응용통계연구 제 **21**권 *3*호, *2008*년, *pp. 523 535*

절대치와 절삭을 이용한 유전자 집단 분석*

이광현[1] 이선호[2]

요 약

본 연구의 목적은 마이크로어레이 자료로부터 암 또는 질병에 유의한 유전자집단을 찾아내는 보다 효과적인 방법을 제안하고자 하는 것이다. 유전자 집단 분석의 대표적 방법인 PAGE와 GSEA의 한계점을 살펴보고, 그것을 보완하기 위한 GSA-AT라는 방법을 제안하였다. 모의실험과 실제자료실험을 통해 분석해 본 결과 본 연구에서 제안한 GSA-AT 방법에서 더 의미있는 결과를 도출하였다.

주요용어: 마이크로어레이, 유전자 집단 분석, PAGE, GSEA, GSA-AT.

1. 서론

마이크로어레이 기술은 동시에 수만 개의 유전자 발현값 측정을 가능하게 함으로써 생물학, 생화학 및 다른 여러 분야에서 사용되어지고 있다. 이 고효율 기술은 측정된 유전자의 발현값을 통해 약이나 질병, 환경 조건 등에 의한 유전자들의 규칙적인 변화 또는 유전자들 간의 상호작용을 발견하여 암 또는 질병의 예방, 진단과 치료 및 신약 개발 등에 크게 기여하고 있다. 하지만 복잡한 측정 과정을 거치면서 다양해진 오차 발생 요인을 줄이기 위해 마이크로어레이 연구는 실험설계부터 시작해서 사용 칩에 대한 수준 분석, 자료의 휠터링과 정규화, 통계분석, 클러스터링 기술 등 여러 분야의 연구로 나누어지기 시작했다.

초기의 마이크로어레이 분석은 정규화한 자료를 통해 질병에 유의한 특이 발현하는 유전자를 개별적으로 찾아내는 단일 유전자 위주의 분석으로 Tusher 등 (2001)의 SAM(Significance Analysis of Microarrays), Tibshirani 등 (2002)의 PAM(Prediction Analysis for Microarrays) 등이 대표적이다. 하지만 이 방법은 특이 발현 유전자가 너무 많거나 적게 나올 경우 해석에 어려움이 있고, 분석 대상 유전자수가 매우 많으므로 다중 검정에 어려움이 발생하며, 같은 질병에 대한 연구라 하더라도 자료가 다르면 결과가 별로 일치하지 않는다는 단점이 있었다.

단일 유전자 분석의 단점을 보완하기 위해 연구된 방법이 유전자 집단 분석이다. 유전자 집단 분석이란, 생물학적 기능이나 염색체 위치가 같은 유전자 집단을 분석하여 질병의 발생

* 이 논문은 2006년 교육인적자원부의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (KRF-2006-C-00020)

1) (143-747) 서울 특별시 광진구 군자동, 세종대학교 대학원 응용통계전공, 석사 과정.
E-mail: gusl-2@nate.com
2) (143-747) 교신저자. 서울 특별시 광진구 군자동, 세종대학교 응용수학전공, 교수.
E-mail: leesh@sejong.ac.kr

에 유의한 역할을 하는 집단을 찾아내는 분석 방법으로 GSEA(Gene Set Enrichment Analysis) (Mootha 등, 2003; Curtis 등, 2005; Subramanian 등, 2005), PAGE(Parametric Analysis of Gene set Enrichment) (Kim과 Volsky, 2005) 등이 대표적이며 이와 관련한 많은 연구가 계속되고 있다 (Goeman 등, 2004; Manoli 등, 2006).

본 논문에서는 대표적 유전자 집단 분석인 GSEA와 PAGE를 소개하고 기존 방법의 단점인 상대적 위치에 의존하는 부분과 유의한 유전자들 간의 상반되는 특성에 의한 상쇄작용이 발생하는 부분, 그리고 질병에 매우 유의한 특이 발현 유전자가 같은 집단에 소속된 대부분의 유의하지 않은 유전자에 묻혀버리는 부분 등을 보완할 수 있는 새로운 방법을 모색하고 제안하려 한다. 그 방법으로는 유전자 발현값의 방향성은 무시한 채 크기만을 보는 방법과 유전자들의 유의성에 따라 절삭시키는 방법을 시도해 보았다. 그리고 각 방법의 통계적 유의성을 보이기 위해 설계를 통한 모의실험과 실질적인 자료를 훈련군(training set)과 시험군(test set)으로 나누어 훈련군의 표본으로 분석하고 그 결과를 시험군의 표본으로 검증하는 방법을 사용하였다.

2절에서는 기존 분석 방법을 통해 유전자 집단 분석의 중요성을 보여주고, 3절에서는 기존 방법의 단점을 보완할 수 있는 새로운 방법을 제안하며, 4절에서는 데이터 소개 및 분석 방법을 설명하고, 실질적 데이터 분석을 통한 결과 등을 살펴볼 것이며, 5장에서는 분석 결과를 토대로 결론을 맺는다.

2. 유전자 집단 분석

기존의 단일 유전자 분석의 단점은 1) 경우에 따라 특이 발현 유전자가 많이 나올 때도 있고 반대로 거의 나오지 않을 때도 있으며, 2) 분석 대상 유전자 수가 너무 많아 다중 검정에 어려움이 있으며, 3) 같은 주제에 관한 연구라 해도 자료에 따른 결과가 매우 달라 일치성이 떨어진다는 것이었다.

이에 대안으로 제시된 유전자 집단 분석(Gene Set Analysis)은 생물학적 경로(pathway)나 염색체 위치가 같은 유전자 집단을 분석하여 질병의 발생에 유의한 집단을 찾아내는 분석 방법으로, 단일 유전자 분석의 단점을 개선했다는 평가를 받고 있다. 분석 방법과 조건이 다른 경우, 단일 유전자 분석에서는 결과에 많은 차이를 보였지만 유전자 집단 분석에서는 큰 차이가 없었다. 그리고 수만 개의 유전자를 공통기능을 가진 몇 개의 집단으로 묶어 줌으로써 검정 대상 수를 줄여 다중 검정이 용이해 졌으며, 단일 유전자 분석에 비해 같은 연구 주제이지만 서로 독립인 자료들의 결과에 대한 일치성도 높아졌다. 또한 개별적으로는 특이 발현 유전자가 없는 집단이라 해도 유전자들의 분포에 의해 질병에 영향을 미치는 집단도 찾을 수 있다는 장점도 있다.

이러한 유전자 집단 분석에 대한 활발한 연구가 진행되면서 여러가지 방법들이 제안되어지고 있다. 2장에서는 유전자 집단 분석의 대표격인 GSEA와 PAGE의 우수성과 문제점을 자세히 알아보도록 하겠다.

표 2.1: GSEA 방법의 분석 결과 (Mootha 등, 2003)

유전자 집합	MES	p-value
OXPHOS_HG-U133A_probes.grp	346.88	0.003
human_mitoDB_6_2002_HG-U133A_probes.grp	215.94	0.091
mitochondr_HG-U133A_probes.grp	207.93	0.087
c20_U133_probes.grp	181.15	0.062
MAP00190_Oxidative_phosphorylation.grp	148.90	0.084
c22_U133_probes.grp	142.90	0.028
c29_U133_probes.grp	131.47	0.026
MAP00960_Alkaloid_biosynthesis_II.grp	110.94	0.022
c23_U133_probes.grp	110.90	0.087

2.1. GSEA(Gene Set Enrichment Analysis)

Mootha 등 (2003)이 제안한 GSEA는 질병과의 연관성을 대표할 수 있는 어떤 기준에 의해 전체 유전자를 순서대로 나열한 후, pathway에 속한 유전자들의 위치가 균일 분포를 따르는가를 보는 방법으로, http://www.broad.mit.edu/gsea/에서 프로그램을 다운로드할 수 있다. GSEA 방법은 pathway 내의 특이발현 유전자들을 통해 유의한 유전자 집단을 판단할 뿐만 아니라, 유전자들의 분포를 파악하여 판단하기도 한다.

GSEA의 구체적인 방법은 1단계에서 염색체 위치나 생물학적 기능에 따라 pathway를 수집하고, 2단계에서는 전체 유전자를 각 표본의 표현형(phenotype: 종양 또는 정상, 재발여부, 병기 등) 간 발현의 차이를 대표할 수 있는 값(상관계수, t-통계량 등)을 기준으로 정렬한 후 정렬된 리스트에서 pathway에 속한 유전자의 위치를 찾아낸다. 3단계에서는 ES(Enrichment Score)를 구한다. N이 전체 유전자 수, G가 pathway 내 유전자 수이며, X_i가 아래와 같이 정의될 때, $\mathrm{ES} = \sum_{i=1}^{j} X_i$이 된다 (단, $j = 1.2, \ldots, N$).

$$X_i = \begin{cases} -\sqrt{\dfrac{G}{N-G}}, & i\text{번째 유전자가 pathway에 속하지 않을 때,} \\ \sqrt{\dfrac{N-G}{G}}, & i\text{번째 유전자가 pathway에 속할 때.} \end{cases}$$

4단계에서는 구해진 ES 중 가장 큰 값인 MES(Maximum Enrichment Score)를 구하고 5단계에서는 permutation을 반복 시행하여 MES의 분포를 알아낸다. 즉, 표본을 랜덤으로 phenotype에 분배하여 2단-4단계를 반복한다. 마지막 6단계에서는 5단계에서 만들어진 MES의 분포를 통해 pathway의 MES값에 대한 p-value를 찾아낸다.

Mootha 등 (2003)에서는 GSEA 방법과 단일 유전자 분석을 비교하기 위하여 골격 근육 생검 자료 (정상 표본 17개. 당뇨 표본 18개)를 이용해 분석한 결과, 단일 유전자 분석의 경우에는 유의한 유전자가 전혀 나오지 않았지만, GSEA 분석의 경우에는 표 2.1과 같은 결과가 나타났다.

그림 2.1은 GSEA 분석을 통해 유의하다고 추정된 pathway 중 OXPHOS_HG에 속하는 유전자들에 대한 산점도로서, 가로축은 정상표본의 평균 발현값, 세로축은 당뇨표본의 평

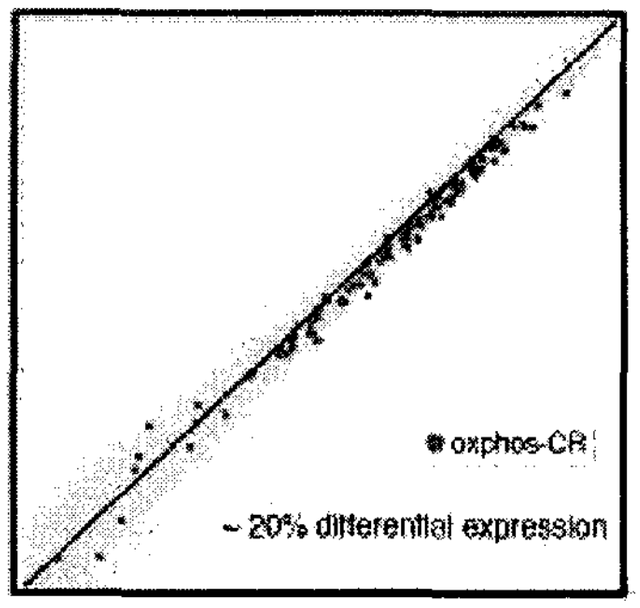


그림 2.1: OXPHOS_HG의 산점도 (Mootha 등, 2003)

균 발현값을 나타내고, 회색 점은 전체 유전자를, 붉은 점은 OXPHOS_HG에 속한 유전자를 나타낸다. OXPHOS(Oxidative Phosphorylation)는 인산화 작용에 영향을 미치는 106개의 유전자들로 구성된 pathway로 특이 발현 유전자의 비율이 낮지만, 약 90%의 유전자가 두 phenotype의 발현값의 동일함을 가정한 선보다 아래에 몰려있는 경향을 보여 대부분이 특정phenotype에서 더 낮은 발현값을 갖고 있다는 것을 알 수 있다. 이것은 개개의 유전자를 살펴보면 질병에 영향을 미치는 유전자가 없더라도, 그런 유전자들이 모여 일정한 흐름을 이루며 질병에 영향을 미칠 수 있다는 것을 나타낸다.

또한, Subramanian 등 (2005)에서는 독립인 두 자료를 통해 단일 유전자 분석과 GSEA의 일치성에 대해 비교하였다. Harvard 대학교와 MIT 연구진이 주축이 된 Boston 연구에 쓰인 폐암 자료의 62개 표본과 Michigan 대학교를 중심으로 하는 Michigan 연구에 쓰인 폐암 자료의 86개 표본을 GSEA와 단일 유전자 분석을 통해 각각 분석 후, 두 자료의 결과가 어느 정도 일치하는지 살펴보았다. 먼저, 단일 유전자 분석의 경우, 두 자료 모두 다중 검정 후, adjusted p-value < 0.05를 만족하는 유전자는 하나도 없었으며, 두 결과의 유의성이 큰 상위 50개 유전자 군은 0%, 상위 100개의 유전자 군은 12%의 일치성을 보였다. 반면, GSEA의 경우, Boston 연구와 Michigan 연구에서 각각 8개, 11개의 pathway가 유의하다고 나왔으며, 그 중 3개의 똑같은 pathway와 2개의 유사한 pathway가 발견되어 약 50%에 가까운 일치성을 보였다.

2.2. PAGE(Parametric Analysis of Gene set Enrichment)

Kim과 Volsky (2005)에 의해 제안된 PAGE는 중심 극한 정리를 이용하여 pathway 분석의 모수적 접근을 시도하였다. 각 유전자의 두 군 간 발현 차이의 대푯값(두 군 간 발현값의 차 또는 t-통계량, fold-change 값 등)의 평균이 μ, 분산이 σ^2일 때, 크기가 m인 pathway에 속한 유전자들의 대푯값들의 평균은, m이 충분히 크면 평균이 μ이고 분산은 σ^2/m인 정규분포에 근사한다는 가정 아래 검정을 실시하는 방법이다.

Kim과 Volsky (2005)에서는 정상인과 당뇨환자 사이의 차이를 나타내는 대푯값으로 fold-change 값을 사용하여 골격 근육 생검 자료을 PAGE 분석한 결과, 두 군간 발현의 차이를 보

표 2.2: PAGE 방법의 분석 결과 (Kim and Volsky, 2005)

유전자 집합	Z - scorep	p-value
OXPHOS_HG_U133A	−10.5835	< 1.0E−11
human_mitoDB_6_2002_HG_U133A	−6.7213	1.81E−11
mitochondr_HG_U133A	−6.4761	9.46E−11
MAP00190_Oxidative_phosphorylation	−4.5745	4.78E−05
c20_U133	−3.7461	0.0002
c25_U133	−2.7617	0.0058
c21_U133	−2.1116	0.0347

이는 pathway들을 찾아내었고(표 2.2의 결과), 유의한 유전자가 하나도 발견되지 않았던 단일 유전자 분석에 비해 우수함을 보였다. 또한, 같은 연구에 대한 서로 독립인 자료들의 결과에 대한 일치성에 관한 문제도, 남녀 근육 기능 분석을 통해 단일 유전자 분석보다 우수함을 보였다. 두 개의 독립적인 마이크로어레이 자료, GDS 287(남자의 근육 기능 자료)과 GDS 472(여자의 근육 기능 자료)를 단일 유전자 분석과 PAGE를 통해 각각 분석한 후 그 결과의 일치성을 살펴본 결과, 단일 유전자 분석에서는 질병에 대해 양의 반응을 보이는 유전자의 경우에는 12.4% 일치하였으며, 음의 반응을 보이는 유전자의 경우에는 4.4% 일치함을 보였다. 반면, PAGE에서는 질병에 대해 양의 반응을 보이는 pathway의 경우에 62.5%의 일치성을, 음의 반응을 보이는 pathway의 경우에는 49.6%을 보이며, 단일 유전자 분석의 결과보다 일치성이 매우 높음을 보였다.

3. GSA-AT(Gene Set Analysis: Absolute and Trimmed) 방법의 제안

유전자 집단 분석을 통해 많은 의학적, 생물학적 발전을 이루었지만 아직도 부족함은 있고, 질병의 발생에 유의한 영향을 미치는 유전자 집단을 더욱 정확하게 판단할 수 있는 강력한 방법을 찾기 위한 연구는 계속 되고 있다.

GSEA 방법의 단점은 유전자의 특이발현 여부와는 관계없이 유전자들의 상대적 위치에 의해 유의성이 결정될 수 있다는 것이다. Pathway에 속한 유전자들의 위치가 순위대로 배열된 전체 유전자 리스트 내에서 균일 분포를 따르는지, 즉, 고르게 분포되어 있는지를 기준으로 판단하기 때문에 유의하지 않은 유전자들을 가진 pathway라 해도 리스트 내에 고르게 분포되지 않는다면 유의하다고 판단될 수 있다. 게다가 이러한 성질 때문에 유전자를 많이 가진 pathway 일수록 유전자를 적게 가진 pathway 보다 유의하다고 판단될 확률이 상대적으로 커지게 된다.

각 유전자들의 특성을 반영하는 대부분의 대푯값들은 양적 특이발현(+)과 음적 특이발현(−)으로 나뉘어 지는데, PAGE는 대푯값들의 평균을 검정통계량으로 사용하기 때문에 각 유전자간 상쇄작용이 발생한다는 단점이 있다. 즉, pathway 내에 양적 특이 발현 유전자와 음적 특이 발현 유전자가 공존하게 되면, 그 평균은 서로 상쇄되어 0에 가깝게 되면서 유의한 pathway임에도 불구하고 유의하지 않다고 판단될 수 있다. 그리고 중심극한 정리를 이용

하여 모수적 접근을 시도한 방법이기 때문에 충분한 유전자 수를 보유하지 않은 pathway는 PAGE를 통한 분석이 불가능하다는 단점도 있다.

마지막으로, 대다수의 유전자 집단 분석 방법에서, 전체 유전자에 초점을 맞추다보니, 개별 유전자의 특이 발현성이 무시되는 단점이 있다. 즉, 질병의 발생에 매우 큰 영향을 주는 유전자를 가지고 있는 pathway라 해도 같은 pathway 내 유의하지 않은 유전자들이 너무 많게 되면 그 비중이 작아져서 유의하지 않은 많은 유전자에 의해 묻혀버릴 수 있다는 것이다.

이 논문에서 제시하는 GSA-AT는 이러한 단점들을 보완하기 위해 설계되었다. 먼저 각 유전자의 중요도가 유전자간의 상대적 위치에 의해 결정되는 GSEA의 단점을 보완하기 위해 각 유전자들의 대표값을 분석에 직접 반영하는 방법을 찾았다. 또한 PAGE의 단점인 각 유전자들의 대푯값의 평균을 사용함으로써 발생하는 상쇄작용은 대푯값의 절대치를 사용함으로써 보완하였다. 또, 특이 발현 유전자를 부각시키기 위해 각 pathway에서 유의하지 않은 유전자를 미리 제거하는 방법을 시도하였다. 정리하면, GSA-AT 방법은 특이발현 유전자의 방향성을 제외시키는 대신 발현 정도의 크기에 가중을 둔 방법으로, 다음의 단계를 밟는다.

1 단계: 주어진 유전자 집합을 바탕으로 pathway 정보를 수집한다.

2 단계: 원자료를 통해 각 표본의 표현형(phenotype) 간의 차이를 나타내는 대푯값을 구하고 절대값이 큰 순서대로 나열한다.

3 단계: 나열된 값 중 하위 $x\%$에 해당되는 유전자들을 지워 자료 및 pathway내 의미없는 유전자를 절삭한다.

4 단계: 절삭된 자료를 통해 pathway 별 대푯값의 절대값 평균을 구한다.

5 단계: 표본을 임의로 phenotype에 분배하여 2단계–4단계를 반복하는 permution을 시행한다.

6 단계: 5단계를 통해 얻어진 절대값 평균의 분포를 통해 p값을 구한다.

4. 자료 분석 및 결과

기존의 PAGE, GSEA와 새로 제안한 GSA-AT를 비교하기 위해 설계된 모의실험을 통해 각 분석의 결과를 비교해보는 방법과 실제 자료를 훈련군(training set)과 시험군(test set)으로 나누어 훈련군에 속한 표본을 분석하고 시험군에 속한 표본을 통해 각 분석의 결과를 확인하는 방법을 사용하였다.

실험에 쓰인 자료로는 www.bioconductor.org에서 다운로드 가능한 111개의 표본으로 구성된 ALL(Acute Lymphoblastic Leukemia: 급성 림프구성 백혈병) 자료를 사용하였다. ALL은 정상 표본 74개와 만성골수성 백혈병 표본 37개로 나뉘어 있고, 12625개의 유전자 정보와 그와 관련된 189개의 pathway로 구성되어 있다.

ALL 자료를 다음의 6가지 방법을 사용하여 분석하였다.

방법 1: phenotype간의 t 통계량값을 사용한 PAGE.

방법 2: phenotype과 발현값과의 상관계수를 사용한 GSEA.

방법 3: phenotype간의 t 통계량의 절대값을 사용한 GSA-AT1.

방법 4: phenotype간의 t 통계량의 절대값을 사용한 GSA-A1(Gene Set Analysis - Absolute).

방법 5: log(정상 표본의 발현값/백혈병 표본의 발현값)의 절대값을 사용한 GSA-AT2.

방법 6: log(정상 표본의 발현값/백혈병 표본의 발현값)의 절대값을 사용한 GSA-A2.

GSA-AT의 경우, 임의로 절삭 정도(x)를 25%. 50%, 75% 각각에 대하여 사용하였는데 전체유전자의 50%와 75%를 절삭할 경우 구성 유전자수가 5개 이내인 pathway의 수가 많아지는 문제점이 발생하였다. 그래서, GSA-AT의 절삭 정도(x)는 최종적으로 25%로 하였고 GSA-A는 GSA-AT와 비교하기 위하여 절삭 정도를 0%, 즉 절삭을 하지 않은 방법이다. 또한 GSEA는 1000번, GSA-A와 GSA-AT는 10000번의 permutation을 실시하였다.

4.1. 모의실험

ALL의 자료를 이용하여 다음의 서로 다른 특징을 만족하는 pathway들을 각 방법이 얼마나 잘 판단하는지 알아보기 위하여 두 가지의 모의실험을 하였다.

실험1: 구성 비율은 작지만. 특이 발현 정도가 매우 큰 유전자들이 포함된 pathway를 각 방법이 유의하다고 판단할 수 있는가?

실험2: 특이 발현 정도가 크지는 않지만, 특이 발현 유전자의 구성 비율이 높은 pathway를 각 방법이 유의하다고 판단할 수 있는가?

또한, 모의실험을 위한 설계는 다음과 같다.

1단계 : 유전자를 발현값과 phenotype간 상관계수를 기준으로 정렬한다.

2단계 : 상관계수에 따라 집단을 만든다.

A 집단 - 상관계수가 큰 유전자들의 집단

B 집단 - 상관계수가 작은 유전자들의 집단

3단계 : A 집단에서 $n*a$개, B 집단에서 $n*(1-a)$개를 랜덤하게 뽑아 n개의 유전자를 가진 pathway를 만든다 $(0 < a < 1)$.

4단계 : 3단계에서 만들어진 pathway를 각 방법을 통해 유의 수준 0.05에서 유의한지 확인한다.

5단계 : 3단계–4단계를 N번 반복해 만들어진 N개의 pathway를 각 방법이 얼마나 유의하다고 판단하는지 비교한다.

표 4.1: 실험1에 대한 결과

포함비율(a)	PAGE	GSEA	A1	AT1	A2	AT2
0.05	0%	16%	0%	68%	0%	62%
0.10	6%	38%	18%	100%	40%	100%
0.15	48%	68%	100%	100%	96%	100%
0.20	80%	92%	100%	100%	100%	100%
0.25	90%	98%	100%	100%	100%	100%

표 4.2: 실험2에 대한 결과

포함비율(a)	PAGE	GSEA	A1	AT1	A2	AT2
0.4	28%	20%	100%	100%	100%	100%
0.5	26%	42%	100%	100%	100%	100%
0.6	40%	44%	100%	100%	100%	100%
0.7	34%	30%	100%	100%	100%	100%
0.8	56%	50%	100%	100%	100%	100%

실험1을 확인하기 위하여 A 집단은 ALL 자료 중 상관계수의 절대값이 0.4 이상(상위 1.5%)인 유전자의 집단으로, B 집단은 상관계수의 절대값이 0.1 이하(하위 40%)인 유전자의 집단으로 지정하였다. Pathway의 유전자 보유수(n)는 GSEA와 중심극한정리를 사용하는 PAGE, 그리고 포함 비율(a)을 고려하여 40으로 하였으며, 총 시행횟수(N)는 50번 그리고 A 집단의 포함비율(a)은 0.05, 0.1, 0.15, 0.2, 0.25 로 하였다.

실험1은 A 집단의 유전자를 포함한 유의한 pathway임에도 불구하고 표 4.1의 N번 반복 실험 중 각 pathway가 유의하다고 판단된 상대비율을 보면 PAGE나 GSEA의 경우 특이발현 유전자의 구성 비율이 작으면, 유의하다고 판단될 확률 또한 작아지는 것을 볼 수 있다. 이 결과를 통해 평균을 사용하는 PAGE의 경우, 비율이 작아짐에 따라 그 유전자가 차지하는 비중도 작아지는 것을 확인할 수 있었고, 상대적 위치에 의존하는 GSEA는 상대적으로 우선 순위에 있는 유전자의 비율이 높을수록 유의한 pathway로 판단될 확률이 높아짐을 확인할 수 있었다. 또한, 절삭을 하지 않은 방법인 GSA-A1 방법과 GSA-A2 방법은 A 집단 비율이 0.15 이상일 때는 기존 방법에 비해 매우 좋은 결과를 얻었지만, 0.05나 0.1일 때처럼 너무 낮은 경우에는 GSEA 보다도 좋지 않은 결과를 얻었다. 하지만, GSA-AT1과 GSA-AT2를 살펴보면, A 집단 내의 유전자의 비율과는 상관없이 다른 4가지 방법에 비해 월등히 좋은 결과를 나타냈다.

실험2에서는 A 집단은 ALL 자료 중 상관계수의 절대값이 0.22 이상(상위 20%)인 유전자의 집단으로, B 집단은 상관계수의 절대값이 0.12 이하(하위 50%)인 유전자의 집단으로 지정하였다. Pathway의 유전자 보유수(n)은 GSEA와 중심극한정리를 사용하는 PAGE를 고려하여 30으로 하였으며, 총 시행횟수(N)는 50번 그리고 A 집단의 포함비율(a)은 0.4, 0.5, 0.6, 0.7, 0.8로 하였다.

표 4.2의 결과를 보면, PAGE는 A 집단의 유전자가 증가함에 따라 유의하다고 판단하는 경우도 증가하는 경향을 보였지만 예측의 정확성이 많이 낮았고, GSEA는 예측의 정확성이

많이 낮을 뿐 아니라 그 어떤 연관성도 찾아볼 수 없었다. 반면, 나머지 네 개의 방법은 절삭과 상관없이 모두 유의한 결과를 보여 절대치를 이용한 permutation의 우수성을 입증하였다.

4.2. 실제 자료를 이용한 실험

모의실험과 더불어 실제 자료를 이용한 실험도 시행해 보았다. 실제 자료를 훈련군과 시험군으로 일정한 비율로 랜덤하게 나누어, 훈련군의 표본을 사용해 각 pathway의 유의성을 분석한 결과를 시험군의 표본을 사용해 상호검증(cross validation)을 실시하였다.

실제 자료 분석 또한 ALL 자료를 사용했으며, 훈련군은 정상 표본 50개와 백혈병 표본 24개로 이루어졌고, 시험군은 정상 표본 24개와 백혈병 표본 13개로 이루어졌다. 유전자는 12625개 모두를 사용했으며, pathway는 189개 중 유전자 보유수가 25개에서 250개 사이인 107개의 pathway만을 사용하였다. Pathway는 계층적 구조를 띄고 있기 때문에 큰 pathway 안에 작은 pathway들이 포함되어 있는 경우가 많아 보유 유전자가 많은 pathway는 큰 의미를 가지지 못하므로 그 상한을 250으로 제한하였다. 또한, 유전자 보유수가 너무 작은 pathway는 유전자 집단 분석의 의미를 살리지 못하며, 기존의 GSEA 분석 시스템으로 25개 미만인 것을 분석하지 못하므로 각 방법 간의 비교를 위해 유전자 보유수가 25개 미만인 pathway를 제외하였다.

결과를 검증하기 위해서는 시험군의 표본을 이용해서 집단을 판별한 후 실제와 얼마만큼 일치하는가를 보았으며, 집단을 판별하는 방법으로는 Tibshirani 등 (2002)에서 사용한 LDA(Linear Discriminant Analysis) 방법을 사용하였다. LDA 방법은 시험군의 표본에 대하여 각 집단별 판별점수를 구하고 그 중 가장 작은 점수를 가진 집단을 선택하는 방법으로, 시험군 표본의 유전자 발현값을 $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)$라 할 때, k 집단의 판별점수는 다음과 같다.

$$\delta_k(\mathbf{x}^*) = \sum_{i=1}^{p} \frac{(\mathbf{x}_i^* - \bar{x}_{ik})^2}{(s_i + s_0)^2} - 2\log \pi_k, \tag{4.1}$$

여기서, $\bar{x}_{ik}$는 훈련군 표본의 k 집단 내 i 유전자 발현값의 평균이고, s_i는 훈련군 표본의 i 유전자 발현값의 표준편차이다. π_k는 k 집단의 초기 확률이며, s_0은 분산이 너무 작아 판별점수가 커지는 것을 방지하기 위한 보정값이다.

표 4.3는 ALL자료의 훈련군 표본 74개를 대상으로 각 방법별 분석을 통해 백혈병의 발생에 영향을 미치는 pathway를 찾아본 결과와 시험군의 표본 37개를 LDA를 사용하여 집단을 분류하였을 때, 예측 오류를 측정한 것이다.

PAGE에서는 모두 19개의 pathway가 유의하다고 판단되었으며, 선택된 pathway에 속한 유전자들만 이용하여 시험군의 37개 표본의 백혈병 발생 여부를 예측해본 결과, 평균 6.63개의 표본을 잘못 예측하였다. GSEA에서는 Butanoate metabolism pathway만 유의하다고 판단되었으며, 이 pathway에 속한 유전자들만 이용하여 시험군의 37개 표본의 백혈병 발생 여부를 예측해본 결과, 6개의 표본을 잘못 예측 하였다. GSA-AT1에서는 모두 19개의 pathway가 유의하다고 판단되었으며, 선택된 pathway에 속한 유전자들만 이용하여 시험군의 37개 표본의 백혈병 발생 여부를 예측해본 결과, 평균 6.42개의 표본을 잘못 예측하였다. GSA-AT2에서는 모두 17개의 pathway가 유의하다고 판단되었으며, 선택된 pathway에 속한 유전

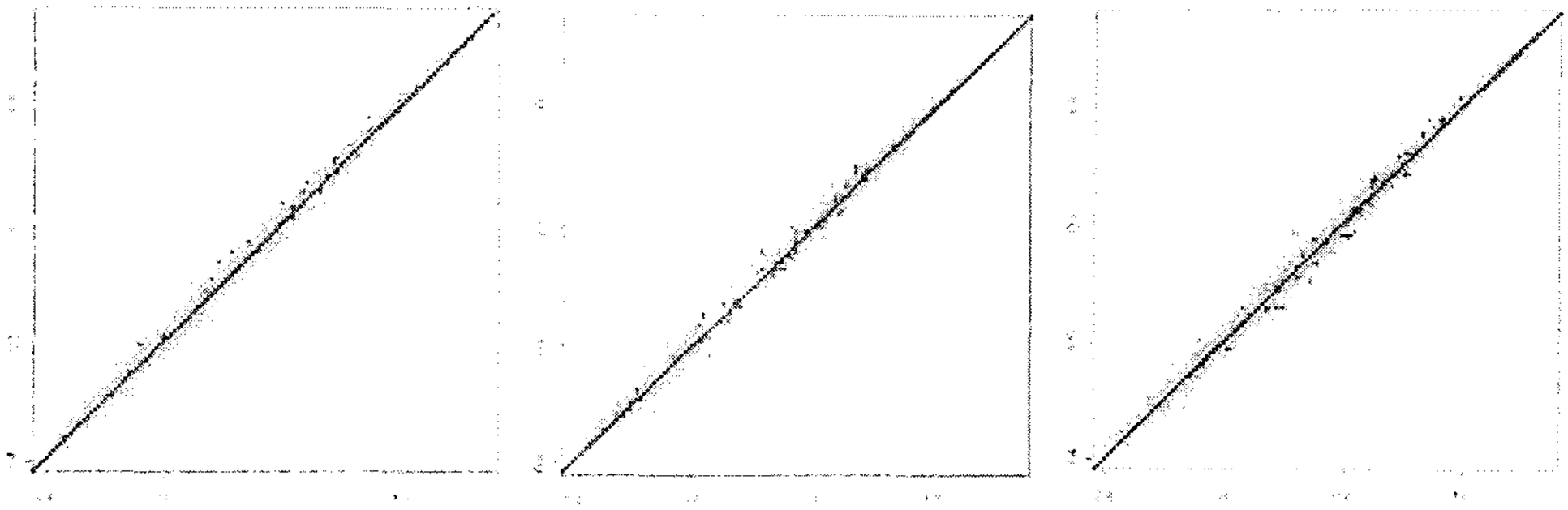
표 4.3: 각 방법에서 선택된 pathway의 비교

유전자 집합	유전자 수	p-value (Bonferroni 보정)				예측 오류
		PAGE	GSEA	GSA-AT1	GSA-AT2	
TGF-beta signaling pathway	136	0.0013	.	.	.	3/37
Cell cycle	185	0.0006	.	0.0107	0.0107	4/37
ECM-receptor interaction	121	0.0073	.	.	.	4/37
Galactose metabolism	29	.	.	0.0107	0.0107	4/37
Notch signaling pathway	44	.	.	0.0107	<0.001	4/37
Tryptophan metabolism	72	.	.	0.0321	.	4/37
Glycolysis / Gluconeogenesis	67	0.0412	.	0.0107	0.0214	5/37
Lysine degradation	39	0.0184	.	<.0001	0.0107	5/37
Antigen processing and presentation	97	0.0086	.	0.0214	.	6/37
Butanoate metabolism	39	0.0012	0.043	<.0001	0.0214	6/37
Glutamate metabolism	32	.	.	<.0001	.	6/37
Jak-STAT signaling pathway	203	0.0179	.	.	.	6/37
Pyruvate metabolism	43	.	.	0.0107	0.0321	6/37
Citrate cycle (TCA cycle)	29	0.0203	.	.	.	7/37
Colorectal cancer	149	.	.	.	<.0001	7/37
Fatty acid metabolism	53	0.0001	.	0.0107	<.0001	7/37
Folate biosynthesis	32	.	.	<.0001	0.0107	7/37
Purine metabolism	152	6.0E−7	.	<.0001	<.0001	7/37
Type I diabetes mellitus	68	1.9E−5	.	.	.	7/37
Valine, leucine & isoleucine degradation	52	5.7E−7	.	<.0001	<.0001	7/37
Alanine and aspartate metabolism	31	0.0087	.	.	<.0001	8/37
Cell Communication	129	0.0237	.	.	.	8/37
Oxidative phosphorylation	98	7.6E−9	.	0.0107	0.0214	8/37
Pyrimidine metabolism	82	<.0001	.	<.0001	<.0001	8/37
Selenoamino acid metabolism	29	.	.	0.0428	0.0214	8/37
Propanoate metabolism	31	0.0215	.	<.0001	<.0001	9/37
Proteasome	37	0.0137	.	0.0214	.	11/37

자들만 이용하여 시험군의 37개 표본의 백혈병 발생 여부를 예측해본 결과, 평균 6.47개의 표본을 잘못 예측하였다. 일반적으로 유한 집단에서 표본을 추출하는 permutation에 의존하는 검정은 모수적 검정에 비해 통계적 유의성을 높이기에 상대적으로 불리함에도 불구하고 GSA-AT는 PAGE보다 높은 유의성으로 두 군간 발현의 차이를 보이는 pathway들을 검색하였다는 것을 표 4.3에서 볼 수 있었다.

또한, 기존의 방법으로는 선택되지 않았었고 GSA-AT1과 GSA-AT2의 두가지 방법에서 추가로 유의하다고 판단된 8개 pathway 각각을 이용하여 시험군의 37개 표본을 예측해본 결과, 평균 5.75개의 표본을 잘못 예측하였으며, 그림 4.1은 새로운 방법에서만 유의하다고 판단한 pathway 중 예측 오류가 가장 작은 3개의 pathway에 대한 산점도이다. 산점도에서 가로축은 log(정상표본의 평균 발현값), 세로축은 log(당뇨표본의 평균 발현값)을 나타낸다.

먼저, 갈락토오스(젖당의 성분)의 대사활동에 영향을 미치는 유전자로 구성된 Galactose



(a) (b) (c)

그림 4.1: Pathway들의 산점도 ((a) Galactose metabolism, (b) Tryptophan metabolism, (c) Notch signaling)

metabolism은 시험군 표본으로 예측한 결과 37개의 표본 중 33개를 정확히 예측하였으며, 산점도를 통해, 특이 발현의 정도가 강하지는 않지만 특이 발현 유전자가 약 40–50% 정도로 많은 비중을 차지하는 것을 볼 수 있다. 다음으로, 트립토판(동물의 영양에 필요한 필수 아미노산)의 대사활동에 영향을 미치는 유전자로 구성된 tryptophan metabolism은 시험군 표본으로 예측한 결과 37개의 표본 중 33개를 정확히 예측하였으며, 산점도를 통해서도 특이 발현 유전자를 다수 포함하고 있는 것을 알 수 있다. 하지만, 전체적으로 유전자 수가 너무 많아 특이 발현 유전자의 비중이 작아져, 기존 방법에서는 발견되지 못했다.

위의 두 pathway는 새로운 방법에서 시도한 절삭의 효과를 통해 새로이 발견된 pathway라 한다면, 세 번째 그림의 Notch signaling pathway는 절대값의 효과를 통해 새로이 발견된 pathway라 할 수 있다. 이 pathway 역시 시험군 표본으로 예측한 결과 37개의 표본 중 33개를 정확히 예측하였으며, 산점도를 통해 특이 발현 유전자가 많음에도 불구하고 양 옆으로 고르게 퍼져 있는 것을 볼 수 있다. 즉, 양적 특이 발현 유전자와 음적 특이 발현 유전자가 공존하여, 그 합이 상쇄가 됨으로써 기존의 방법에서는 발견되지 못한 경우라 할 수 있다.

5. 결론

새로 제시한 유전자 집단 분석 방법인 GSA-AT는 비교대상인 두 군 중 어느 군에서 더 발현이 많이 되었는지를 반영하는 방향성(부호)은 무시하고 특이 발현 정도가 큰 유전자들에 가중을 둔 대푯값을 사용한 것이 특징으로, 모의실험과 실제 자료 실험에서 모두 좋은 결과를 보였다.

여러 특성의 유의한 pathway를 잘 검색할 수 있는지 알아보기 위한 모의실험에서 GSA-AT 방법은 설계 조건에 상관없이 훨씬 안정적이며 정확한 결과를 보였으며, 실제 자료를 이용한 상호검증에서도 기존 방법에서 찾아내지 못한 유의한 pathway를 찾아냄을 확인하였다. 확실한 생물학적 검증없이 pathway에 속한 유전자들의 산포도와 예측오류만을 통하여 pathway의 유의성을 단정 지을 수 없지만, 이 실험들을 통해 GSA-AT 방법이 기존의 방법에 비해

서도 경쟁력 있는 방법 중 하나라고 생각한다.

계산 과정에서 매번 permutation 마다 자료 절삭을 하기 때문에 시간이 더 걸린다는 단점이 있지만 GSA-AT의 특성을 살리면서 계산을 용이하게 할 수 있는 방법에 관하여 계속 연구하고자 한다.

참고문헌

Curtis, R. K., Oresic, M., Vidal-Puig, A. (2005). Pathways to the analysis of microarray data, *Trend in Biotechnology*, **23**, 429–435.

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

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

Manoli, T., Gretz, N., Grone, H. J., Kenzelmann, M., Eils, R. and Brors, B. (2006). Group testing for pathway analysis improves comparability of different microarray datasets, *Bioinformatics*, **22**, 2500–2506.

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 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes, *Nature Genetics*, **34**, 267–273.

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 of the United States of America*, **102**, 15545–15550.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression, In *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567–6572.

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

[2008년 3월 접수, 2008년 5월 채택]

Gene Set Analysis - Absolute and Trim*

Kwang-Hyun Lee[1] Sunho Lee[2]

ABSTRACT

Initial work of microarray data analysis focused on identification of differentially expressed genes, and recently, the focus has moved to discovering significant sets of functionally related genes. We describe some problems of GSEA and PAGE, and propose a modified method to identify significant gene sets. The results based on a simulated experiment and real data analysis using a set of publicly available data show the superiority of the newly proposed method, GSA-AT, in detecting significant pathways with the accurate prediction.

Keywords: Microarray, gene-set, PAGE, GSEA, GSA-AT.

* This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(KRF-2006-C-00020)

1) Graduate student, Dept. of Applied Statistics, Sejong University, Gunja-dong, Kwangjin-gu, Seoul 143-747, Korea.
E-mail: gusl-2@nate.com

2) Corresponding author. Professor, Dept. of Applied Statistics, Sejong University, Gunja-dong, Kwangjin-gu, Seoul 143-747, Korea.
E-mail: leesh@sejong.ac.kr