응용통계연구 (2009)
22(3), 617-626

cDNA 마이크로어레이에서 유전자간 상관 관계에 대한 보고

김병수[1] · 장지선[2] · 김상철[3] · 임요한[4]

[1]연세대학교 응용통계학과, [2]한국 경제 연구원, [3]연세대학교 응용통계학과, [4]서울대학교 통계학과

(2009년 2월 접수, 2009년 4월 채택)

요 약

최근에 보고되는 일련의 연구는 Affymetrix 마이크로어레이 자료에서 유전자간 상관관계가 강하고 長範圍(long-ranged)로 나타나고 있으며, 기존의 "편한" 가정, 즉 유전자간 상관관계가 매우 약하며, 따라서 유전자간 유사 독립성을 가정할 수 있다는 주장이 비현실적이라는 것을 보고하고 있다. Qui 등 (2005b)은 각 유전자의 검정통계량을 병합하여 통계적 추론을 하는 이른바 비모수적 경험적 베이즈 방법을 적용하면 검색된 특이발현 유전자수의 분산이 커진다는 것을 보고하고 있고, 이러한 분산의 불안전성 이유로서 유전자간 강한 상관관계를 지적하고 있다. 또한 Klebanov와 Yakovlev (2007)는 유전자간 상관관계가 통계적 분석을 어렵게 하는 요인이라기 보다는 유용한 정보의 원천이고 적정한 변환을 통하여 근사 독립을 유지할 수 있는 급수를 만들 수 있으며 이 급수를 δ-급수라고 불렀다. 본 보고에서는 국내에서 생산된 2조의 cDNA 마이크로어레이 자료에서 유전자간 상관관계가 비교적 강하며, 長範圍로 나타나는 것을 확인하며, 유사 독립성을 전제할 수 있는 δ-급수가 cDNA 마이크로어레이에서도 발견되는 것을 보고하고자 한다. 동 보고는 추후 cDNA 마이크로어레이 자료의 분석에서도 유전자간 상관관계를 고려하여야 함을 강조하고 있다.

주요용어: cDNA마이크로어레이, 비모수적 경험적 베이즈 방법, 상관관계, 유사 독립성, 특이 발현.

1. 서론

지난 10여년간 꾸준히 발전하여온 마이크로어레이 기술은 특이발현 유전자를 검색하는 과제를 통계학계에 안겨다 주었다. 동 과제는 이른바 "작은 n, 큰 p"의 문제를 구성하면서 통계학의 새로운 패러다임을 구성하고 있다. "작은 n, 큰 p"의 문제는 소위 "차원의 저주"로 인한 난제를 구성하므로 이를 우회하는 한 가지 방법은 많은 수의 유전자가 있는 것을 이용하여 동 유전자 정보를 공유하는 방법이다. 유전자 정보를 공유하는 방법의 통계적 이론은 Efron 등 (2001)과 Efron (2003, 2004)이 비모수적 경험적 베이즈 방법에 기초하여 소수의 어레이를 이용한 반복된 실험에서 특이발현 유전자를 검색하는 이론적 틀을 제시함으로써 공식화되었다.

위의 비모수적 경험적 베이즈 방법은 각 유전자마다 이표본 t통계량을 계산하고, p개의 이표본 t통계량이 서로 독립이거나, 혹은 대수의 법칙이 성립할 정도로 약한 의존도를 나타낸다는 것을 대전제로 하고 있다. 그리고, 이러한 전제는 그동안 실증적 검증이 없이 학계에서 널리 수용되어 왔다. 그러나, 통계학계에서는 검색된 특이발현 유전자 개수의 불안정성이 꾸준하게 보고되었으며 (Qui 등, 2006; Qui와

이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(KRF-2008-312-C00055).

[1]교신저자: (120-749) 서울특별시 서대문구 성산로 262, 연세대학교 응용통계학과, 교수.
E-mail: bskim@yonsei.ac.kr

Yakovlev, 2006; Stolovitzky, 2003), 생물학계에서는 선별 실험인 마이크로어레이 실험이 결국 추후 실험을 위한 올바른 단서를 제공하지 못하는 현상을 보고하면서 마이크로어레이 실험 자체에 회의를 제기하기도 하였다 (Frantz, 2005; Marshall, 2004). 이러한 맥락에서 Qui 등 (2005b)은 마이크로어레이 자료의 통계적 방법이 현실 자료와 어긋나는 가정에 기초하고 있음을 우려하면서, "현 시점에서 마이크로어레이 자료의 복잡한 구조를 무리하게 단순화시켜 새로운 방법이나 이론을 제시하는 것보다, 현재 통용되는 통계적 추론 절차를 일단 뒤돌아 보고 동 절차의 한계를 고찰하는 것이 더욱 중요한 일이다"라고 피력하고 있다.

최근 들어 Klebanov와 Yakovlev (2007)는 일련의 보고를 통하여 유전자간 독립성, 혹은 약의존성이 실제 Affymetrix 마이크로어레이 실험 자료에서 성립하지 않는 것을 보고하고 있고, 다음의 결론을 내렸다.

가) 유전자간 상관관계는 매우 강하다. 모든 가능한 두 개 유전자들을 구성하여 상관계수를 계산한 결과 상관계수는 매우 크며, 長範圍(long-ranged)하다. 장범위한 상관관계란 한 유전자가 짝을 구성하여 높은 상관관계를 나타내는 상대 유전자의 개수가 많음을 의미한다.

나) 표준화 절차는 불필요하다. 흔히 통용되는 이른바 표준화 절차는 유전자간 상관관계를 왜곡시키며, 표준화를 한다고 하여도 유전자간 상관 관계가 완전히 제거되지는 않는다 (Qui 등, 2005a). 또한 Affymetrix 마이크로어레이에서 배경 강도를 빼주는 것은 유전자간 상관 관계를 더욱 크게 할 뿐이므로 불필요하다 (Klebanov와 Yakovlev, 2007).

다) 대수의 법칙은 더 이상 적용되지 않는다. 유전자간 상관 관계가 약하거나 독립적일 경우 어레이 수보다 유전자 수를 늘림으로써 대수의 법칙을 적용하는 것이 가능하지만, 실제 마이크로어레이 자료는 유전자간 상관 관계가 강하여서 대수의 법칙이 성립되지 않는다 (Klebanov와 Yakovlev, 2006).

라) 유전자간 상관 관계는 유용한 정보를 제공한다. 유전자간 상관은 특이발현 유전자 검색에 장애요인이라기 보다는 유용한 정보의 원천이며, 적정한 변환을 통하여 근사 독립을 유지하는 확률변수의 급수를 도출할 수있으며, 유전자간 의존적 구조를 밝힘으로써 특정 의존적 구조를 나타내는 표현형을 새롭게 검색할 수 있다 (Klebanov 등, 2006).

본고에서는 국내에서 생산된 두조의 cDNA 마이크로어레이 실험 자료에 기초하여 위에서 언급한 가)~라)항이 cDNA 마이크로어레이 실험 자료에서도 그대로 성립하는 것을 확인하며, 추후 cDNA 마이크로어레이 분석에서 유전자간 상관을 고려하는 것이 중요하다는 것을 지적하고자 한다. 2절에서는 두조의 cDNA 마이크로어레이 자료를 설명하고, Kim 등 (2005)에서 제안한 t_3통계량을 소개하며, Klebanov와 Yakovlev (2007)에서 사용한 통계적 방법을 소개한다. 3절에서는 동 통계적 분석을 두조의 자료에 적용한 결과를 보고하고 있으며, 4절에서는 결론과 토의 내용을 다루고 있다.

2. cDNA 마이크로어레이 자료와 분석방법

2.1. 두 조의 cDNA 마이크로어레이 자료

본 연구에서 사용한 첫 번째 cDNA 마이크로어레이 자료는 Kim 등 (2005)에서 사용한 대장암 87예에 기초한 cDNA 마이크로어레이 실험 자료이고 두 번째 cDNA 마이크로어레이 자료는 109예의 위암을 기초로 한 cDNA 마이크로어레이 실험자료이다. 위암 cDNA 마이크로어레이 실험 자료의 일부는 Yang 등 (2007a)과 Yang 등 (2007b)에서 보고된 바 있다. 두 마이크로어레이 실험은 모두 17,000여 개의 인간 유전자가 점적된 같은 cDNA 마이크로어레이를 사용하였고, 암 세포주를 공통 준거로 사

표 **2.1.** cDNA 마이크로어레이에서 부분적으로 짝을 이룬 자료 구조

보합			
(공통준거대 정상조직)	(공통준거대 암조직)	대장암 例數	위암 例數
X	Y	36	85
U	결측	19	13
결측	V	32	11
		87	109

용한 간접설계로 이루어 졌다. 각 환자의 수술시 종양과 종양 근처의 정상 조직을 짝으로 채취하였다. 그러나 경우에 따라 종양이나 정상조직으로부터 cDNA 마이크로어레이 실험을 수행하기에 충분한 mRNA를 추출할 수 없는 경우가 발생하여 일부 예에서는 결측치가 발생하였다. 결과적으로 표 2.1에서 보는 바와 같이 부분적으로 짝을 이룬 자료 구조를 구성하게 되었다. 표 2.1에서 X, Y는 각각 공통준거대 정상조직, 공통 준거대 종양의 보합시 발현 강도를 나타내고, U와 V는 각각 X와 Y의 독립적 복제 확률변수이다.

두 마이크로어레이 실험 자료는 각각 Kim 등 (2005)에서 보고한 선별절차, 표준화 그리고 결측치 대치 등의 사전 처리 절차를 통하여 대장암 cDNA 마이크로어레이 실험 자료는 최종적으로 14886×123의 자료 행렬을 구성하였고, 위암 cDNA 마이크로어레이 실험 자료는 15177×194의 자료 행렬을 구성하였다. 표준화의 필요성 여부를 다루는 추후 논의에서는 상기 사전 처리 절차에서 표준화만 생략한 채로 사전 처리과정을 거친 자료를 "표준화 이전" 자료로 부르기로 한다.

2.2. 통계적 방법

Kim 등 (2005)은 표 2.1처럼 부분적으로 짝을 이룬 자료에서 특이발현 유전자를 검색하는 t_3통계량으로서 다음 식 (2.1)의 통계량을 제안하였다.

$$t_3 = \frac{n_1\overline{D} + n_H\left(\overline{V} - \overline{U}\right)}{\sqrt{n_1 S_D^2 + n_H^2\left(\frac{1}{n_2}S_U^2 + \frac{1}{n_3}S_V^2\right)}} \sim N(0,1), \tag{2.1}$$

단, n_1, n_2, n_3는 각각 완전하게 짝을 이룬 例數, Y가 결측이 된 例數, X가 결측이 된 例數를 나타내며, n_H는 n_2와 n_3의 조화 평균이고, $\overline{D}$, $\overline{U}$, $\overline{V}$는 각각 $D \equiv X - Y$, U, V의 표본 평균이며, S_D^2, S_U^2, S_V^2는 각각 D, U, V의 표본 분산을 나타낸다. 그리고 식 (2.1)의 분포는 귀무가설인 "검사 유전자가 특이발현 유전자가 아니다"에서 성립한다. 식 (2.1)의 t_3통계량은 짝을 이룬 t통계량과 이표본 t통계량을 조합하여 표 2.1의 자료를 모두 활용하면서 특이 발현 유전자를 검색하고 있다.

Qui 등 (2005a)은 표준화가 유전자간 상관관계를 완전하게 제거하여 주는지를 파악하기 위하여 St.Jude Children's Research Hospital(SJCRH, 미국 테네시주 멤피스 소재)에서 실시한 마이크로어레이 실험자료(http://www.stjuderesearch.org/data/ALL1)를 대상으로 표준화 이전과 표준화 이후, 그리고 독립적인 관찰치를 N(0,1)에서 생성한 모의실험 자료 각각에 대하여 모든 가능한 유전자 쌍의 상관계수 분포를 조사하였다. 표준화 이후의 유전자간 상관계수의 분포는 표준화 이전보다 훨씬 대칭에 가까운 분포를 보이고 있으나, 모의실험 자료 보다는 두꺼운 꼬리를 나타내고 있음을 보고하면서 표준화가 유전자간 상관관계를 완전하게 제거하지 못하는 증거로 제시하고 있다. Qui 등 (2005a)의 표준화 효과를 검색하는 일련의 계산과정을 표 2.1의 자료에 적용할 경우는 t_3통계량을 사용하여야 한다. 본고에서는 동 계산과정을 표준화효과 검색과정으로 부르기로 하겠다.

Klebanov와 Yakovlev (2006)는 표준화를 실시하여 유전자간 상관관계를 "충분히" 제거할 수 있다면 표준화 이후의 마이크로어레이 자료는 대수의 법칙을 만족해야 하는 것을 주목하여 SJCRH 마이크로어레이 자료를 대상으로 대수의 법칙 성립 여부를 조사하였다. 동조사에 적용된 일련의 계산 과정을 본고에서는 대수의 법칙 계산 과정이라고 부르기로 하겠다.

Klebanov와 Yakovlev (2007)는 유전자간 강한 상관 관계를 가지고 있다 하더라도 표준화 하기 이전 자료에 적정한 변환을 통하여 독립성이 유지되는 급수를 발견할 수 있었으며, 동 급수에다 기존의 독립성을 전제로 하는 통계적 방법을 적용할 수 있음을 보고하였다. 저자들은 동 급수를 δ 급수라고 불렀으며 다음과 같이 정의하였다. 우선 유전자를 분산의 크기 순서대로 오름차순으로 정열하고 짝수 행에서 홀수 행을 빼서 다음 식 (2.2)와 같이 δ 급수를 정의한다.

$$\delta = x_{2i} - x_{2i-1}, \quad i = 1, \ldots, \frac{p}{2}. \tag{2.2}$$

원자료인 x_i는 강하고 長範圍한 상관을 보이고 있지만, δ_i는 근사 독립관계를 나타내는 것을 보고하였다. δ 급수가 과연 대장암자료와 위암자료 각각에서도 존재하는지를 확인하기 위하여 표 2.1의 각 자료에서 완전하게 짝을 이룬 자료만 사용하기로 한다. δ 급수의 독립성 여부를 파악하는 일련의 계산과정을 본고에서는 δ 급수 계산과정이라고 부르기로 하겠다.

다음의 논의에서는 표준화 효과 검색과정과 대수의 법칙 계산과정, 그리고 δ 급수 계산과정을 대장암 cDNA 마이크로어레이 자료와 위암 cDNA 마이크로어레이 자료 각각에 적용한 과정을 기술하고 있다. 각 계산을 두 組의 cDNA 마이크로어레이에 적용한 과정은 사례수와 유전자갯수의 차이만 제외하고는 유사하므로 대장암 cDNA 마이크로어레이 자료의 경우를 대상으로 기술하기로 하고, 위암 cDNA 마이크로어레이 자료의 경우는 특이 사항만 기술하기로 하겠다.

2.2.1. 표준화 효과 검색과정(대장암의 경우) 유전자간 상관관계가 존재한다면 개별 유전자별로 계산한 t_3통계량 간에도 상관관계가 존재할 것이다. t_3통계량간에 상관관계가 존재하는지를 파악하기 위하여 모든 가능한 t_3통계량의 쌍에 대하여 상관계수를 계산하여 동 상관계수의 분포를 살펴보았다.

1) 표 2.1의 대장암 자료를 6개의 副群으로 랜덤하게 나누되, 부분적으로 짝을 이룬 구조가 유지되도록 나눈다. 즉, 완전하게 짝을 이룬 36예로부터 각 副群의 크기가 6이 되도록 랜덤하게 나누고, Y가 결측인 19예로부터 각 副群의 크기가 3이나 4가 되도록 랜덤하게 나누며, X가 결측인 32예로부터 각 副群당 크기가 5나 6이 되도록 랜덤하게 나누어 6개 각 副群마다 부분적으로 짝을 이룬 구조를 유지하도록 한다.

   Y가 결측인 19예에 기초하여 각 유전자 쌍마다 t_3 통계량의 상관계수를 안정적으로 계산하기 위하여 19例를 가급적 많은 갯수의 副群으로 나눌 필요가 있다. 한편 식 (2.1)의 t_3 통계량을 계산하기 위하여 표준편차 S_u를 계산하여야 하고 동 계산을 수행하기 위하여는 副群의 크기가 2이상이어야 하므로 副群 크기를 3으로 잡았다. 19예를 가지고 크기 3(혹은 4)의 副群을 구성하여 6개의 副群을 얻게 되었다.

2) 각 유전자 마다 6개의 t_3통계량을 계산한다.

3) 모든 가능한 유전자 쌍에 대하여 t_3통계량의 상관계수를 계산한다.

표준화 효과 검색과정을 표준화 하기 이전과 표준화 한 이후 각각에 대하여 적용하고, 또 $N(0,1)$으로부터 난수를 생성하여 14886×123 자료 행렬을 구성하고 동 자료 행렬에다 표준화 효과 검색과정을 적용함으로써, 표준화 이전, 표준화 이후 그리고 독립적인 자료를 대상으로 계산한 t_3통계량간의 상관계수 분포를 비교하여 볼 수 있다. 위암 자료의 경우도 기본적으로 표준화 효과 검색과정을 따르되 副群의 개

수를 5개로 한다. 이는 표 2.1에서 보듯이 X가 결측인 예가 11예가 있고, 동11예를 가지고 副群을 결정할 때 각 副群의 크기가 2이상이 되어야지 표본분산의 계산이 가능하므로 5개의 副群으로 결정한다.

대수의 법칙이 적용되는지 여부를 조사하기 위하여서는 표 2.1의 부분적으로 짝을 이룬 자료중 완전하게 짝을 이룬 자료만 사용하기로 한다. 이 경우 대장암은 36예(72개 마이크로어레이)를 대상으로 사전처리를 하였으며 최종적으로 14987×72 자료 행렬을 얻었다. 위암 자료의 경우 85예(170개 마이크로어레이)를 대상으로 사전 처리를 하였고 15192×170 자료 행렬을 얻을 수 있었다.

2.2.2. 대수의 법칙 계산과정(대장암의 경우)

1) 유전자를 처음 1,000개에서 매번 1,000개씩 더해가면서 14,000개까지 14개의 副群 ($i = 1, \ldots, 14$)으로 구분한다. $\hat{a}_i$, $\hat{b}_i$은 각각 i번째 副群의 유전자 평균과 표준편차로 정의한다. 즉 X_{lk}를 k번째 어레이의 l번째 유전자의 발현 강도라고 할 때 $X_{lk} = Y_r$, $r = 1, \ldots, np$로 재 정의한다. Y_r을 사용하여 $\hat{a}_i$, $\hat{b}_i$을 다음과 같이 정의한다.

$$\hat{a}_i = \frac{1}{np_i}\sum_{r=1}^{np_i}\log(Y_i), \quad \hat{b}_i = \left(\frac{1}{np_i}\sum_{r=1}^{np_i}\log(Y_i)^2 - \hat{a}_i^2\right)^{\frac{1}{2}}, \quad \text{단 } p_i\text{는 } i\text{번째 부군의 크기.}$$

2) 36예에서 크기 10의 붓스트랩 표본 200개 ($j = 1, \ldots, 200$)를 추출한다.

3) 1)과 2)에서 유전자 副群(i)과 붓스트랩 표본(j)이 결정되면 모든 (i, j)에 대하여 유전자들의 평균($\hat{a}_{i,j}$)과 표준편차($\hat{b}_{i,j}$)를 구한다.

4) i를 고정시켰을 때 200개 붓스트랩 표본 ($j = 1, \ldots, 200$)을 통하여 3)에서 구해진 평균, 표준 편차 각각의 표준편차를 계산한다.

5) 4)에서 구한 평균의 표준편차와 표준편차의 표준편차를 모든 i에 대하여 계산한다.

6) i를 증가시킴에 따라 평균의 표준편차와 표준편차의 표준편차가 0으로 수렴하는지를 관찰한다.

대수의 법칙 계산과정을 표준화 이전과 표준화 이후, 그리고 $N(0, 1)$에서 14,000개 난수를 생성하여 적용한 경우를 비교하여 봄으로써 표준화 이후에 유전자간 독립성이 성립하는지를 파악할 수 있다. 위암의 경우는 유전자 갯수만 다를 뿐 대장암의 경우와 유사하게 대수의 법칙 계산과정을 적용할 수 있다.

2.2.3. δ 급수의 계산과정(대장암의 경우)

1) 표준화하기 전 14987×36 자료(X-Y를 반응변수로 사용함)로 행별 분산이 작은 순서부터 오름차순으로 유전자들 배열한다.

2) 총 유전자수가 홀수개이므로 맨 마지막 행을 삭제하여 14,986개로 만들어 두 행씩 짝지어 급수를 구성한다.

3) 각 어레이 마다 k×100개 유전자의 평균을 구한다($k = 1, \ldots, 35$).

4) k가 고정되었을 때 k×100개 평균의 36개 어레이간 표준편차를 계산한다.

5) 4)의 표준편차를 Y축에 k를 X축에 그려 k가 커짐에 따라 표준편차가 0으로 수렴하는지를 확인한다.

만일 원자료에서 유전자간 독립성이 유지된다면 3)-5)만 실시하여도 k×100개 유전자 평균의 표준편차는 k가 커짐에 따라 0으로 수렴하게 될 것이다. 따라서 δ 급수의 계산과정을 전부 실시한 δ 급수의 평

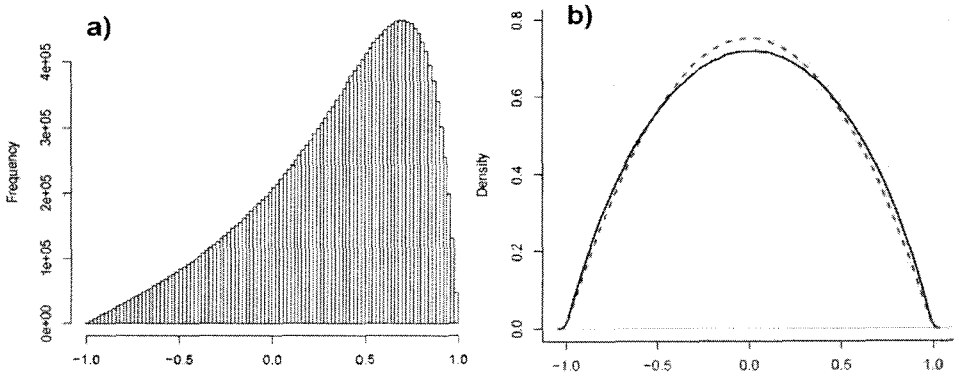

그림 3.1. 대장암 자료에서 t_3 통계량간 상관계수의 분포(a) 표준화 이전, b) 표준화 이후(실선)와 모의실험을 통한 독립적 자료(점선)에 기초한 상관계수)

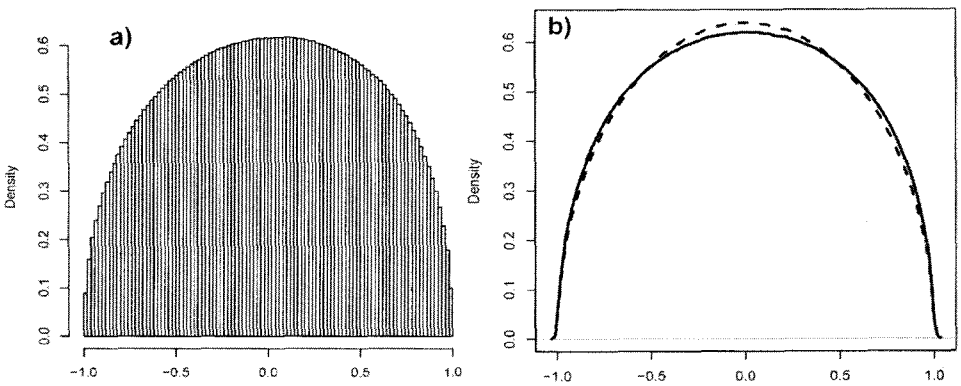

그림 3.2. 위암 자료에서 t_3 통계량간 상관계수의 분포(a) 표준화 이전, b) 표준화 이후(실선)와 모의실험을 통한 독립적 자료(점선)에 기초한 상관계수)

균의 표준편차와 3)–5)의 과정만 거친 급수의 평균의 표준편차가 각각 0으로 수렴하는지 여부를 파악함으로써 δ 급수의 독립성 뿐 아니라 유전자간 독립성 여부도 아울러 판단할 수 있다. 위암의 경우도 유전자 개수와 완전하게 짝을 이룬 예 수만 달라질 뿐 유사한 절차를 적용할 수 있다.

3. 결과

대장암 자료에서 유전자간 상관관계를 파악하기 위하여 표준화효과 검색과정을 적용하였다. 즉 모든 가능한 유전자쌍에 대하여 통계량의 상관계수를 표준화 이전, 표준화 이후 그리고 모의실험을 통한 독립적 자료 각각에 대하여 계산하여 다음 그림 3.1의 결과를 얻었다. 위암 자료의 경우는 그림 3.2와 같이 얻어졌다.

그림 3.1a)의 경우 표준화 이전의 상관계수의 분포는 우측으로 왜곡되어 있고, 중앙값이 0.4402이고 전체 쌍의 25%가 0.7 이상의 강한 상관관계를 보이고 있다. 그림 3.1b)의 실선은 표준화 이후의 상관계수

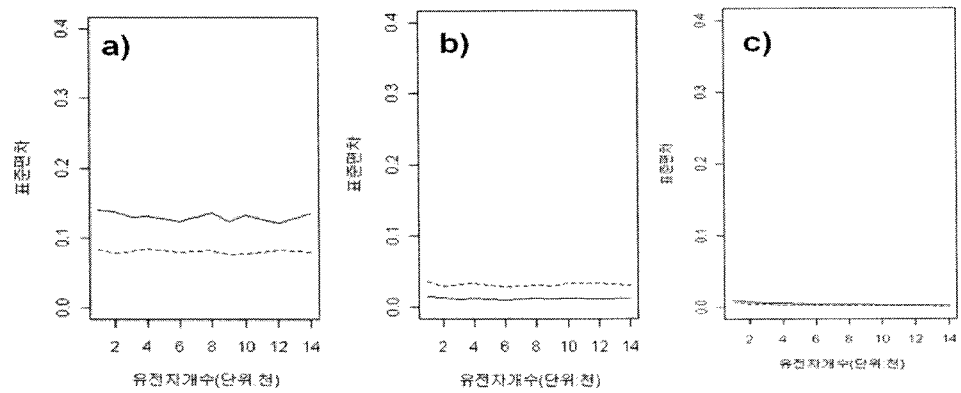

그림 3.3. 대장암 자료의 a) 표준화 이전, b) 표준화 이후, c) 모의실험 자료에 기초하여 얻은 평균($\hat{a}_i$)의 표준편차, 표준편차($\hat{b}_i$)의 표준편차를 유전자 갯수(i)의 함수로써 표시하고 있다(실선: 평균들의 표준편차, 점선: 표준편차의 표준편차)

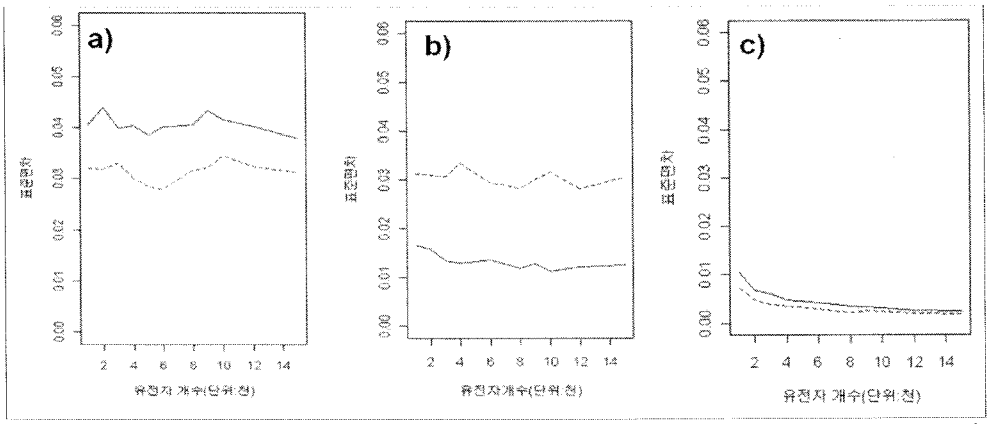

그림 3.4. 위암 자료의 a) 표준화 이전, b) 표준화 이후, c) 모의실험 자료에 기초하여 얻은 평균($\hat{a}_i$)의 표준편차, 표준편차($\hat{b}_i$)의 표준편차를 유전자 갯수(i)의 함수로써 표시하고 있다(실선: 평균들의 표준편차, 점선: 표준편차의 표준편차)

의 분포인데 대칭적 분포(평균 0.0048, 표준편차 0.4594)를 보이지만 독립적인 모의실험 자료에 기초한 상관계수의 분포(점선)과 비교하여 볼 때 표준화 이후 상관계수의 분포가 어깨 부분이 더 두꺼운 것을 알 수 있다. 이는 표준화를 통하여 유전자간 의존성을 완전하게 제거하지는 못한다는 것을 보여주고 있다. 위암 자료의 경우도 비슷한 양상을 나타내고 있다. 표준화를 실시한 이후에 유전자간 독립성이 유지되는지에 대한 논의는 대수의 법칙이 성립되는지 여부로 확실하게 알 수 있다. 대장암 자료의 표준화 이전, 표준화 이후 그리고 모의실험을 통한 독립적 자료 각각에 대수의 법칙 계산과정을 적용하여 다음 그림 3.3을 얻었다. 위암 자료의 경우도 유사한 결과를 얻었으며 그림 3.4와 같다.

그림 3.3과 그림 3.4를 살펴보면 표준화를 한 후에 $\hat{a}_i$, $\hat{b}_i$ 각각의 표준 편차가 현저하게 줄어들기는 하였지만 유전자 개수가 늘어남에 따라 동 표준편차가 0으로 수렴하는 현상은 나타나지 않는다. 반면에 모의실험의 독립적 자료에 기초한 표준편차는 0으로 수렴하는 현상이 나타남을 알 수 있다. 이는 표준화를 하여도 대수의 법칙이 성립할 정도로 유전자간 독립성이 유지되지 않는다는 것을 보여주고 있다. δ

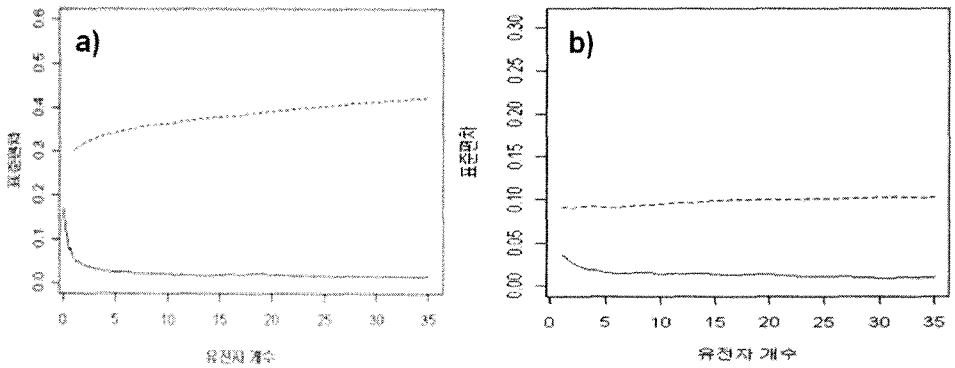

그림 3.5. δ 급수(실선)와 보통 급수(점선) 각각의 평균의 표준편차를 유전자 개수의 함수로 표시한 그림(a) 대장암 자료, b) 위암자료)

급수의 존재를 확인하기 위하여 δ 계산과정을 표준화 하기 이전의 대장암 자료와 위암 자료 각각에 적용한 결과는 다음 그림 3.5와 같이 얻었다.

그림 3.5에서 점선의 보통 급수의 경우 유전자 개수가 늘어남에 따라 평균의 표준편차가 0에 수렴하는 현상을 보이지 않는 반면, 실선으로 표시된 δ 급수의 경우 평균의 표준편차가 0으로 급속히 수렴하는 현상을 볼 수 있다.

4. 결론 및 토의

본고에서는 국내에서 생산된 두조의 cDNA 마이크로어레이 실험 자료에서 유전자간 상관 관계가 강하고 장범위하다는 것을 확인하고 있으며, 아울러 표준화과정이 유전자간 상관관계를 완전하게 제거하지 못한다는 것을 여러 측면에서 보여주고 있다. 비모수적 경험적 베이즈 방법은 어느 한 유전자에 대한 추론을 실시할 때 다른 유전자 정보를 "빌려 오는", 이른바 유전자 정보의 병합을 실시하고 있다. 그러나, 유전자간 상관관계가 강할 때 유전자간 정보의 병합이 추가로 제공할 수 있는 정보는 독립적인 자료의 경우와는 판이하게 다를 수 있다. 이미 Qiu 등 (2005b)과 Qiu 등 (2006)에서 보고 하고 있듯이 유전자간 강한 상관 관계를 무시한 채 비모수적 경험적 베이즈 방법이나 FDR 절차를 적용하면 검색되는 특이발현 유전자 개수가 불안정하게 되고, FDR 추정량의 분산이 커지는 현상이 나타나고 있다.이러한 맥락에서 Efron (2007)은 유전자간 상관을 FDR 추정에 반영하여 이른바 조건부 FDR을 제안하였다. 유전자간 상관관계는 δ 급수와 같은 적정한 변환을 통하여 독립적인 급수로 변환이 가능하고, 동 급수에 기존의 통계적 방법을 적용할 수 있으리라 믿는다. 그러나, δ 급수를 원자료로 사용한 통계적 분석의 결과를 어떻게 해석할지와 유전자간 상관구조를 파악하여 표현형을 검색하는 과제 등은 아직 관련 학계의 과제로 남아 있다.

참고문헌

Efron, B. (2003). Robbins, empirical Bayes and microarrays, *The Annals of Statistics,* **31**, 366–378.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null hypothesis, *Journal of the American Statistical Association,* **99**, 96–104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing, *Journal of the American Statistical Association,* **102**, 93–103.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment, *Journal of the American Statistical Association,* **96**, 1151–1160.

Frantz, S. (2005). An array of problems, *Nature Reviews Drug Discovery,* **4**, 302–303.

Kim, B. S., Kim, I., Lee, S., Kim, S., Rha, S. Y. and Chung, H. C. (2005). Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer, *Bioinformatics,* **21**, 517–528.

Klebanov, L., Jordan, C. and Yakovlev, A. (2006). A new type of stochastic dependence revealed in gene expression data, *Statistical Applications in Genetics and Molecular Biology,* **5**, Ariticle 7.

Klebanov, L. and Yakovlev, A. (2006). Treating expression levels of different genes as a sample in microarray data analysis: Is it worth a risk?, *Statistical Applications in Genetics and Molecular Biology,* **5**, Ariticle 9.

Klebanov, L. and Yakovlev, A. (2007). Diverse correlation structures in gene expression data and their utility in improving statistical inference, *The Annals of Applied Statistics,* **1**, 538–559.

Marshall, E. (2004). Getting the noise out of gene arrays, *Science,* **306**, 630–631.

Qui, X., Brooks, A. I., Klebanov, L. and Yakovlev, A. (2005a). The effects of normalization on the correlation structure of microarray data, *BMC Bioinformatics,* **6**, 120.

Qui, X., Klebanov, L. and Yakovlev, A. (2005b). Correlation between gene expression levels and limitations of the empirical Bayes methodology for finding differentially expressed genes, *Statistical Applications in Genetics and Molecular Biology,* **4**, Ariticle 34.

Qui, X., Xiao, Y., Gordon, A. and Yakovlev, A. (2006). Assessing stability of gene selection in microarray data analysis, *BMC Bioinformatics,* **7**, 50.

Qui, X. and Yakovlev, A. (2006). Some comments on instability of false discovery rate estimation, *Journal of Bioinformatics and Computational Biology,* **4**, 1057–1068.

Stolovitzky, G. (2003). Gene selection in microarray data: The elephant, the blind men and our algorithm, *Current Opinions in Structural Biology,* **13**, 370–376.

Yang, S., Jeung, H. C., Jeong, H. J., Choi, Y. H., Kim, J. E., Jung, J. J., Rha, S. Y., Yang, W. I. and Chung, H. C. (2007a). Identification of genes with correlated patterns of variations in DNA copy number and gene expression level in gastric cancer, *Genomics,* **89**, 451–459.

Yang, S., Shin, J., Park, K. H., Jeung, H-C., Rha, S. Y., Noh, S. H., Yang, W. I. and Chung, H. C. (2007b). Molecular basis of the difference between normal and tumor tissues of gastric cancer, *Biochimica et Biophysica Acta,* **1772**, 1033–1040.

A Report on the Inter-Gene Correlations in cDNA Microarray Data Sets

Byung Soo Kim[1] · Jee Sun Jang[2] · Sang-Cheol Kim[3] · Johan Lim[4]

[1]Department of Applied Statistics, Yonsei University; [2]Korea Economic Research Institute;
[3]Department of Applied Statistics, Yonsei University;
[4]Department of Statistics, Seoul National University

(Received February 2009; accepted April 2009)

Abstract

A series of recent papers reported that the inter-gene correlations in Affymetrix microarray data sets were strong and long-ranged, and the assumption of independence or weak dependence among gene expression signals which was often employed without justification was in conflict with actual data. Qui *et al.* (2005) indicated that applying the nonparametric empirical Bayes method in which test statistics were pooled across genes for performing the statistical inference resulted in the large variance of the number of differentially expressed genes. Qui *et al.* (2005) attributed this effect to strong and long-ranged inter-gene correlations. Klebanov and Yakovlev (2007) demonstrated that the inter-gene correlations provided a rich source of information rather than being a nuisance in the statistical analysis and they developed, by transforming the original gene expression sequence, a sequence of independent random variables which they referred to as a δ-sequence. We note in this report using two cDNA microarray data sets experimented in this country that the strong and long-ranged inter-gene correlations were still valid in cDNA microarray data and also the δ-sequence of independence could be derived from the cDNA microarray data. This note suggests that the inter-gene correlations be considered in the future analysis of the cDNA microarray data sets.

Keywords: cDNA microarray, nonparametric empirical Bayes method, correlation, independence, differential expression.

This work was supported by the Korea Research Foundation Grant funded by the Korean Government(KRF-2008-312-C00055).

[1]Corresponding author: Professor, Department of Applied Statistics, Yonsei University, 262 Seongsanno, Seodaemun-Gu, Seoul 120-740, Korea. E-mail: bskim@yonsei.ac.kr