응용통계연구 (2009)
22(4), 781–792

유전자 연관성이 랜덤검정 P값과 유의 유전자군의 탐색에 미치는 영향

이미성[1] · 송혜향[2]

[1]가톨릭대학교 의과대학 의학통계학과, [2]가톨릭대학교 의과대학 의학통계학과

(2009년 5월 접수, 2009년 7월 채택)

요 약

유전체 초기단계 연구에서는 비교적 소수의 마이크로어레이 샘플자료로서 실험을 진행하여 심도 깊게 연구해야 할 유전자 부분군(subsets)을 탐색하게 된다. 이러한 과정에서 요구되는 부분군 탐색에 사용되는 분석방법은 다수 샘플자료 분석의 경우와는 매우 다른 방법들이다. 유전자 극소수 샘플자료의 분석에 매우 적절한 방법인 랜덤검정법을 적용하여 정확한 P값(exact P value)의 이산형 분포가 얻어지고, 일양분포 귀무가설의 검정으로 유의 유전자가 존재하는지를 파악할 수 있다. 한 단계 더 나아가 Fuchs와 Kenett (1980)이 제시한 M검정을 이용하여 이산형 P값 다항분포에서 이상범주군(outlier cells)을 찾을 수 있으며 이로써 유의 유전자로서의 가능성이 있는 유전자군을 선정한다. 대다수의 마이크로어레이 유전체 연구에서 수 천 또는 수 만개의 유전자가 서로 독립이라고 가정하고 분석하는 것이 문제점이다. 그러나 본 논문에서는 유전자 연관성을 그대로 유지하는 순열에 기초한 랜덤검정법과 M검정법으로서 유전자 연관성이 분석에 미치는 영향을 모의실험으로 알아보았으며, 그 영향이 결코 미약하지 않음을 확인할 수 있었다.

주요용어: 랜덤검정법, 정확한 P값, 유의 유전자, 이상범주군.

1. 서론

최근에 와서 마이크로어레이 실험비용이 크게 감소하였으나 이러한 저가 현상은 극소수 샘플(array or chip)자료, 다시 말하면 대조와 처리의 각 군에 셋, 넷 또는 다섯 개의 샘플로서 단시일에 가설을 설립하고자 하는 목적으로 실험되는 경우를 더욱 증가시켰다고 Hu와 Wright (2007)는 언급하였고, 또한 Fierro 등 (2008)의 마이크로어레이 자료의 메타분석 논문에서는 적게는 두, 세 개의 샘플로 현재도 실험이 진행되고 있음을 보고하고 있다. 극소수 마이크로어레이 샘플자료의 경우에, 두 처리를 비교하는 모수적 t 검정법은 부적절하다. 다시 말하면, 극소수 마이크로어레이 샘플자료의 특성상, 다수의 마이크로어레이 샘플자료의 경우와는 다른 분석방침이 채택되어야 한다. 본 논문에서는 이러한 극소수 샘플의 경우에 두 처리군을 비교하는 검정법으로서 Fisher (1935)가 제안한 랜덤검정법(randomization test)을 고려하며, 이 검정법은 샘플자료의 모든 가능한 배열을 고려하여 P값을 구하는 정확한(exact) 검정법으로서 순열검정법(permutation test)의 특별한 경우에 해당한다 (Welch, 1990). 계산상의 어려움으로 사용이 제한되었던 랜덤검정법은 컴퓨터의 보편화로 최근에 와서는 그 한계를 찾아볼 수 없이 여러 분야에서 활발히 사용되고 있다 (Lambert, 1985). 한편, Murie와 Nadon (2008)은 극소수 마이크로

[2]교신저자: (137-701) 서울시 서초구 반포동 505, 가톨릭대학교 의과대학 의학통계학과, 교수.
E-mail: hhsong@catholic.ac.kr

어레이 샘플자료의 분석법으로서 수정된 t 검정법을 제안하고 있어 본 논문의 비모수적인 랜덤검정법과는 다르다. 극소수 샘플의 분석 문제로서 Hu와 Wright (2007)는 올리고 뉴클레오타이드어레이 샘플의 경우를 다루고 있고, Dondrup 등 (2009)은 한 샘플에 동일 유전자의 최대 400회 반복까지를 실험에서 고려하여 극소수 샘플의 문제를 해결하는 계획과 이에 따른 분석법을 다루고 있다.

마이크로어레이 자료에서 각 유전자의 검정결과 P값은 대조군과 비교한 처리군의 발현이 유의한 정도를 표현한 것으로서 이에 근거하여 처리에 의해 발현이 유의한(differentially expressed: DE) 유전자와 유의하지 않은 유전자, 즉 대등한 발현(equally expressed: EE)을 보이는 유전자를 탐색하는데 도움이 된다. 다수, 즉 수백 개의 마이크로어레이 샘플자료의 경우에는 P값이 연속적인 값으로 표현되어 유의 유전자를 밝히는데 직접 사용되지만, 본 논문에서 고려하는 극소수 샘플자료의 경우에는 P값이 이산형 값을 가지며 또한 이 P값이 예를 들어서 1/10이하의 값을 가질 수 없는 경우도 있으므로 유의 유전자로서의 가능성이 있는 범주, 따라서 장차 더욱 연구되어야 하는 유전자군을 알려주는데 그치게 된다. 유전체 초기단계 연구에서는 비교적 소수의 샘플자료로서 DE유전자로서의 가능성이 있는 유전자 부분군(subsets)을 밝히는 이러한 작업이 매우 중요한 과정임을 Parmigiani 등 (2002)은 설명하고 있다.

최근 Gadbury 등 (2003)은 랜덤검정법을 유전자 극소수 샘플자료에 적용하였고 또한 유전자 자료의 분석결과를 제시하면서 더욱 연구되어야 하는 문제점이 남아있음을 시사한 바 있다. 본 논문에서는 저자들의 논문에 대해 두 가지를 더욱 발전시킨다. 첫째는 여러 유전자가 실제 상황에서와 같이 서로 연관되었을 때 랜덤검정의 결과인 P값에 어떤 영향을 미칠 것인지를 알아보고자 한다. Gadbury 등 (2003)은 영향을 미치는 현상을 지적하였으나 이에 대한 원인은 설명하지 않았다. 둘째로 몇 천개 또는 몇 만개 유전자 자료에서 각 유전자의 랜덤검정 P값이 동일 수치를 가지는 여러 범주로 분류되었을 때 유의 유전자로서의 가능성이 있는 유전자군을 대략적인 추측에 그치지 않고 구체적인 검정으로서 밝히는 것이며, 이러한 결정에 여러 유전자간의 연관성에 의한 영향을 알아볼 것이다.

2. 방법

마이크로어레이 유전체 연구에서 대조와 처리의 각 군에 n개의 실험단위, 즉 샘플로 실험이 진행된 균형자료의 경우를 고려하는데 그 이유는 곧 설명하게 되는 랜덤검정의 P값이 정확한 값을 가지기 때문이다. 두 군의 실험단위수가 다른 경우에 대해서는 곧 언급하게 된다. 랜덤검정법은 통계량의 선택이 우선되며 극소수 샘플의 경우에는 단순히 대조군과 처리군의 평균차(mean difference)가 적절하다. 두 처리간에 차이가 없다는 귀무가설 하에서 이 평균차 통계량의 분포는 총 유전자로 구성된 샘플자료의 모든 가능한 순열 후마다 계산된 평균차의 분포를 이용하여 구한다. 이와 같은 순열방식은 유전자간의 연관성을 그대로 유지하면서도 샘플자료가 대조 또는 처리군으로 나뉘는 것과 유전자 발현수치가 서로 무관하도록 진행하는 것이다.

2.1. 랜덤검정법에 의한 정확한 P값의 계산

총 유전자의 개수가 G일 때 Y_{ij}^k $(i = 1, \ldots, G,\ j = 1, \ldots, N,\ k = t, c)$는 i번째 유전자에서 k실험군의 j번째 실험단위의 반응변수라 하자 (Gadbury 등, 2003). 각 군의 실험단위수가 n개이고, 총 실험단위수 N은 $2n$이다. Gadbury 등 (2003)은 i번째 유전자에 대한 반응변수 (Y_{ij}^t, Y_{ij}^c)를 이변량, 즉 $N \times 2$ 행렬로 표현하여 j번째 실험단위에서의 Y_{ij}^t, Y_{ij}^c 중 하나는 반응변수를 갖고 나머지 하나는 결측값을 가진다고 표현한다. k실험군의 평균 반응변수를 $\bar{Y}_{i\cdot}^k = 1/N \sum_{j=1}^{N} Y_{ij}^k$라 두었을 때, i번째 유전자의 처리효과는 다음과 같은 평균차, 즉 $\bar{D}_i = \bar{Y}_{i\cdot}^t - \bar{Y}_{i\cdot}^c$이 된다.

가법모형 $Y_{ij}^t = Y_{ij}^c + \tau_i$ 하에서 랜덤검정법의 가설은 각 유전자의 경우에

$$H_0^{(i)} : \tau_i = 0 \quad vs. \quad H_1^{(i)} : \tau_i \neq 0 \tag{2.1}$$

이며, τ_i는 각 유전자마다 다르게 정의되는 상수이다. 랜덤검정법은 랜덤순열을 이용하여 N개의 실험단위를 대조군과 처리군에 각각 n개씩 랜덤할당하여 검정하며, 여기서 랜덤할당의 경우수는 구체적으로 $B = \binom{N}{n}$이다. 이를 하나의 확률변수를 이용하여 표현하면 다음과 같다. j번째 실험단위가 처리군에 속할 때 $Z_j = 1$이고, 대조군에 속할 때 $Z_j = 0$이라 하면, $Z = (Z_1, \ldots, Z_N)$는 길이가 N이고 n개의 1과 n개의 0개로 이루어져 있으며, 서로 다른 Z벡터의 경우수는 B개이고 각각의 확률은 동일하여 $P(Z = z) = 1/B$이 된다.

Z를 이용한 i번째 유전자의 평균 처리효과 추정량 $\bar{d}_i$는 다음과 같이 표현된다.

$$\bar{d}_i = \frac{1}{n}\sum_{j=1}^{2n} Y_{ij}^t Z_j - \frac{1}{n}\sum_{j=1}^{2n} Y_{ij}^c (1 - Z_j) \tag{2.2}$$

그러므로 $\bar{d}_i$에 근거한 귀무가설에 대한 양측검정 P값은 각 Z의 확률이 $1/B$을 가지는 이산분포함을 이용하여 구한다. 실제 i번째 유전자로부터 구한 평균 처리효과가 d_i^*일 때 이 확률은 다음과 같다.

$$p_i = 2\min\left\{\Pr(\bar{d}_i \geq d_i^*),\ \Pr(\bar{d}_i \leq d_i^*)\right\}, \tag{2.3}$$

즉, B개의 결과 중에서 $|d_i^*|$보다 크거나 같은 $|\bar{d}_i|$의 개수를 세어 B로 나누면 i번째 유전자의 처리효과에 대한 정확한 P값(exact P value)이 된다. 이산형 결과로 표현되는 정확한 P값은 양측검정의 경우 $B/2$ 경우수에 한정되므로 총 G개 유전자로부터 구한 정확한 P값인 $p_1^*, \ldots, p_G^*$은 다시 다음의 경우수로서 요약된다.

$$O_v = \sum_{i=1}^{G} I_{\{\frac{2v}{B}\}}(p_i^*), \quad v = 1, \ldots, \frac{B}{2}, \tag{2.4}$$

여기서 $I_{\{b\}}(a)$는 $a = b$일 때에 1의 값을 가진다. 즉, O_v는 총 G개 유전자 중 p_i^*가 $2v/B$와 동일한 값을 가지는 유전자수를 나타낸다.

이제 관측도수 O_v를 이용하여 $(O_1, \ldots, O_{B/2-1})$가 일양분포하는가에 대한 귀무가설을 적합도 카이제곱검정으로 알아볼 수 있다. 유의한 발현수치를 나타내는 유전자가 존재치 않는다면 일양분포 양상을 보이게 된다. 귀무가설 하에서 O_v의 기대값은 $2G/B$이므로 $(O_1, \ldots, O_{B/2-1})$는 모수가 G와 $(\pi_1, \pi_2, \ldots, \pi_{B/2-1})$인 다항분포이며 여기서 $\pi_1 = \pi_2 = \cdots = \pi_{B/2-1}$이다. 따라서 귀무가설에 대한 검정은 다음 통계량

$$\chi^2 = \sum_{v=1}^{\frac{B}{2}} \frac{(O_v - 2G/B)^2}{2G/B} \tag{2.5}$$

이 자유도가 $B/2 - 1$인 카이제곱분포를 따름을 이용하여 검정한다.

Gadbury 등 (2003)이 언급하였듯이 적합도 카이제곱검정법은 총 유전자를 대상으로 실시한 총체적인 검정(global test)으로서 귀무가설은 $H_0 : \pi_1 = \pi_2 = \cdots = \pi_{B/2-1}$으로 일양분포 가정이며, 이 카이제곱검정이 기각될 때 하나 또는 그 이상 범주의 확률이 $2/B$에서 벗어남을 뜻한다. 한편 랜덤검정법의 귀무가설은 가법모형 $Y_{ij}^t = Y_{ij}^c + \tau_i$ 하에서 각 유전자에 대해 $H_0^{(i)} : \tau_i = 0$으로서 이 가설이 기각될 때 i번째 유전자는 DE유전자로 결론내린다. 따라서 카이제곱검정법과 랜덤검정법의 귀무가설에는 큰 차이를 보이고 있다. 그러나 유전자 극소수 샘플자료의 경우에는 서론에서도 언급하였듯이 각 유전자에 대한 랜덤검정으로 일반적으로 채택하는 유의수준인 0.05이하의 확률, 또는 한 단계 더 나아가 여러 유전자를 함께 검정하므로 다중비교를 적용한 0.05보다 훨씬 작은 확률을 얻을 수가 없다. 예를 들어서 각

실험군에 샘플자료가 2개씩 할당된다면 $B = 6$으로서 각 랜덤순열의 최소확률은 $2/B = 1/3$이 되며, 각 실험군에 샘플자료가 3개씩 할당된다면 $B = 20$으로서 각 랜덤순열의 최소확률은 $2/B = 1/10$이 되어 유의수준 0.05이하의 확률을 나타내는 유전자를 밝힐 수 없는 상황이다. 따라서 총체적인 검정으로서 DE유전자로서의 가능성을 지닌 유전자군을 찾게 되고 이 군에 속한 유전자는 분자생물학적 실험, 또는 다른 방법으로서 추후 연구하게 된다.

대조군과 처리군에 동일 실험단위가 배정된 균형자료의 경우에는 랜덤검정 P값이 정확한 값을 가지는 반면에 두 군에 서로 다른 실험단위가 배정되는 경우에는 랜덤검정에서의 평균차가 치우친 이산형 분포양상을 나타내므로 랜덤검정 P값을 더욱 커지는 방향으로, 또는 더욱 작아지는 방향으로 결정할 수가 있다 (Gibbons와 Pratt, 1975). 따라서 랜덤검정법은 균형자료의 경우에만 추천되는 방법이다.

2.2. 이상범주군 탐색

랜덤검정법에서 각 실험군에 예를 들어서 4개(또는 5개)의 실험단위가 배정된다면 랜덤순열의 최소확률은 $2/B = 1/35(1/126)$가 되며 이러한 경우에는 35개(126개)의 범주에 나타난 랜덤검정 P값의 양상으로부터 유의 유전자군을 찾는 작업이 요구되며 다음의 M검정법이 유용하다. Fuchs와 Kenett (1980)에 제시된 M검정법은 다항분포의 확률이 균일한가의 검정에서 한 단계 더 나아가 최대 수정잔차(Maximum adjusted residual)를 이용하여 이상범주군(outlier cell)을 찾는다. 다항분포를 따르는 확률변수를 $\mathbf{n}$이라 할 때 $\mathbf{n} = \{n_v : 1 \le v \le B/2\}$은 모수가 G와 π인 다항분포를 따른다 ($G = \sum n_v$, $\pi_v \ge 0$, $\sum \pi_v = 1$).

귀무가설은 $H_0 : \pi = \pi^{(0)}$, 대립가설은 $H_1 : \pi > \pi^{(0)}$이며 $\pi^{(0)}$는 임의의 정해진 빈도벡터이다. 귀무가설 하에서 n_v는 평균이 $G\pi_v^{(0)}$, 분산이 $G\pi_v^{(0)}(1-\pi_v^{(0)})$인 정규분포를 근사적으로 따른다.

수정잔차 Z_v는 다음과 같다.

$$Z_v = \frac{n_v - G\pi_v^{(0)}}{\sqrt{G\pi_v^{(0)}\left(1 - G\pi_v^{(0)}\right)}}, \quad v = 1, \ldots, \frac{B}{2}. \tag{2.6}$$

유의수준이 α인 단측검정에서 다음을 만족하면 H_0를 기각한다.

$$\max Z_v > M^*, \quad \text{단, } \Pr\{\max Z_v > M^* | H_0\} = \alpha. \tag{2.7}$$

M^*의 값으로는 Bohrer 등 (1981)에 의해 제시된 Bonferroni 수정을 이용한 Z_M 또는 Šidák (1968)이 제시한 Z_S 중에서 선택할 수 있다.

$$\begin{aligned} Z_M &= \Phi^{-1}\left\{1 - \frac{\alpha}{k}\right\}, \\ Z_S &= \Phi^{-1}\left\{\frac{1 + (1-\alpha)^{\frac{1}{k}}}{2}\right\}, \quad \text{단, } k = \frac{B}{2}, \end{aligned} \tag{2.8}$$

여기서 $Z_v > M^*$인 v번째 범주를 이상범주군으로 결정하게 된다.

3. 모의실험

3.1. 모의실험 계획

Gadbury 등 (2003)을 참고하여 유전자 연관성을 나타내는 다변량 정규분포를 따르는 G개의 유전자를 생성한다. Gadbury 등 (2003)은 한 가지 상관행렬 구조(correlation matrix structure)와 자료로부터

추출된 한 가지 기댓값 벡터에 대해서만 검토하였으나 본 논문에서는 여러 상관행렬 구조와 여러 기댓값 벡터를 생성하여 이들 각각의 영향을 알아본다. 처리군과 대조군의 각 군에서 n개의 실험단위로 실험이 진행되어 각 실험단위에서 관측된 발현수치는 구체적으로 G차원의 다변량 정규분포를 따르며, 이 때 상관행렬은 구형성 가정을 만족하여 대각원소가 1이고 비대각원소가 ρ인 $G \times G$ 행렬 또는 소수의 유전자간 연관성을 갖는 유전자블록을 블록대각행렬로 둔 행렬이다. 유전자블록 크기가 50인 경우를 예로 들면 50개의 유전자가 서로 상관관계를 갖고 다른 유전자블록과는 독립이며, 본 논문의 모의실험 계획에서 500개의 유전자를 대상으로 연구하기에 서로 독립인 10개의 유전자블록으로 구성된다.

두 처리간에 차이가 없다는 귀무가설 하에서는 총 500개의 유전자 모두가 EE유전자이고, 대립가설 하에서는 총 500개의 유전자 중 250개의 유전자는 DE유전자이고, 나머지 250개의 유전자는 EE유전자로 정한다. 여기서 EE유전자의 발현수치 분포는 대조군과 처리군에서 각각 $Y_{ij}^c \sim N(\mu_c, \sigma_c^2)$, $Y_{ij}^t \sim N(\mu_c, \sigma_c^2)$으로 생성하며, DE유전자의 경우에 대조군의 발현수치 분포는 $Y_{ij}^c \sim N(\mu_c, \sigma_c^2)$로, 처리군의 발현수치 분포는 $Y_{ij}^t \sim N(\mu_t, \sigma_t^2)$로 생성한다. 여기서 $\mu_c = 0$, $\sigma_c^2 = 1$로 정하였고 이러한 대조군 분포의 결정은 모의실험 결과에 영향을 미치지 않으며, 단지 처리군의 μ_t와 ρ, 상관행렬의 유전자블록 크기에 의해 영향을 받으며 이를 변화시켜 랜덤검정법의 결과를 검토한다. 구체적으로 μ_t는 1, 2, 3, 4로, 다변량 정규분포에서의 유전자간 상관계수 ρ는 0, 0.3, 0.6으로 그리고 상관행렬의 유전자블록 크기는 10, 50, 250, 500으로 변화시켜 검토한다.

랜덤검정 P값의 결과의 검토는 각 군에 3개의 실험단위로 배정된 경우를 고려한다. 이 경우에 랜덤검정 P값은 $1/10, 2/10, \ldots, 10/10$ 중 하나의 값을 가진다.

이상범주군 탐색의 M검정에서는 각 군에 5개의 실험단위가 배정된 경우를 고려하고, 이때 랜덤검정 P값은 $1/126, 2/126, \ldots, 126/126$ 중 하나의 값을 가진다. M검정에서는 기각치인 M^*의 값이 Bohrer 등 (1981)이 제시한 Z_M인 경우와 Šidák (1968)이 제시한 Z_S인 경우가 거의 일치하여 Z_M을 기각치로 둔 M검정의 결과만을 제시한다. 유의수준 α가 0.05일 때 M검정의 기각치 $Z_M = 3.355$이다.

모의실험의 각 경우에서 귀무가설 하에서는 2,500번, 대립가설 하에서는 500번의 반복수로 시행한다.

3.2. 모의실험 결과

표 3.1은 2,500번의 반복수로 구해진 유전자 연관성, 유전자블록 크기 변화에 따른 각 P값의 평균과 표준편차(괄호안)의 결과이다. 두 처리간에 차이가 없다는 귀무가설 하에서는 모든 유전자가 EE유전자이므로 각 P값에서 기대되는 평균은 0.1이며 모의실험 결과에서도 모두 0.1에 근사한 값을 나타내 보인다. 또한 유전자 연관성과 유전자블록 크기의 변화가 각 P값에 큰 영향을 미치지 않는다. 반면, P값 0.1의 경우를 예로 들어서 유전자블록 크기가 500개일 때의 표준편차를 살펴보면, ρ가 0, 0.3, 0.6로 변함에 따라 표준편차가 각각 0.012, 0.044, 0.093으로 유전자 연관성이 클수록 표준편차가 커지는 것을 알 수 있다. 또한 ρ가 0.6이면서 유전자 블록 크기가 10개, 50개, 250개, 500개로 변할 때 표준편차가 각각 0.020, 0.040, 0.058, 0.093으로 유전자 블록이 클수록, 특히 500개 유전자의 상관행렬이 구형성을 만족하는 경우에 표준편차가 특히 커지는 것을 알 수 있다. 그리고 매우 작은 P값의 경우와 매우 큰 P값의 경우에는 표준편차의 증가폭이 크지만, P값 0.4, P값 0.5, P값 0.6인 경우에는 증가폭이 작음을 볼 수 있다.

이제 2,500번의 반복에서 적합도 카이제곱검정 결과 유의한 경우의 확률을 살펴보면 ρ가 0일 때에는 전체 반복 중 4%만이 검정결과 유의하였다. 그러나 귀무가설하인데도 불구하고 유전자 연관성과 유전자블록 크기가 커질수록 유의한 경우가 5%보다도 크게 증가하였다. 이러한 결과는 다수의 유전자가 실제로 연관되어 있을 때, 독립인 경우에 적절한 방법인 적합도 카이제곱검정을 적용한다면 위양성율(false

표 **3.1.** 귀무가설 하에서 유전자 연관성, 유전자블록 크기에 따른 각 P값의 평균과 표준편차

P값	ρ	0	0.3			0.6			
	블록크기		50개	250개	500개	10개	50개	250개	500개
0.1		0.101	0.100	0.097	0.097	0.102	0.097	0.085	0.091
		(0.012)	(0.023)	(0.031)	(0.044)	(0.020)	(0.040)	(0.058)	(0.093)
0.2		0.098	0.100	0.099	0.098	0.099	0.098	0.095	0.099
		(0.012)	(0.016)	(0.021)	(0.027)	(0.014)	(0.023)	(0.038)	(0.050)
0.3		0.101	0.099	0.099	0.099	0.101	0.101	0.102	0.103
		(0.014)	(0.014)	(0.017)	(0.019)	(0.013)	(0.021)	(0.033)	(0.037)
0.4		0.101	0.101	0.102	0.100	0.098	0.102	0.102	0.103
		(0.013)	(0.012)	(0.013)	(0.014)	(0.015)	(0.014)	(0.022)	(0.025)
0.5		0.101	0.101	0.101	0.099	0.100	0.101	0.103	0.103
		(0.012)	(0.013)	(0.013)	(0.014)	(0.012)	(0.015)	(0.018)	(0.023)
0.6		0.103	0.100	0.101	0.102	0.102	0.101	0.103	0.101
		(0.013)	(0.014)	(0.014)	(0.015)	(0.013)	(0.018)	(0.019)	(0.024)
0.7		0.098	0.101	0.101	0.103	0.097	0.100	0.103	0.102
		(0.014)	(0.015)	(0.016)	(0.020)	(0.015)	(0.017)	(0.025)	(0.031)
0.8		0.098	0.099	0.100	0.100	0.099	0.100	0.103	0.100
		(0.016)	(0.015)	(0.017)	(0.022)	(0.015)	(0.020)	(0.028)	(0.036)
0.9		0.100	0.099	0.100	0.100	0.100	0.099	0.101	0.099
		(0.015)	(0.015)	(0.018)	(0.021)	(0.016)	(0.022)	(0.034)	(0.043)
1.0		0.099	0.101	0.101	0.103	0.102	0.101	0.103	0.099
		(0.013)	(0.015)	(0.020)	(0.024)	(0.015)	(0.020)	(0.035)	(0.046)
χ^2검정의 검정력		0.040	0.150	0.328	0.489	0.167	0.499	0.778	0.840

표 **3.2.** 대립가설 하에서 처리군 평균 변화에 따른 각 P값의 평균과 표준편차

P값	DE$\sim N(1,1)$		DE$\sim N(2,1)$		DE$\sim N(3,1)$		DE$\sim N(4,1)$	
0.1	0.182	(0.016)	0.365	(0.017)	0.496	(0.013)	0.542	(0.009)
0.2	0.131	(0.014)	0.134	(0.015)	0.087	(0.014)	0.055	(0.009)
0.3	0.114	(0.014)	0.094	(0.013)	0.061	(0.012)	0.051	(0.010)
0.4	0.099	(0.013)	0.072	(0.011)	0.054	(0.010)	0.051	(0.009)
0.5	0.089	(0.012)	0.063	(0.010)	0.053	(0.009)	0.052	(0.009)
0.6	0.084	(0.013)	0.060	(0.011)	0.052	(0.010)	0.051	(0.010)
0.7	0.078	(0.012)	0.054	(0.010)	0.049	(0.009)	0.049	(0.009)
0.8	0.075	(0.012)	0.052	(0.010)	0.049	(0.011)	0.047	(0.010)
0.9	0.073	(0.012)	0.052	(0.010)	0.049	(0.010)	0.050	(0.010)
1.0	0.075	(0.011)	0.054	(0.009)	0.051	(0.008)	0.050	(0.009)
χ^2검정의 검정력	1		1		1		1	

positive rate)이 매우 크다는 것을 의미한다.

표 3.2에 제시된 결과는 DE유전자의 경우로서 유전자 연관성은 0인 경우 처리군의 평균 μ_t가 1, 2, 3, 4로 변화함에 따른 500번의 반복수로서 구해진 랜덤검정 P값의 평균과 표준편차(괄호안)이다. 유전자 연관성이 0이 아닌 경우의 결과는 표 3.3에 제시한다.

표 3.2의 결과를 보면 대립가설 하에서 μ_t가 커짐에 따라 각 P값이 기대값인 0.1보다 큰 값을 보인다. 구체적으로 μ_t가 1, 2, 3, 4로 증가할수록 P값 0.1의 평균은 각각 0.182, 0.365, 0.496, 0.542로 크게 증가한다. 즉, 이들 P값 0.1의 평균이 모두 0.1보다 커진다는 것은 DE유전자로서의 가능성을 지닌 유전

표 **3.3.** 대립가설하(DE$\sim N(2,1)$)에서 유전자 연관성, 유전자블록 크기에 따른 각 P값의 평균과 표준편차

P값	ρ / 블록크기	0	0.3			0.6			
			50개	250개	500개	10개	50개	250개	500개
0.1		0.365	0.362	0.367	0.364	0.363	0.358	0.359	0.374
		(0.017)	(0.040)	(0.083)	(0.085)	(0.033)	(0.063)	(0.140)	(0.141)
0.2		0.134	0.137	0.135	0.134	0.139	0.138	0.129	0.128
		(0.015)	(0.018)	(0.029)	(0.026)	(0.019)	(0.029)	(0.058)	(0.050)
0.3		0.094	0.093	0.093	0.093	0.094	0.094	0.090	0.090
		(0.013)	(0.015)	(0.026)	(0.023)	(0.016)	(0.023)	(0.039)	(0.038)
0.4		0.072	0.070	0.072	0.070	0.070	0.070	0.070	0.070
		(0.011)	(0.012)	(0.016)	(0.018)	(0.012)	(0.016)	(0.028)	(0.028)
0.5		0.063	0.063	0.061	0.060	0.063	0.063	0.062	0.062
		(0.010)	(0.012)	(0.014)	(0.013)	(0.011)	(0.014)	(0.023)	(0.021)
0.6		0.060	0.059	0.057	0.058	0.058	0.058	0.062	0.060
		(0.011)	(0.012)	(0.013)	(0.013)	(0.010)	(0.014)	(0.023)	(0.018)
0.7		0.054	0.056	0.055	0.057	0.053	0.054	0.057	0.056
		(0.010)	(0.010)	(0.013)	(0.014)	(0.011)	(0.012)	(0.023)	(0.019)
0.8		0.052	0.055	0.055	0.055	0.054	0.055	0.058	0.054
		(0.010)	(0.012)	(0.014)	(0.014)	(0.010)	(0.014)	(0.024)	(0.021)
0.9		0.052	0.052	0.053	0.054	0.054	0.055	0.055	0.053
		(0.010)	(0.012)	(0.016)	(0.013)	(0.012)	(0.016)	(0.026)	(0.021)
1.0		0.054	0.052	0.051	0.054	0.052	0.054	0.057	0.052
		(0.009)	(0.010)	(0.015)	(0.015)	(0.011)	(0.016)	(0.028)	(0.024)
χ^2검정의 검정력		1	1	0.998	1	1	1	0.982	0.988

자군의 비율이 0.1보다 증가하는 것을 의미한다. 표 3.2의 가장 오른쪽 DE유전자의 처리군의 발현수치 분포가 $N(4,1)$인 경우의 결과를 보면 다음과 같다. 랜덤검정의 결과로서 P값의 분포는 (0.54, 0.06, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05)로 나타났으며 이는 절반인 250개의 유전자는 EE유전자로서 발현수치로 인한 P값은 균일분포, 즉 각 P값이 모두 0.05에 가깝고, 나머지 250개(50%)의 유전자는 DE유전자로서 모두가 가장 작은 P값 0.10에 속한 결과이며, 이와 같이 P값이 0.10인 경우의 비율이 0.54로 다른 비율보다 현저히 높아져서 DE유전자로서의 가능성을 지닌 유전자군을 쉽게 가릴 수 있는 것이다. 그리고 500번의 반복에서 적합도 카이제곱검정 결과 유의한 경우의 확률을 표에 제시하였는데 μ_t가 1, 2, 3, 4일 때 모든 반복에서 적합도 카이제곱검정 결과가 유의하다. 즉, 모든 경우에 일양분포 귀무가설을 기각하여 하나 또는 그 이상범주의 확률이 2/B, 즉 0.1에서 벗어남을 알 수 있다.

다음으로, 표 3.3은 유전자 연관성이 0, 0.3, 0.6으로 변화하고 유전자블록 크기가 10, 50, 250, 500으로 변화할 때의 각 P값의 평균과 표준편차(괄호안) 결과이다. 이 때 DE유전자의 처리군의 발현수치 분포는 $Y_{ij}^t \sim N(2,1)$이다. 대립가설 하에서 각 P값의 평균은 유전자 연관성이나 유전자블록 크기에 큰 영향을 받지 않는다. 그러나 유전자 연관성과 유전자블록 크기가 클수록 각 P값의 표준편차는 커진다. 특히 P값 0.1, P값 0.2, P값 0.3의 경우에는 표준편차의 증가폭이 비교적 크지만, P값 0.4와 그 이상 P값의 경우에는 표준편차가 미미한 증가를 보인다. 적합도 카이제곱검정 결과가 대부분의 경우에 일양분포 귀무가설을 기각하여 유의하다. 특히 유전자블록 크기가 500일 때는 샘플의 모든 유전자가 연관된 경우로서 이때에는 P값 표준편차가 독립된 유전자 부분군이 있는 경우보다도 훨씬 커짐을 알 수 있다.

이제 이상범주군 탐색의 M검정의 모의실험 결과인 표 3.4~3.6에 대해 살펴본다. 각 군의 실험단위가 5인 경우의 랜덤검정 P값은 $1/126, 2/126, \ldots, 126/126$ 중 하나의 값을 가지며 이러한 126개 P값 범

표 **3.4.** 귀무가설 하에서 유전자 연관성, 유전자 블록 크기 변화에 따른 2,500번 반복 중 M검정으로 유의한 확률

범주	ρ	0	0.3				0.6			
	블록크기		10개	50개	250개	500개	10개	50개	250개	500개
1–7		0.0044	0.0053	0.0070	0.0614	0.0791	0.0077	0.0436	0.1129	0.1363
8–14		0.0030	0.0026	0.0014	0.0216	0.0357	0.0030	0.0043	0.0766	0.1147
15–21		0.0014	0.0040	0.0040	0.0120	0.0166	0.0036	0.0084	0.0217	0.0804
22–28		0.0010	0.0019	0.0077	0.0063	0.0047	0.0020	0.0041	0.0121	0.0326
29–35		0.0059	0.0024	0.0043	0.0046	0.0070	0.0014	0.0034	0.0080	0.0154
36–42		0.0053	0.0000	0.0014	0.0014	0.0027	0.0034	0.0013	0.0024	0.0059
43–49		0.0000	0.0014	0.0027	0.0077	0.0020	0.0027	0.0066	0.0023	0.0039
50–56		0.0014	0.0029	0.0031	0.0017	0.0034	0.0000	0.0011	0.0056	0.0033
57–63		0.0013	0.0011	0.0013	0.0031	0.0017	0.0023	0.0034	0.0047	0.0027
64–70		0.0013	0.0016	0.0011	0.0033	0.0029	0.0016	0.0000	0.0000	0.0073
71–77		0.0034	0.0060	0.0034	0.0044	0.0014	0.0027	0.0091	0.0033	0.0100
78–84		0.0014	0.0014	0.0000	0.0037	0.0106	0.0040	0.0011	0.0123	0.0117
85–91		0.0029	0.0033	0.0000	0.0000	0.0063	0.0041	0.0000	0.0091	0.0149
92–98		0.0014	0.0031	0.0000	0.0013	0.0041	0.0076	0.0067	0.0073	0.0164
99–105		0.0043	0.0000	0.0000	0.0013	0.0031	0.0031	0.0049	0.0036	0.0081
106–112		0.0046	0.0036	0.0021	0.0039	0.0041	0.0040	0.0034	0.0160	0.0177
113–119		0.0000	0.0011	0.0073	0.0020	0.0016	0.0030	0.0000	0.0081	0.0169
120–126		0.0057	0.0026	0.0026	0.0036	0.0054	0.0017	0.0016	0.0157	0.0210

주군의 결과로부터 이상범주군을 탐색하게 된다. 표 3.4의 결과는 균일 다항분포 귀무가설 하에서 반복수 2,500번 중 M검정결과 유의한 셀의 확률을 7개씩의 범주의 평균으로 제시하였다. 귀무가설 하에서는 Bonferroni 수정을 이용한 M검정결과 유의한 셀의 총 확률은 반복 2,500번 중 5%에 가까운 값이 나올 것이라 예상된다. 실제로 모든 유전자가 독립인 ρ가 0인 경우에는 유전자 연관성과 유전자블록의 영향을 받지 않고 7개 범주의 평균합이 4.87%로서 5%에 가까운 값이다. 그러나 유전자 연관성과 유전자 블록 크기의 영향을 살펴보면, 1–7 범주에서 ρ가 0.6이면서 유전자 블록크기가 10개, 50개, 250개, 500개로 변화하면서 M검정결과 유의한 셀의 확률이 0.0077, 0.0436, 0.1129, 0.1363으로 커지는 것을 알 수 있다. 또한 유전자 블록 크기가 500일 때는 샘플의 모든 유전자가 연관된 경우이며 결과를 살펴보면, ρ가 0, 0.3, 0.6으로 변함에 따라 M검정결과 유의한 셀의 확률이 0.0044, 0.0791, 0.1363으로 커지는 것을 알 수 있다. 이는 표 3.1에서 유전자 연관성과 유전자 블록 크기가 클수록 각 P값의 표준편차가 크게 나타난 것과 연결지어 이해할 수 있다.

표 3.5에 제시된 결과는 유전자 연관성이 0일 때 DE유전자의 경우로서 처리군의 평균인 μ_t가 1, 2, 3, 4로 변화함에 따른 500번의 반복수로서 구한 범주군에서 M검정 결과 유의한 셀의 확률이다. 유전자 연관성이 0이 아닌 경우는 표 3.6에 제시한다.

표 3.5에서의 이상범주군 1셀은 랜덤검정 P값이 $1/126 = 0.008$인 범주만이 M검정으로 유의한 경우의 확률이며 이상범주군 1~2셀은 P값이 $1/126 = 0.008$과 $2/126 = 0.016$일 때가 동시에 M검정으로 유의하고 나머지 124개의 셀은 유의하지 않은 경우의 확률이다. 다른 셀에 대해서도 마찬가지이며 이 모든 경우의 셀이 상호배반적이다. 이와 같은 방법으로 126개의 셀 가운데 유의한 유형의 셀의 서로 다른 가지수는 2^{126} 즉, 8.51×10^{37}가지가 되며, 본 논문의 관심은 작은 P값의 범주들이 동시에 유의하다고 결론내려지는 결과에 한정된다하겠다.

표 **3.5.** 대립가설 하에서 처리군 평균 변화에 따른 500번 반복 중 M검정으로 유의한 확률

이상 범주군	DE$\sim N(1,1)$	DE$\sim N(2,1)$	DE$\sim N(3,1)$	DE$\sim N(4,1)$
1셀	0	0	0	0.620
1~2셀	0.008	0	0.118	0.376
1~3셀	0.026	0	0.736	0.002
1~4셀	0.060	0.060	0.132	0
1~5셀	0.034	0.220	0.002	0
1~6셀	0.022	0.186	0	0
1~7셀	0.010	0.062	0	0

표 **3.6.** 대립가설하(DE$\sim N(2,1)$)에서 유전자 연관성, 유전자 블록 크기 변화에 따른 500번 반복 중 M검정으로 유의한 확률

이상 범주군	ρ	0	0.3				0.6			
	블록크기		10개	50개	250개	500개	10개	50개	250개	500개
1셀		0	0	0	0	0	0	0	0.008	0
1~2셀		0	0	0	0.002	0	0	0.002	0.026	0.024
1~3셀		0	0	0.014	0.040	0.026	0.002	0.038	0.112	0.116
1~4셀		0.060	0.072	0.090	0.098	0.126	0.066	0.120	0.120	0.104
1~5셀		0.220	0.220	0.206	0.168	0.162	0.178	0.170	0.142	0.118
1~6셀		0.186	0.212	0.174	0.110	0.122	0.200	0.102	0.078	0.084
1~7셀		0.062	0.072	0.088	0.080	0.082	0.068	0.092	0.054	0.068
1~8셀		0.006	0.016	0.016	0.044	0.030	0.012	0.014	0.018	0.022
1~9셀		0	0.002	0.008	0.012	0.016	0.002	0.020	0.028	0.024
1~10셀		0	0	0	0.002	0.006	0	0	0.004	0.010
1~11셀		0	0	0	0.012	0	0	0	0.004	0.002
1~12셀		0	0	0	0.002	0.004	0	0	0.006	0.008
1~13셀		0	0	0	0.002	0	0	0	0.004	0
합		0.534	0.594	0.596	0.572	0.574	0.528	0.558	0.604	0.580

표 3.5에서 μ_t가 4인 경우를 살펴보면, 1셀, 1~2셀, 1~3셀의 총 3개의 범주유형으로 전체 500번 반복에서 99.8%가 M검정의 결과 이상범주군으로 결론 내려진다. μ_t가 3인 경우에는 1~2셀, 1~3셀, 1~4셀, 1~5셀의 4개의 범주유형으로 98.8%가 이상범주군으로 결론 내려진다. μ_t가 2인 경우 P값이 $7/126 = 0.056$로 유의수준 0.05에 가까운 7개 셀까지를 고려해 보면 52.8%가 이상범주군으로 결론 내려진다. 즉, μ_t가 작을수록 소수의 셀이 이상범주군으로 판단될 확률이 훨씬 낮아지며, 따라서 μ_t가 클수록 이상범주군을 탐색하기가 매우 용이한 것이다. 구체적으로 μ_t가 4인 경우에는 3개의 범주유형만을 더욱 연구해야하는 유전자군으로 선정하였다.

다음으로, 표 3.6은 유전자 연관성이 0, 0.3, 0.6으로 변화하고 유전자블록 크기가 10, 50, 250, 500으로 변화할 때의 M검정결과 유의한 셀의 확률이다. 이 때 DE유전자의 경우에 처리군의 발현수치 분포는 $Y_{ij}^t \sim N(2,1)$이다.

유전자 블록 크기가 500이면서 유전자 연관성이 0, 0.3, 0.6으로 변화할 때 1~5셀과 1~6셀의 이상범주군 확률합을 살펴보면, 40.6%, 28.2%, 20.2%로 작아지는 것을 알 수 있다. 또한, 유전자 연관성이 0.6이면서 유전자 블록크기가 10개, 50개, 250개, 500개로 변화할 때 1~5셀과 1~6셀의 이상범주군 확률합을 살펴보면, 37.8%, 27.2%, 22.0%, 20.2%로 작아지는 것을 알 수 있다. 이는 유전자 연관성이 크고, 유전자 블록이 클수록 소수의 이상범주군을 장차 연구해야하는 유전자군으로서 선정하기 어렵다는 것을 의미한다.

4. 적용 사례

Gadbury 등 (2003)에 제시된 마이크로어레이 자료는 관절염 세포주로부터 얻어진 것으로서 총 유전자수는 12,625개이고 각 실험군의 실험단위는 3개이다. 여기서 가능한 랜덤할당의 경우수 B는 $\binom{6}{3} = 20$이고, 양측랜덤검정으로 10개의 정확한 P값의 범주에 속한 관측도수는 다음과 같다.

$$(O_1, O_2, \ldots, O_{10}) = (2975,\ 1457,\ 1261,\ 1137,\ 1058,\ 1002,\ 934,\ 931,\ 957,\ 913). \qquad (4.1)$$

두 처리간에 차이가 없다는 귀무가설 하에서 각 범주의 기대값은 1262.5이며, 이를 이용한 적합도 카이제곱검정 결과는 χ^2가 2795.41로 매우 큰 값이다. 자유도가 9이고 유의수준 0.05인 기각치는 $\chi^2(0.95, 9) = 16.919$이므로 적합도 카이제곱검정 결과가 매우 유의하여, 하나 또는 그 이상 범주의 확률이 0.1에서 벗어나는 것을 알 수 있다.

또한 M검정으로 어떤 범주가 확률 0.1에서 벗어나는지 검정해 본 결과, 첫째 범주와 둘째 범주의 Z값이 각각 50.80과 5.77로서 유의수준이 0.05일 때 Z_M인 2.58보다 커서 1~2셀이 이상범주군이다. 이로서 DE유전자의 가능한 군은 1~2셀임을 알 수 있다. 따라서 Gadbury 등 (2003)과는 달리 P값이 0.1과 0.2인 유전자군 모두가 이상범주군으로서 두 군에 속한 유전자 모두를 추후 더욱 연구해야 한다고 결론내린다.

5. 결론 및 고찰

본 연구에서는 유전자 극소수 샘플자료, 즉 대조와 처리의 각 군에 세 개 또는 다섯 개의 샘플로 실험되고 실제 상황에서와 같이 여러 유전자가 연관된 경우에 구체적으로 랜덤검정의 결과인 정확한 P값에 미치는 영향을 모의실험으로 알아본 결과, 유전자 연관성과 유전자블록의 크기의 변화는 각 P값의 평균에는 큰 영향을 미치지 않으나 유전자 연관성과 유전자블록의 크기가 커질수록 각 P값의 표준편차가 커져 신뢰성이 크게 떨어짐을 볼 수 있다. 이는 P값이 작은 값일 때에 더욱 뚜렷이 나타난다. 대립가설 하에서, 즉 발현 유전자가 존재하는 경우에 처리군의 평균이 비발현 유전자의 평균값과 차이가 클수록 특히 P값 0.1의 평균이 커지며, 이는 유의 유전자로서의 가능성이 있는 유전자군의 빈도가 증가하는 것을 의미한다.

각 유전자의 랜덤검정 P값이 동일 수치를 가지는 여러 범주로 분류되었을 때 유의 유전자로서의 가능성이 있는 유전자 이상범주군을 구체적으로 M검정으로서 탐색할 수 있다. 본 연구에서는 여러 유전자가 연관되었을 때 M검정에 미치는 영향을 알아본 결과, 유전자 연관성이 크고 유전자 블록이 클수록, 즉 더욱 많은 유전자수가 서로 연관될수록 이상범주군이 낮은 빈도의 여러 범주로 흩어져 있어 소수의 이상범주군을 선정하기 어렵다는 것을 알 수 있다. 그러나 자연스럽게 발현 유전자의 처리군의 발현수치 평균이 비발현 유전자군의 평균과 차이가 클수록 소수의 이상범주군을 선정할 수 있는 확률이 높아져 탐색이 수월해진다.

장차, 유전자 연관성에 대한 연구를 다수 샘플자료의 경우로도 확대하여 유의 유전자 탐색에 사용되는 여러 통계적 검정결과에 미치는 영향을 알아보는 과제가 남아 있다. 이러한 여러 유전자 연관성의 통계적 검정결과에 미치는 영향에 대한 연구결과는 구체적으로 연관성을 감안한 통계적 분석법의 개발을 촉진하는 계기가 될 것이다.

참고문헌

Bohrer, R., Chow, W., Faith, R., Joshi, V. and Wu, C. F. (1981). Multiple three-decision rules for factorial simple effects: Bonferroni wins again!, *Journal of the American Statistical Association*, **76**, 119–124.

Dondrup, M., Hüser, A. T., Mertens, D. and Goesmann, A. (2009). An evaluation framework for statistical tests on microarray data, *Journal of Biotechnology*, **140**, 18–26.

Fierro, A. C., Vandenbussche, F., Engelen, K., Van de Peer, Y. and Marchal, K. (2008). Meta analysis of gene expression data within and across species, *Current Genomics*, **9**, 525–534.

Fisher, R. A. (1935). *The Design of Experiments*, Oliver and Boyd, Edinburgh.

Fuchs, C. and Kenett, R. (1980). A test for detecting outlying cells in the multinomial distribution and two-way contingency tables, *Journal of the American Statistical Association*, **75**, 395–398.

Gadbury, G. L., Page, G. P., Heo, M., Mountz, J. D. and Allison, D. B. (2003). Randomization tests for small samples: An application for genetic expression data, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **52**, 365–376.

Gibbons, J. D. and Pratt, J. W. (1975). P-values: Interpretation and methodology, *The American Statistician*, **29**, 20–25.

Hu, J. and Wright, F. A. (2007). Assessing differential gene expression with small sample sizes in oligonucleotide arrays using a mean-variance model, *Biometrics*, **63**, 41–49.

Lambert, D. (1985). Robust two-sample permutation tests, *The Annals of Statistics*, **13**, 606–625.

Murie, C. and Nadon, R. (2008). A correction for estimating error when using the Local Pooled Error Statistical Test, *Bioinformatics*, **24**, 1735–1736.

Parmigiani, G., Garrett, E. S., Anbazhagan, R. and Gabrielson, E. (2002). A statistical framework for expression-based molecular classification in cancer, *Journal of The Royal Statistical Society. Series B*, **64**, 717–736.

Šidák, Z. (1968). On multivariate normal probabilities on rectangles: Their dependence on correlations, *The Annals of Mathematical statistics*, **39**, 1425–1434.

Welch, W. J. (1990). Construction of permutation tests, *Journal of the American Statistical Association*, **85**, 693–698.

Effect of Genetic Correlations on the P Values from Randomization Test and Detection of Significant Gene Groups

Mi-Sung Yi[1] · Hae-Hiang Song[2]

[1]Department of Biostatistics, Medical College, The Catholic University of Korea;
[2]Department of Biostatistics, Medical College, The Catholic University of Korea

(Received May 2009; accepted July 2009)

Abstract

At an early stage of genomic investigations, a small sample of microarrays is used in gene expression experiments to identify small subsets of candidate genes for a further accurate investigation. Unlike the statistical analysis methods for a large sample of microarrays, an appropriate statistical method for identifying small subsets is a randomization test that provides exact P values. These exact P values from a randomization test for a small sample of microarrays are discrete. The possible existence of differentially expressed genes in the sample of a full set of genes can be tested for the null hypothesis of a uniform distribution. Subsets of smaller P values are of prime interest for a further accurate investigation and identifying these outlier cells from a multinomial distribution of P values is possible by M test of Fuchs *et al.* (1980). Above all, the genome-wide gene expressions in microarrays are correlated, but the majority of statistical analysis methods in the microarray analysis are based on an independence assumption of genes and ignore the possibly correlated expression levels. We investigated with simulation studies the effect that correlated gene expression levels could have on the randomization test results and M test results, and found that the effects are often not ignorable.

Keywords: Randomization test, exact P value, significant gene groups, outlier cells.

[2]Corresponding author: Professor, Department of Biostatistics, Medical College, The Catholic University of Korea, Seoul 137-701, Korea. E-mail: hhsong@catholic.ac.kr