 수평선의 위쪽에 위치한 380개의 유전자는 F_1에 의해 유의수준 $\alpha = 0.000005$에서 특이 유전자로 검색된 것이고 수직선 오른쪽에 위치한 188개 유전자는 $\alpha = 0.002$일 때 F_3를 사용하여 검색된 유전자이다. 통계량 F_2를 사용하여 $\alpha = 0.002$일 때 검색된 148개 유전자는 $\otimes$로 표시하였다.

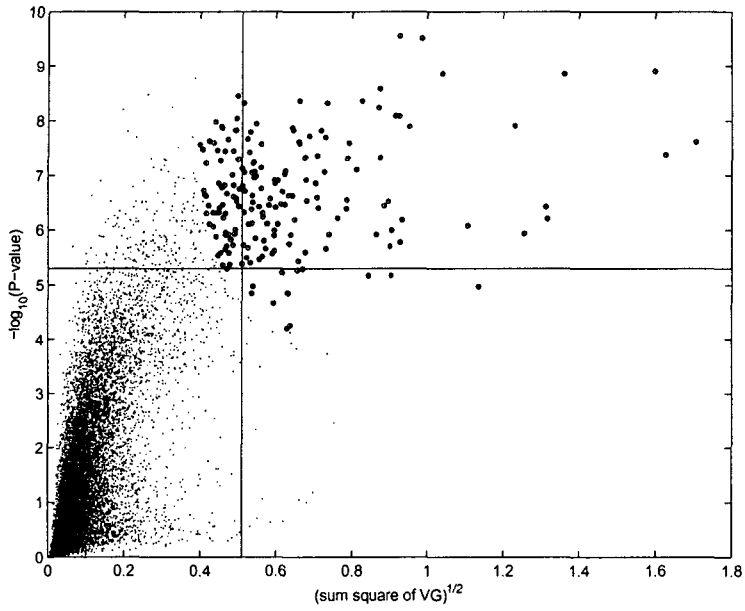

그림 4.2: 세가지 검정통계량을 이용한 특이유전자 검색

4.4. 군집분석

F_1, F_2와 F_3의 세 가지 검정통계량 모두에 의해서 특이 유전자로 검색된 115개 유전자를 대상으로 k-평균 군집화(k-means clustering)를 실시하여 표본에 따라 발현의 형태가 유사한 군집으로 분류해 보았다. 붓스트랩방법(bootstrap method)을 500회 반복하여 115개 유전자를 5개의 군집으로 분류하고 같은 군집에 400회 이상 분류되었던 유전자만 그 군집에 속한다고 인정하였다(Kerr et al.(2001b)). 그림 4.3은 5개의 군집으로 분류된 유전자들과 어느 군집에도 속하지 못한 나머지 유전자들의 발현 형태를 나타낸 것이다.

군집분석은 병의 진전에 따라 분자 유전학적인 해석이나 최소한의 모형 설정을 가능하게 한다. 예를 들어 첫번째 군집은 N과 A에서는 평형을 유지하다가 L에서는 증가된 기능을 나타내는 형태를 보이는 24개 유전자군으로 형성되어 있다. 단백질의 성질상 모든 유전자가 치료를 목적으로하는 인위적 조작이 가능한 것은 아니므로 비슷한 기능을 갖는 후보 유전자군을 찾는다는 것은 의미있는 일이다.

백혈병의 치료에 기여하기 위해서 각 군집에 속한 유전자들의 특성을 연구하고 질병과의 관계를 밝히는 것이 의학자와 생물학자들의 몫일 것이다.

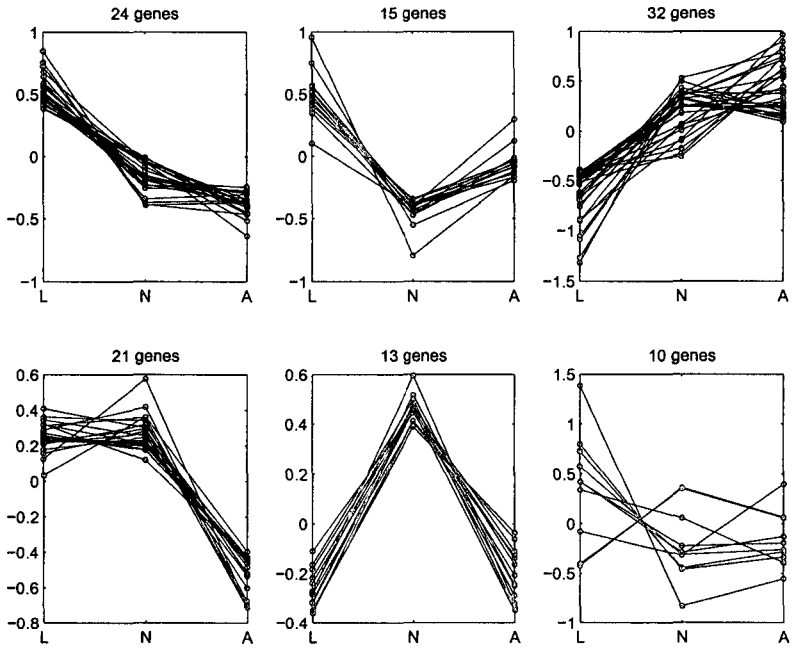

그림 4.3: 군집 분석에 의한 특이 유전자들의 분류

5. 결론 및 토의

마이크로어레이 실험에서 사용하는 어레이 숫자가 동일할 경우 흔히 쓰이는 준거설계보다는 되돌림설계가 자료 손실이 없어 훨씬 효율적이라는 것은 확실하다. 두 설계의 장단점을 제대로 비교하기 위해서는 같은 표본을 이용하여 두가지 설계의 결과를 비교하는 것이 바람직하나 마이크로어레이 실험이 워낙 경비가 많이 들고 환자의 골수에서 채취할 수 있는 RNA 양이 적기 때문에 본 연구에서는 되돌림 설계를 이용한 마이크로어레이 실험 자료만 분석 가능하였다.

특이 유전자를 검색하는 통계적인 방법은 제일 단순한 형태의 fold change rule을 시작으로 이를 개선한 Newton et al.(2001)의 방법, Dudoit et al.(2002)의 t-통계량을 이용한 방법, Tusher et al.(2001)가 제안한 SAM (Significance Analysis of Microarrays)이 많이 쓰이는데 이들 모두 두 시료의 발현강도 상대비를 분석에 사용한다. 되돌림설계에서는 상대발현비보다는 두 가지 시료 각각의 발현강도가 중요하므로 ANOVA 모형을 이용한 분산분석이 최적의 통계분석법이라 하겠다.

앞에서 언급한 것과 같이 백혈병 자료의 실험은 각 처리당 2명의 RNA를 합쳐서 사용하였다. 실험에서 생물학적 복제 관찰치(biological replicate)를 두는 것은 분석에 개인간의 변동을 반영하기 위해서지만 생물학적 변동의 효과를 줄이기 위해 일부러 RNA를 합치기도 한다. 이번 실험처럼 RNA가 부족한 경우에는 합치는 것이 유일한 수단이지만 그렇지

않은 경우 RNA를 합치는 것과 합치지 않는 것이 상황에 따라 결과에 어떤 차이를 보이는지 연구 발표된 결과들이 있다(Kendziorski et al.(2003), Pheiffer et al.(2002)). 또한 각 처리당 2명의 표본만 사용되었기 때문에 검색된 유전자들이 일반적인 특성보다는 개인의 특정 성향을 반영하여 결과에 대한 신뢰도가 높지 않다. 그렇지만 본 논문이 급성 백혈병, 재생 불량성 빈혈과 정상인 사이에 발현의 차이를 보이는 유전자를 검색하기 보다는 되돌림 설계를 이용한 마이크로어레이 실험 자료를 ANOVA 모형을 사용하여 통계적으로 분석하는 방법을 제시하는 것이 목적이었다는 것을 다시 한번 강조한다.

감사의 글

본 연구를 위해 마이크로어레이 실험자료를 제공해 주신 연세대학교 의과대학 암전이 연구센터(Cancer Metastasis Research Center)에 감사드린다. 또한 논문이 개선될 수 있는 좋은 지적을 주신 편집위원과 두 분 심사위원께도 감사드린다.

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., 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., Botstein, D., Brown, P.O., Staudt, L.M.(2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature Genetics,* **403**, 503-511.

Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M. (2000). Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Research,* **10**, 2022-2029.

Cui,X.Q., Hwang, J.T., Qiu,J., Blades, N.J. and Churchill,G.A. (2004). Improved statistical tests for differential gene expression by shrinking variance components, submitted. Posted on http://www.jax.org/staff/churchill/labsite/pubs.

Cui, X.Q., Kerr, K. and Churchill, G.A. (2003). Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology,* **2**, article 4.

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

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science,* **286**, 531-537.

Huang, E., Ishida, S., Pittman, J., Dressman, H., Bild, A., Kloos, M., Amico, M.D., Pestell, R.G., West, M., Nevins, J.(2003). Gene expression phenotypic models that predict the activity of oncogenic pathways. *Nature Genetics,* **34**, 226-230.

Kendziorky, C.M., Zhang, Y., Lan, H., Attie, A.D. (2003). The efficienty of pooling mRNA in microarray experiment. *Biostatistics,* **4**, 465-77.

Kerr, K., Churchill, G.A.(2001a). Experimental design for gene expression microarrays. *Biostatistics,* **2**, 183-201.

Kerr, K., Churchill, G.A.(2001b). Bootstrapping cluster analysis : Assessing the reliability of conclusion from microarray experiments. *Proceedings of the National Academy of Sciences,* **98**, 8961-8965.

Kerr, K., Martin, M., Churchill, G.A.(2000). Analysis of variance for gene expression microarray data. *Journal of Computational Biology,* **7**, 819-837.

Kim, B.S., Lee, S., Rha, S.Y., Chung, H.C.(2003). cDNA microarray experiment : Design issues in early stage and the need of normalization. *Cancer Research and Treatment,***35**, 533-40.

Newton, M.N., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology,* **8**, 37-52.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E., Batchelor, T.T., Black, P.M., Deimling, A.V., Pomeroy, S.L., Golub, T.R., Louis, D.N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research,* **63**, 1602-1607.

Oleksiak, M.F., Churchill, G.A., Crawford, D.L.(2002). Variation in gene expression within and among natural populations. *Nature Genetics,* **32**, 261-266.

Pheiffer, R.M., Rutter, J.L., Gail, M.H., Struewing, J. and Gastwirth, J.L.(2002). Efficiency of DNA pooling to estimate joint allete frequencies and measure linkage disequilibrium. *Genetic Epidemiology,* **22**, 94-102.

Sapir, M. and Churchill, G.A. (2000). Estimating the posterior probability of differential gene expression from microarray data, poster, www.jax.org/research/churchill/pubs /index.html.

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

Tusher, V.,Tibshirani, R., Chu, G.(2001). Significance analysis of microarray applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences,* **98**, 5116-5121.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Benett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.(2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology,* **8**, 625-37.

Wu, H., Kerr, K. and Churchill, G.A.(2003). MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments, *Chapter of the analysis of gene expression data: methods and software,* Springer.

Yang, Y.H., Buckley, M.J., Dudoit, S., Speed, T.P. (2000). Comparison of methods for image analysis on cDNA microarray data. *Technical Report,* **581,** Dept. of Statistics, Univ. of California at Berkeley.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngal, J. and Speed, T. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research,* **30** , 4: e15.

[2004년 3월 접수, 2004년 8월 채택]

Statistical Analysis of a Loop Designed Microarray Experiment Data *

Sunho Lee [1)]

ABSTRACT

Since cDNA microarray experiments can monitor expression levels for thousands of genes simultaneously, the experimental designs and their analyzing methods are very important for successful analysis of microarray data. The loop design is discussed for selecting differentially expressed genes among several treatments and the analysis of variance method is introduced to normalize microarray data and provide estimates of the interesting quantities. MA-ANOVA is used to illustrate this method on a recently collected loop designed microarray data at Cancer Metastasis Research Center, Yonsei University.

Keywords: Microarray experiment; Reference design; Loop design; Analysis of variance, Cluster analysis.

* This work was supported by grant No. R04-2003-000-10145-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

1) Associate Professor, Dept. of Applied Mathematics, Sejong University. 98, Gunjadong, Kwangjinku, Seoul 143-747, Korea.
E-mail: leesh@sejong.ac.kr