Hotelling의 T^2 통계량을 이용한 cDNA 마이크로어레이 분석

김병수[1], 이선호[2], 김인영[3], 김상철[4], 라선영[5], 정현철[6]

요약

본 논의에서는 cDNA 마이크로어레이 분석에서 다변량 분석의 한 방법인 Hotelling의 T제곱 통계량을 이용하여 유의적 유전자군을 검색하고, 이 유전자군을 사용하여 검사 자료를 두군으로 분류하는데 단변량 t통계량에 기초한 접근보다 얼마나 효율적인지를 평가하고자 한다.

주요 용어: 다중검정, 분류, cDNA 마이크로어레이, 유의적 유전자, Hotelling의 T제곱 통계량

1. 실험자료 및 실험의 목적

대장암 환자 58예 각각을 대상으로 대장암과 근처의 정상조직을 채취하여 급속 냉각하였다. 각 환자로부터 채취한 암조직과 정상조직으로부터 total RNA를 추출하고 간접설계법을 사용하여 두장의 마이크로어레이에다 실험을 하였다. 즉, 11개 암세포주를 섞어서 공통 준거(Ref)를 구성하였으며 Ref에는 cy3-dUTP를 표지하고, 첫 번째 마이크로어레이는 암조직에다 cy5-dUTP를 표지하여 경쟁적으로 보합시켰으며, 두 번째 마이크로어레이는 정상조직에다 cy5-dUTP를 표지하여 보합시켰다. 동 실험에 사용한 마이크로어레이는 대장암관련 유전자를 포함하여 17075개 유전자가 점적되었다.

모두 58예의 환자들을 대상으로 실험을 실시하였으나 환자에 따라서 채취한 조직의 양이 충분치 못하여 2장의 마이크로어레이 실험중 (Ref, 정상조직), 혹은 (Ref, 암조직)의 실험을 마칠수 없는 경우가 있었고, 결과적으로 다음 표1과 같은 세가지 유형의 자료가 얻어졌다. 본 실험의 목적은 상기 58예의 마이크로어레이 실험자료에 기초하여 대장암과 정상조직을 구별하여 주는 소수의 유전자를 검색하고, 동 유전자군을 이용하여 진단의 기준을 마련하고자 함이다. 본 분석에서는 짝을 이른 자료를 훈련자료 (training data)로 간주하고, 동 훈련자료로부터 유의적 유전자를 찾고, 두 개의 독립적 자료인 16예의 정상조직과 22예의 암조직 자료를 검사

1) 연세대학교 상경대학 응용통계학과 교수, 서울 서대문구 신촌동 134번지 우:120-749. 본 연구는 보건복지부 보건의료기술진흥사업의 지원에 의하여 이루어 진 것임 (02-PJ1-PG3-10411-0003).
2) 세종대학교 자연대학 응용수학과, 부교수, 서울시 광진구 군자동 98번지, 우: 143-747
3) 연세대학교 의과대학, 암전이연구센터, SRC박사후 연구원. 서울 서대문구 신촌동 134 우:120-749.
4) 연세대학교 의과대학, 암전이연구센터, BK21 Project for Medical Science, 석사과정
5) 연세대학교 의과대학, 암전이연구센터, BK21 Project for Medical Science, 조교수. 본 연구는 보건복지부 IMT 2000 출연금 기술개발사업 지원에 의하여 이루어진 것임(01-PJ11-PG9-01BT00A-0028).
6) 연세대학교, 의과대학, 내과학교실, BK21 Project for Medical Science, 부교수, 본 연구는 과학기술부 우수연구센터 지원에 의하여 이루어진 것임

자료 (test data)로 간주하여 분류를 하고자 한다.

표1: 짝을 이룬 자료, 두 개의 독립적인 자료와 표기법

(Ref, 정상조직)	(Ref, 암조직)	例數
X	Y	20
U	결측	16
결측	V	22

Cy5와 cy3의 발현 강도를 각각 R, G로 나타낼 때 유전자의 발현 강도는 log(R/G)로서 나타내기로 한다. 모두 40개의 마이크로어레이로 구성된 훈련자료와 38개의 마이크로어레이로 구성된 검사자료 각각에 대하여 표준화를 실시하였고, 비결측율이 80%(검사자료의 경우는 70%)이상되는 유전자 13466개를 분석에 사용하였다. 결측치는 KNN (K=10)방법으로 대치하였고, 중복 점적된 유전자들에 대하여서는 평균을 취하였다. 이렇게 하여 최종적으로 12312 x 78 행렬의 자료가 얻어졌다.

2. 단변량 t통계량에 기초한 분석

X_{gh}를 h 번째 例의 정상조직 마이크로어레이에서 g번째 유전자의 발현강도를 나타낸다고 하고, Y_{gh}는 암조직에서 g번째 유전자의 발현강도를 나타낸다고 하자. $D_{gh}=Y_{gh}-X_{gh}$ 로 정의한다, g=1,...,12312, h=1,...,20. 단변량 t통계량에 기초한 분석은 다음 식(1)의 귀무가설 對 대립가설을 검정하는 문제를 구성한다.

$$H_0^{(g)}: E[D_{gh}]=0 \quad 對 \quad E[D_{sh}]\neq 0, \ g=1,...,12312 \qquad (1)$$

식(1)의 검정통계량인 짝을 이룬 t 통계량을 사용하고, 재표본 방법에 기초한 Westfall and Young (1993)의 step-down 절차를 이용하여 다중검정에 따른 P값을 보정할 수 있다. (Dudoit et al., 2002). 몇가지 족별유의수준(FWER)에 따라 검색된 유전자 수는 다음 표2와 같다.

표2: 족별유의수준에 따른 유의적 유전자의 개수

P값 보정 방법	족별유의수준		
	0.01	0.005	0.001
Westfall and Young의 step-down 절차 (보정한 p값)	722	559	346
짝을 이룬 t 검정(보정안한 p값)	3748	3230	2199

Westfall and Young의 방법으로 계산된 (보정된) P값을 작은 순서대로 나열하고, 그중 처음 n개에 기초하여 검사자료를 대상으로 대각항 선형 판별 분석(DLDA; Dudoit, Fridlyand and Speed, 2002)을 적용한 결과는 다음 표3과 같다.

표3: 단변량 t 통계량에 기초한 38例의 판별분석 결과

n	2	4	6-12	14	≥16
예측율	0.947	0.947	0.974	1.0	1.0

3. Hotelling의 T^2통계량에 기초한 분석

우선 두 개의 유전자로 구성된 확률벡타를 구성하고 Hotelling의 T^2통계량이 가장 큰 100쌍의 유전자를 찾고, 이를 기초로 대각항 선형 판별분석을 실시한 결과는 다음 표4와 같다.

표4: Hotelling의 T^2통계량에 기초한 판별분석 결과

n	2	4	6-10	12	14	≥16
예측율	0.947	0.947	0.974	1.0	0.974	1.0

4. 토의

Hotelling의 T^2통계량의 값이 가장 큰 처음 10쌍의 유전자에서 개별 t값과 보정된 P값을 살펴보면 단변량 t 검정 절차를 통하여서는 검색되지 않는 유전자를 8개 포함하고 있음을 알 수 있다. 이처럼 두군의 유전자가 서로 상이함에도 불구하고 비슷한 정도의 예측율을 보이고 있는 것을 주목하며, 이는 Hotelling의 T^2통계량이 분석도구로서 유용하게 활용될 수 있음을 시사하고 있다.

참고문헌

Dudoit S, Yang YH, Callow MW, Speed TP. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12:111-139.

Dudoit S, Fridyland J, Speed TP. (2002). Comparison of discrimination methods for the classification of tumors using gene expressin data. *J. Amer. Stat. Assoc.* 97:77-87.

Westfall PH, Young. S. (1993). *Resampling-based multiple testing*, New York:Wiley.