The Korean Journal of Applied Statistics (2016) DOI: http://dx.doi.org/10.5351/KJAS.2016.29.3.437
29(3), 437–448

Identifying statistically significant gene sets based on differential expression and differential coexpression

Sunho Lee[a,1]

[a]Division of Mathematics and Statistics, Sejong University

(Received December 22, 2015; Revised February 12, 2016; Accepted February 29, 2016)

Abstract

Gene set analysis utilizing biologic information is expected to produce more interpretable results because the occurrence of tumors (or diseases) is believed to be associated with the regulation of related genes. Many methods have been developed to identify statistically significant gene sets across different phenotypes; however, most focus exclusively on either the differential gene expression or the differential correlation structure in the gene set. This research provides a new method that simultaneously considers the differential expression of genes and differential coexpression with multiple genes in the gene set. Application of this NEW method is illustrated with real microarray data example, p53; subsequently, a simulation study compares its type I error rate and power with GSEA, SAMGS, GSCA and GSNCA.

Keywords: microarray experiment, gene set analysis, differential expression, differential coexpression

1. 서론

마이크로어레이 기술의 개발은 수만개 유전자들의 발현을 동시에 관찰할 수 있게 하였고, 통계학계에서는 '작은 n, 큰 p'의 문제 안에서 질병과 관계있는 유전자들을 찾는 방법 연구를 활발히 진행하여 질병의 진단과 치료에 큰 기여를 하고 있다.

초기 유전자 분석은 질병의 발생에 영향을 미치는 유전자를 찾는 단일유전자 분석으로 표현형(질병군 또는 정상군)에 따라 발현 차이가 큰 유전자를 개별적으로 찾아내는 이표본 t 검정과 이를 변형한 Significance Analysis of Microarray(SAM) (Tusher, 2001) 등의 방법이 있다. 그러나 분석 대상 유전자가 많아 결과 해석이 어렵고 생물학적 의미 도출을 위해서는 별도의 분석과정이 필요하며 분석 표본이 달라짐에 따라 결과에 차이가 생기기도 한다. 이런 문제점들을 보완하기 위한 시도로 각 유전자들의 발현 자료와 함께 생물학적 정보를 분석에 반영하는 방법을 연구하게 되었고 공통의 생물학적 요소를 지닌 유전자 집단을 대상으로 서로 다른 표현형 사이에 발현의 차이가 유의한지 검색하는 집단분석(gene set analysis)이 대두되었다. 이미 구축된 Gene Ontology(GO, http://www.geneontoloty.org)나 Kyoto Encyclopedia of Genes and Genomes(KEGG, http://www.genome.jp./kegg) 등의 데이터베이스를 이용하여 생물학적 경로(pathway)나 염색체 위치 등이 같은 다양한 유전자 집단을 구성할 수 있고 통계적 방법으로는 Gene Set Enrichment Analysis(GSEA) (Mootha 등, 2003; Subramanian 등, 2005),

[1]Division of Mathematics and Statistics, Sejong University, 209, Neungdongro, Kwangjinku, Seoul 05006, Korea. E-mail: leesh@sejong.ac.kr

Significance Analysis of Microarray-Gene Set(SAMGS) (Dinu 등, 2007)와 Parametric Analysis of Gene set Enrichment(PAGE) (Kim과 Volsky, 2005) 등이 쉽게 쓰였다.

이런 초기 방법들은 마이크로어레이 자료를 무조건 단순화시키기 위하여 특별한 검증없이 독립성을 가정하고 그들의 특이발현(differential expression)에 초점을 둔 것이다. 점차 분석 방법론에 대한 연구가 활발해지면서 실제 자료들이 이런 독립성 가정과 어긋난다는 지적들이 제기되었고 (Qui 등, 2005; Klebanov와 Yakovkev, 2007; Kim 등, 2009) 집단에 속한 유전자들이 서로 다른 표현형에서 보이는 상호관계의 차이를 관찰하는 특이공발현(differential coexpression) 분석법이 대두되었다. Gene Set Coexpression Analysis(GSCA) (Choi와 Kendziorski, 2009)와 Gene Sets Net Correlations Analysis(GSNCA) (Rahmatallah 등, 2014) 등이 있다.

유전자들의 특이발현과 특이공발현의 양상은 서로 다르기 때문에 질병의 발생에 영향을 미치는 유전자 집단이라 하여도 검정 방법에 따라 유의성을 진단할 수 없는 경우가 있다. 제 2장에서는 특이발현과 특이공발현을 각각 이용하여 유전자집단의 유의성을 검색하는 방법들 중 일부를 소개하고, 또 여러 방법을 한꺼번에 비교할 수 있는 새 방법을 제시하였다. 제 3장과 4장에서 실자료와 모의자료를 이용하여 방법들의 성능을 비교하였다.

2. 관심 유전자 집단의 유의성 진단 방법

g개의 유전자로 이루어진, 표본의 수가 각 n_1, n_2인 서로 다른 표현형을 갖는 두 군의 마이크로어레이 발현 자료가 있다고 하자. 이 때 k번째 군($k = 1, 2$)에 속한 l번째($l = 1, \ldots, n_k$) 표본의 i번째 유전자의 발현값을 x_{ilk}, i번째 유전자와 j번째 유전자 사이의 상관계수를 $r_{ij}^{(k)}$라 하자($i, j = 1, \ldots, g$). 전체 유전자 중 p개 유전자로 구성된 집단 P가 질병의 발생에 유의한 영향을 미치는지 알아보는 것이 유전자 집단 분석의 목적이다.

2.1. 특이발현에 기초한 유전자 집단 검정법

유전자 집단의 유의성 검정을 특이발현에 초점을 두어 검색한 방법에 대하여는 이미 여러 논문에서 이야기된 바 있다 (Lee 등, 2009; Maciejewski, 2014).

유전자 집단 분석의 시초격인 GSEA (Mootha 등, 2003; Subramanian 등, 2005)는 전체 유전자를 특이발현 정도에 따라 순위를 매기고, Kolmogorov-Smirnov 통계량을 사용하여 관심 집단에 속한 유전자들의 상대적 특이발현 순위를 분석하였다.

전체 유전자 자료로부터 특이발현 유전자 집단을 만든 후 이 집단과 관심 유전자 집단 사이의 독립성 여부를 Fisher의 정확성 검정이나 초기하분포를 이용한 검정 (Draghici 등, 2003; Khatri 등, 2004)을 실시하기도 하였는데, 이 방법들은 특이발현 유전자 집단에 대한 정의가 애매하고, 관심 집단에 속한 유전자들의 특이발현 유전자 집단 포함 여부만 중요할 뿐, 그들의 특이발현 정도는 고려되지 않는 단점이 있다.

Global test (Goeman 등, 2004)는 score 검정을 이용하여 질병과 관련있는 대사경로를 찾는 방법을 제시하였고, 환자의 생존율과 관계있는 대사경로를 찾는 방법으로까지 확장하였다 (Goeman 등, 2005).

대부분의 방법들이 비모수적 배경에서 유도되었으므로 p-value 계산을 위한 permutation이 수반되어야 하는 단점이 있고 이를 극복하기 위한 시도로 Kim과 Volsky (2005)는 관심 집단에 속한 유전자들의 수가 충분히 크고 서로 독립이라는 가정 아래 중심극한정리를 적용하는 PAGE를 제안하였다. 그러나 이는 과다발현 유전자와 발현억제 유전자들이 공존할 때 특이발현성이 상쇄되는 결함이 존재한다.

Dinu 등 (2007)은 단일 유전자 검정을 위해 t-통계량을 보정한 SAM (Tusher, 2001)의 방법을 확장하여 유전자 집단 검정을 위한 SAMGS를 선보였다. s_i는 i번째 유전자의 합동표준편차, s_0는 분산 보정을 위한 상수라 할 때 통계량은 아래의 D_{SAMGS}로 정의된다.

$$D_{SAMGS} = \sum_{i=1}^{p} \frac{(\bar{x}_{i.1} - \bar{x}_{i.2})^2}{s_i + s_0}, \quad \text{단 } \bar{x}_{i.k} = \frac{\sum_{l=1}^{n_k} x_{ilk}}{n_k}.$$

이 외에도 Efron과 Tibshirani (2007)의 Gene Set Analysis(GSA), Newton 등 (2007)의 random-set enrichment scoring 등이 있다.

2.2. 특이공발현에 기초한 유전자 집단 검정법

암 발생의 원인을 밝히기 위해서는 정상에서 암의 상태로 넘어가는 세포과정에서의 분자변화를 탐구하는 것이 중요하다. 세포과정에서 유전자들은 서로 상호작용을 하기 때문에 암의 발생과 관련된 유전자들 사이의 상호발현 양상은 정상상태에서의 발현 양상과 다르다. 그러므로 어떤 유전자 집단이 질병 발생에 유의한 영향을 미치는지 알아보기 위하여 유전자들 사이의 상호발현에 대해 관심을 갖게 되었다 (Lai 등, 2004).

Choi와 Kendziorski (2009)의 GSCA는 집단에 속한 모든 유전자들의 쌍이 서로 다른 표현형에서 나타내는 상관계수 차이를 측정한 통계량으로 아래의 D_{GSCA}를 제시하였는데 이 방법은 미미한 상관관계의 변화에 민감하고 집단에 속한 유전자 수에 따라 영향을 받는다는 문제점도 있다 (Jung과 Kim, 2014).

$$D_{GSCA} = \sqrt{\frac{1}{p(p-1)/2} \sum_{i=2}^{p} \sum_{j=1}^{i-1} \left(r_{ij}^{(1)} - r_{ij}^{(2)} \right)^2}.$$

Rahmatallah 등 (2014)는 집단에 속한 임의의 두 유전자 사이의 상관계수 대신 집단에 속하는 전체 유전자들의 상호관계 구조를 비교하는 Gene Sets Net Correlations Analysis(GSNCA)로 확장하였다. 집단에 속한 각 유전자들의 가중치는 다른 모든 유전자들과의 전반적인 상호관계에 비례한다고 가정하였고, k번째 군($k = 1, 2$)의 i번째 유전자의 가중치를 $w_i^{(k)}$라 할 때 Perron-Frobenius 정리 (Meyer, 2001)를 이용하여 $\boldsymbol{w}^{(k)} = (w_1^{(k)}, \ldots, w_p^{(k)})'$의 값을 구하였다. 서로 다른 두 군에서 구한 가중치를 이용하여 집단 P의 유의성을 판단하는 통계량 D_{GSNCA}를 사용하였다.

$$D_{GSNCA} = \sum_{i=1}^{p} \left| w_i^{(1)} - w_i^{(2)} \right|. \tag{2.1}$$

Tesson 등 (2010)은 집단에 속한 유전자뿐만 아니라 전체 유전자들의 상호 관계를 규명하는 Differentially Coexpressed gene modules(DiffCoEx)를 제시하였다.

이외에 유전자들 사이의 네트워크에 바탕을 두고 집단에 속한 유전자들의 표현형에 따른 종속 패턴 변화를 검색한 Evaluation of Dependency Differentiality(EDDY) (Jung과 Kim, 2014)와 Condition Specific sub Network identification(COSINE) (Ma 등, 2011) 방법도 있다.

2.3. 새로운 방법의 제시

특이발현 또는 공발현에 기초하여 질병의 발생에 유의한 영향을 미치는 유전자 집단을 찾는 방법은 많지만 이중 어느 방법이 최고인지 단정지을 수 없으며, 특이발현에 기초한 검색 방법으로 특이공발현을 보이는 집단은 찾아낼 수 없고 반대의 경우도 마찬가지이다. 그러므로 관심 유전자 집단의 유의성 검정을

Table 2.1. Algorithm for NEW

Input	서로 다른 표현형을 갖는 두 군의 마이크로어레이 자료
Output	$M_1, \ldots, M_m$을 적용했을 때의 p-value들과 이를 종합한 NEW의 p-value
Step 1.	방법 $M_1, \ldots, M_m$를 적용하여 관심 유전자군의 통계량값 $D_{M_1}, \ldots, D_{M_m}$을 구한다.
Step 2.	sample permutation을 b회 반복하여 관심 유전자군의 $D^{(i)}_{M_1}, \ldots, D^{(i)}_{M_m}$ $(i = 1, \ldots, b)$을 구한다. 방법 $M_1, \ldots, M_m$을 적용했을 때의 각 p-value들을 구할 수 있다.
Step 3.	각 방법별로 원자료를 이용한 D_{M_j}와 sample permutation을 실시한 $D^{(1)}_{M_j}, \ldots, D^{(b)}_{M_j}$를 평균 0, 분산 1이 되도록 표준화하여 $D^S_{M_j}$와 $D^{S(1)}_{M_j}, \ldots, D^{S(b)}_{M_j}$ $(j = 1, \ldots, m)$를 구한다.
Step 4.	$D_{NEW} = \max(D^S_{M_1}, \ldots, D^S_{M_m})$와 $D^{(i)}_{NEW} = \max(D^{S(i)}_{M_1}, \ldots, D^{S(i)}_{M_m})$ $(i = 1, \ldots, b)$를 계산, 방법 NEW의 p-value를 구한다.

위하여 서로 다른 배경에서 유도된 여러 방법을 함께 사용하게 되지만 각기 다른 결과가 나왔을 때 이것들을 서로 비교하거나 하나의 결론으로 도출하기에는 어려움이 있다. 본 연구에서는 어떤 유전자 집단이 질병의 발생에 유의한 영향을 미치는지 판단하기 위하여 여러 가지 검정법을 사용하고 여기서 하나의 결론을 찾는 새로운 방법, NEW를 제시하였다 (Table 2.1).

m가지 서로 다른 방법 $M_1, \ldots, M_m$을 사용하여 유전자 집단의 유의성을 검정하려 할 때 대부분의 통계량들은 거의 비모수적인 통계량들이므로 p-value를 계산하기 위해서 많은 반복의 permutation을 실시해야 한다. 그러므로 각 방법별로 원자료에서 구한 통계량과 permutation을 통하여 얻어진 통계량값들을 평균 0, 분산 1이 되도록 모두 표준화 변환을 할 수 있다. 이 때 D_{NEW}는 m개의 표준화된 통계량 중 최대값이라 정의하면 추가 permutation 없이도 p-value 계산이 가능하고 사용된 방법 중 관심 유전자군의 유의성을 찾아낸 것이 있다면 이것이 새 통계량 NEW에 반영이 될 것이다.

NEW를 이용한 유의성 검정을 하는데 몇 가지 방법을 사용할지, 또 어떤 방법을 사용할지는 분석자의 연구 환경과 상황에 따라 마음대로 정할 수 있으며 발현의 차이를 찾는 방법과 공발현의 차이를 찾는 서로 다른 성격의 방법들을 함께 사용하는 것이 바람직하겠다. NEW를 사용하면 여러 분석 결과로부터 하나의 결론을 도출할 수 있고, 서로 다른 방법에 의해 유의성이 검증된 두 유전자 집단 사이의 우열을 가릴 수 있는 큰 장점이 있다.

3장의 실제 자료 분석과 4장의 모의실험에서는 특이발현 통계량과 특이공발현 통계량들 중 대표적인 D_{SAMGS}와 D_{GSNCA}의 두 가지를 사용하여 D_{NEW}를 계산하였다.

3. 실제 자료 분석을 이용한 방법 비교

새로 제시한 방법 NEW와 GSEA, SAMGS, GSCA, GSNCA를 비교하기 위하여 이들을 처음 다룬 논문들 (Subramanian 등, 2005; Dinu 등, 2007; Rahmatallah 등, 2014)에서 공통적으로 다룬 33예의 돌연변이군과 17예의 정상군의 p53 자료(http://software.broadinstitute.org/gsea/datasets.jsp)를 분석하였다.

NCI-60 암 세포 라인에서 다양한 자극 신호에 따라 유전자 발현을 조절하는 전사인자인 p53의 표적을 규명하기 위한 이 자료는 서로 다른 10,010개 유전자의 마이크로어레이자료이며 Molecular Signatures Database(MSigDB)에서 다운받은 905개의 pathway 정보를 사용하여 유전자 집단 분석을 실시하였다.

각 pathway의 통계치를 계산하고 10,000번의 permutation을 실시하여 p-value를 구한 후 돌연변이군과 정상군 사이에 $\alpha = 0.05$를 기준으로 유의한 차이를 보이는 pathway를 검색한 결과, Table 3.1을 얻었다.

Table 3.1. Number of significant pathways

	Number(diagonal and upper diagonal) and percentage(lower diagonal) of significant pathways detected using both methods in column and row				
	GSEA	SAMGS	GSCA	GSNCA	NEW
GSEA	52	19	25	5	16
SAMGS	**36.5**	128	27	8	80
GSCA	**48.1**	**21.1**	195	39	31
GSNCA	**9.6**	**6.3**	**20.0**	89	37
NEW	**30.8**	**62.5**	**15.9**	**41.6**	111

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

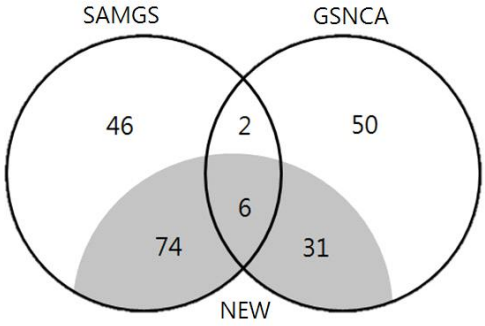


Figure 3.1. Venn diagram of significant pathways detected by SAMGS, GSNCA and NEW (SAMGS = significance analysis of microarray-gene set, GSNCA = gene sets net correlations analysis).

GSEA로는 52개의 pathway가, SAMGS로는 128개 pathway가 유의하다고 검색되었다. 두 가지 모두 특이발현에 기초한 방법이지만 GSEA는 집단에 속한 유전자들의 돌연변이군과 정상군 사이의 특이발현 방향성이 일관되었는지를 보았고, SAMGS는 특이발현의 정도에 초점을 둔 차이가 있기 때문에 공통적으로 유의하다는 결과가 나온 pathway는 19개 뿐이었다. Dinu 등 (2007)의 Table 4에서도 이미 두 방법의 분석 결과 사이에 큰 차이가 있음을 논하였다.

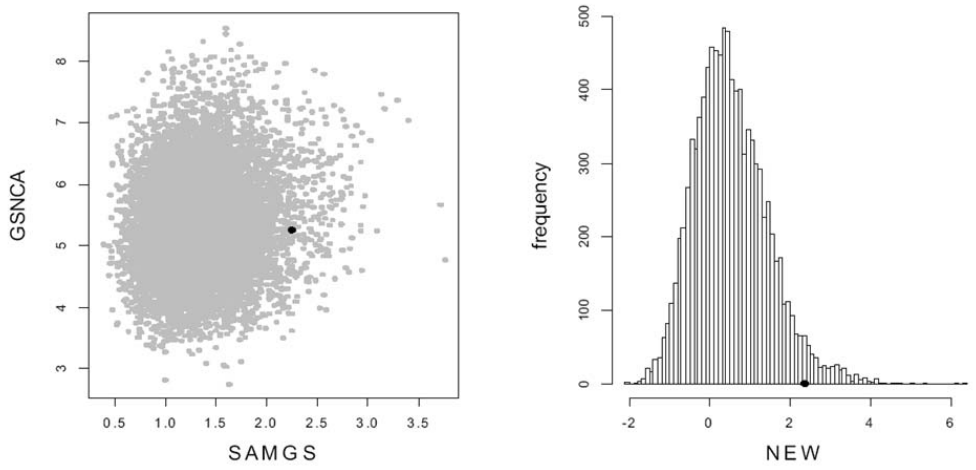
GSCA와 GSNCA의 두 방법에서 모두 유의한 차이를 보인 pathway의 수는 39개였고, 각각에서만 유의한 차이를 보인 pathway의 수는 156개와 50개였다. Rahmatallah 등 (2014)의 지적과 같이 본 연구에서도 GSCA는 신진대사(metabolism)와, 그리고 GSNCA는 신호기전(signaling)과 관계된 pathway들이 주로 검색되었다. 두 방법 모두 유전자간의 상관관계를 중심으로 돌연변이군과 정상군 사이에 유의한 차이를 보이는 pathway를 찾는다는 목적은 같지만, 실제 GSCA의 귀무가설은 집단에 속한 모든 유전자 쌍의 상관계수 평균이 0이라는 것인 반면에 GSNCA는 유전자들 사이의 상관관계에 기인한 가중치가 돌연변이군과 정상군에서 서로 같다는 것이기 때문에 결과에는 차이가 있을 수 밖에 없다.

Figure 3.1은 SAMGS와 GSNCA, 그리고 이 두 방법을 결합하여 만든 NEW가 찾아낸 유의한 pathway 사이의 포함관계이다. SAMGS나 GSNCA를 사용할 때 비하여 NEW는 일부만 찾아낸 것으로 보일 수도 있지만, 실제로 NEW로는 찾을 수 없던 pathway 대부분은 유의성의 순위가 떨어지는 것들로 다중검정을 실시하면 유의하지 않다는 결론이 나오는 것들이다. 또한 SAMGS만 사용하면 찾아낼 수 없는 31개, GSNCA만 사용하면 찾아낼 수 없는 74개의 pathway 들이 NEW를 사용함으로서 찾아낸다는 큰 장점이 있다. 그러나 SAMGS와 GSNCA에서는 찾아냈던 2개의 유의한 pathway가 NEW에서는 유의

Table 3.2. p-values of 2 pathways

	GSEA	SAMGS	GSCA	GSNCA	NEW
REACTOME SIGNALING BY ERBB4	0.2360	0.0360	0.0287	0.0273	0.0602
BIOCARTA FAS PATHWAY	0.3426	0.0475	0.0995	0.0479	0.0776

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.



(a) Scatter plot of SAMGS and GSNCA (b) Histogram of NEW

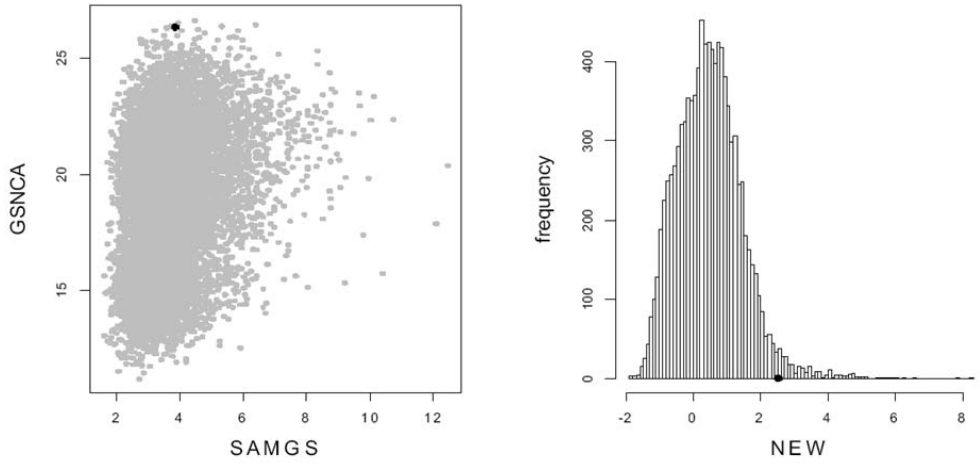
Figure 3.2. ST_INTERLEUKIN_4 PATHWAY (SAMGS = significance analysis of microarray-gene set, GSNCA = gene sets net correlations analysis).

하지 않다는 결과가 나왔다. 이것은 SAMGS와 GSNCA가 서로 비슷한 정도의 유의성을 보이는 상황에서는 두 통계량 중의 큰 값을 취하는 NEW의 유의성은 떨어지기 때문이다 (Table 3.2).

Figure 3.2(a)는 24개 유전자로 구성이 된 'ST_INTERLEUKIN_4_PATHWAY'의 SAMGS와 GSNCA 통계치의 산포도이고 (b)는 NEW의 분포도이다. 검은 점은 실제 자료의 통계치이고 회색 점은 10,000번 permutation하여 얻은 통계치들이다. 이 pathway에 SAMGS를 적용하면 유의한 집단이라는 결론($p = 0.0004$)을 얻지만 GSNCA는 전혀 유의하지 않다($p = 0.5308$)는 상반된 결론이 나옴은 (a)의 산포도를 통하여 알 수 있다. 그러나 NEW를 적용하면 이 집단이 유의하다($p = 0.0001$)는 결론을 얻게 된다. Figure 3.3은 97개 유전자로 구성이 된 'KEGG_LYSOSOME'의 결과인데 'ST_INTERLEUKIN_4_PATHWAY'와는 정반대로 SAMGS로는 유의하지 않지만($p = 0.3701$) GSNCA 결과는 유의($p = 0.0005$)하며 NEW를 적용하였을 때는 $p = 0.0261$을 얻었다.

4. 모의실험을 이용한 방법 비교

p개의 유전자로 구성된 집단이 질병의 발생에 유의한 영향을 미치는지 알아보기 위하여 표본의 수가 각 n개인 비교군과 처리군의 모의자료를 생성하였다. 비교군에 속한 l번째 ($l = 1, \ldots, n$) 표본의 유전자 발현값 분포는 $\boldsymbol{X_l^c} = (x_{1l}^c, \ldots, x_{il}^c, \ldots, x_{pl}^c)' \sim N(\boldsymbol{0}, \boldsymbol{I_p})$, 처리군의 l번째 표본은 $\boldsymbol{X_l^t} = (x_{1l}^t, \ldots, x_{il}^t, \ldots, x_{pl}^t) \sim N(\boldsymbol{\mu_p}, \boldsymbol{\Sigma_p})$를 가정하였다. 처리군의 분포를 5가지로 나누어 앞에서 논한 SAM-GS, GSCA, GSNCA와 NEW의 유의수준과 검정력을 비교하기 위해 각 유형별로 유전자 집



(a) Scatter plot of SAMGS and GSNCA (b) Histogram of NEW

Figure 3.3. KEGG_LYSOSOME (SAMGS = significance analysis of microarray-gene set, GSNCA = gene sets net correlations analysis).

Table 4.1. Size comparison ($\alpha = 0.05$) (type I)

n	$p = 20$				$p = 40$				$p = 60$			
	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW
15	0.052	0.066	0.052	0.046	0.051	0.039	0.047	0.055	0.042	0.060	0.050	0.045
25	0.058	0.053	0.037	0.047	0.063	0.045	0.044	0.063	0.045	0.050	0.053	0.049
40	0.037	0.042	0.067	0.045	0.053	0.044	0.039	0.053	0.042	0.043	0.062	0.047
60	0.039	0.053	0.051	0.050	0.054	0.041	0.058	0.056	0.037	0.045	0.052	0.040

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

단을 생성하여 500번의 permutation을 실시하여 p-value를 구하고, 이 작업을 1,000번 반복하였다. GSEA는 집단의 유의성 판단을 위해 전체 유전자 자료가 필요하기 때문에 모의실험에서 제외하였다.

유형 I. 유의수준을 알아보기 위한 모의실험

처리군 표본의 분포도 비교군과 동일하게 모든 유전자가 서로 독립인 $\boldsymbol{X_l^t} \sim N(\mathbf{0}, \boldsymbol{I_p})$ $(l = 1, \ldots, n)$를 가정하였고 표본의 수 n은 15, 25, 40, 60으로, 집단에 속한 유전자 수 p는 20, 40, 60으로 변화시켜 Table 4.1의 결과를 얻었다.

유의수준 $\alpha = 0.05$의 95% 신뢰구간이 (0.0365, 0.0635)인 것을 고려하였을 때 집단에 속한 유전자 수가 작은 경우의 GSCA와 GSNCA만 제외하고 안정적으로 유의수준을 제어하고 있음을 볼 수 있다.

유형 II. 검정력을 비교하기 위한 모의실험

처리군의 $\boldsymbol{\mu_p} = (\mu_1, \ldots, \mu_p)'$와 $\boldsymbol{\Sigma_p} = (\sigma_{ij})_{i=1}^{p}{}_{j=1}^{p}$의 값에 여러 가지 변형을 주어 집단에 속한 유전자들 중 특이발현을 보이는 정도와 유전자들 사이의 상관관계 형태를 표현해 보았다.

각 군의 표본의 수 n은 15와 40, 집단에 속한 유전자 수 p는 20과 60, 집단에 속한 유전자 중 특이발현 또는 특이공발현을 보이는 비율 γ는 0.3과 0.5의 값을 주어 유의수준 $\alpha = 0.05$일 때 검정력을 비교했다.

Table 4.2. Power comparison when some genes are only DE, no DC* (type II-1)

n	p	γ	$\mu=0.3$				$\mu=0.5$				$\mu=1.0$			
			SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW
15	20	0.3	0.157	0.064	0.047	0.125	0.452	0.052	0.051	0.378	0.994	0.047	0.033	0.985
		0.5	0.262	0.052	0.043	0.201	0.707	0.051	0.044	0.628	1	0.059	0.020	1
	60	0.3	0.270	0.049	0.035	0.199	0.785	0.052	0.049	0.725	1	0.054	0.006	1
		0.5	0.464	0.053	0.049	0.381	0.985	0.053	0.037	0.969	1	0.113	0	1
40	20	0.3	0.415	0.036	0.036	0.345	0.940	0.045	0.042	0.915	1	0.058	0.032	1
		0.5	0.698	0.051	0.052	0.621	0.999	0.059	0.037	0.998	1	0.064	0.016	1
	60	0.3	0.765	0.055	0.039	0.697	0.999	0.041	0.035	0.999	1	0.059	0.001	1
		0.5	0.963	0.052	0.041	0.952	1	0.046	0.026	1	1	0.108	0	1

∗ DE: differentially expressed, DC: differentially correlated.
GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

II-1) 일부의 유전자가 특이발현만 하는 경우

$$\mu_i = \begin{cases} \mu, & \forall\ i \le \gamma p, \\ 0, & \text{o.w.}, \end{cases} \qquad \sigma_{ij} = \begin{cases} 1, & i = j, \\ 0, & \text{o.w.}, \end{cases} \quad (i = 1, \ldots, p,\ j = 1 \ldots, p).$$

집단에 속한 p개 유전자 중 γ 만큼만 특이발현하고 이들의 평균 발현값은 μ라 하였다. μ는 0.3, 0.5, 1.0로 변화시키며 Table 4.2의 결과를 얻었다.

네 가지 방법 중 SAMGS의 검정력이 제일 크고 그 다음이 NEW인 것은 당연하고, μ가 커질수록, 표본의 수와 특이발현 유전자수가 커질수록 검정력이 커지는 것을 볼 수 있다. 반면 GSCA와 GSNCA는 이런 현상과 무관하며 심지어 $\mu = 1$인 경우 GSNCA의 검정력은 0에 가까워짐을 볼 수 있다.

II-2) 일부의 유전자가 특이공발현만 하는 경우

$$\mu_i = 0, \qquad \sigma_{ij} = \begin{cases} 1, & i = j, \\ r, & i \ne j,\ \forall\ i, j \le \gamma p, \\ 0, & \text{o.w.}, \end{cases} \quad (i = 1, \ldots, p,\ j = 1 \ldots, p).$$

집단에 속한 p개 유전자 중 γ 만큼만 서로 상관관계가 있다 가정하였고, γp개 유전자 중 임의의 두 유전자 간의 공분산은 r이라 가정하였다. r을 0.3, 0.45, 0.6의 값으로 변화시켜 Table 4.3의 결과를 얻었다.

특이발현에 기초한 SAMGS는 유의수준에 머무는 검정력을 보였고 GSCA와 GSNCA는 서로 우열을 가리기 힘들지만, 표본의 수나 공발현하는 유전자 수가 커질수록 검정력이 커졌고 NEW와의 검정력 차이도 커짐을 볼 수 있다.

II-3) 일부의 유전자가 특이발현인 동시에 특이공발현하는 경우

$$\mu_i = \begin{cases} \mu, & \forall\ i \le \gamma p, \\ 0, & \text{o.w.}, \end{cases} \qquad \sigma_{ij} = \begin{cases} 1, & i = j, \\ r, & i \ne j,\ \forall\ i, j \le \gamma p, \\ 0, & \text{o.w.}, \end{cases} \quad (i = 1, \ldots, p,\ j = 1 \ldots, p).$$

집단에 속한 유전자 중 γp개 유전자는 특이발현을 하는 동시에 특이공발현 유전자를 가정하였다. 처리군에서의 특이발현 유전자의 평균 발현값은 μ, 특이공발현하는 정도는 r이라 가정하며 μ는 0.3, 0.5,

Table 4.3. Power comparison when some genes are only DC, no DE (type II-2)

n	p	γ	$r = 0.3$				$r = 0.45$				$r = 0.6$			
			SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW
15	20	0.3	0.060	0.122	0.101	0.085	0.062	0.252	0.245	0.186	0.040	0.457	0.563	0.432
		0.5	0.058	0.296	0.172	0.127	0.052	0.638	0.478	0.351	0.039	0.902	0.785	0.627
	60	0.3	0.052	0.348	0.413	0.322	0.051	0.730	0.833	0.786	0.049	0.942	0.987	0.984
		0.5	0.064	0.783	0.752	0.684	0.049	0.971	0.981	0.957	0.053	0.996	0.997	0.996
40	20	0.3	0.050	0.320	0.390	0.293	0.068	0.748	0.894	0.837	0.046	0.974	0.998	0.995
		0.5	0.050	0.873	0.726	0.605	0.053	0.999	0.991	0.969	0.056	1	0.999	0.998
	60	0.3	0.046	0.899	0.962	0.954	0.038	1	1	1	0.043	1	1	1
		0.5	0.048	1	0.999	0.998	0.057	1	1	1	0.055	1	1	1

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

Table 4.4. Power comparison when some genes are both DE and DC (type II-3)

n	p	γ	r	$\mu = 0.3$				$\mu = 0.5$				$\mu = 1$			
				SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW
15	20	0.3	0.3	0.060	0.122	0.101	0.085	0.062	0.252	0.245	0.186	0.040	0.457	0.563	0.432
			0.6	0.058	0.296	0.172	0.127	0.052	0.638	0.478	0.351	0.039	0.902	0.785	0.627
		0.5	0.3	0.052	0.348	0.413	0.322	0.051	0.730	0.833	0.786	0.049	0.942	0.987	0.984
			0.6	0.064	0.783	0.752	0.684	0.049	0.971	0.981	0.957	0.053	0.996	0.997	0.996
	60	0.3	0.3	0.050	0.320	0.390	0.293	0.068	0.748	0.894	0.837	0.046	0.974	0.998	0.995
			0.6	0.050	0.873	0.726	0.605	0.053	0.999	0.991	0.969	0.056	1	0.999	0.998
		0.5	0.3	0.046	0.899	0.962	0.954	0.038	1	1	1	0.043	1	1	1
			0.6	0.048	1	0.999	0.998	0.057	1	1	1	0.055	1	1	1
40	20	0.3	0.3	0.060	0.122	0.101	0.085	0.062	0.252	0.245	0.186	0.040	0.457	0.563	0.432
			0.6	0.058	0.296	0.172	0.127	0.052	0.638	0.478	0.351	0.039	0.902	0.785	0.627
		0.5	0.3	0.052	0.348	0.413	0.322	0.051	0.730	0.833	0.786	0.049	0.942	0.987	0.984
			0.6	0.064	0.783	0.752	0.684	0.049	0.971	0.981	0.957	0.053	0.996	0.997	0.996
	60	0.3	0.3	0.050	0.320	0.390	0.293	0.068	0.748	0.894	0.837	0.046	0.974	0.998	0.995
			0.6	0.050	0.873	0.726	0.605	0.053	0.999	0.991	0.969	0.056	1	0.999	0.998
		0.5	0.3	0.046	0.899	0.962	0.954	0.038	1	1	1	0.043	1	1	1
			0.6	0.048	1	0.999	0.998	0.057	1	1	1	0.055	1	1	1

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

1.0로, r은 0.3과 0.6 의 값으로 변화시켰다. Table 4.4의 결과를 보면 모든 방법들은 표본의 수, 특이발현과 특이공발현을 보이는 유전자의 수와 정도가 커질수록 검정력이 커지는 것을 볼 수 있고, 이 유형에서 NEW의 검정력이 다른 방법들에 비해 훨씬 우수함을 볼 수 있다.

II-4) 일부의 유전자는 특이발현을, 다른 일부의 유전자는 특이공발현하는 경우

$$\mu_i = \begin{cases} \mu, & \forall\, i \le \dfrac{\gamma p}{2}, \\ 0, & \text{o.w.}, \end{cases} \qquad \sigma_{ij} = \begin{cases} 1, & i = j, \\ r, & i \ne j,\ \dfrac{\gamma p}{2} \le i, j < \gamma p, \\ 0, & \text{o.w.}, \end{cases} \quad (i = 1, \ldots, p,\ j = 1 \ldots, p).$$

집단에 속한 유전자 중 $\gamma p/2$개 유전자는 특이발현만 하고, 또 다른 $\gamma p/2$개 유전자는 특이공발현만 한다고 가정하였다. 처리군에서 특이발현 유전자의 평균 발현값은 μ, 특이공발현하는 정도는 r이라 가정하였다. μ는 0.3, 0.5, 1.0로, r은 0.3, 0.6의 값으로 변화시켰다. Table 4.5의 결과를 보면 μ와 r의

Table 4.5. Power comparison when some genes are DE or DC, not both (type II-4)

n	p	γ	r	$\mu = 0.3$				$\mu = 0.5$				$\mu = 1.0$			
				SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW	SAMGS	GSCA	GSNCA	NEW
15	20	0.3	0.3	0.099	0.051	0.053	0.078	0.246	0.054	0.049	0.187	0.798	0.062	0.051	0.743
			0.6	0.091	0.099	0.088	0.105	0.209	0.107	0.095	0.188	0.795	0.107	0.078	0.725
		0.5	0.3	0.146	0.090	0.072	0.122	0.310	0.106	0.076	0.277	0.974	0.105	0.083	0.959
			0.6	0.119	0.293	0.381	0.324	0.346	0.305	0.372	0.414	0.972	0.297	0.351	0.968
	60	0.3	0.3	0.151	0.094	0.097	0.144	0.390	0.107	0.103	0.336	0.995	0.096	0.076	0.988
			0.6	0.104	0.348	0.595	0.515	0.366	0.375	0.617	0.662	0.995	0.348	0.520	0.993
		0.5	0.3	0.210	0.250	0.304	0.339	0.665	0.255	0.274	0.670	1	0.242	0.178	1
			0.6	0.166	0.867	0.964	0.954	0.594	0.851	0.959	0.985	1	0.840	0.916	1
40	20	0.3	0.3	0.202	0.075	0.077	0.173	0.580	0.083	0.061	0.518	1	0.098	0.063	1
			0.6	0.203	0.278	0.359	0.361	0.592	0.246	0.371	0.621	1	0.249	0.333	1
		0.5	0.3	0.360	0.234	0.263	0.398	0.883	0.203	0.225	0.848	1	0.224	0.224	1
			0.6	0.338	0.874	0.976	0.974	0.858	0.862	0.981	0.990	1	0.870	0.973	1
	60	0.3	0.3	0.375	0.272	0.380	0.498	0.938	0.279	0.363	0.928	1	0.265	0.285	1
			0.6	0.373	0.934	0.999	0.998	0.942	0.943	0.999	0.999	1	0.950	1	1
		0.5	0.3	0.654	0.774	0.899	0.942	1	0.763	0.902	1	1	0.766	0.756	1
			0.6	0.562	1	1	1	0.997	1	1	1	1	1	1	1

GSEA = gene set enrichment analysis, SAMGS = significance analysis of microarray-gene set, GSCA = gene set coexpression analysis, GSNCA = gene sets net correlations analysis.

값의 크기에 따라 SAMGS와 GSNCA 사이의 검정력 우열이 달라졌고 전반적으로 표본의 수가 작을 때는 SAMGS와 GSNCA가 NEW보다 강세를 보였으나 표본의 수가 많아지면 NEW의 검정력이 SAMGS나 GSNCA보다 커짐을 볼 수 있다.

5. 결론

관심 유전자 집단이 표현형에 따라 유의한 차이가 있는지 알아보려 할 때, '차이'를 어떻게 정의하는가에 따라 검정통계량을 유도하는 배경이 달라지는 것은 당연하다. 유전자 집단의 유의성을 검정하는 방법이 많이 있지만 사용하는 방법에 따라 완전히 서로 다른 결과가 나오는 것도 이러한 이유 때문이다.

본 연구에서는 표현형에 따른 차이로 평균발현값에 차이가 있는 경우와 유전자들 사이의 공발현 구조에 차이가 있는 경우로 나누어 생각하였고 그에 따른 검정 방법들을 살펴 보았다. 그러나 어떤 방법이 유의한 유전자군을 찾는데 제일 좋다고 단정할 수 없으므로 결국 여러 방법을 사용하여 각각 다른 결과들을 얻게 된다. 새로 제시한 검정방법 NEW는 특이발현과 특이공발현의 차이를 찾아내는 여러 통계량들을 표준화시킨 값 중에서 최대값을 이용하여 유전자집단의 유의성을 검색하는 방법이다. NEW에서 사용하는 방법의 수와 어떤 방법을 사용할지는 분석자가 결정할 수 있고 사용 방법에 따른 여러 가지 분석결과를 하나로 정리한다는 것이 큰 장점이다.

특이발현과 특이공발현의 차이를 찾아내는 통계량으로 각각 대표적인 SAMGS와 GSNCA를 사용하여 실제 자료 분석과 모의실험을 실시하였고, NEW를 이용하여 두 방법을 통합한 p-value 계산이 가능하고 검정력도 좋음을 확인하였다. 다만, p53의 자료 분석에서 비슷한 정도의 특이발현과 특이공발현이 공존하는 경우, NEW의 검정력이 떨어짐을 볼 수 있었다.

References

Choi, Y. and Kendziorski, C. (2009). Statistical methods for gene set co-expression analysis, *Bioinformatics*, **25**, 2780–2786.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2007). Improving gene set analysis of microarray data by SAM-GS, *BMC Bioinformatics*, **8**, 242.

Draghici, 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, *Annals of Applied Statistics*, **1**, 107–129.

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

Goeman, 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.

Jung, S. and Kim, S. (2014). EDDY: a novel statistical gene set test method to detect differential genetic dependencies, *Nucleic Acids Research*, **42**, e60.

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, B. S., Jang, J. S., Kim, S. C., and Lim, J. (2009). A report on the inter-gene correlations in cDNA microarray data sets, *The Korean Journal of Applied Statistics*, **22**, 617–626.

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

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.

Lai, Y., Wu, B., Chen, L., Zhao, H. (2004). A statistical method for identifying differential gene-gene co-expression patterns, *Bioinformatics*, **20**, 3146–3155.

Lee, S. H., Lee, S. K., and Lee, K. H. (2009). Developing a parametric method for testing the significance of gene sets in microarray data analysis, *Communications for Statistical Applications and Methods*, 397–408.

Ma, H., Schadt, E. E., Kaplan, L. M., and Zhao, H. (2011). COSINE: condition-specific sub-network identification using a global optimization method, *Bioinformatics*, **27**, 1290–1298.

Maciejewski, H. (2014). Gene set analysis methods: statistical models and methodological differences, *Briefings in Bioinformatics*, **15**, 504–518.

Meyer, C. (2001). *Matrix Analysis and Applied Linear Algebra*, Society for industrial and applied mathematics (SIAM), Philadelphia.

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-1-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. (2007). Random set methods identify distinct aspects of the enrichment signal in gene-set analysis, *Annals of Applied Statistics*, **1**, 85–106.

Qui, X., Klebanov, L., and Yakovlev, A. (2005). 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.

Rahmatallah, Y., Emmert-Streib, F. and Glazko, G. (2014). Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets, *Bioinformatics*, **30**, 360–368.

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, In *Proceedings of the National Academy of Sciences*, **102**, 15545–15550.

Tesson, B. M., Breitling, R., and Jansen, R. C. (2010). DiffCoEx: a simple and sensitive method to find differentially coexpressed gene modules, *BMC Bioinformatics*, **11**, 497.

Tusher, V. G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, In *Proceedings of the National Academy of Sciences*, **98**, 5116–5121.

특이발현과 특이공발현을 고려한 유의한 유전자 집단 탐색

이선호[a,1]

[a]세종대학교 수학통계학부

(2015년 12월 22일 접수, 2016년 2월 12일 수정, 2016년 2월 29일 채택)

요 약

서로 상관있는 유전자들의 발현조절이 질병이나 종양의 발생에 영향을 미치기 때문에 단일유전자 분석 대신 공통의 생물학적 요소를 지닌 유전자 집단 분석이 각광을 받게 되었고 생물학적으로 좀더 설명하기 쉬운 결과를 얻게 되었다. 표현형에 따라 유의한 차이를 보이는 유전자 집단을 찾는 여러 방법들이 있지만, 대부분의 방법들이 집단에 속한 유전자들의 표현형에 따른 발현의 차이를 탐색하거나 유전자들 사이의 공발현 구조가 다른지 탐색하는 것이다. 본 연구에서는 특이발현과 특이공발현의 차이를 모두 고려하는 탐색방법을 제시하였고 p53이란 유전자 자료와 모의 자료를 이용하여 제시한 방법의 성능을 알아 보았다.

주요용어: 마이크로어레이 자료, 유전자 집단 분석, 특이발현, 특이공발현

[1](05006) 서울시 광진구 능동로 209, 세종대학교 수학통계학부. E-mail: leesh@sejong.ac.kr