한국통계학회논문집
2009, 16권, 3호, 397–408

마이크로어레이 자료분석에서 모수적 방법을 이용한 유전자군의 유의성 검정

이선호[1,a], 이승규[a], 이광현[a]

[a]세종대학교 응용통계학 전공

요 약

마이크로어레이 기술은 수만 개 유전자의 발현 패턴을 동시에 관찰하는 것을 가능하게 하였고, 이들을 하나씩 검정하여 찾아낸 특이발현 현상을 보이는 유전자를 중심으로 질병의 진단, 치료법 정립과 신약 개발을 위한 기본 정보를 확립하였다. 그러나 개별 유전자 분석의 여러 문제점이 발견되면서 유전자들을 생물학적 대사경로나 염색체 위치가 같은 것끼리 묶은 집단을 분석하여 질병의 발생이나 생존에 영향을 미치는 집단을 찾는 방법이 제시되었다. 이러한 유전자 집단의 유의성에 대한 연구는 2002년에 MIT에서 비롯되어 GSEA, SAM-GS와 중심극한 정리의 개념을 이용한 모수적 방법인 PAGE 등이 사용되고 있다. 본 논문에서는 이들 통계량의 구조적 한계를 극복하고 계산이 간단한 새로운 모수적 방법을 제안하고 자료 분석을 통하여 효율성을 보였다.

주요용어: 마이크로어레이 실험, 개별 유전자 분석, 유전자 집단 분석, 정규점수.

1. 서론

1995년 미국 Stanford 대학교에서 동시에 수천 개 이상의 유전자를 관찰할 수 있는 마이크로어레이 기술이 개발된 이후로 종양에 대한 유전학적 특성과 기전 연구가 더욱 활발해졌고 질병의 진단과 치료를 하는데 큰 기여를 하였다.

초기의 마이크로어레이 자료 분석은 각 유전자를 대상으로 서로 다른 표현형(phenotype: 질병군/정상군, 종양의 분류나 전이 여부, 병기 등) 사이에 발현 차이를 보이는 특이발현 유전자(differentially expressed genes)를 찾는 작업에서부터 시작되었다 (Tibshirani 등, 2002; Tusher 등, 2001). 개별 유전자분석으로 검색된 특이발현 유전자가 소수일 때는 그들의 기능이나, 어떤 대사경로와 관련 있는지 조사가 가능하다. 그러나 일일이 조사하기에 많은 수의 유전자가 검색되거나 또는 하나도 검색되지 않은 경우에는 후속 연구에 어려움이 생긴다. 또한 개별 유전자분석에서 동일한 자료라도 분석 방법을 다르게 하거나, 동일 질병에서 다른 자료를 사용하여 분석을 하였을 때 각기 얻어진 결과들 사이에 공통점을 찾기 힘들었다 (Curtis 등, 2005; Subramanian 등, 2005). 이러한 문제점을 보완하기 위해 유전자들을 개별적으로 분석하는 대신 기능이나 대사작용이 비슷한 것끼리 묶어 집합 단위로 분석(gene set analysis)하는 것으로 관심이 옮겨졌는데 유전자군 분석은 수만 개 유전자를 대상으로 개별분석을 하였을 때 생기는 다중검정의 문제를 완화시키는 장점도 있다. 유전자군 분석 방법으로 GSEA(Gene Set Enrichment Analysis) (Mootha 등, 2003; Subramanian 등, 2005), GSA(Gene Set Analysis) (Efron과 Tibshirani, 2007), PAGE(Parametric Analysis of Gene set Enrichment) (Kim과 Volsky, 2005)와 SAM-GS(Significance Analysis of Microarrays to Gene-Set analysis) (Dinu 등, 2007) 등이 개발되어 사용되고

이 논문은 2008년 교육인적자원부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(2008-531-C00019).
[1] 교신저자: (143-747) 서울시 광진구 군자동, 세종대학교 응용통계학 전공, 교수. E-mail: leesh@sejong.ac.kr

있는데 본 논문에서는 이러한 분석방법을 연구하여 단점을 보완하고 더 정확하고 빠른 방법으로 표현형 사이에 발현 차이를 보이는 유전자군을 찾는 방법을 개발하는 것이 목적이다.

2장에서는 기존 방법의 특징과 문제점을 알아보고 이를 보완할 수 있는 효율적인 새 방법을 3장에서 제안한다. 4장에서는 실제 자료를 이용하여 새 방법의 모수적 가정에 대한 적합성과 검정의 정확성을 검증하였다.

2. 기존의 분석방법

마이크로어레이 자료 분석에서 '어떤 기능, 또는 어떤 대사경로를 수행하는 유전자군이 가장 중요한가?'라는 문제에 제일 먼저 접근한 사람들은 미국 Columbia 대학교의 연구진들이었다. 그들은 기능이 같거나 생물학적 정보를 공유하는 유전자들로 class를 구성한 후 class에 속한 유전자를 대상으로 모든 짝진 유전자들의 상관계수 평균이나, 유전자별 분산분석에서 얻은 p-value를 이용하는 방법으로 class의 점수를 매기고 permutation을 사용하여 유의한 class를 선택하는 functional class scoring 방법을 제시하였다 (Pavlidis 등, 2002). 그러나 상관계수를 이용하는 경우, 동일한 대사경로에 참여하는 유전자들의 상관계수가 항상 높은 것이 아니며, 반대로 유의한 class에 속한 유전자들이 모두 상관계수가 높다고 하여 특이발현 유전자는 아니므로 적용에 문제가 있다.

Goeman 등 (2004)은 score 검정을 이용하여 관심 유전자군이 환자들의 임상 결과(종양/정상, 전이 여부 등)와 관련 있는지 검정하는 global test를 유도하였고, 이 아이디어를 각 대사경로가 환자들의 생존율과 관계있는지 검정가능한 통계량으로 영역을 확대하였다 (Goeman 등, 2005). Comparative p-value의 개념으로 관심 유전자군에 속한 유전자 수를 고려하여 p-value를 비교할 수 있는 장점이 있으나 이 방법은 관심 유전자군에 속한 유전자들의 발현정보만 사용하고 기타 유전자들의 발현 정보는 사장시킨다는 아쉬움이 있다.

전체 유전자중 특이발현 유전자들을 검색한 후, 관심 유전자군과 특이발현 유전자군 사이의 관련 여부를 검정하는 카이제곱 검정 (Khatri 등, 2004), Fisher의 Exact test (Drăghici 등, 2003)와 초기하분포를 이용한 z-score 검정 (Doniger 등, 2003) 등이 있다. 그런데 이 방법들은 관심 유전자군에 속한 유전자들이 특이발현 유전자군에 속하는지 포함 여부만을 따지고 그 정도는 고려하지 않을 뿐 아니라 특이발현 유전자를 정하는 기준이 달라짐에 따라 전체 검정 결과가 달라진다는 문제가 있다.

Mootha 등 (2003)은 특이발현 유전자의 포함 여부에서 한걸음 더 나아가 상위(top-ranking) 특이발현 유전자가 관심 유전자군에 얼마나 포함되었는지를 분석에 반영하는 GSEA를 개발하였고 Subramanian 등 (2005)은 이를 발전시켜 GSEA-P라는 software tool을 공개하였다. 전체 유전자 중 관심 유전자군에 속한 유전자들의 상대적인 특이발현 순위를 Kolmogorov-Smirnov 통계량을 사용하여 검정하는 GSEA는 기존의 방법에 비해 알고리즘이 탄탄하여 현재 많이 사용되고 있지만 특이발현 유전자가 존재하지 않는 경우에도 상대적으로 높은 순위의 유전자로 구성된 유전자군은 유의하다는 결론이 나오며, 낮은 순위의 특이발현 유전자로 구성된 관심 유전자군이라도 군에 속한 유전자의 수가 클 경우 높은 순위의 특이발현 유전자로 구성된 관심 유전자군보다 종양에 더 유의한 영향을 준다는 결론이 나오는 허점이 있다 (Damian과 Gorfine, 2004). Barry 등 (2005)는 두 가지의 표현형 사이에서 차이를 보이는 유전자군을 찾는 GSEA를 확장하여 표현형이 3개 이상, 연속적 또는 생존 자료에 대하여도 적용가능 하도록 하였다.

Dinu 등 (2007)이 제시한 SAM-GS는 개별 유전자분석 방법인 SAM을 유전자군 분석으로 확장한 것으로 유전자군에 속한 유전자들의 SAM-t 통계량의 재곱합을 이용하여 유전자군의 유의성을 검증하였다. 이광현과 이선호 (2008)는 집단에 속한 유전자중 환자의 표현형에 따라 어느 정도 이상의 발현 차이를 보이는 유전자들을 대상으로 그들의 발현 방향성은 무시하고 크기만 사용하여 분석하는 절대치

와 절삭을 이용한 방법을 제시하였다.

그런데 이러한 방법들은 모두 비모수적인 방법으로 유전자군의 유의성을 판단하기 위하여는 permutation을 실시하여야 하는데 워낙 유전자 자료가 방대하므로 시간이 많이 걸리는 큰 문제가 있다.

Kim과 Volsky (2005)는 관심 유전자군에 속한 유전자들의 수가 충분히 클 때 그들의 로그발현비(log(정상군에서의 평균발현값/질병군에서의 평균발현값))의 평균은 중심극한 정리에 의하여 정규분포를 따른다는 것을 이용하는 PAGE라는 방법을 제안하였다. 그런데 PAGE는 모수적 방법으로 즉시 유의성을 검정할 수 있다는 장점이 있지만 사용된 로그발현비의 평균은 특이발현 유전자가 많이 속한 유전자군이라도 질병에 의해 과다발현하는 유전자와 발현이 억제된 유전자들이 공존할 경우 유의하지 않다는 결론이 나오게 되는 큰 취약점이 있다.

3. 새로운 방법의 제안

유전자군의 유의성을 분석하는 기존의 방법들을 분석한 결과, 효과적인 통계량을 이용한 모수적 방법 개발의 필요성을 발견하고 다음의 세 단계로 나누어 접근하였다.

i) 유전자군에 속한 각 유전자들의 특성을 나타내는 대표값 정하기.

ii) 각 유전자의 대표값을 종합하여 유전자군을 나타낼 수 있는 통계량 찾기.

iii) 통계량의 모수적 분포 구하기.

각 유전자들의 특성을 나타내는 대표값으로는 개별유전자분석에서 쓰이는 로그발현비, t-통계량, SAM의 보정된 t-통계량 등이 있고 개별 유전자의 대표값을 종합하는 방법으로 합을 이용하는 것을 고려할 수 있다. 그러나 합으로 과다발현 유전자와 발현이 억제된 유전자들이 공존하는 유전자군의 유의성을 판단할 경우 그들의 특이발현성이 서로 상쇄되어 유전자군의 유의성을 찾아내지 못하는 문제가 생긴다. 그러므로 대표값의 합보다는 모수적 분포를 가정할 수 있는 대표값의 제곱합으로 이루어진 통계량을 찾는 시도를 하였다.

개별유전자 검색에서 특이발현 여부를 잘 반영하는 t-통계량을 각 유전자의 대표값으로 사용하였을 때 $\sum t^2$이 만족하는 모수적 분포를 찾을 수 없었다. 대신 통계량의 제곱합이 모수적 분포를 만족한다고 할 때 제일 먼저 떠 올릴 수 있는 형태는 표준정규분포를 따르는 통계량들의 제곱합이 카이제곱분포를 만족한다는 것이다. 만약 유전자들의 여러 가지 대표값 중 표준정규분포를 따르는 통계량을 찾아낼 수 있다면 유전자군에 속한 유전자들의 통계량의 제곱합은 카이제곱분포를 따를 것이고 유전자군의 유의성여부에 대한 논의는 모수적 방법으로 접근 가능하게 된다.

대부분의 마이크로어레이자료에서 유전자들의 t-통계량은 정규분포를 만족하지는 않지만 중심에 몰려 있고 분포의 모양이 서로 대칭이기 때문에 평균 0, 분산 1로 만드는 표준화(standardization) 과정을 거치면 표준정규분포와 흡사한 분포를 따른다. 그러므로 아래의 절차에 따라 t-통계량을 약간 수정하여 표준정규분포를 만족하는 새로운 통계량을 제안하게 되었다.

i) 표본의 수 n, 유전자의 수 g인 마이크로어레이 발현자료로부터 표현형이 다른 두 군의 차이를 나타내는 t-통계량, $t_j (j = 1, \ldots, g)$를 구한다.

ii) t_j를 표준화하여 $t_j^S = (t_j - \bar{t})/s(t), (j = 1, \ldots, g)$를 구한다. 여기서 $s(t)$는 t값의 표준편차임.

iii) t_j^S의 순위를 이용하여 정규점수 $t_j^N = \Phi^{-1}(t_j^s$의 순위 $- 0.375)/(n + 0.25), (j = 1, \ldots, g)$를 구한다. 이때 $\Phi^{-1}(\cdot)$는 Blom (1958)의 방식을 이용한 역누적 표준정규분포를 의미한다.

iv) 각 유전자의 대표값으로 $t_j^M = (t_j^S + t_j^N)/2, (j =, \ldots, g)$를 사용한다.

t^M은 각 유전자의 표준화한 t-통계량 t^S와 정규점수 t^N의 평균으로, 정규점수를 이용하여 정규분포에 가깝게 보정한 통계량이다. 그러므로 t^S가 정규분포에서 많이 벗어나지 않는다면 t^M은 유전자의 t-통계량 순위는 유지하면서 정규분포를 만족하고, 유전자군에 속한 k개 유전자의 t^M의 제곱합 MTS(modified t square)는 아래와 같이 카이제곱분포를 따를 수 있다.

$$\text{MTS} = \sum_{j=1}^{k} \left(t_j^M\right)^2 \sim \chi^2(k).$$

분석 대상 유전자군이 표현형에 따라 유의한 차이를 나타낸다면 유전자군에 속한 유전자들 중 특이 발현 유전자가 많아 MTS의 값이 커질 것이다. 이 방법은 표현형에 따라 나타나는 각 유전자들의 발현 차이의 평균을 통계량으로 사용한 PAGE의 문제점을 보완하였고 모수적 방법으로 p-value를 구할 수 있어 permutation에 의존하는 GSEA와 SAM-GS 보다 훨씬 빠르고 간단히 분석을 할 수 있게 되었다.

4. 새 통계량의 검증

실제 마이크로어레이 자료를 이용하여 새로 제안한 통계량 MTS의 모수적 가정이 올바른지, 또한 기존의 다른 분석법들과 비교하여 유의한 유전자군을 잘 찾아내는지 알아보았다.

4.1. 자료 집합과 개별 유전자분석

GSEA의 홈페이지(www.broad.mit.edu/gsea/)에 공개되어 대표적으로 인용되는 다음의 세 가지 마이크로어레이 실험 자료를 사용하였다.

p53

NCI-60 collection of cancer cell lines으로부터 세포 자극의 신호에 반응하는 유전자 발현을 제어하는 p53 전사인자의 표적을 찾아내는 실험에 사용된 33예의 돌연변이군(MUT)과 17예의 정상군(NORM)의 자료로서 12625개의 유전자와 440개의 pathway 정보로 구성되어 있다. Subramanian 등 (2005)과 Dinu 등 (2007) 등에서 분석 결과를 찾아볼 수 있다.

DM2

제2형 당뇨병(Type 2 diabetes mellitus, DM2)은 인슐린 분비 저하와 인슐린 저항성으로 인해 고혈당이 발생하는 인슐린 비의존성 당뇨병을 말한다. 정상군(NGT:normal glucose tolerance) 17예와 질병군(DM2) 18예의 22283개 유전자, 323개 pathway의 정보로 이루어져 있다. Mootha 등 (2003)은 이 자료를 분석하여 당뇨에 대한 특이발현 형태를 보이는 유전자는 찾지 못하였으나 OXPHOS_HG와 관련된 pathway에 속한 유전자들은 일관된 발현 형태를 갖는다는 것을 보이면서 유전자군 분석의 필요성을 제시하였고 Kim과 Volsky (2005)의 논문에서도 이 자료를 인용하였다.

Leukemia

급성 백혈병 환자중 림프구성 백혈병(acute lymphocytic leukemia: ALL) 환자 24예와 골수구성 백혈병(acute myelocytic leukemia: AML) 환자 24예를 비교하는 자료로 Golub 등 (1999)과 Subramanian 등 (2005) 등에서 분석을 하였다. 12564개 유전자와 크로모좀 위치에 따른 267개의 유전자군의 정보가 있다.

표 1: 세가지 자료 집합을 이용한 개별 유전자분석 결과 비교

자료	유전자수	특이발현 유전자 수			
		FDR* < 0.10		FDR < 0.25	
		t검정	SAM	t검정	SAM
p53	22283	0	0	2	0
DM2	12625	0	0	0	0
Leukemia	12564	8376	4965	11577	7525

*FDR = False discovery rate estimate

위의 세 가지 자료집합을 대상으로 개별 유전자분석을 한 결과 (표 1), p53과 DM2는 특이발현 유전자가 거의 존재하지 않는 반면 Leumekia는 FDR(false discovery rate) 기준 0.25 이하인 특이발현 유전자가 전체의 92%나 차지하는 특별한 자료임을 보였다.

4.2. 새 통계량의 분포 검증

크기 k인 유전자군의 유의성검정을 하는 통계량 MTS = $\sum_{j=1}^{k}(t_j^M)^2$에 대한 카이제곱분포의 가정이 맞는지 검증하기 위하여 각 유전자의 대표값 t^M이 정규분포를 따르는지 먼저 알아보고, 그들의 제곱합이 카이제곱분포를 따르는지 확인하였다.

4.2.1. t^M이 표준정규분포를 만족하는가

위의 세 가지 자료를 이용하여

i) t-통계량을 표준화한 t^S가 표준정규분포를 따르는지 (그림 1(a)),

ii) t^S를 보정한 t^M이 표준정규분포를 잘 따르는지 (그림 1(b)),

iii) 정규점수를 이용한 보정이 각 유전자의 t값에 얼마나 큰 변화를 주는지 (그림 1(c))

알아보았다.

그림 1의 (a), (b)와 Kolmogorov-Smirnov의 정규성 검정결과를 보면 DM2는 정규점수를 사용하여 보정하기 전에도 이미 t^S가 표준정규분포를 따르지만 p53과 Leukemia는 보정의 영향으로 t^M이 표준정규분포를 만족한다는 것을 알 수 있다. 또한 그림 1(c)에서는 특이발현과 무관한 $|t^S| < 0.5$를 제외한 범위에 있는 유전자들의 상대오차를 계산하였는데 극소수의 유전자가 최대 11%의 변화율을 보였고 대부분은 5% 이내였다. 이는 정규점수를 이용한 보정이 정규분포를 벗어난 대표값들의 분포를 정규분포로 변화시키지만 실제 유전자들의 대표값을 크게 변화시키지는 않았다는 것을 보인다.

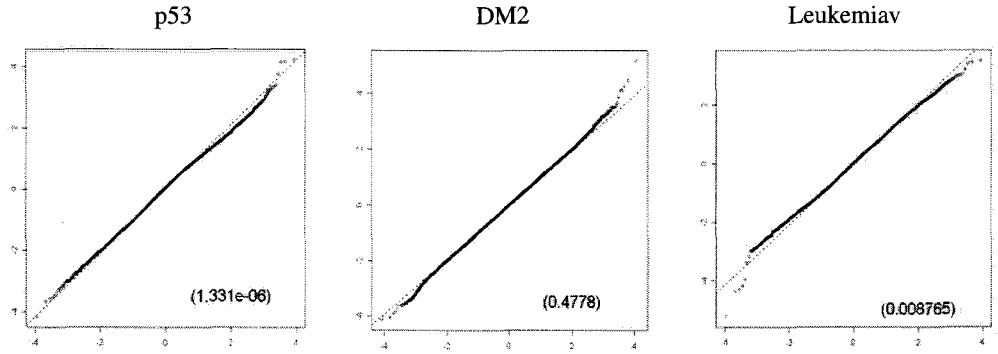
4.2.2. MTS에 대한 카이제곱분포의 가정이 올바른가?

카이제곱분포로부터 계산한 MTS의 p-value와, permutation을 이용하여 구한 MTS의 근간이 되는 $\sum(t^S)^2$의 p-value를 비교하기 위하여 다음과 같은 모의실험을 하였다.

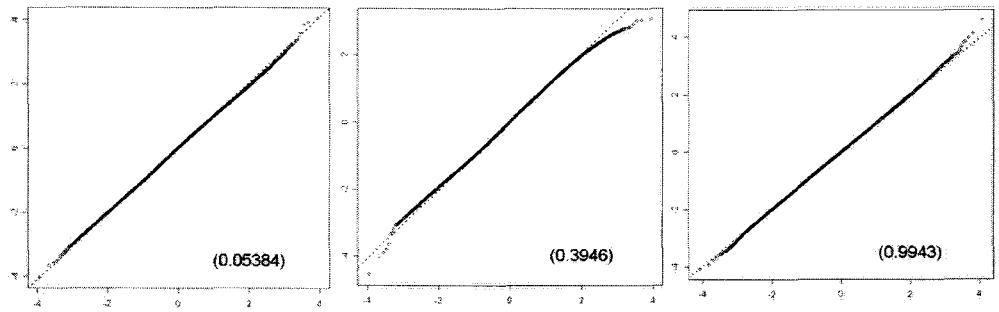
i) 전체 유전자를 대상으로 t^S와 t^M값을 구한다.

ii) 전체 유전자로부터 크기가 k(40, 80, 120)인 유전자군을 임의로 구성하여 MTS 값을 계산하고 자유도 k인 카이제곱분포를 이용하여 p-value를 구한다. 또한 k개 유전자의 t^S의 제곱합을 구한 후 phenotype에 대한 permutation을 1000번 실시하여 p-value를 구한다.

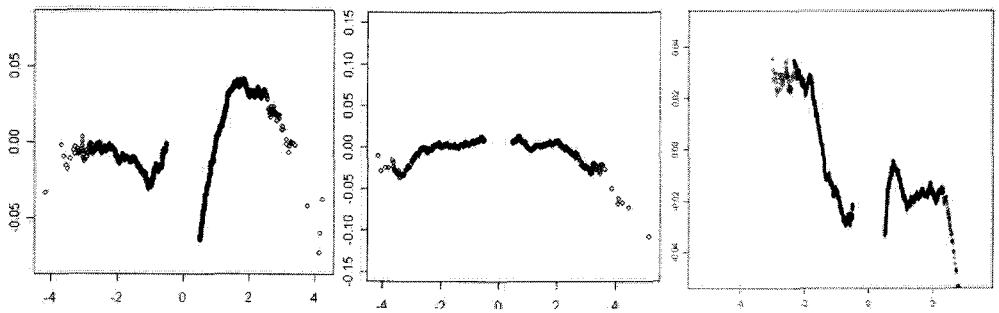
iii) (ii)를 100번 반복하였다.



(a) t^S의 Normal Q-Q plot (Kolmogorov Smirnov 정규성 검정 p-value)



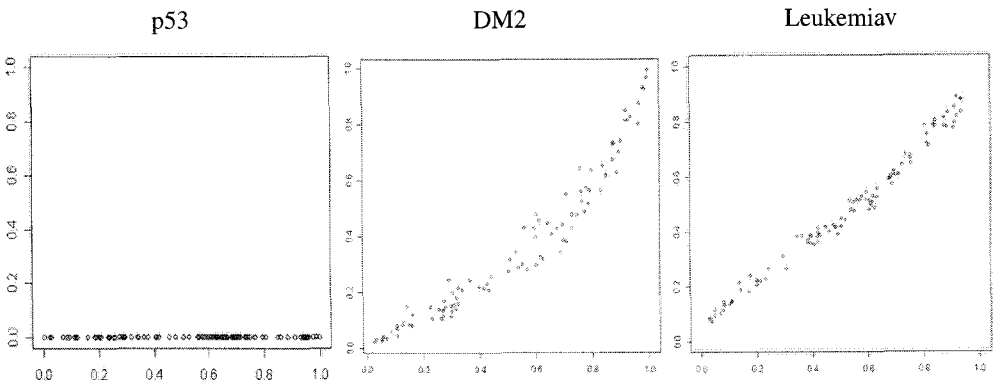
(b) t^M의 Normal Q-Q plot (Kolmogorov Smirnov 정규성 검정 p-value)



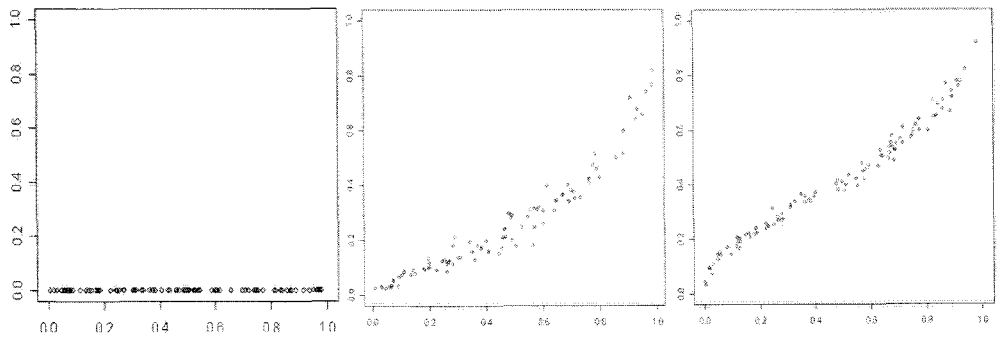
(c) $(t^M - t^S)/t^S$의 산점도 (단, $|t^S| > 0.5$)

그림 1: t^M의 정규성 검정

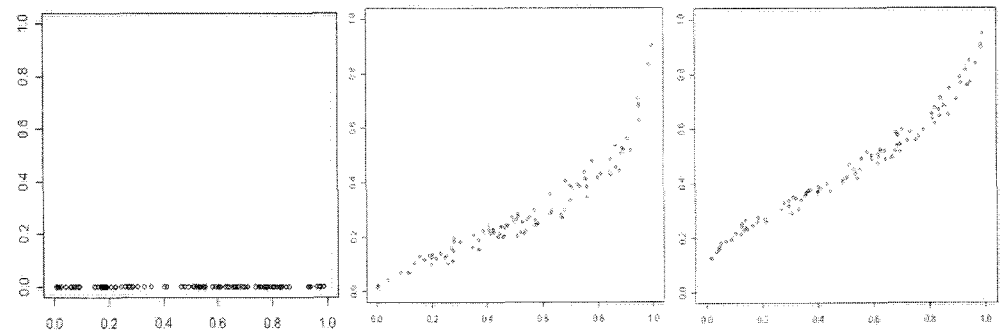
임의로 구성한 크기가 k인 100개 유전자군의 MTS와 $\sum(t^S)^2$의 p-value를 비교(그림 2)한 결과, p53과 DM2의 경우 카이제곱분포로부터 구한 p-value가 permutation을 사용하여 구한 p-value에 비해 조금 작고 k가 커질수록 차이가 커지는 경향이 있는데 이것은 두 통계량을 계산하는데 쓰인 유전자들의 대표값이 서로 다르기 때문에 당연한 결과이다. 그러나 그림 2에서 100개 군의 두 p-value를 비교한 결과 각 군의 순위나 유의성 정도에는 거의 차이를 주지 않음을 알 수 있다. 이러한 결과는 마이크로어레이 실험의 특성인 대용량 자료의 permutation을 위한 장시간의 컴퓨터 작업 대신 순간에 답을 얻을 수 있는 간편한 MTS의 활용가능성을 충분히 보여주고 있다.



(a) k = 40



(b) k = 80



(c) k = 120

그림 2: *크기 k인 유전자군의 MTS (x축)와* $\sum(t^S)^2$ *(y축)의 p-value 비교*

전체 유전자의 92%가 특이발현 현상을 보인 Leukemia 경우는 모든 유전자군의 permutation에 의한 p-value가 0인 특이한 현상을 보인 것이 당연하다. 그러나 MTS은 각 통계량의 대표값으로 정규점수를 이용하여 t-통계량을 보정하여 사용하였기 때문에 특이발현유전자가 많을 경우, 각 유전자들의 특이발현 정도의 순위는 보존되지만 특이발현 여부는 희석되어 다양한 p-value가 나온 것이다. 실질적

표 2: 세가지 자료 집합을 이용한 개별 유전자분석 결과 비교

자료	군의 수*	유의한 pathway 수의 비율(%)							
		FDR < 0.10				FDR < 0.25			
		GSEA	SAMGS	PAGE	MTS	GSEA	SAMGS	PAGE	MTS
p53	249	0.8	8.8	7.2	5.6	4.0	100	47.4	25.3
DM2	323	0	0	2.5	1.9	3.4	0	18.3	5.0
Leukemia	169	0	100	1.2	0	0.6	100	12.4	7.1

*유전자수가 25개~500개인 군으로 제한(단, p53은 250개 이하)

표 3: p53 자료중 31개 pathway의 유의성 비교

Pathway	유전자수	*p*-value			오류
		GSEA	SAM-GS	MTS	
ATM Pathway	40	0.21	<0.001	0.000	18
BAD Pathway	41	0.04	<0.001	<0.001	15
Calcineurin pathway	36	0.13	<0.001	0.188	21
Cell cycle regulator	42	0.29	<0.001	0.002	18
Mitochondria pathway	33	0.32	<0.001	<0.001	16
p53 signalling	153	0.01	<0.001	<0.001	18
Raccycd pathway	49	0.56	<0.001	0.010	18
SA_TRKA receptor	28	0.34	<0.001	0.001	10
bcl2family and reg network	59	0.42	0.001	<0.001	12
Cell cycle arrest	46	0.49	0.001	0.152	18
Ceramide pathway	47	0.30	0.001	<0.001	16
DNA Damage Signalling	139	0.23	0.002	<0.001	12
SIG IL4 receptor in B lyphocytes	48	0.27	0.002	0.098	25
Cell cycle pathway	130	0.72	0.003	0.003	16
G2 pathway	44	0.50	0.003	<0.001	8
Chemical pathway	44	0.04	0.005	<0.001	18
Drug resistance and metabolism	198	0.08	0.005	0.062	14
G1 pathway	63	0.37	0.005	0.337	16
Breast cancer estrogen signalling	180	0.85	0.006	<0.001	19
Ca nf at signalling	175	0.08	0.007	0.012	18
Cytokine pathway	29	0.05	0.007	0.341	34
ST Interleukin 4 pathway	39	0.07	0.007	0.146	26
CR DEATH	125	0.31	0.008	<0.001	25
Porphyrin&chlorophyll metabolism	21	0.29	0.010	0.052	23
Ck1 pathway	26	0.02	0.011	0.020	17
Hivnef pathway	111	0.48	0.011	<0.001	14
Ets pathway	29	0.45	0.012	0.022	18
ST Wnt Ca2 cyclic GMP pathway	29	0.13	0.012	0.007	16
Chrebp pathway	25	0.42	0.013	0.009	14
GPCRs Class A Rhodopsin-like	148	0.04	0.013	0.338	33
ST Fas signaling pathway	98	0.52	0.013	0.003	18

으로 마이크로어레이 자료분석에서 Leukemia 처럼 특이발현 유전자가 많은 경우는 극히 드물지만 특이발현 유전자가 많을수록 Permutaion에 의한 결과와 MTS의 결과 사이에는 큰 차이가 있는 문제점이 발생한다. 하지만 permutation 방법으로는 모든 유전자군의 p-value가 0이지만 MTS가 이들 사이에 변별력을 부여한 장점도 있다.

4.3. 자료 분석 및 결과

통계량 MTS가 유전자군의 유의성 분석에 적당한지 알아보기 위하여 GSEA, SAM-GS, PAGE의 결과와 비교하여 보았다(표 2). 각 방법마다 '유의한 유전자군'에 대한 정의에 조금씩 차이가 있기 때문에 분석 결과가 다른 것은 당연하며, 생물학적 검증없이 통계적 결론만으로 어떤 방법이 좋다고 단정할 수 없지만 MTS의 결과는 다른 방법들에 비하여 무난하다고 보인다.

더 자세한 비교를 위하여 p53 자료를 분석하였을 때 GSEA와 SAM-GS의 결과 사이에 큰 차이(GSEA FDR $\geq$ 0.49, SAM-GS FDR $\leq$ 0.01)가 있는 31개 pathway (Dinu 등 (2007)의 Table 4)를 대상으로 MTS를 이용하여 유의성 검정을 하였다(표 3). MTS는 전체의 75%정도에 대하여 GSEA보다는 검정통계량의 맥락이 같은 SAM-GS와 같은 결론을 보였으나 Cytokine, G1, GPCRs Class A Rhodopsin-like, ST Interleukin 4 와 같은 pathway에 대하여는 SAM-GS와 뚜렷한 차이를 보였다. 31개 pathway가 돌연변이군과 정상군의 차이를 얼마나 잘 반영하는지 알아보기 위하여 p53의 50예 자료를 대상으로 leave one out cross validation(LOOCV)을 실시하여 잘못 분류된 sample의 수를 조사하였더니 평균 18.19개의 오류를 보였다. 그러나 MTS의 결과를 FDR = 0.01 기준으로 20개와 11개의 두 군으로 나누어 비교한 결과 각각 평균 15.95개와 22.27개로 두 군 사이에 분명한 차이가 있었다(p-value < 0.01).

더 자세한 비교를 위하여 p53 자료를 분석하였을 때 GSEA와 SAM-GS의 결과 사이에 큰 차이(GSEA FDR0.49, SAM-GS FDR0.01)가 있는 31개 pathway (Dinu 등 (2007)의 Table 4)를 대상으로 MTS를 이용하여 유의성 검정을 하였다(표3). MTS는 전체의 75%정도에 대하여 GSEA 보다는 검정통계량의 맥락이 같은 SAM-GS와 같은 결론을 보였으나 Cytokine, G1, GPCRs Class A Rhodopsin-like, ST Interleukin 4 와 같은 pathway에 대하여는 SAM-GS와 뚜렷한 차이를 보였다. 31개 pathway가 돌연변이군과 정상군의 차이를 얼마나 잘 반영하는지 알아보기 위하여 p53의 50예 자료를 대상으로 leave one out cross validation(LOOCV)을 실시하여 잘못 분류된 sample의 수를 조사하였더니 평균 18.19개의 오류를 보였다. 그러나 MTS의 결과를 FDR = 0.01 기준으로 20개와 11개의 두 군으로 나누어 비교한 결과 각각 평균 15.95개와 22.27개로 두 군 사이에 분명한 차이가 있었다(p-value < 0.01).

5. 토론

실질적으로 같은 군에 속한 유전자들은 서로 비슷한 기능을 수행하므로 공동조절(coregulation)에 의한 유전자간 상관관계가 존재할 수 있지만, 유전자군의 유의성을 다룰 때 상관관계는 부수적인 문제이기 때문에 많은 연구들이 유전자간 유사독립성을 전제로 진행되었고 (Kim과 Volsky, 2005; Newton 등, 2007; Efron과 Tibshirani, 2007), 본 논문에서 제시한 MTS도 유전자들이 서로 독립이라는 가정 아래에서 유도되었다. 그러나 MTS와 비교한 값은 유전자간의 종속적인 관계를 고려하여 유전자 대신 sample 간의 permutation을 실시하여 얻은 결과이며 두 값이 서로 일관된 양상을 띠우고 있음을 보였다.

MTS 통계량의 근간을 고려할 때, 특이 발현 유전자가 차지하는 비율이 높은 자료분석에서 MTS 사용은 극히 제한적일 수 있다. 그러나 permutation을 이용한 검정에서 대상 유전자군들이 모두 거의 같은 수준으로 유의하다는 변별력없는 결과를 보이는데 비해 MTS는 유전자군 사이의 순위를 정하여 준다는 장점이 있다.

확실한 생물학적 검증 없이 몇 가지 자료 분석과 자료 일부의 예측오류를 갖고 MTS가 다른 방법보다 더 좋다고 단정하기는 어렵지만 통계량의 의미와 방법의 간편성을 볼 때 MTS는 충분한 경쟁력이 있다고 생각한다.

생물학자들은 유전자의 기능을 고려한 대사경로가 같은 유전자들 집단(pathway)의 중요성을 강조하고 있어 'pathway 분석'의 필요성이 절실하지만 pathway 자체가 지닌 구조적 특성 때문에 이를 뒷

받침하기 위한 통계적 기법은 아직 그 수준이 미미하다. Gene Ontology의 biological process의 범주에는 약 4100개의 pathway가 있어 여전히 다중검정이 문제가 되며, 각 pathway 사이에 존재하는 parent/child의 구조 때문에 독립성이 결여되기 때문이다. 또한 단변량적 접근을 거쳐 유의하다고 판명된 pathway들 사이에는 강한 상관관계가 존재하는 경우가 많다. 그러므로 질병에 내재된 메카니즘을 규명하기 위하여 이런 구조적 문제와 중복성을 해결할 수 있는 Pathway 사이의 상관관계 연구가 필수적이라 생각한다.

참고 문헌

이광현, 이선호 (2008). 절대치와 절삭을 이용한 유전자 집단 분석, <응용통계연구>, **21**, 523–535.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: A structured permutation approach, *Bioinformatics*, **21**, 1943–1949.

Blom, G. (1958). *Statistical Estimates and Transformed Beta- Variables*, John Wiley & Sons, New York.

Curtis, R. K., Oresic, M. and Vidal-Puig, A. (2005). Pathways to the analysis of microarray data, *Trends in Biotechnology*, **23**, 429–435.

Damian, D. and Gorfine, M. (2004). Statistical concerns about the GSEA procedure, *Nature genetics*, **36**, 663.

Dinu, I., Potter, J. D., Mueller, T., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2007). Improving GSEA for analysis of biologic pathways for differential gene expression across a binary phenotype, COBRA Preprint Series, Article 16.

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2003). MAPPFinder: Using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data, *Genome Biology*, **4**, R7.

Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003). Global functional profiling of gene expression, *Genomics*, **81**, 98–104.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes, *The Annals of Applied Statistics*, **1**, 107–129

Goeman, J. J., van de Geer, S. A., de Kort, F. and van Houwelingen, H. C. (2004). A global test for groups of genes: Testing association with a clinical outcome, *Bioinformatics*, **20**, 93–99.

Goeman, J. J., Oosting, J., Cleton-Jansen, A. M., Anninga, J. K. and van Houwelingen, H. C. (2005). Testing association of a pathway with survival using gene expression data, *Bioinformatics*, **21**, 1950–1957.

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. and Lander, E. S. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring, *Science*, **286**, 531–537.

Khatri, P., Bhavsar, P., Bawa, G. and Draghici, S. (2004). Onto-Tools: An ensemble of web-accessible, ontology- based tools for the functional design and interpretation of high- throughput gene expression experiments, *Nucleic Acids Research*, **32**, 449–456.

Kim, S. Y. and Volsky, D. J. (2005). PAGE: Parametric analysis of gene set enrichment, *BMC Bioinformatics*, **6**, 1471–2105.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., Houstis, N., Daly, M. J., Patterson, N., Mesirov, J. P., Golub, T. R., Tamayo, P., Spiegelman, B., Lander, E. S., Hirschhorn, J. N., Altshuler, D. and Groop, L. C. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, **34**, 267–273.

Newton, M. A., Quintana, F. A., den Boon, J. A., Sengupta, S. and Ahlquist, P. (2007). Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis, *The Annals of Applied Statistics*, **1**, 85–106.

Pavlidis, P., Lewis, D. P. and Noble, W. S. (2002). Exploring gene expression data with class scores, In *Proceedings of the Pacific Symposium on Biocomputing*, 474–485.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge- based approach for interpreting genome-wide expression profiles, *PNAS*, **102**, 15545–15550.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, *PNAS*, **99**, 6567–6572.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *PNAS*, **98**, 5116–5121.

2008년 12월 접수; 2009년 2월 채택

Developing a Parametric Method for Testing the Significance of Gene Sets in Microarray Data Analysis

Sunho Lee[1,a], Seungkyu Lee[a], Kwanghyun Lee[a]

[a]Department of Applied Statistics, Sejong University

Abstract

The development of microarray technology makes possible to analyse many thousands of genes simultaneously. While it is important to test each gene whether it shows changes in expression associated with a phenotype, human diseases are thought to occur through the interactions of multiple genes within a same functional category. Recent research interests aims to directly test the behavior of sets of functionally related genes, instead of focusing on single genes. Gene set enrichment analysis(GSEA), significance analysis of microarray to gene-set analysis(SAM-GS) and parametric analysis of gene set enrichment(PAGE) have been applied widely as a tool for gene-set analyses. We describe their problems and propose an alternative method using a parametric analysis by adopting normal score transformation of gene expression values. Performance of the newly derived method is compared with previous methods on three real microarray datasets.

Keywords: Microarray experiment, single gene analysis, gene set analysis, normal score.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund) (KRF-2008-531-C00019).

[1] Corresponding author: Professor, Department of Applied Statistics, Sejong University, Gunjadong, Kwangjingu, Seoul 143-747, Korea. E-mail: leesh@sejong.ac.kr