The Korean Communications
in Statistics Vol. 11 No. 2, 2004
pp. 247-254

A Study on Two Group Comparison in Gene Expression Data

Kyung Ha Seok[1], Sangfeel Lee[2], and Whasoo Bae[3]

Abstract

Tusher, Tibshirani and Chu (2001) suggested SAM (Significance Analysis of Microarrays) to compare two groups under different conditions for each gene, using microarray data. They used two sample t-statistic adding fudge factor in the denominator to prevent the value of statistic from being inflated by large sample variance, which might result in significant difference despite of a small value in the numerator. This paper aims at finding robust fudge factor and replacing it in two-sample t-statistic used in SAM, which we call Modified SAM (MSAM). Using the simulated data and data used in Dudoit et al.(2002), it is shown that MSAM find significant genes better and has less error rate than SAM.

Keywords : error rate, fudge factor, robust statistic, SAM

1. 서론

최근 유전학 분야에서 가장 관심이 되는 것 중의 하나는 질병 및 암과 관련된 유전자를 찾는 작업이라 볼 수 있으며 이는 1995년과 1996년에 걸쳐 마이크로 어레이 기법이 도입되면서 더욱 활성화되었다. 마이크로어레이는 기술적 측면에서 cDNA 마이크로어레이와 oligonucleotide 칩 등 두 가지로 분류된다. 두 경우 모두 mRNA (messenger RNA)의 발현 량을 측정한다는 점은 동일하고 발현된 유전자의 양은 일반적으로 $p \times n$ 행렬 X로 표현되는데 흔히 X를 유전자 표현행렬 (gene expression matrix)이라고 부른다. 이러한 자료를 유전자 발현 자료 (gene expression data)라 하며, 이는 실험에서 얻어진 특정 유전자들의 발현 변화 정도를 비교하는 유전자의 기능을 밝히는데 사용된다.

Tusher, Tibshirani and Chu (2001)는 유전자 발현 자료를 이용해 서로 다른 두 실험조건 (처리군과 대조군)에서 유전자들의 유의한 차이를 알아보기 위해 각 유전자별로 이 표본 t-통계량을 이용하되, 두 그룹간의 차이는 작음에도 불구하고 분산이 너무 작아 검정 통계량의 값이 커지는 것을 막기 위하여 보완 인자 (fudge factor)를 구하여 분모에 더해주는 방법인 SAM을 제안

1) Professor, Department of Data Science, Inje University, Kimhae, 621-749, Korea.
E-mail : skh@stat.inje.ac.kr
2) Graduate Student, Department of Data Science, Inje University, Kimhae, 621-749, Korea.
3) Associate Professor, Department of Data Science, Inje University, Kimhae, 621-749, Korea.

했다.

SAM에서 보완 인자를 구하는데 사용된 주된 통계량인 변동계수로 이는 이상치에 민감한 평균과 분산이 사용되고 있다. 유전자 발현 자료는 변동이 심할 뿐 아니라 이상치가 존재하는 경우가 빈번하므로, 본 논문에서는 보완인자를 구하는 절차에서 이러한 자료의 특성을 고려하여 강건한 (robust) 통계량으로 대체하여 보완 인자를 구하는 방법으로 수정된 SAM (MSAM)을 제안한다.

제 2장에서는 MSAM을 제안하고 제 3장에서는 모의실험을 통한 자료와 실제자료를 이용하여 SAM과 MSAM을 적용시킨 결과에 대한 비교를 하고, 결론 및 향후과제에 대해 제 4장에 다루었다.

2. 수정된 SAM (MSAM)

2.1 SAM (Significance Analysis of Microarrays)

Tusher, Tibshirani and Chu (2001) 에 의해 제안되었으며, 서로 다른 두 실험조건에서 어떠한 유전자들이 차이가 있는지를 알아보기 위한 통계학적인 기술로서 유전자 발현과 반응 변수 사이의 관계의 정도를 측정하는 방법이다. 일반적으로 SAM에서의 입력에는 마이크로어레이 실험에서 나온 반응 변수나 각종 실험에서 나온 유전자의 발현 척도들이 사용되며, 이때의 반응 변수들은 2개 혹은 3개의 그룹으로 나눌 수 있는 변수가 될 수 있고 양적 변수 혹은 중도 절단된 생존시간이 될 수 있다.

유전자 발현과 반응 변수 사이의 관계의 정도를 측정할 때, SAM은 일반적인 t-검증에서 분산이 작은 경우에 t-통계량의 값이 커지는 것을 막기 위하여 보완인자 (fudge factor)를 구하여 분모에 더하여 통계량 d를 식(1)에서와 같이 계산한다.

$$d(i)=\frac{\overline{x_T}(i)-\overline{x_C}(i)}{s(i)+s^{\alpha}} \tag{1}$$

여기서 는 처리 군에서 (i)번째 유전자에 대한 평균 수준으로, $\overline{x}_C(i)$는 대조군에서의 각 유전자 (i)별 평균 수준으로 정의되며, $s(i) = \sqrt{a\left\{\sum_m[x_m(i)-\overline{x}_T(i)]^2+\sum_n[x_n(i)-\overline{x}_C(i)]^2\right\}}$ 로 정의되는데 여기서, $\sum_m$ 과 $\sum_n$은 각각 처리 군과 대조 군에서의 발현 척도의 총합이며, 또한 $a=\frac{(1/n_1+1/n_2)}{(n_1+n_2-2)}$ 으로, n_1 과 n_2는 각각 처리 군과 대조 군에서의 발현 척도들의 수를 나타낸다. 또한 s_0는 보완인자(fudge factor)로써 작은 $s(i)$값으로 인해 $d(i)$의 값이 커지는 것을 방지하기 위한 값으로 다음과 같이 계산된다.

1. $s(i)$값들의 α백분 위수를 s^{α} 라 둔다.
2. $d^{\alpha}(i)=\frac{r(i)}{s(i)+s^{\alpha}}$, $r(i)=\overline{x_T}(i)-\overline{x_C}(i)$ 로 둔다.

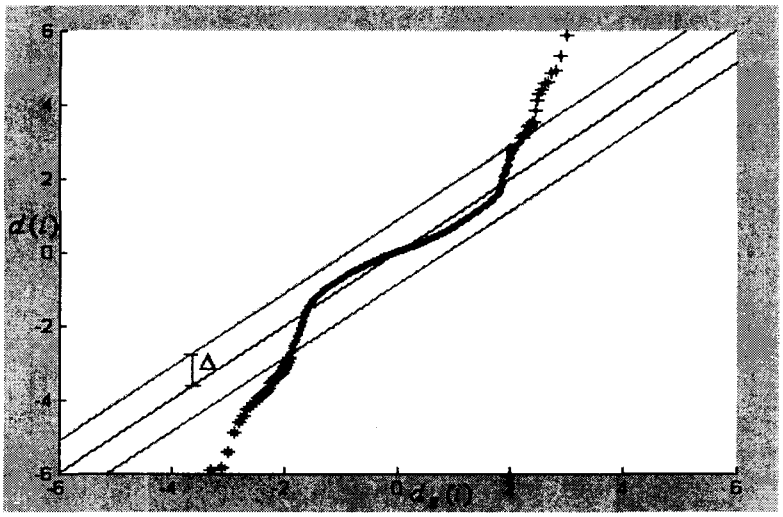
3. $s(i)$값들의 백분 위수들을 계산하여 $q(1)<q(2)<\cdots<q(100)$로 정의한다.
4. $\alpha\in(0,0.05,0.1,\cdots,1.0)$에 대하여
   a. $v(j)=mad\{d^{\alpha}(i)\mid s(i)\in[q(j),q(j+1)]\}$ 를 계산한다.
   여기서 $mad(x_i)=\dfrac{median\mid x_i-median(x_i)\mid}{0.64}$ 이다.
   b. $v(j)$값들의 변동 계수인 $cv(\alpha)$를 계산한다.
5. $cv(\alpha)$를 최소로 하는 $\hat{\alpha}$ 를 구한다.
6. $\widehat{s_0}=s^{\hat{\alpha}}$ 를 구하여 s_0의 값으로 사용한다.

p개 유전자 각각에 대해 식 (1)을 이용하여 $d(i)$ 값들을 다시 크기 순으로 나열하여 $d(1)\le d(2)\le\cdots\le d(p)$로 둔다. j번째 데이터의 치환 (permutation)을 통해 구한 d 통계량, $d_j(i)$로 하여 값을 구해 크기 순으로 나열하면, $d_j(i)$는 치환 j에 대한 d의 i번째 순서통계량의 값인데, 총 n번의 치환을 하여 얻은 d 통계량들의 평균을 $d_E(i)=\sum_{j=1}^{n}d_j(i)/n$로 나타내면, $d_E(1)\le d_E(2)\le\cdots\le d_E(p)$ 가 된다. 각 유전자 i에 대해 정해진 Δ를 이용하여 $\mid d(i)-d_E(i)\mid>\Delta$ 이면 그 유전자는 유의 하다고 할 수 있다. <그림 1>은 $d(i)$ 와 $d_E(i)$의 산점도를 나타내는데, 여기서 (+) 또는 (*)으로 표시된 유전자가 정해진 Δ에 대해 유의한 유전자들이다.

2.2 수정된 SAM (MSAM)

2.1절의 SAM에서 보완인자를 구할 때, 단계4와 5에서 변동계수인 cv 를 사용하는데 이는 평균과 표준편차를 사용하므로 이상치에 영향을 받을 수 있고, 유전자 발현 자료는 이상치가 존재하는 경우가 빈번하므로 이러한 상황에 보다 강건한 보완인자를 구하는 것이 필요하다.

본 연구에서는 보완인자를 구하는 SAM의 기존의 단계 1, 2, 3은 그대로 두고, 단계 4와 5의 절차에서 사용되는 cv의 분모인 표준편차대신 사분위수 제곱근을, 분자의 평균대신 중앙값을 사용하여 이상치에 덜 영향을 받는 보완인자를 사용하는 검정통계량을 주는 MSAM을 제안한다. 본 연구에선 MSAM을 적용시키기 위하여 사용된 프로시저는 Chu 등(2001)에 의해 지원된 SAM의 기술 지원 문서를 바탕으로 하여, Matlab v6.1을 사용하여 프로그래밍 하였다.



<그림 1> $d(i)$ 와 $d_E(i)$의 산점도

SAM의 경우, 유의 하다고 판단된 유전자에 대한 신뢰를 부여하기 위하여 유의한 유전자들(significant genes)의 수와 잘못 분류된 유의한 유전자들(falsely significant genes)의 수의 비로 계산되어지는 오분류율, 즉 False Discovery Rates (FDR)를 사용하였다. 본 연구에서 제안된 MSAM과 SAM의 효율을 비교하기 위해 다음 장에서 Dudoit 등 (2002)에서 사용된 Apo AI 자료와 모의실험 자료를 이용하여 FDR을 비교하고 있다.

3. SAM과 MSAM의 비교

3.1 Apo AI 자료를 이용한 비교

Apo AI 자료는 Dudoit 등(2002)에서 사용된 cDNA 마이크로어레이 실험에서 얻어진 쥐의 유전자 발현 자료로 16마리의 쥐들 중 8 마리는 HDL 신진대사에서 중요한 역할을 하는 유전자인 Apo AI (apolipoprotein AI)를 없앤 쥐로 처리군을 형성하고 나머지 8마리는 정상의 쥐로 대조군을 형성하여 이들로부터 얻어진 16개의 열과 각 열이 5,548개의 cDNA 프로브(probes)로 구성되어 있는 마이크로어레이 자료다.

$\Delta = 0.55$로 하여 SAM을 적용시킨 결과가 <표 1>에 나타나 있으며 이 표에서 볼 때, $|d(i) - d_E(i)| > 0.55$ 를 만족하는 유전자의 수와 잘못 판단된 유전자의 수의 비인 FDR의 값이 <표 1>에 나타나 있다. Apo AI 실험 데이터에서 100개의 유의한 유전자를 식별하였으며 그에 따른 FDR 값은 27% 란 것을 알 수 있다. MSAM을 적용시킨 결과가 <표 2>에 나타나 있다.

<표 1>과 <표 2>로부터 전반적인 FDR값이 MSAM가 SAM보다 낮음을 알 수 있으며 SAM에서의 유의한 유전자 100개와 MSAM에서 유의한 유전자로 식별된 98개중 92개가 일치하여 MSAM에서의 프로시저 내의 특별한 문제점은 없는 것으로 나타났다.

<표 1> Apo AI 자료 SAM 적용 결과($\Delta = 0.55$)

#significant genes	#falsely significant genes	FDR(%)
130	29.2	22.46
118	20.8	17.63
112	19.3	17.23
98	14.0	14.29
95	12.1	12.60
92	10.1	10.98
89	9.3	10.45
87	7.9	9.08
84	6.4	7.62

<표 2> Apo AI 자료 MSAM 적용결과

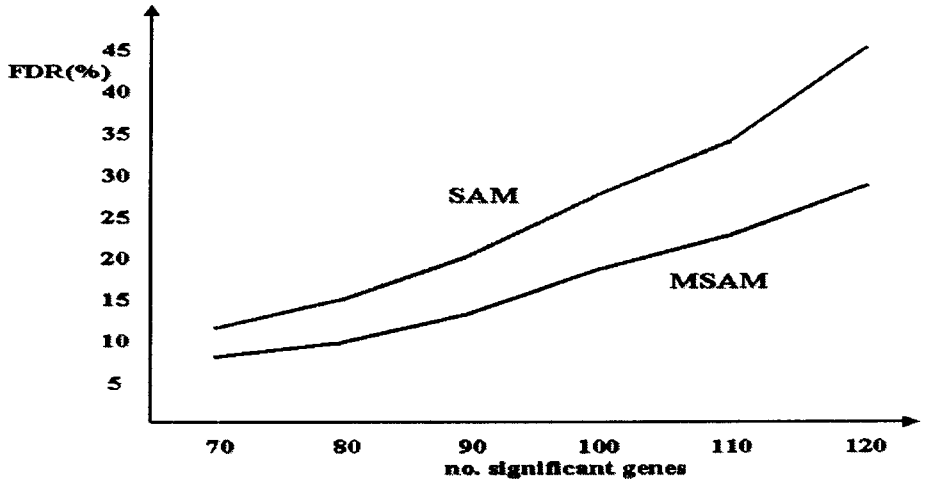
#significant genes	#falsely significant genes	FDR(%)
116	51.1	44.05
114	48.9	42.89
109	36.0	33.03
105	33.3	31.71
100	26.8	26.80
97	20.7	21.34
87	18.1	20.80
81	12.7	15.85

SAM과 MSAM을 비교하기 위하여 유의한 유전자의 수와 FDR 값에 대한 그래프가 <그림 2>에 나타나 있다. <그림 2>에서 볼 때 산출된 유의한 유전자의 수가 비슷한 경우 기존의 SAM 보다 MSAM에서 더 낮은 FDR값을 보여줌을 알 수 있다.

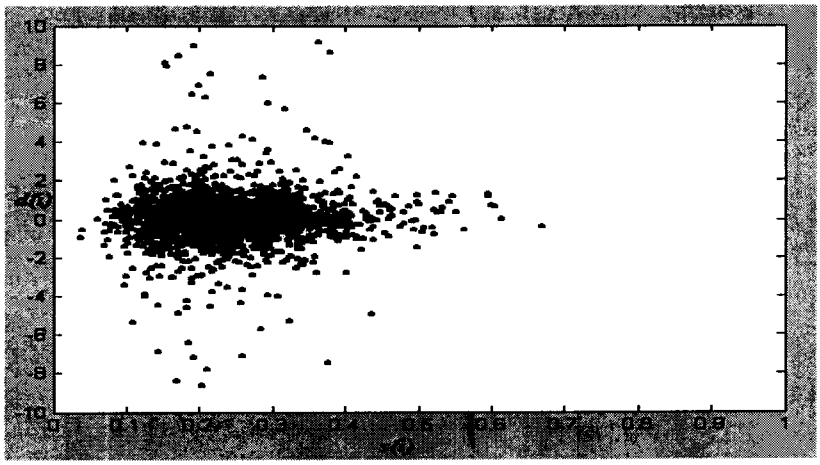
3.2 모의실험

좀 더 정확한 검증을 위해 모의실험 자료를 다음과 같이 만들어서 분석해 보았다.

총 2000개의 가상 유전자에 대한 데이터를 4개의 대조군과 4개의 처리군으로 2000행, 8열의 행렬로 생성하였다. 실제자료와 비슷한 분포를 가진 자료를 생성하기 위하여 각 셀은 $\{\chi^2(1) + N(0,1)\}/4$ 의 분포를 따르는 난수 값으로 이루어졌다.



<그림 2> SAM 과 MSAM의 FDR 비교



<그림 3> 모의 실험 자료의 $d(i)$ 와 $s(i)$의 산점도

그리고 유의한 유전자로 엄선하기 위하여 1행에서 100행까지의 처리군에 한해서는 대조군의 표준편차에 $N(0,1)\times 4$의 값을 곱하여 처리군의 평균이 대조군의 평균보다 큰 경우에는 더해주고 처리군의 평균이 대조군의 평균보다 작은 경우에는 빼줌으로써 다른 행과는 확실히 차이를 주었다. 따라서 예상되어지는 유의한 유전자의 행 번호는 1에서 100까지의 값이다. 먼저, 생성되어진 모의 데이터와 실제 유전자 발현 데이터와의 유사성을 살펴보기 위하여 <그림 3>과 같은 $d(i)$ 와 $s(i)$의 산점도를 그려보았다.

실험을 10번씩 반복하면서 전체적중률을 비슷하게 하는 Δ를 선택하였더니 0.15, 1.0이 나왔다. 이를 이용하여 FDR 값을 비교하기 위해 다음의 <표 3>을 얻었다.

<표 3> 모의 실험데이터의 SAM 과 MSAM의 비교

	SAM($\Delta = 0.15$)				MSAM($\Delta = 1.0$)			
	# significant genes	# false genes	FDR(%)	1:100적중률(%)	# significant genes	# false genes	FDR(%)	1:100적중률(%)
1	50	6.2	12.4	90.0	50	2.0	4.0	96.0
2	58	4.7	8.1	91.4	55	4.8	8.7	85.4
3	59	7.0	11.9	89.8	57	5.8	10.2	84.2
4	60	5.3	8.8	88.3	57	6.6	11.7	82.5
5	64	3.1	4.8	92.2	60	4.8	8.0	88.3
6	67	8.8	13.1	85.1	60	5.2	8.7	81.7
7	68	12.1	17.8	82.4	61	5.5	8.9	82.0
8	71	14.8	20.9	83.1	63	8.5	13.4	92.1
9	71	11.3	15.9	81.7	64	5.6	8.8	84.4
10	72	14.5	20.1	76.4	68	9.1	13.4	86.8
평균	64.0	8.79	**13.4**	**86.0**	57.3	5.45	**9.6**	**86.3**

각 실험에서 치환은 36번으로 하였고, 각 치환에서 구하여진 유의한 유전자 수의 평균을 나타내었다(number of false genes). 또한 <표 3>에서 산출된 유의한 유전자와 실험배경에서 유의한 유전자로 사전에 정의한 1행에서 100행까지의 유전자중에서 일치하는 데이터의 수를 1:100 적중수로 정의하였다. 그리고 1:100 적중 수를 유의한 유전자의 수로 나눈 백분율을 1:100 적중률(%)로 정의하였다.

유의한 유전자가 50개 일 때 SAM의 FDR값과 적중률은 각각 12.4%, 90%이고 MSAM의 FDR값과 적중률은 각각 4.0%, 96%로 MSAM이 더 좋은 결과를 보였다. 그리고 유의한 유전자가 68개 일 때 SAM의 FDR값과 적중률은 17.8%, 82.4%이고 MSAM의 FDR값과 적중률은 13.4%, 86.8%로써 MSAM에서 더 좋은 결과를 보였다.

<표 3>을 전체적으로 정리하면, SAM과 MSAM의 적중률의 평균은 86%와 86.3%로 비슷한 수준을 이루고 있으며 SAM의 평균 FDR값은 13.4%이었으며 MSAM의 평균 FDR값은 9.6%로 이들 평균값에 대한 t-검정을 한 결과 유의확률이 0.033으로 MSAM의 FDR이 SAM의 FDR보다 평균적으로 낮다고 할 수 있어 MSAM의 적용하는 경우, 더 좋은 결과를 보인다고 할 수 있다.

4. 결론 및 향후과제

SAM과 MSAM의 비교에서 볼 때 이상치에 민감한 평균과 분산을 써서 변동계수를 최소화시켜 보완인자로 사용한 SAM의 경우보다 이들을 중앙값과 사분위수 범위로 대체하여 사용한 MSAM의 경우 오류율이 더 낮게 나타난다는 사실을 자료에 적용시켜 본 결과에서<표 1>, <표 2> 그리고 <표 3>에서 알 수 있었다.

SAM의 경우, 일반적인 t-검증의 경우와는 달리 한번에 수천 번 이상의 통계량을 계산해내고 데이터 분산에 의한 통계량의 변동을 줄이기 위하여 보완 인자를 산출하여 분모에 더해줌으로써 유의한 유전자를 추정한다. Fold change를 비롯한 기존의 통계적 방법들에 비해 SAM이 유전자 발현 분석에서 한층 좋은 결과를 보이긴 하였으나, 본 논문에서 제안하였던 MSAM의 FDR값이 SAM의 FDR값보다 유의한 유전자 산출 시에 더 낮은 값을 얻은 것으로 미루어, 현재의 SAM에

도 보완할 부분이 있을 것으로 추정된다.

SAM과 MSAM을 비교하는 근거로 사용된 FDR, 즉 오류율에 대한 연구도 중요하며, 이에 대해서 Benjamini과 Hochberg (1995) , Efron 등(2001) 과 Storey (2001)등의 연구에서 보여진 오류율에 대한 적용을 해 보고 비교하는 것도 의미가 있을 것이다. Westfall and Young (1993)에서 볼 수 있는 것처럼 수천 번 이상의 통계량을 계산하여 검정하는 다중비교에서 발생하는 오류에 대한 통제에 대한 부분을 포함하는 비교 연구도 되어질 필요가 있다.

유전자에 관한 관심이 더 한층 높아지는 이때, DNA칩에 의한 실험이 더욱더 확산되고 그에 따라 얻어지는 유전자 발현 데이터와 이들에 대한 분석 연구는 시스템 적인 변동이 줄여진 데이터를 얻고 있지만 아직도 얻어진 자료를 정제하는 부분은 데이터에 중요한 영향을 미친다. Yang 등(2001)에서 연구된 Normalization에 대한 데이터의 정제 방법을 분석전반부에 도입하여 적용시키는 것도 중요할 것이다.

참고문헌

[1] Benjamini, Y. and Hochberg, Y.(1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Ser. B., 57, 289-300.

[2] Chu, G., Narasimhan, B., Tibshirani, R.,Tusher, V.(2001). SAM("Significance Analysis of Microarrays") *Users guide and technical document.*

[3] Dudoit, S., Yang, Y.H., Callow, M.J.(2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, 12, 111-140.

[4] Efron, B., Tibsirani, R., Storey, J.D., and Tusher, V.(2001). Empirical Bayes analysis of microarray experiment, *Journal of the American Statistical Association*, 96, 1151-1160.

[5] Storey, J.D.(2001). A direct approach to false discovery rates, *Journal of the Royal Statistical Society*, Ser. B., 64, 479-498.

[6] Tusher V.G., Tibsirani, R., and Chu, G.(2001). Significance analysis of microarrays pplied to the ionizing radiation response, *Proceedings of the National Academy of Science*, 98, 5116-5121.

[7] Yang, Y.H., Dudoit, S., Luu P.D.M., and Speed, T.(2001). Normalization for cDNA Microarray Data, Technical Report 584, Department of Statistics, UC at Berkeley.

[8] Westfall, P. and Young, S.(1993). *Resampling-based multiple testing,* Wiley, New York.

[2003년 8월 접수, 2004년 4월 채택]