유전자 발현 메트릭에 기반한 모수적 방식의 유의 유전자 집합 검출 비교 연구

(A Comparative Study of Parametric Methods for Significant Gene Set Identification Depending on Various Expression Metrics)

김 재 영 † 신 미 영 ††
(Jaeyoung Kim) (Miyoung Shin)

요 약 최근 마이크로어레이 데이터를 기반으로 두 개의 샘플 그룹간에 유의한 발현 차이를 나타내는 생물학적 기능 그룹을 검출하기 위한 유전자 집합 분석(gene set analysis) 연구가 많은 주목을 받고 있다. 기존의 유의 유전자 검출 연구와는 달리, 유전자 집합 분석 연구는 유의한 유전자 집합과 이들의 기능적 특징을 함께 검출할 수 있다는 장점이 있다. 이러한 이유로 최근에는 PAGE, GSEA 등과 같은 다양한 통계적 방식의 유전자 집합 분석 방법들이 소개되고 있다. 특히, PAGE의 경우 두 샘플 그룹간의 유전자 발현 차이를 나타내는 스코어의 분포가 정규 분포임을 가정하는 모수적 접근 방식을 취하고 있다. 이러한 방법은 GSEA 등과 같은 비모수적 방식에 비해 계산량이 적고 성능이 비교적 우수한 장점이 있다. 하지만, PAGE에서 유전자 발현 차이를 정량화하기 위한 메트릭으로 사용하고 있는 AD(average difference)의 경우, 두 그룹간에 절대적 평균 발현 차이만을 고려하기 때문에 실제 유전자의 발현값 크기나 분산의 크기에 따른 상대적 중요성을 반영하지 못하는 문제가 있다. 본 논문에서는 이를 보완하기 위해 실제 유전자의 발현값 크기나 그룹 내 샘플들의 분산 정보 등을 스코어 계산에 함께 반영하는 WAD(weighted average difference), FC(Fisher's criterion), 그리고 Abs_SNR(Absolute value of signal-to-noise ratio)을 모수적 방식의 유전자 집합 분석에 적용하고 이에 따른 유의 유전자 집합 검출 결과를 실험을 통해 비교 분석하였다.

키워드 : 마이크로어레이, 유전자 집합 분석, 모수적 방식

Abstract Recently lots of attention has been paid to gene set analysis for identifying differentially expressed gene-sets between two sample groups. Unlike earlier approaches, the gene set analysis enables us to find significant gene-sets along with their functional characteristics. For this reason, various novel approaches have been suggested lately for gene set analysis. As one of such, PAGE is a parametric approach that employs average difference (AD) as an expression metric to quantify expression differences between two sample groups and assumes that the distribution of gene scores is normal. This approach is preferred to non-parametric approach because of more effective performance. However, the metric AD does not reflect either gene expression intensities or variances over samples in calculating gene scores. Thus, in this paper, we investigate the usefulness of several other expression metrics for parametric gene-set analysis, which consider actual expression intensities

· 이 논문은 2009년 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구임(No. 20090058662)

† 학생회원 : 경북대학교 전자전기컴퓨터학부
widebrowboy@gmail.com
†† 종신회원 : 경북대학교 전자전기컴퓨터학부 교수
shinmy@knu.ac.kr
(Corresponding author임)

논문접수 : 2009년 4월 28일
심사완료 : 2009년 11월 1일

정보과학회논문지 : 소프트웨어 및 응용 제37권 제1호(2010.1)

of genes or their expression variances over samples. For this purpose, we examined three expression metrics, WAD (weighted average difference), FC (Fisher's criterion), and Abs_SNR (Absolute value of signal-to-noise ratio) for parametric gene set analysis and evaluated their experimental results.

Key words : Microarray, gene set analysis, parametric methods

1. 서 론

최근 생명과학 및 의학 분야에서는 DNA 칩과 같은 대용량 바이오 실험 기술의 발달로 질병 진단이나 신약 개발 등에 관한 새로운 패러다임의 연구들이 진행되고 있다. 특히 수십만 개의 유전자들의 발현 양상을 동시에 관찰하는 것을 가능하게 하는 마이크로어레이 기술의 발달은 주어진 조건에서 전체 지놈 수준의 반응을 snap-shot의 형태로 한 번에 모니터링할 수 있게 함으로써 특정 질병과 관련된 마커 유전자 발굴이나 중요 유전자의 기능적 특징 및 상호 작용 규명 등의 연구에 많은 기술적 발전을 이루어왔다.

최근들어 마이크로어레이를 기반으로 하는 연구 중 많은 주목을 받고 있는 분야 중의 하나는 유전자 집합 분석(Gene set analysis, GSA)[1]에 관한 연구이다. 이는 두 개의 샘플 그룹(실험군과 대조군, 정상세포와 암세포)을 가지는 마이크로어레이 실험 데이터와 생물학적 지식 데이터 베이스의 유전자 기능 정보, 패스웨이 정보 등을 이용하여 두 그룹 간에 유의한 발현 차이를 나타내는 유전자 집합과 이들의 기능적 특징을 함께 검출하는 방법이다[2]. 이 방법은 두 샘플 그룹 간에 유의한 발현 차이를 보이는 소수의 개별 유전자를 검출하고 이에 대한 생물학적 의미 해석이 후처리 형태로 이루어졌던 기존의 방식과는 달리, 검출 결과에 대한 생물학적 의미 해석 과정이 유의 유전자 집합을 검출하는 과정에 내포되어 있다. 특히, 기존의 개별적 유의 유전자 검출에 관한 연구는 전체 유전자를 발현 차이에 따라 정렬하고 정해진 임계값 이상의 발현 차이를 보이는 개별 유전자들을 유의 유전자로 선정하였기 때문에 사용된 임계값에 의해 그 결과가 달라질 뿐만 아니라 개별 유전자의 발현 차이가 크지는 않더라도 이러한 유전자들을 포함하는 특정 기능에 관여하는 유전자 그룹이 발현 차이에 있어 통계적 유의성을 나타내는 경우에 이를 찾아내지 못하는 문제가 있다. 이러한 이유로 최근에는 다양한 통계적 방법을 활용하여 두 샘플 그룹 간에 유의한 발현 차이를 보이는 '유전자군(즉, 유전자 집합)'을 분석하는 연구가 많은 주목을 받고 있다[1]. 특히 이러한 연구에서는 마이크로어레이 실험 데이터뿐만 아니라 Gene Ontology[2], Pathway DB[3] 등과 같은 알려진 여러 생물학적 리소스들을 함께 활용하고 있다.

유전자 집합 분석(Gene set analysis, GSA)은 일반적으로 경쟁적(competitive) 방식과 자기 완결적(self-contained) 방식으로 구분될 수 있다[1,4]. 경쟁적 방식의 유전자 집합 분석은 생물학적 리소스를 이용하여 기능적 특징에 따라 분류된 여러 유전자 집합들에 대해 각 유전자 집합에 속한 유전자들과 그렇지 않은 유전자들 간에 나타나는 발현 차이의 통계적 유의성을 판단함으로써 유의한 유전자 집합을 검출하는 방법이다. 이러한 방법으로는 GSEA[1], PAGE[5] 등이 있다. 한편, 자기 완결적 방식의 유전자 집합 분석은 생물학적 리소스를 이용하여 생성된 유전자 집합에 속한 유전자들만을 대상으로 발현값의 통계적 유의성을 판단하여 유의한 유전자 집합을 검출하는 방법이다. 이러한 방법으로는 SAFE[6], SAM-GS[7], GlobalANCOVA[8], GSEAlm[9] 등이 있다. 이러한 두 가지 유전자 집합 분석 방법 중 경쟁적 방식이 현재 많이 사용되고 있는 추세이다[1].

경쟁적 방식의 유전자 집합 분석 방법 중 대표적인 방법의 하나인 PAGE[5]는 주어진 두 샘플 그룹 간의 fold change를 이용하여 각 유전자 집합의 Z-스코어를 계산하고 이들의 정규 분포를 가정하는 모수적 (parametric) 접근 방식에 의해 각 유전자 집합의 통계적 유의성을 추론하고 있다. 그러나, 이처럼 fold change만을 스코어 계산에 이용하는 경우 실제 발현값의 크기나 그룹 내 샘플들의 분산 정보 등이 스코어에 반영되지 않아 유의 유전자 집합 검출 과정에 고려되지 못하는 문제가 있다.

본 논문에서는 이러한 문제점을 보완하기 위해 모수적 방식의 유전자 집합 분석을 위한 몇 가지 새로운 유전자 발현 메트릭을 제안하고 이들을 기반으로 통계적 검정을 통해 유의한 유전자 집합을 검출하는 연구를 수행하였다. 제안한 방법의 유용성 검증을 위해 1999년 발표된 Golub et al.의 급성 백혈병(Acute Leukemina) 마이크로어레이 실험데이터[10]와 2002년 발표된 Singh et al.의 전립선암(Prostate cancer) 마이크로어레이 실험데이터[11]를 이용하여 유의한 유전자 집합 검출 실험을 수행하였으며, 검출 결과의 생물학적 검증을 위해 관련 문헌 및 공개 데이터베이스를 통해 급성 백혈병과 전립선암에 관련하여 이미 알려진 선행지식들을 각각 수집하여 결과 분석에 이용하였다.

2. 모수적(Parametric) 방식에 의한 유전자 집합 분석

PAGE와 같은 모수적 방식에 의한 유전자 집합 분석 방법은 마이크로어레이 실험에 사용된 전체 유전자들을 fold change 기반의 유전자 발현 메트릭에 의해 스코어를 계산하고, 이들을 모집단으로 하여 정규 분포(평균 μ, 분산 σ^2)를 따른다고 가정한다. 또한, 생물학적 리소스를 기반으로 특정 기능에 관여하는 것으로 알려진 유전자들 중 실험에 사용된 유전자만을 추출하여 각 유전자 집합을 생성하고 이러한 유전자 집합을 위의 모집단으로부터 무작위로 샘플링된 m개의 샘플로 구성된 확률 표본 s로 간주한다. 이 때, 확률 표본 s의 평균 스코어는 μ_s라 한다. 이러한 경우, 아래의 식 (1)의 확률변수 $Z(s)$는 중심 극한 정리(central limit theorem)에 의해 모든 m에 대해 평균 0, 분산 1인 표준 정규 분포를 따르게 된다[12,13].

$$Z(s)=\frac{(\mu_s-\mu)\times\sqrt{m}}{\sigma} \quad (1)$$

그리하여 각 유전자 집합 s에 속한 유전자들의 발현 메트릭 스코어 X_1, …, X_m이 주어질 때, 식 (1)과 같이 계산된 유전자 집합의 Z-스코어 $Z(s)$는 표준 정규 분포를 따르며, 이를 통해 유전자 집합 s의 평균 스코어 μ_s와 모집단의 평균 스코어 μ와의 차이에 대한 통계적 유의성을 테스트함으로써 주어진 유전자 집합 s가 유의한지를 판단할 수 있다. 구체적으로 각 유전자 집합 s에 대한 통계적 유의성은 아래와 같이 결정될 수 있다. 예를 들어, 유의 수준 α를 가정할 때, 유전자 발현 메트릭에 의한 전체 스코어 분포에서 스코어가 양끝으로 갈수록 중요한 의미를 가진다면, 그림 1에서와 같이 양측 검정(two-tailed test)[12,13]에 의해 $|Z(s)| \geq z_{\alpha/2}$을 만족하는 경우 통계적 유의성이 있다고 결정한다.

반면에, 전체 스코어 분포에서 스코어가 양의 방향이나 음의 방향 중 어느 한쪽에서만 중요한 의미를 가진다면, 아래 그림 2에서와 같이 단측 검정(one-tailed test)[12,13]에 의해 각각 $Z(s) \geq z_\alpha$을 만족하거나 $Z(s) \leq -z_\alpha$을 만족하는 경우 통계적으로 유의성이 있다고 결정한다.

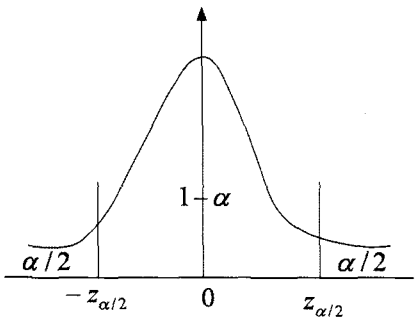
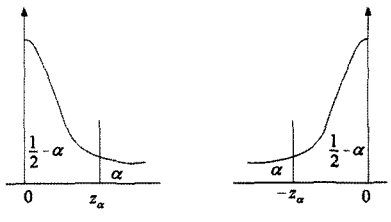


그림 1 양측 검정에서 유의수준 α에 따른 유의성 결정 방법



(a) (b)

그림 2 단측 검정에서 유의 수준 α에 따른 유의성 결정 방법: (a) 양의 방향에 관한 단측 검정, (b) 음의 방향에 관한 단측 검정

3. 유전자 발현 메트릭

앞서 기술한 바와 같이, 모수적 방식에 의한 유전자 집합 분석 방법에서 유전자 집합의 통계적 유의성을 결정하기 위해서는 각 유전자 집합 s에 대한 Z-스코어 $Z(s)$를 필요로 한다. 이를 위해 각 유전자 집합 s에 속한 유전자들의 스코어인 X_1, …, X_m을 계산할 필요가 있다. 본 장에서는 두 샘플 그룹 A, B에 대해, 각 유전자의 그룹 간 발현 차이를 측정하기 위한 발현 메트릭으로 fold change 기반의 그룹간 평균 차이만을 이용하는 AD(average difference), 그룹간 평균 발현 차이 뿐만이 아니라 발현값의 크기를 가중치로 함께 고려하는 WAD(weighted average difference), 그리고 그룹간 평균 발현 차이와 그룹 내의 샘플들의 분산을 같이 고려하는 FC(Fisher's Criterion)와 Abs_SNR(Absolute value of signal-to-noise ratio)을 모수적 방식의 유전자 집합 분석에 적용함으로써 기존의 PAGE 방법에서의 문제점을 보완하고 다양한 발현 메트릭에 따른 분석 결과의 차이와 그 의미를 살펴보고자 한다.

3.1 AD

AD[5,14]는 두 개의 샘플 그룹을 가지는 마이크로 어레이 실험 데이터에서 그룹별 평균 발현값의 차이가 얼마나 나는지를 정량화하는 방법이다. 식 (2)는 AD를 계산하는 식이다.

$$AD(i)=\overline{x_i^B}-\overline{x_i^A} \quad (2)$$

즉, 특정 유전자 i에 대한 AD의 스코어는 각 그룹 A와 B에서 로그 변환된 발현값들의 그룹별 샘플 평균인 $\overline{x_i^A}$와 $\overline{x_i^B}$간에 차이로 계산된다. 그러나 AD를 발현 메트릭으로 사용할 경우 유전자의 발현값이 매우 작아 그 자체로서 의미가 거의 없는 경우에도 두 그룹간의 평균 차이만을 고려하는 메트릭의 특징 때문에 스코어가 높게 나타나는 문제가 있다.

3.2 WAD

WAD[14]는 두 그룹간의 발현 차이를 정량화하기 위하여 각 유전자의 그룹별 평균 발현값의 차이뿐만이 아니라 실제 발현값의 크기를 함께 고려하는 방법이다. 이를 위해 전체 유전자의 평균 발현값이 분포하는 범위 중 주어진 유전자의 평균 발현값이 어느 정도 위치에 있는가에 따라 가중치를 부여하는 방식을 취하고 있다. 즉, 특정 유전자의 평균 발현값이 높아 전체 유전자들 중 최소 평균 발현값과 많은 차이가 나면 높은 가중치를 부여하고, 그렇지 않고 최소 평균 발현값에 가깝게 되면 낮은 가중치를 부여하게 된다. WAD를 계산하는 방법은 식 (2)와 같다.

$$WAD(i) = \left(\overline{x_i^A} - \overline{x_i^B}\right) \times \left(\frac{\left((\overline{x_i^A} + \overline{x_i^B})/2\right) - \min}{\max - \min} \right) \quad (3)$$

식 (3)에서 $\overline{x_i^A}$와 $\overline{x_i^B}$은 특정 유전자 i에 대한 각 그룹 A와 B에서 로그 변환된 발현값들의 그룹별 샘플 평균을 나타내며, max와 min은 전체 유전자들의 로그 변환 후 평균값($\overline{x_1}, \overline{x_2}, ..., \overline{x_N}$ 여기서 $\overline{x_i} = \left(\overline{x_i^A} + \overline{x_i^B}\right)/2$, N = 전체 유전자의 수) 중에서 가장 큰 값과 가장 작은 값을 나타낸다.

3.3 FC

FC[15-17]는 두 그룹 간에 발현값의 평균과 분산의 차이를 고려하여 정량화하는 방법이다. 앞에서 설명한 AD나 WAD의 경우 그룹별 평균 발현값만을 사용하여 스코어를 결정하는 반면, FC는 그룹별 분산을 함께 사용함으로써 궁극적으로 그룹 간에 평균 차이가 같더라도 각 그룹내의 분산이 작을수록 높은 스코어를 주게 된다. 아래 식 (4)은 특정 유전자 i에 대한 FC를 계산하는 식이다.

$$FC(i) = \frac{\left(\overline{x_i^A} - \overline{x_i^B}\right)^2}{\left(\sigma_i^A\right)^2 + \left(\sigma_i^B\right)^2} \quad (4)$$

식 (4)에서 $\overline{x_i^A}$와 $\overline{x_i^B}$은 특정 유전자 i에 대한 각 그룹 A와 B에서 로그 변환된 발현값들의 그룹별 샘플 평균을 구한 것이고, $\overline{\sigma_i^A}$와 $\overline{\sigma_i^B}$는 각 그룹 내의 표준편차를 나타낸다.

3.4 Abs_SNR

Abs_SNR 방법은 SNR 비(signal-to-noise ratio)에 절대값을 취한 것으로서 그룹 간에 평균 발현값의 차이를 그룹별 표준편차의 합에 상대적인 값으로 정량화하는 방법이다. 특정 유전자 i에 대한 Abs_SNR의 계산식은 아래 식 (5)와 같다.

$$Abs\_SNR(i) = \left| \frac{\overline{x_i^A} - \overline{x_i^B}}{\sigma_i^A + \sigma_i^B} \right| \quad (5)$$

식 (5)에서 $\overline{x_i^A}$와 $\overline{x_i^B}$은 특정 유전자 i에 대하여 서로 다른 두 그룹 A와 B에 속한 샘플들의 발현값 평균을 각각 나타내고, $\overline{\sigma_i^A}$와 $\overline{\sigma_i^B}$는 각각 표준편차를 나타낸다.

4. 실험데이터

본 논문에서 제시한 유전자 발현 메트릭을 모수적 방식의 유전자 집합 분석에 적용하여 유의한 유전자 집합을 검출하고 성능을 평가하기 위해 1999년의 Golub et al.의 급성백혈병 데이터와 2002년의 Singh et al.의 전립선암 데이터를 사용하였다. 이 두 데이터는 두 개의 샘플 그룹을 가지는 실험데이터로 유의 유전자 분석과 관련하여 많이 사용되는 검증된 자료[18]이며 이와 관련된 자료나 참고 문헌들이 존재하여 실험 결과들의 생물학적인 내용들을 이해하기가 쉽다. 본 논문에서는 각 실험데이터의 유전자 집합을 생성하기 위해 생물학적 지식 리소스로 KEGG pathway를 사용하였으며, 특정 기능을 수행하는 패스웨이에 속해 있는 유전자들로 구성된 유전자 집합을 생성하였다. 또한, KEGG pathway를 통해 급성백혈병과 관련된 패스웨이 정보, 전립선암과 관련된 패스웨이 정보들과 관련 다른 패스웨이 정보들을 이용하여 결과를 분석하였다.

4.1 급성백혈병 실험데이터

본 실험에 사용한 급성백혈병 마이크로어레이 실험데이터는 1999년 Golub et al.이 급성백혈병과 관련하여 급성 임파구성 백혈병(acute lymphoblastic leukemia, ALL)과 급성 골수성 백혈병(acute myeloid leukemia, AML)의 서브타입을 분석하기 위해 생성한 7129개의 유전자들로 구성된 데이터이다. 이 데이터는 ALL 47개와 AML 25개로 구성된 총 72개의 샘플들로 이루어져 있다. 여기에 사용된 7129개의 유전자 정보를 이용하여 생물학적 지식 리소스인 KEGG pathway를 기반으로 특정 기능을 수행하는 패스웨이에 포함된 유전자들 중에 급성백혈병 실험에 사용된 유전자 정보와 일치하는 유전자들로 구성된 유전자 집합을 생성하였다. 유전자 집합의 생성을 위해 [2]에서와 마찬가지로 유전자 집합을 구성하는 유전자의 수가 9개 이하이면 분석에서 제외시켰으며, KEGG pathway를 이용하여 총 153개의 유전자 집합을 생성하여 유의한 유전자 집합을 검출하는 데에 사용하였다.

한편, 유전자 집합 분석 결과의 생물학적 검증을 위해 문헌이나 공개 데이터베이스를 통해 확보한 현재 알려져 있는 급성 백혈병 관련 패스웨이를 수집하여 gold

standard로서 성능 분석에 활용하였다. 먼저 문헌[19-21]을 참고하여 AML과 ALL간에 차이를 나타낼 것이라 예측되는 5개의 KEGG pathway(cell cycle, Apotosis, T cell receptor signaling pathway, B cell receptor signaling pathway, hematopoietic cell lineage)를 검출하였다. cell cycle과 apotosis는 암과 관련된 중요한 패스웨이[19,21]로 알려져 있으며, T cell receptor signaling pathway, B cell receptor signaling pathway, hematopoietic cell lineage는 AML과 ALL을 구분하는 골수와 혈액에 관련된 중요한 패스웨이[20,21]이다. 또한, 질병과 관련된 유전자 정보를 가지고 있는 데이터베이스인 GAD(Genetic Association Database)[18,22]로부터 급성 백혈병과 관련된 MeSH 용어[22]를 질의문에 사용하여 AML과 ALL에 관계되는 유전자를 찾고, 이 유전자들과 관련된 KEGG pathway 중에서 Fisher's Extract test[23]를 이용하여 p-value가 0.05이하인 2개의 KEGG pathway (Glutathione metabolism, Meta-bolism of xenobiotics by cytochrome P450)를 검출하였다. 마지막으로 KEGG pathway 중에서 Acute myeloid leukemia pathway에 속해 있는 7개의 KEGG pathway(Hematopoietic cell lineage, MAPK signaling pathway, Jak-STAT signaling pathway, Apoptosis, mTOR signaling pathway, Cell cycle, Acute myeloid leukemia pathway)을 검출하였다. 그리하여 총 11개의 급성백혈병과 관련된 KEGG pathway를 검출하였고, 이것을 이용하여 본 논문에서 제시한 방법의 성능을 평가하였다.

4.2 전립선암 실험데이터

전립선암 마이크로어레이 실험데이터는 2002년 Singh et al.[10]이 정상세포의 유전자 발현값과 전립선암 유전자의 발현값을 분석하여 전립선암과 관련된 유의 유전자를 검출하기 위해 사용된 마이크로어레이 실험 데이터이다. 이 데이터는 12,600개의 유전자에 대해, 전립선암 관련 샘플 52개와 정상 샘플 50개로 이루어진 총 102개의 샘플로 구성되어 있다. 본 논문의 실험을 위해 유전자 집합에 포함된 유전자의 수가 최소 10개 이상이 되는 유전자 집합 169개를 KEGG pathway를 기반으로 생성하였다. 한편, 유전자 집합 분석 실험에 대한 결과 분석을 위해 2007년 Huang D. et al.[9]이 발표한 논문에서 사용된 전립선암 관련 13개의 패스웨이인 Neurodegenerative Diseases, Amyotrophic lateral sclerosis (ALS), Apoptosis, MAPK signaling pathway, Cell cycle, Focal adhesion, Regulation of actin cytoskeleton, Wnt signaling pathway, Tight junction, Toll-like receptor signaling pathway, Adipocytokine signaling pathway, TGF-beta signaling pathway, Prostate cancer)을 gold standard로서 활용하였다. 다만 Huang D. et al.의 논문에는 포함되어 있지만 현재 KEGG pathway 데이터베이스에서 제외된 "Dorso-ventral axis formation pathway" 대신에 "Prostate cancer"[24]를 추가하여 분석에 이용하였다.

5. 실험 결과 분석

유의한 유전자 집합 검출에 관한 실험 방법은 4개의 유전자 발현 메트릭 AD, WAD, FC, 그리고 Abs_SNR을 모수적 방식의 유전자 집합 분석에 적용하고 유의수준 0.05를 이용하여 유의한 유전자 집합들을 추출하였다.

표 1은 Golub의 급성백혈병 데이터에 관한 분석 결과로서 4개의 유전자 발현 메트릭 AD, WAD, FC, Abs_SNR 각각을 이용하여 앞서 기술한 급성백혈병 관련 11개의 패스웨이(gold standard) 중에서 실제 어떤 패스웨이가 유의한 것으로 검출되었는지를 나타낸 것이다. 표 1에서 나타난 바와 같이, 급성 백혈병 데이터의 경우, 유전자의 그룹별 분산 정보를 스코어 계산에 사용한 FC와 Abs_SNR이 분산 정보를 사용하지 않은 AD와 WAD에 비해 상대적으로 좋은 결과를 보여주고 있다.

표 2는 Singh et al.의 전립선암 데이터에 관한 분석 결과로서 4개의 유전자 발현 메트릭 AD, WAD, FC, Abs_SNR 각각을 이용하여 검출한 유의 유전자 집합 결과 중에서 앞서 기술한 전립선암 관련 13개의 패스웨이들과 일치하는 부분을 나타낸 것이다. 표 2에서 나타난 바와 같이, 전립선암 데이터의 경우, AD와 WAD를 적용했을 때에 전립선암과 관련된 패스웨이 3개를 실제 검출한 반면에, 분산 정보를 고려한 FC나 Abs_SNR를 적용했을 때에는 급성 백혈병의 경우와 달리 별 다른 좋은 결과를 보여주지 못하고 있다.

한편, 상기 표 1과 2의 결과는 유의수준 0.05를 기준으로 검출된 결과이기 때문에 각 메트릭에 대해 검출된 최종 유의 유전자집합의 수가 서로 다른 문제가 있다. 그리하여 precision과 recall을 고려한 결과분석을 위해 앞의 두 실험데이터의 결과들을 F_1 측정치[25]를 이용하여 표 3에 나타내었다. F_1 측정치는 precision과 recall의 조화평균으로서 그 값이 높을수록 좋은 성능을 나타낸다. F_1을 구하는 수식은 아래 식 (6)와 같다.

$$F_1 = \frac{2rp}{r+p} = \frac{2\times TP}{2\times TP + FP + FN} \quad (6)$$

여기서 TP는 true positive로 유의 유전자집합에서 gold standard와 공통으로 일치하는 패스웨이의 수를 나타낸 것이다. FN은 false negative로 gold standard 중에서 유의한 유전자집합으로 검출되지 않은 패스웨이의 수를 나타내고, FP는 false positive로 유의한 유전자 집합들로 검출된 것 중에서 gold standard에 속하지

않은 패스웨이의 수이다. 표 3의 결과들을 살펴보면, 표 1과 2에서와 같이, 급성백혈병 데이터에서는 유전자 발현 메트릭 FC와 Abs_SNR이 분산 정보를 사용하지 않는 AD와 WAD에 비해 상대적으로 좋은 결과들을 보여주고 있으며, 전립선암 데이터의 경우에는 분산 정보를 사용하지 않는 AD와 WAD를 사용했을 때가 FC를 사용한 것보다는 상대적으로 좋은 결과들을 나타내고 있다. 다만, F_1 측정치의 측면에서는 Abs_SNR이 가장 높은 수치를 보여주고 있어 두 실험 데이터 모두에서 좋은 결과를 나타내고 있다.

이처럼, 표 1, 2, 3에서와 같이, 실험 결과가 데이터에 따라 다른 이유를 알아보기 위하여 두 데이터의 분산 정보를 계산하였다. 표 4에 나타난 바와 같이, 급성백혈병 데이터의 경우 유전자 발현값의 샘플간 표준편차가 클래스와 상관없이 전반적으로 높게 나타났으며, 전립선암 데이터의 경우에는 샘플간 표준편차가 상대적으로 낮은 값을 보이고 있다. 따라서 유전자 발현값의 샘플간 표준편차가 상대적으로 큰 데이터의 경우 분산정보를 고려하는 FC와 Abs_SNR이 유리할 것으로 판단되며, 데이터의 샘플간 표준편차가 그다지 크지 않은 경우 평

표 1 급성백혈병 데이터 분석 결과: AD, WAD, FC, Abs_SNR을 적용한 후 153개의 유전자 집합에서 유의한 유전자 집합을 검출하여 급성백혈병 관련 패스웨이들과 일치하는 것을 나타낸 표

Pathway Names	AD	WAD	FC	Abs_SNR
Glutathione metabolism			●	●
Metabolism of xenobiotics by cytochrome P450			●	●
MAPK signaling pathway				●
Cell cycle	●	●	●	●
mTOR signaling pathway				
Apoptosis				
Jak-STAT signaling pathway				●
Hematopoietic cell lineage	●		●	●
T cell receptor signaling pathway				
B cell receptor signaling pathway	●	●	●	●
Acute myeloid leukemia				

표 2 전립선암 데이터 분석 결과: AD, WAD, FC, Abs_SNR을 적용한 후 169개의 유전자 집합에서 유의한 유전자 집합을 검출하여 전립선암 관련 패스웨이들과 일치하는 것을 나타낸 표

Pathway Names	AD	WAD	FC	Abs_SNR
Neurodegenerative Diseases				
Amyotrophic lateral sclerosis (ALS)				
Apoptosis		●		
MAPK signaling pathway				
Cell cycle				
Focal adhesion	●	●	●	●
Regulation of actin cytoskeleton	●	●		
Wnt signaling pathway				
Tight junction				
Toll-like receptor signaling pathway				
Adipocytokine signaling pathway				
TGF-beta signaling pathway				
Prostate cancer	●			●

표 3 실험 데이터별 F_1 측정값을 이용한 결과 분석(TP: True positive, FN: False negative, FP: False positive, TN: True negative)

(a) 급성백혈병

	TP	FN	FP	TN	F_1 measure
AD	3	8	23	119	0.162162162
WAD	2	11	27	113	0.095238095
FC	5	6	12	130	0.357142857
ABS_SNR	7	4	23	119	0.341463415

(b) 전립선암

	TP	FN	FP	TN	F_1 measure
AD	3	10	52	104	0.088235294
WAD	3	10	54	102	0.085714286
FC	1	12	28	128	0.047619048
ABS_SNR	2	11	25	131	0.1

표 4 실험 데이터별 유전자의 샘플간 표준편차 정보

Data-set	$\overline{\sigma_A}$	$\overline{\sigma_B}$	$\left(\overline{\sigma_A}+\overline{\sigma_B}\right)/2$	$\overline{\sigma_{A+B}}$
급성백혈병	2.092393	2.112612	2.102503	2.16509
전립선암	1.482079	1.180006	1.331043	1.361984

균 발현 차이에 집중하는 AD와 WAD를 사용하는 것이 바람직할 것으로 보인다.

6. 결론 및 향후 계획

본 논문에서는 마이크로어레이를 이용하여 생성된 두 개의 샘플 그룹으로 이루어진 유전자 발현 데이터로부터 유의한 유전자 집합을 검출하는 방법에 관한 연구를 수행하였다. 이를 위해, 두 그룹 간에 유전자 발현 차이를 정량화하기 위한 메트릭으로서 현재 모수적 방식의 분석에 주로 사용하고 있는 AD 이외에 새로운 발현 메트릭 WAD, FC, Abs_SNR을 사용하는 것을 제안하였으며, 이를 통해 계산된 스코어를 모수적 방식의 통계적 검정에 의해 유의성을 판단함으로써 유의한 유전자 집합을 검출하는 방법을 제안하였다. 이러한 방법은 현재 공개되어 있는 Golub의 급성백혈병 데이터와 Singh의 전립선암 데이터에 적용하여 각 질병에 관련된 것으로 알려진 중요한 패스웨이들을 얼마나 잘 검출하는지를 살펴봄으로써 그 방법의 유용성을 검증하고자 하였다. 실험결과에 의하면, 본 논문에서 제안한 방법들은 데이터의 특성에 따라 다른 결과를 보여주었다. 즉, 유전자 발현값의 샘플간 표준편차의 정도가 상대적으로 크게 나타난 급성백혈병 데이터의 경우 이러한 표준편차의 정보를 스코어 계산시 반영할 수 있는 FC와 Abs_SNR을 사용하는 것이 좋은 결과를 보여준 반면에, 유전자의 샘플간 표준편차가 상대적으로 크지 않은 전립선암 데이터의 경우 이러한 표준편차 정보를 스코어 계산에 반영하지 않는 AD, WAD을 사용하는 것이 전반적으로 좋은 결과를 보여줌을 확인할 수 있었다. 또한, Abs_SNR은 표준편차의 크기에 상관없이 전반적으로 좋은 결과를 나타내었다. 이러한 관찰결과는 유전자 집합 분석에 유용한 정보를 제공해 줄 수 있으리라 판단된다. 다만, 어떤 발현 메트릭을 사용할지를 결정짓는 표준편차의 기준에 대한 문제는 추가적인 연구가 필요한 부분이다.

참 고 문 헌

[1] Nam, D., Kim, S. Y., "Gene-set approach for expression pattern analysis," *Briefings in bioinformatics*, vol.9, no.3, pp.189-197, May 2008.

[2] 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., Mesirov, J. P., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol.102, no.43, pp.15545-12250, Oct. 2005.

[3] Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A., "The KEGG databases at GenomeNet," *Nucleic Acids Res.*, vol.30, no.1, pp.42-46, Jan. 2002.

[4] Efron, B. and Tibshirani, R. On testing the significance of sets of genes. *Stanford tech report rep*, Available: http://www-stat.stanford.edu/ tibs/ftp/GSA.pdf, 2006.

[5] Kim, S. Y., Volsky, D. J., "PAGE: parametric analysis of gene set enrichment," *BMC Bioinformatics*, vol.8, no.6, pp.144, Jun. 2005.

[6] Barry, W. T., Nobel, A. B., Wright, F. A., "Significance analysis of functional categories in gene expression studies: a structured permutation approach," *Bioinformatics*, vol.21, no.9, pp.1943-1949., May 2005.

[7] Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P., Yasui, Y., "Improving gene set analysis of microarray data by SAM-GS," *BMC Bioinformatics*, vol.5, no.8, pp.242, Jul. 2007.

[8] Hummel, M., Meister, R., Mansmann, U., "Global-ANCOVA: exploration and assessment of gene group effects," *Bioinformatics*, vol.24, no.1, pp.78-85, Jan. 2008.

[9] Oron, A. P., Jiang, Z., Gentleman, R., "Gene set enrichment analysis using linear models and diagnostics," *Bioinformatics*, vol.24, no.22, pp.2586-2591, Nov. 2008.

[10] Golub, T. R., Slonim, D. K., Tamayo, P., Huard., C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., Lander, E. S., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol.286, no. 5439, pp.531-537, Oct. 1999.

[11] Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E.S., Loda, M., Kantoff, P. W., Golub, T. R., Sellers, W. R., "Gene expression correlates of clinical prostate cancer behavior," *Cancer Cell*, vol.1, no.2, pp.203-209, Mar. 2002.

[12] Hogg, R. V., Craig, A. T., Mckean, J., *Introduction to Mathematical Statistics*, 6th ed., Pearson Education, 2005.

[13] Alberto, L. G., *Probability, Statistics, and Random Processes for Electrical Engineering*, 3rd Ed., Pearson Education, 2009.

[14] Kadota, K., Nakai, Y., Shimizu, K., "A weighted

average difference method for detecting differentially expressed genes from microarray data," *Algorithms for molecular biology*, vol.26, no.3, pp.8, Jun. 2008.
[15] Bishop, C., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.
[16] Blum, A., Langley, P., "Selection of relevant features and example in machine learning," *Artificial intelligence*, vol.97, pp.245-271, Dec. 1997.
[17] Bradley, P., Mangasarian O., Street W., "Feature selection via mathematical programming," *Technical report to appear in INFORMS Journal on computing*, vol.10, no.2, pp.209-217, Feb. 1998.
[18] Trajkovski, I., Lavrac, N., Tolar, J., "SEGS: search for enriched gene sets in microarray data," *Journal of biomedical informatics*, vol.41, no.4, pp.588-601, Aug. 2008.
[19] Potten, C., Wilson J., *Apoptosis*, Cambridge University Press, 2005.
[20] Knudsen, S., *Cancer Diagnostics with DNA Microarrays*, John Wiley & Sons, Inc., 2006.
[21] Weinberg, R. A., *The biology of CANCER*, Carland Science, 2007.
[22] The Genetic Association Database, Available: http://geneticassociationdb.nih.gov/
[23] Huang, D., Chow, T. W., "Identifying the biologically relevant gene categories based on gene expression and biological data: an example on prostate cancer," *Bioinformatics*, vol.23, no.12, pp.1503-1510. Jun. 2007.
[24] "KEGG(Kyoto Encyclopedia of Genes and Genomes) PATHWAY Database," Available: http://www.genome.ad.jp/kegg/pathway.html
[25] Tan, P. N., Steinbach, M., Kumar, V., INTRODUCTION TO DATA MINING, Pearson Education, Inc., 2006.

김 재 영
경북대학교 전자전기컴퓨터학부. (School of Electrical Engineering and Computer Science, Kyungpook National University). 2006년 위덕대학교 컴퓨터공학과학사 졸업(학사). 2009년 경북대학교 정보통신학과 졸업(석사). 2009년~경북대학교 전자전기컴퓨터학부 박사과정. 관심분야는 생물정보학, 데이터마이닝, 패턴인식

신 미 영
경북대학교 전자전기컴퓨터학부. (School of Electrical Engineering and Computer Science, Kyungpook National University). 1991년 연세대학교 전산과학과 졸업(학사). 1993년 연세대학교 전산과학과 졸업(석사). 1998년 미국 Syracuse Univ., EECS Dept., 졸업(박사). 1999년~2005년 한국전자통신연구원 선임 연구원. 2005년~경북대학교 전자전기컴퓨터학부 부교수. 관심분야는 패턴인식, 바이오인포매틱스, 데이터마이닝