

cDNA 마이크로어레이 자료에서 유의한 유전자  
선택을 위한 통계적 방법의 비교

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

방 정 숙

# cDNA 마이크로어레이 자료에서 유의한 유전자 선택을 위한 통계적 방법의 비교

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2004년 6월 일

연세대학교 대학원

의학전산통계학협동과정

의학통계학전공

방 정 숙

## 감사의 글

새로운 학문에 대한 호기심과 기대를 안고 대학원에 진학한지 엇그제 같은데, 어느새 시간이 흘러 졸업의 시간이 다가오고 있습니다. 2년 남짓한 시간 동안 낯선 환경에 고생한 적도 많았지만 부족한 제게는 많은 것을 알게 해준 소중한 시간으로 기억될 것입니다.

의학통계학의 길을 열어주시고 석사과정을 무사히 마칠 수 있도록 인도해주신 김동기 선생님께 감사드립니다. 유전통계라는 새로운 분야를 알게 해 주시고 지속적인 관심과 도움을 주신 임길섭 선생님께도 진심으로 감사드립니다. 바쁘신 가운데도 부족한 제 논문이 완성되기까지 도와주시고 통계이론과 적용에 대해 깊이 알게 해주신 이학배 선생님께도 감사드립니다. 지금까지 통계학을 공부할 수 있는 계기가 되어주시고 도움의 손길을 주신 박철용 선생님, 김동건 선생님께도 감사드립니다.

대학원 생활동안 항상 든든한 기둥이 되어주시고 익숙하지 않은 타지 생활에 적응하는데 많은 배려와 도움을 주신 기준오빠, 이 논문이 완성되기까지 많은 것을 알려주시고 지속적인 관심으로 부족한 저를 돌봐주신 성민오빠에게 이 자리를 빌어 감사의 마음을 전합니다. 힘들거나 곤란한 일이 있을 때 도움을 주시고 따뜻한 격려와 질책을 아끼지 않은 무영오빠, 미영언니, 부족한 저를 따뜻한 시선으로 지켜봐 주시고 술잔을 같이 하며 고민을 덜어주시고 항상 제게 힘이 되어주신 민지언니에게도 감사의 말을 전하고 싶습니다. 늘 밝은 모습으로 따뜻하게 대해주신 찬미언니, 우선오빠, 원열오빠, 봉섭오빠에게도 감사드립니다. 웃음과 재치로 늘 즐겁게 해 준 수옥씨, 신영씨, 은혜, 혜리씨, 많은 것을 함께 하지 못해서 미안한 소연, 민진에게도 감사의 마음을 전합니다. 멀리서도 언제나 관심과 격려를 아끼지 않는 사랑하는 친구 은숙, 소현, 타지에서 다시 만나 동지애를 느끼게 해준 은영, 서로의 고민을 덜어주고 위안을 주는 언니 같은 현주, 선영, 재현선배, 재훈선배, 종무선배에게 고마움의 뜻을 전하고 싶습니다.

지금의 저를 세상에 있게 해주시고 항상 관심과 걱정의 손길을 놓지 않으시는 부모님께 깊이 감사드립니다. 언제나 힘이 되어주고 의지할 수 있게 해 주는 오빠, 사랑스러운 조카 한석이 건강하게 자라기를 기원하며 언니와 형부에게도 감사의 말을 전합니다. 사회인으로 새로운 출발을 하려는 지금 제 삶의 버팀목이 되어주신 모든 분들께 머리 숙여 감사드리며 늘 최선을 다하고 노력하는 모습을 약속드립니다.

2004년 7월  
방 정 숙 올림

# 차 례

표 차례 .....	iii
그림 차례 .....	iv
국 문 요 약 .....	v
제1장 서 론 .....	1
1.1 연구 배경 .....	1
1.2 연구 목적 및 방법 .....	2
제2장 cDNA 마이크로어레이 .....	4
2.1 cDNA 마이크로어레이에 대한 소개 .....	4
2.2 표준화 (Normalization) .....	6
2.3 실험설계 .....	9
제3장 유의한 유전자 선택방법 .....	11
3.1 유의한 유전자선택 .....	11
3.2 통계량을 이용한 유전자 선택방법 .....	12
3.2.1 T-통계량 .....	13
3.2.2 B-통계량 .....	16
3.2.3 D-통계량 .....	18
3.3 점수를 이용한 유전자 선택방법 .....	22
3.3.1 TNoM 점수 .....	23
3.3.2 INFO 점수 .....	25
3.3.3 Separation 점수 .....	28
3.4 요약 .....	30
제4장 실제자료에 적용 .....	31
4.1 실험 배경 .....	31
4.2 실제 자료 (ApoAI 실험자료) .....	32
4.3 실험 결과 .....	33

제5장 모의 실험을 통한 비교 .....	41
5.1 실험 배경과 실험 자료 .....	41
5.2 실험 결과 .....	42
제6장 결론 및 향후과제 .....	45
참 고 문 헌 .....	46
ABSTRACT .....	49

## 표 차례

표 1-1. 준거 설계 .....	9
표 1-2. 염료 교체 실험 .....	10
표 2. 유의한 유전자의 비율 과 오류율(FDR)의 관계 .....	21
표 3. 유전자 선택방법들의 요약 .....	30
표 3-1. 유의한 유전자들에 해당하는 통계량과 유전자 순위 .....	35
표 3-2. 유의한 유전자들에 해당하는 점수와 유전자 순위 .....	36
표 4. 모의실험 자료(정규분포)의 평균과 표준편차들 .....	41
표 5. 각 방법의 유의한 유전자 평균 순위 비교 .....	44

## 그림 차례

그림 1. cDNA 마이크로어레이 실험과정 .....	5
그림 2-1. M-A 그림과 Box 그림 (표준화하기 전의 자료) .....	7
그림 2-2. M-A 그림과 Box 그림 (표준화 후의 자료) .....	8
그림 3-1. ApoAI 실험자료의 2-표본 동일분산 T-통계량의 히스토그램 .....	14
그림 3-2. ApoAI 실험자료의 2-표본 동일분산 T-통계량의 정규 Q-Q 그림 .....	14
그림 4. ApoAI 실험자료의 B통계량과 M값에 대한 그림 .....	17
그림 5. ApoAI 실험자료의 $d(i)$ 와 $d_E(i)$ 의 산점도 .....	21
그림 6. ApoAI 실험자료의 Info방법에 의해 관측된 유의한 유전자들의 수와 임의로 표시되어진 자료에서의 기대유전자수의 비교 .....	27
그림 7. 정상 쥐들(위)과 ApoAI 녀아웃 쥐들(아래)의 M값들의 히스토그램 .....	33
그림 8. 2-표본 동일분산 T-통계량의 정규 Q-Q그림에서 유의한 유전자들의 위치 .....	37
그림 9. B-통계량과 M값 그림에서 유의한 유전자들의 위치 .....	37
그림 10. TNoM 방법에서 유의한 유전자수와 임의로 표시되어진 자료에서의 기대유전자수의 비교. 유의한 유전자들의 위치(아래) .....	38
그림 11. 각 방법에 의해 선택된 유전자들을 이용한 분류분석의 정확성 비교 .....	39
그림 12. 유전자 선택방법들의 ROC 곡선과 B-통계량방법의 ROC 곡선 .....	43

## 국 문 요 약

### cDNA 마이크로어레이 자료에서 유의한 유전자 선택을 위한 통계적 방법의 비교

유의한 유전자는 특정한 실험 조건의 특성을 나타내주는 발현수준의 유전자를 의미한다. 이 유전자들은 여러 집단 간의 발현수준에서 유의한 차이를 보여주며, 실제로 집단 간의 차이를 유발하는 유전자일 확률이 높아 특정 생물학적 현상과 관련 있는 정보적 유전자를 찾는 연구에 이용될 수 있다. 그리고 유전자 발현자료를 분석할 때에는 특수한 자료 구조와 DNA 마이크로어레이 실험의 특성상 많은 오차요인들로 인하여 기존의 통계학 방법들을 적용하기 어려운데, 이 때 자료의 차원을 축소시키기 위하여 유전자선택을 이용할 수 있다. 유의한 유전자를 선택하는 방법으로는 T-통계량, 로그 사후 우도비 B-통계량, SAM를 이용하는 D-통계량, TNoM점수, Info점수, Separation점수를 이용하는 방법이 있다. 본 논문에서는 다양한 유전자 선택방법들에 대해 비교 분석하고, 실제 자료와 모의실험 자료를 이용하여 각 자료에 적합한 방법에 대하여 알아보고자 한다.

---

핵심되는 말 : 유전자 발현자료, 유의한 유전자, T-통계량, B-통계량, D-통계량, TNoM점수, Info점수, Separation점수



# 제1장 서론

## 1.1 연구 배경

세계의 관심이 집중되었던 인간유전체(Human Genome) 프로젝트의 목표는 인간이 보유한 22쌍의 상염색체와 X, Y로 표현되는 성염색체 등에 포함된 유전자의 염기서열 해독과 유전체지도의 작성이었다. 성공적인 연구의 결과로, 2002년 2월 인간유전체지도가 완성되었다. 이는 2000년 6월에 발표되었던 초안 위에 생물학적 표지를 심어 특정 유전자의 위치를 표시한 99% 수준의 상세 지도로, 당시 누락되었던 부분에 대해서도 대폭 보완된 것이다. 지도의 완성으로, 인간유전체로부터 각 유전자들의 생체기능을 밝히고 개인, 인종, 생물간 유전체 정보를 비교하여 그 차이점으로 인한 생체 기능의 차이를 규명하는 포스트 유전체(Post-Genome) 시대가 열리게 되었다.

인간유전체 프로젝트를 통해 여러 개체의 완전한 유전자 서열정보 등 방대한 양의 자료들을 밝혀내면서 이제는 자료 자체보다는 얻어지는 방대한 양의 자료를 어떻게 해석하고 이용하는가가 더 어렵고 당면한 문제로 부각되었다. 기존의 전통적인 분자생물학에 의한 연구방법은 단일 유전자에 대한 실험에 근거하여 진행되기 때문에 그 결과물 또한 매우 제한적이고 유전자의 전체 움직임을 관찰하기에는 한계가 있다. 반면에, DNA 마이크로어레이(microarray 또는 microchip)는 하나의 칩(chip)상에서 전체 유전체(genome)의 발현양상을 탐색할 수 있고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰할 수 있다. 또한 짧은 시간에 엄청난 분량의 자료를 만들어낼 수 있는 가능성을 지니고 있어, 최근 주목받고 있다.

DNA 마이크로어레이 실험과정을 거쳐 생성되는 유전자 발현자료들을 이용하여, 생물학, 의학 및 약학 분야의 연구자들은 유전자의 기능, 유전자의 발현 및 통제 방식, 그리고 유전체구조의 비밀 등과 같은 문제를 해결하기 위해 노력하고 있다. 그 중에서도, 관심 있는 생물학적 현상과 관련 있는 정보를 제공해 주는 유전

자들을 찾는 것은 중요한 과제라 할 수 있다. DNA 마이크로어레이 실험의 특성상 원하는 만큼의 반복실험을 할 수 없기 때문에 관측치의 개수보다 변수의 개수가 더 많은, 기존의 통계학 방법들을 적용할 수 없는 새로운 유형의 자료구조를 갖게 된다. 판별분석, 군집분석, 분류분석과 같은 통계방법들을 이용하기 위해서는 자료의 차원을 축소시켜주어야 하는데, 이 때 수천 개의 유전자들 중에서 각 집단간의 발현양상에서 유의한 차이를 보이는 유전자들만을 선택하여 자료로 이용하게 된다. 이 유의한 유전자들은 정상 조직과 이상 조직의 차이, 서로 다른 질병의 차이, 질병의 소분류들의 차이 등을 밝히거나 새로운 표본조직의 생물학적 현상을 예측하는데 이용할 수 있어, 이들을 찾기 위해 다양한 유전자 선택(gene selection)방법들이 개발되었고 새로운 방법에 대한 연구들도 활발히 이루어지고 있다.

## 1.2 연구 목적 및 방법

본 논문에서는 DNA 마이크로어레이 실험의 유전자 발현자료들을 이용한 유전자 선택방법(Gene-Selection)들에 대해 비교 분석하여 평가하고자 한다. 이 방법들은 특정 생물학적 현상과 관련 있는 정보적 유전자를 찾거나 기존의 통계학 방법들을 이용하기 위해 자료의 차원을 감소시키는 데 이용할 수 있다. 유의적인 유전자들을 찾기 위해 여러 방법들이 개발되었는데, 그 중 대표적인 것으로는 TNoM, INFO, Separation 점수 방법과 T-검정, Bayes, SAM 방법 등이 있다. 본 논문에서는 이를 크게 통계량을 이용하는 방법과 점수를 이용하는 방법으로 구분하였다.

유전자 선택을 할 경우에는 유전자 발현자료의 많은 오차요인들과 관측치의 개수보다 변수의 개수가 더 많은 자료구조의 특성상, 한 가지 방법이 아니라 여러 통계적 방법을 이용하는 게 일반적이다(Kaminski and Friedman, 2002). 따라서, 본 논문에서는 여러 유전자 선택방법들을 고려하고 있으며, 실제 자료와 모의실험 자료를 이용하여 각 자료에 적합한 선택방법에 대하여 알아보하고자 한다.

제 1장에서는 cDNA 마이크로어레이에 대해 먼저 소개하고 유전자 발현자료를 정제하는 표준화 방법과 설계방법들을 소개한다. 그리고 제 3장에서 통계량을 이용하는 유전자선택방법과 점수를 이용하는 유전자선택방법들을 소개한다. 4장과 5장에서는 실제자료와 모의실험에 대한 결과를 보여주고 다각적 비교를 시도한다. 기존에 알려진 유의한 유전자들을 이용하여 각 방법의 유의한 유전자식별능력을 비교하고, 선택된 유전자들로 분류분석을 하였을 때 새로운 표본의 집단을 예측력이 좋은 방법들에 대해 살펴본다. 6장에서는 결론 및 향후과제에 대해 논의하기로 한다. 마지막 7장에서 본 논문에서 사용한 프로그램에 대하여 소개한다.

## 제2장 cDNA 마이크로어레이

### 2.1 cDNA 마이크로어레이에 대한 소개

22쌍의 상염색체와 X, Y로 표현되는 성염색체에 있는 DNA(deoxyribonucleic acid)는, 당과 인산분자로 구성된 두 개의 가닥이 염기에 의해 연결되어 꼬여진 이중나선구조를 형성하고 있다. 아데닌(A), 구아닌(G), 시토신(C), 티민(T)의 4 가지 염기의 배열순서에 따라, 유전자는 독특한 기능을 갖게 된다.

DNA는 어떤 생물의 어떤 세포에서도 그 생물의 전체 유전정보 분량을 갖는다. 세포분열이 일어나기 전에 DNA는 복제되므로, 분열 후 2개의 딸세포는 DNA를 똑같이 함유한다. 이 때 DNA의 유전정보는 mRNA(전령 RNA)를 통해서 단백질로 전달된다. 여기서 mRNA로 DNA가 복사되는 과정을 '전사(Transcription)'라고 하고, mRNA로부터 단백질이 합성되는 과정을 '번역(Translation)'이라고 한다.

분자생물학적인 실험기술과 robotics 기술의 발달로 인해 등장하게 된 DNA 마이크로어레이 또는 마이크로칩(Microchip)은 유전자 형별분석이나 유전자의 발현정도를 조사하는데 이용되고 있다. 칩(chip)은 수 천개의 DNA를 작은 공간에 고밀도로 붙여놓은 것이다. DNA 마이크로어레이를 통하여 하나의 칩(chip)상에서 전체 유전체(genome)의 발현양상을 탐색할 수 있게 되었고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰이 가능하게 되었다. DNA 칩은 그 제작과 기능에 cDNA 마이크로어레이 칩과 올리고뉴클레오티드 칩(Oligonucleotide chip)으로 나누어 질 수 있다. 올리고뉴클레오티드(Oligonucleotide)는 한 유전자에서 추출한 15-25개의 염기들로 이루어진 DNA서열(DNA sequence)로, 포토리소그래피(photolithography) 방법을 이용하여 칩에 결합시켜 제작한다. 특정유전자의 돌연변이를 찾거나 유전자의 염기서열을 밝히는데 주로 사용된다. cDNA 마이크로어레이인 경우 DNA 클론들을 PCR에 의해 증폭하고 정제한 후, 슬라이드(array slide) 위에 일정 크기로 찍어서(spotting) 칩을 제작한다. 실험하고자 하는 두 개의 다른 조직으로부터 mRNA

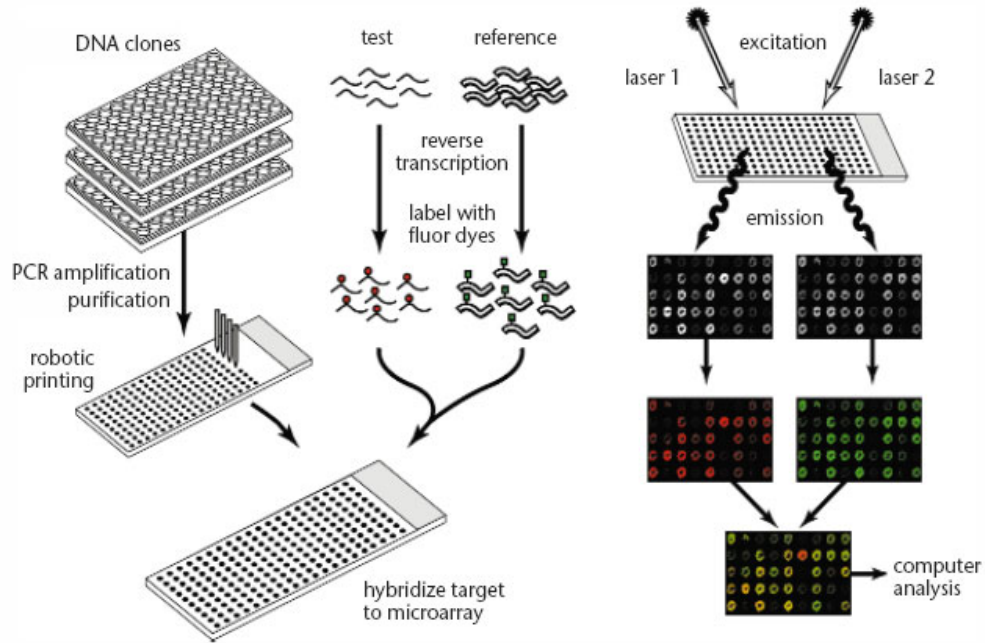


그림 1. cDNA 마이크로어레이 실험과정

를 추출하여, RT-PCR(reverse transcription polymerase chain reaction)방법을 통해 cDNA로 역전사(reverse transcription)시킨다. 이 과정에서 다른 색깔의 형광 시료로 각각 표지하여 빨강색(Cy5)이나 녹색(Cy3)을 띤 cDNA를 합성한다. 이렇게 합성된 두 가지 cDNA를 같은 양으로 섞어서 제작해둔 칩에 결합시키고 결합이 안 된 cDNA는 씻어낸다. 레이저 형광 스캐너(laser fluorescence scanner)를 이용하여 각 유전자의 형광정도를 읽어 들이는데, 이는 그 유전자의 발현정도를 알려주는 것으로 Genepix, Imagen 등 이미지 처리 소프트웨어를 이용하여 수치화한다. 이 모든 과정([그림 1]<sup>(1)</sup>)을 거쳐 얻어진 자료를 유전자 발현자료라고 한다.

이미지 분석(image analysis) 과정을 통해, 각 유전자의 발현수준은 Foreground red, Background red, Foreground green, Background green intensity의 4 가지 수

(1) Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M. (1999)

치로 측정되어지는데, 이 후 분석에서는 바탕보정과 표준화과정을 거친 값을 사용한다(Dudoit et al., 2002).

## 2.2 표준화 (Normalization)

바탕보정(background correction)은 이미지의 품질(quality)을 개선해 편의를 줄이기 위한 것으로, 일반적으로 녹색, 빨강색의 foreground 발현강도에서 background 발현강도를 감하는 방법을 사용한다. 바탕보정을 한 발현강도를 다음과 같이 R, G로 표기한다.

$$\begin{aligned}R &= \log(\text{Foreground Red} - \text{Background Red}) \\G &= \log(\text{Foreground Green} - \text{Background Green})\end{aligned}$$

R, G는 다음과 같은 변환을 통해 M, A로 나타낼 수 있다. 각 유전자의 M값은 녹색과 빨강색의 발현 강도의 비(ratio)를 나타내는데, 군집분석, 판별분석과 같은 통계적인 분석에서는 주로 M을 이용한다.

$$\begin{aligned}M &= \log_2(R/G) = \log_2 R - \log_2 G \\A &= \log_2(\sqrt{R \times G}) = (\log_2 R + \log_2 G)/2\end{aligned}$$

기초적인 분석 단계에서 그래프진단(graphical diagnostic)을 통해 실험 자료의 특성을 파악해야 하는데, 통상적으로 R-G그림, logR-logG그림, M-A그림, Box그림 등이 이용되어진다(Smyth et al., 2003).

유전자 발현수준(gene expression level)를 측정하기 위한 마이크로어레이 실험에서는 실험 준비 또는 실험 과정 중 많은 계통적 변동(systematic variation)이 있을 수 있다. 여러 비생물학적 요인들이 변동을 유발시키는데, 첫째 빨강색과 녹색

염료는 DNA 합성단계에서 똑같은 효율로 결합되지 않을 수 있으며 둘째 자료수집 단계에서 스캐닝 효율도 다를 수 있다. 이들은 두 형광물질 간의 상호작용이나 염료의 물리적인 성질에 의한 것으로 대체로 녹색 염료가 높은 형광강도를 보인다. 셋째, DNA 칩을 제작할 때에는 점적시간을 줄이기 위하여 여러 개의 핀을 동시에 사용한다. 이 때 특정 핀(pin, printing tip)을 사용한 유전자들이 다른 핀으로 점적된 유전자들에 비하여 전체적으로 높거나 낮은 값을 갖거나 이상치들을 많이 포함하는 경우가 발생할 수 있다. 이러한 차이를 핀 효과(pin effect)라고 하는데, 이를 보정해주지 않고 그대로 사용하게 되면 유의한 유전자의 대부분이 특정 핀에서만 선택되는 문제점이 있을 수 있다. 그 외에도 서로 다른 슬라이드의 자료가 있을 경우에는 슬라이드간의 이질성으로 실험상의 편이가 생길 수 있다. 이러한 편이들은 유전자 발현수준을 정확하게 측정하지 못하게 하여 잘못된 결과를 도출하게 될 수도 있다.

대다수(95%이상)의 유전자의 발현수준이 서로 비슷하며 극소수의 유전자만이 유의하게 다른 발현강도를 보일 것이므로, R-G, logR-logG 그림이 대각선(45°) 중심이거나, M-A 그림을 살펴보았을 때 수평축(0) 중심이 되어야 하는데, 실제로 이러한 경향을 보이는 경우는 거의 없다. 일반적으로 곡선 등의 비선형 추세가 나타나거나 점들이 한 쪽으로 치우친 모습을 하고 있다.

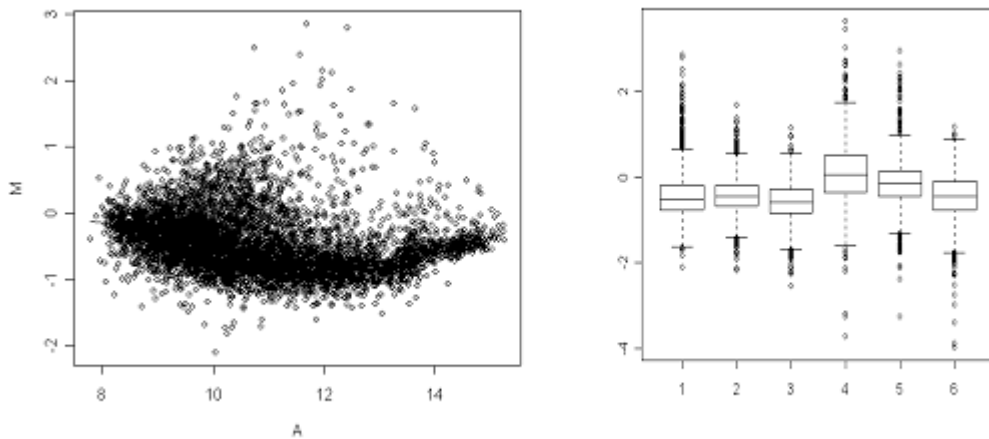


그림 2-1. M-A 그림과 Box 그림 (표준화하기 전의 자료)

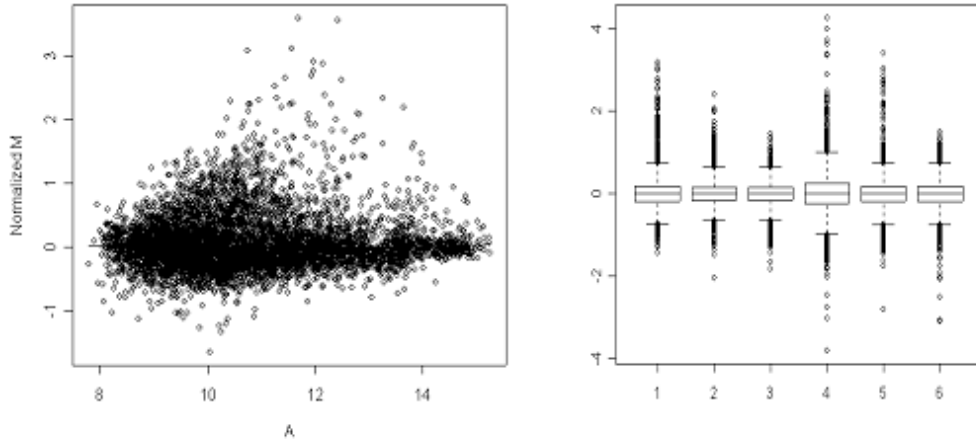


그림 2-2. M-A 그림과 Box 그림 (표준화 후의 자료)

두 염료 형광강도의 균형을 맞추고 슬라이드간의 발현수준 차이, 핀 효과, 스캐닝 효율 차이 등의 변동요인을 제거시켜 자료를 보정해 주어야 하는데, 이를 표준화(Normalization)라고 하며 통계적인 분석에 들어가기 전에 거쳐야 하는 과정이다. DNA 마이크로어레이 결과를 표준화하는 방법은 여러 가지가 있는데, M값의 중앙값(median) 또는 평균(mean)을 이용하여 M-A 그림에서 M의 중심을 0으로 이동시켜 자료를 보정하는 Global normalization, 자료를 핀(pin, print-tip, grid) 별로 보정하는 intensity-dependent normalization, 그 외에 Within-print tip group normalization, scale normalization 등이 있다. 여기서 intensity-dependent normalization은 수평축(0) 중심에서 벗어나는 정도가 유전자 발현강도에 따라 다를 수 있어 이를 lowess 함수를 이용하여 지역적으로 적합하는(robust locally linear fit) 방법이다. 그리고 scale normalization은 print-tip별 유전자 발현강도의 분산이 다른 경우 척도(scale)를 보정하여 편차(deviation)가 같도록 하는 방법인데, 이는 여러 슬라이드의 척도 보정으로 확장시킬 수 있다(Yang et al., 2002).



## 2.3 실험설계

DNA 마이크로어레이 실험에서 두 집단을 비교하고자 하는 경우, 준거 설계와 염료 교체 실험 설계방법을 사용한다.

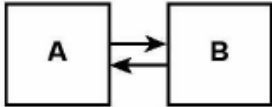
준거 설계(Reference Design)는 보편적으로 많이 사용하는 방법으로, 모든 어레이(array slide)에 대하여 녹색(빨강색)으로 염색한 동일한 준거군(common reference)을 배치하고 비교하려는 집단들을 각각 빨강색(녹색)으로 염색하여 실험한다. 동일한 준거군을 이용하기 때문에 새로운 집단에 관심이 있을 경우 실험을 확장하기 쉽다는 장점이 있다. 그러나 실제 분석에 이용하지 않는 준거군의 자료는 반복하여 관찰되고 반면에 관심 있는 집단들에 대해서는 한 번씩만 관찰되어 정확도가 떨어지고 정보의 손실이 큰 편이다.

표 1-1. 준거 설계

어레이 슬라이드 (Array Slide)			(1)
염료(Dye)	A	B	
녹색(Cy3)	준거군	준거군	
빨강색(Cy5)	실험군(A)	대조군(B)	

(1) Churchill, G.A. (2002) 화살표는 빨강색으로 염색된 집단을 향하고 있다

표 1-2. 염료 교체 실험

어레이 슬라이드 (Array Slide)			(1)
염료(Dye)	1	2	
녹색(Cy3)	실험군(A)	대조군(B)	
빨강색(Cy5)	대조군(B)	실험군(A)	

염료 교체 실험(Dye Swap Experiment)의 경우 첫 번째 어레이는 실험군을 빨강색으로, 대조군을 녹색으로 염색하여 실험하고 두 번째 어레이는 실험군을 녹색으로, 대조군을 빨강색으로 염색하여 실험한다. 각 처리를 2 번씩 관찰하게 되므로 염료의 차이(Dye-bias)로 인한 변동을 줄일 수 있다는 장점이 있다. 이 실험설계를 확장하여 비교하고자 하는 처리가 여러 개일 경우에도 사용할 수 있는데, 이를 루프 설계(Loop Design) 방법이라 한다. 가장 효율적인 방법으로 알려져 있으나 확장이 용이하지 않고 실험하기가 어려워 많이 사용되지는 않는다(Smyth et al., 2003; Kerr and Churchill, 2001).

(1) Churchill GA (2002) 화살표는 빨간 색으로 염색된 집단을 향하고 있다

## 제3장 유의한 유전자 선택방법

### 3.1 유의한 유전자선택

유의한 유전자는 특정한 실험 조건(집단)의 특성을 나타내주는 유전자발현수준의 유전자로 정의된다. 이 유전자들은 여러 집단 간의 발현수준에서 유의한 차이를 보여주며, 실제로 집단 간의 차이를 유발하는 유전자일 확률이 높아 특정 생물학적 현상과 관련 있는 정보적 유전자를 찾는 연구에 이용될 수 있다. 생물학, 의학 및 약학 분야의 연구자들에게 있어서, 관심 있는 생물학적 현상과 관련 있는 정보를 제공해 주는 유전자들을 찾는 것은 중요한 과제라 할 수 있으므로, 유의한 유전자선택의 필요성은 강조되어야 할 것이다. 그리고 유전자 발현자료의 특수한 자료 구조와 DNA 마이크로어레이 실험의 특성상 많은 오차요인들로 인하여 기존의 통계학 방법들을 적용하기 어려우므로, 유전자선택을 통하여 자료의 차원을 축소시켜야 한다.

초기 마이크로어레이 실험에서는 유의한 유전자를 찾기 위하여 배수변화(Fold Change)에 근거한 방법을 이용하였다. 이는 가장 단순한 방법으로 한 슬라이드에서 구한 빨강색과 녹색의 발현강도의 비(ratio)를 계산하여 그 절대값이 임의로 정한 경계값(2-배수나 3-배수)보다 크다면 두 집단 간에서 유의한 차이를 보이는 유전자로 간주하게 된다. 이 방법의 가장 큰 문제점은 유의한 유전자를 결정하기 위한 경계값(cutoff value)을 임의로 정하기 때문에, 자료에 적절치 않을 수 있다는 것이다. 일반적으로 DNA 마이크로어레이 자료는 낮은 발현강도의 유전자들이 높은 발현강도의 유전자들에 비해 분산이 큰 경향을 보이는데, 경계가 일정한 배수변화 방법은 이러한 통계적 변동을 고려하지 않기 때문에 유의하지 않은 유전자가 선택되거나 유의한 유전자를 놓칠 가능성이 커지게 된다. 이미 여러 연구자가 이 방법에 문제가 있음을 밝혔으며, 지금은 신뢰도가 낮은 방법으로 알려져 있다(Pan,

W., 2002; Draghici, S., 2002).

그 이후로 더 정교해지고 복잡한 통계학적 방법들이 제안되기 시작하였는데. 그 중 대표적인 것으로는 전통적인 2-표본 T-검정, B-통계량, SAM, TNoM, Info, Separation 점수를 이용하는 방법 등이 있다. 본 논문에서는 이를 크게 통계량을 이용하는 방법과 점수를 이용하는 방법으로 구분하여 설명한다.

### 3.2 통계량을 이용한 유전자 선택방법

비교하고자 하는 집단 간에서 유의한 차이를 보이는 유전자를 찾는다는 것은, 각 유전자에 대해 집단 간에서 유의한 차이를 보이는 유전자인지 아닌지를 검정한다는 것과 동일한 의미를 가진다. 이러한 통계학적 검정을 할 경우 귀무가설과 대립가설을 다음과 같이 정의한다.

$$H_0 : \mu_{g(T)} - \mu_{g(C)} = 0$$

$$H_1 : \mu_{g(T)} - \mu_{g(C)} \neq 0$$

여기서,  $\mu_{g(T)}$  는  $g$ 번째 유전자에 대한 실험군에서의 평균 발현 수준(Average levels of expression)으로서 정의되며, 마찬가지로  $\mu_{g(C)}$  는  $g$ 번째 유전자에 대한 대조군에서의 평균 발현 수준이다( $g = 1, 2, \dots, N$ ) (Pan, W., 2002).

유의한 유전자를 찾기 위하여, 우선 통계학적 검정을 위한 적절한 통계량을 설정한 후, 그 통계량들을 크기 순으로 나열한다. 그 다음에 할 일은 임계값(Critical-value)를 정하는 것인데, 어떤 유전자의 통계량이 이 값보다 크다면 유의한 유전자로 간주되어진다. 그러나, DNA 마이크로어레이 자료 분석에서 통계적으로 유의한 유전자들 전부를 추가분석 등에 사용하기에는 그 수가 너무 많으므로,

일반적으로 가장 유의할 것 같은 100개의 유전자나 50개의 유전자를 사용한다.

### 3.2.1 T-통계량

T-검정은 서로 다른 두 집단에서 어떠한 유전자들이 차이가 있는지를 알아보기 위하여 통계학적 검정을 이용하는 간단한 방법으로, 표준화 과정을 거친 후에 로그 변환한 유전자의 발현강도 자료들을 사용한다. 이 방법은 비교하려는 각 집단에서의 유전자들의 정규성(Normality) 가정에 근거한다.

염료 교체에 의한 마이크로어레이 자료인 경우, 단일 슬라이드에서 구한  $M = \log_2(R/G)$  값을 이용하여 실험군과 대조군을 직접 비교(Direct comparison)할 수 있다.  $g$ 번째 유전자에 대하여 t-통계량(t-statistics)을 구하는 공식은 다음과 같다.

$$t_g = \frac{\bar{M}_g}{s_g/\sqrt{n}}$$

여기서,  $n$  은 반복 슬라이드(replicate array slide)의 수를 말하는데, 염료 교체 실험설계이므로, 실험군( $n_T$ )과 대조군( $n_C$ )에서의 관측 측도들의 수와 동일하다 ( $n_T = n_C = n$ ).

준거설계에 의한 마이크로어레이 자료인 경우, 동일한 준거군을 이용하여 두 집단을 간접 비교(Indirect comparison)를 할 수 있다. 실험군(Treatment)과 준거군(Reference)을 비교하는 어레이에서 구한  $M_{(T)} = \log_2(T/R)$  값과 대조군(Control)과 준거군을 비교하는 어레이에서 구한  $M_{(C)} = \log_2(C/R)$  값을 이용하여, 다음과 같이 t-통계량을 구할 수 있다.

$$t_g = \frac{\bar{M}_{g(T)} - \bar{M}_{g(C)}}{s_p \sqrt{1/n_T + 1/n_C}}$$

여기서,  $s_p$ 는  $g$ 번째 유전자의 합동표준편차(Pooled standard deviation)를 말하는데, 2-표본 동일분산 T-검정에 사용한다.

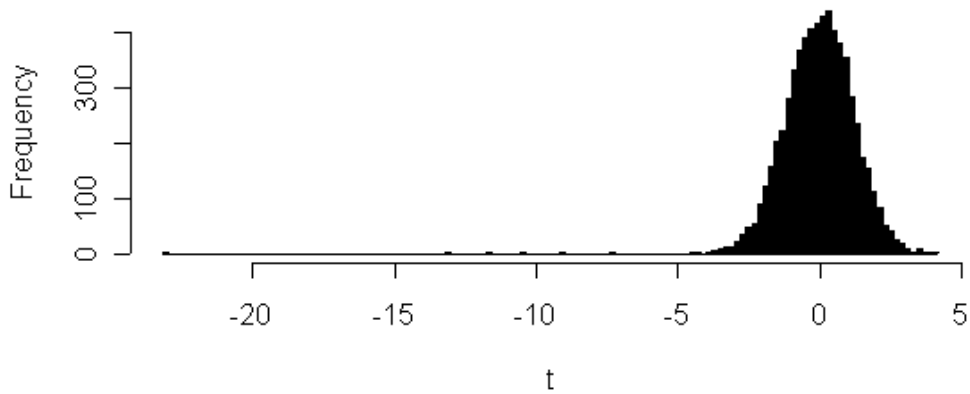


그림 3-1. ApoAI 실험 자료의 2-표본 동일분산 T-통계량의 히스토그램

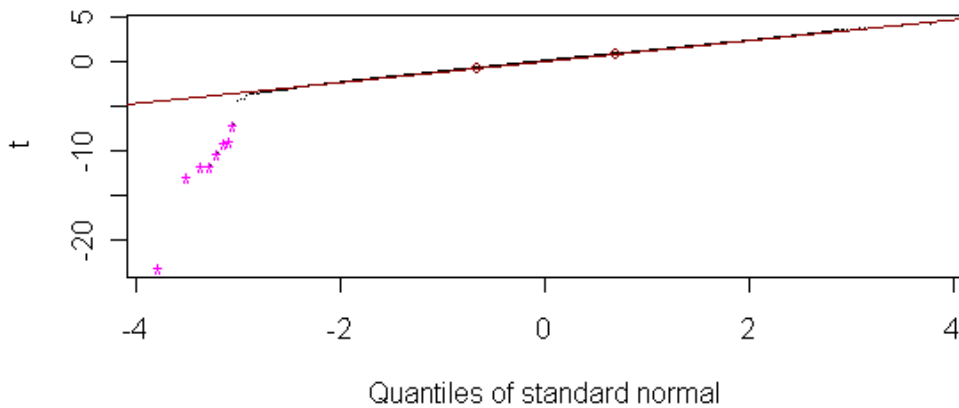


그림 3-2. ApoAI 실험 자료의 2-표본 동일분산 T-통계량의 정규 Q-Q 그림

동일분산을 가정하지 않을 경우에는 t-통계량은 다음과 같이 정의된다.

$$t_g = \frac{\bar{M}_{g(T)} - \bar{M}_{g(C)}}{\sqrt{s_{g(T)}^2/n_T + s_{g(C)}^2/n_C}}$$

여기서,  $s_{g(T)}$ 와  $s_{g(C)}$ 는  $g$ 번째 유전자에 대해 실험군에서의 표준편차와 대조군에서의 표준편차를 말한다.

일반적으로 Q-Q 그림은 두 가지의 통계적 의미를 가지는데, 하나는 자료가 특정한 분포에서 추출되었는가를 분석하는 것이고 또 하나는 두 개의 자료가 같은 분포를 지니고 있는가를 분석하는 것이다. 만약 자료가 특정한 분포에서 추출되거나 2 개의 자료가 같은 분포에서 추출된 것이라면 그림은 직선의 형태를 띠게 될 것이다. 정규 Q-Q 그림은 표준 정규 분포의 분위수(quantile)를 수평축에 자료의 순서통계량을 수직축으로 하는데, 자료의 분포가 표준정규분포와 비슷하다면 직선에 가깝게 나타날 것이다. DNA 마이크로어레이에서는 예외적인(unusual) 통계량을 가지는 유전자들을 시각적으로 선별하는데 도움을 준다. 정규 Q-Q 그림에서 직선에서 벗어난 점들(그림에서 \*)을 발견할 수 있는데, 이러한 예외적인 통계량을 가지는 점일 수록 비교하고자 하는 집단 간에 유의한 차이를 보이는 유전자일 가능성이 높아진다.

이 방법의 정규성을 만족하기 위해서는 충분한 반복 자료(replicate array slides)가 필요하다. 그러나 마이크로어레이 실험을 하려면 많은 비용이 필요하고 이용 가능한 RNA 표본들도 제한적이어서, 사실상 2~8개 정도의 반복 자료가 대부분이다. 특히, 각 집단에 속하는 관측 측도들의 수가 2~3개 정도로 작을 경우에는, 단순한 우연에 의하여 유의한 유전자로 판명될 가능성이 적지 않기 때문에 검정에 대한 신뢰도는 낮아지게 된다. 예를 들어, 매우 작은 분산을 가지는 유전자의 경우 t-통계량 값이 커지므로 유의한 유전자로 선택되어질 수 있다. 게다가 평균을 비교하는 방식은 이상치(Outlier)에 영향을 많이 받는다(Dudoit et al., 2002; Thomas et al., 2001).

분산이 작은 경우 t-통계량이 커지는 문제를 해소하기 위하여, 여러 방법들이 제시되었는데, 예를 들면  $M$ 값이 작으며 그 분산이 하위 1%에 속하는 유전자를 검정대상에서 제외시키는 방식으로  $\bar{M}$  과  $t$  을 절충하여 사용하는 방법들이 있다. 그 외에 각 유전자의 표준편차에 적절한 상수를 더하여 t-통계량을 조정하는 방법, 유의한 유전자에 대한 로그 사후 우도비를 이용하는 베이지안 방법 등이 있는데, 그 구체적인 설명은 다음 절에서 하도록 한다.

### 3.2.2 B-통계량

Lönnstedt et al(2002)에서 제시한 B-통계량은, 매우 작은 분산으로 인해 t-통계량이 커지는 문제를 해소하기 위해 고안된 방법으로, 유의한 유전자(Differential expression)에 대한 로그 사후 우도비(Log Posterior odds)이다.  $n$ 개( $j = 1 \dots n$ )의 반복어레이(Replicated array slide)가 있고, 각 어레이마다  $N$ 개( $i = 1 \dots N$ )의 유전자가 있다고 가정한다. 이 방법에서는, 각 유전자의  $M_{ij}$ 을 평균  $\mu_i$  와 분산  $\sigma_i^2$  을 갖는 서로 독립인 정규 확률변수로 간주하는데, 다음과 같이 표기한다.

$$\text{모든 } j \text{에 대하여, } M_{ij} | \mu_i, \sigma_i \sim N(\mu_i, \sigma_i^2), \quad i = 1 \dots N, \quad j = 1 \dots n$$

지시자 변수  $I$  를 사용하여,  $g$ 번째 유전자가 유의한 차이를 보인다면 1로 정의하고 유의하지 않으면 0으로 정의한다.

$$I_g = \begin{cases} 0 & \text{if } \mu_g = 0 \\ 1 & \text{if } \mu_g \neq 0 \end{cases}$$



$g$ 번째 유전자가 유의한 유전자인지에 대한 로그 사후 우도비  $B_g$  는 다음과 같이 계산할 수 있다.

$$B_g = \log \frac{Pr(I_g = 1 | (M_{ij}))}{Pr(I_g = 0 | (M_{ij}))}$$

이 B-통계량은 유전자들의 평균이 0이 아니며 평균과 분산이 공액 사전 분포 (Conjugate prior distribution)를 따른다는 가정 하에서 다음과 같은 식으로 나타낼 수 있다.

$$B_g = \log \frac{p}{1-p} \frac{1}{\sqrt{(1+nc)}} \left( \frac{a + s_g^2 + M_g^2}{a + s_g^2 + \frac{M_g^2}{1+nc}} \right)^{\nu + \frac{n}{2}}$$

여기서,  $p$ 는  $g$ 번째 유전자가 유의한 차이를 보이는 유전자일 확률,  $Pr(I_g = 1)$  을 말하며,  $s_g^2$  는  $g$ 번째 유전자의 표본분산,  $s_i^2 = \frac{1}{n} \sum_j (M_{ij} - M_{i.})^2$ 을 말한다.

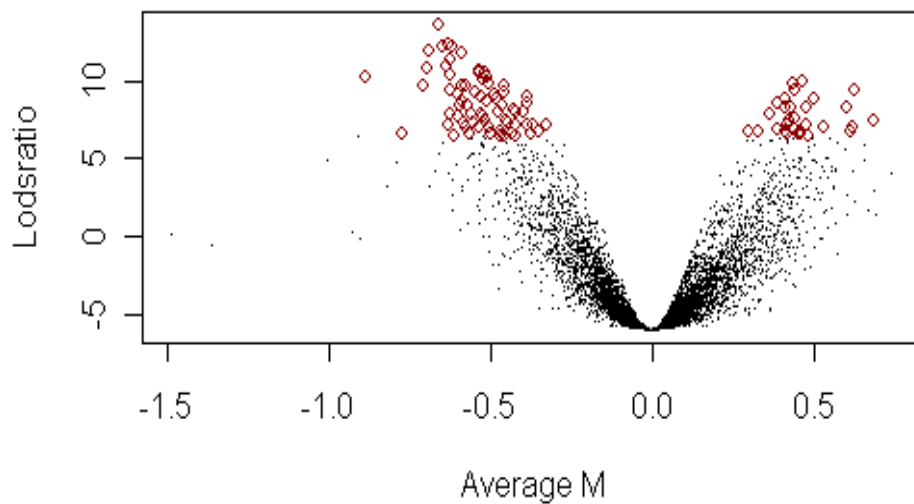


그림 4. ApoAI 실험 자료의 B통계량과 M값에 대한 그림

그리고  $a$  와  $\nu$ 는 분산의 역 감마 사전분포(Inverse Gamma prior)의 초모수(Hyperparameter)이며,  $c$ 는 0이 아닌 평균의 정규사전분포의 초모수를 말하는데, 다음과 같이 정의된다.

$$\text{모든 } i \text{에 대하여, } \tau_j = \frac{na}{2\sigma_j^2} \sim \Gamma(\nu, 1)$$

$$\mu_i | \tau_i \begin{cases} = 0 & \text{if } I_i = 0 \\ \sim N(0, cna/2\tau_i) & \text{if } I_i = 1 \end{cases}$$

B-통계량은 경험적 베이지 접근에 근거하여, 모든 유전자들의 정보들을 결합하여 사전 분포의 모수를 추정하고 유전자별 분산을 고려해 주므로, t-통계량에 비해 안정적인 방법이다(Lönnstedt et al., 2002). 이 방법을 이용하였을 때 [그림 4]와 같이 사용자가 정한 경계값(cutoff) 이상의 B-통계량을 가지는 유전자를 검색할 수 있으며, 유의한 유전자를 사용자가 정한 개수(그림에서 100개)만큼 검색할 수 있다.

### 3.2.3 D-통계량

일반적인 t-검정에서 분산이 작은 경우에 t-통계량이 커지는 것을 막기 위하여, 표준편차에 벌점(Penalty)을 적용하는 벌점 t-통계량을 사용할 수 있는데, Efron et al(2000)에서 제시한 방법과 Tusher et al(2001)에서 제시한 방법이 여기에 속한다.

Efron et al(2000)에서는 각 유전자의 표준편차에 적절한 상수  $a$  를 더하여 t-통계량을 다음과 같이 조정하는 방법을 제시하였다.

$$t_g^* = \frac{\overline{M}_g}{(s_g + a)/\sqrt{n}}$$

여기서,  $a$  는 90 백분위수(90th percentile)를 사용하며, 염료 교체에 의한 마이크로어레이 자료인 경우이다. 준거설계에 의한 자료일 때는 다음과 같은 식으로 구할 수 있다(Efron et al., 2001).

$$t_g^* = \frac{\bar{M}_{g(T)} - \bar{M}_{g(C)}}{(s_g + a) \sqrt{1/n_T + 1/n_C}}$$

Tusher et al(2001)에서 SAM(Significance analysis of microarray)을 통해, t-통계량의 절대치들의 변동계수(Coefficient of variation)를 최소화하는 값  $s_0$ 을 구하여 표준편차에 더해주는 방법을 제시하였다. 상대적 차이(Relative difference)인 d-통계량을 구하는 식은 다음과 같다.

$$d_g = \frac{\bar{M}_{g(T)} - \bar{M}_{g(C)}}{s_g + s_0}$$

여기서,  $\bar{M}_{g(T)}$ 는  $g$ 번째 유전자에 대해 실험군에서의 평균 발현수준(Average levels of expression)으로서 정의되며, 마찬가지로  $\bar{M}_{g(C)}$ 는  $g$ 번째 유전자에 대해 대조군에서의 평균 발현수준이다( $g = 1, 2, \dots, N$ ).

유전자별 산포도(gene-specific scatter)  $s_g$ 는  $g$ 번째 유전자의 표준편차로, 다음과 같이 정의된다.

$$s_g = \sqrt{a \left( \sum_{j(T)} [M_{gj(T)} - \bar{M}_{g \cdot (T)}]^2 + \sum_{j(C)} [M_{gj(C)} - \bar{M}_{g \cdot (C)}]^2 \right)}$$

여기서  $\sum_{j(T)}$ 과  $\sum_{j(C)}$ 는 각각 실험군과 대조군에서의 유전자들의 발현 수준의 총합이며  $a$ 는 다음과 같이 정의된다.

$$a = \frac{(1/n_T + 1/n_C)}{(n_T + n_C - 2)}$$

SAM에서는 잠재적인 중첩효과(Confounding factor)을 최소화하기 위하여 자료의 치환(Permutation)을 이용하는데, 이를 통하여 오류율을 추정할 수 있다. 여기서 오류율(FDR, False discovery rate)은 유의한 유전자들(significant gene) 중 잘못 판단된 유의한 유전자들(falsely significant gene)의 비율(FDR=FN+FP)로 정의되며, 전체 유전자 중 유의한 유전자의 비율  $p(\alpha)=TP+FP$  과 [표 2]과 같은 관계를 가진다.

우선 각 유전자에 대하여 d-통계량을 구한 다음, 크기 순으로 나열하여 다음과 같이 순서통계량을 계산한다.  $d(i)$ 는  $i$ 번째로 큰 상대적 차이, d-통계량을 말한다.

$$d(1) \geq d(2) \geq \dots \geq d(i) \geq \dots \geq d(N)$$

SAM은 중첩효과를 최소화하기 위하여 여러 번의 치환(permutation)을 하는데, 각 치환자료마다 d-통계량을 계산하고 크기 순으로 나열한다. 상대적 차이의 평균  $d_E(i)$ 는 다음과 같이 정의된다.

$$d_E(i) = \frac{\sum_p d_p(i)}{p}$$

여기서,  $d_p(i)$ 는  $p$ 번째 치환자료에서 구한 d-통계량 중  $i$ 번째로 큰 상대적 차이를 말하며,  $p$ 는 치환자료의 수이다.

$g$ 번째 유전자에 대하여 그 상대적 차이( $d$ )가 순수한 변동(Chance variation)에 의해 나타날 수 있는 상대적 차이( $d_E$ )보다 크다면, 유의한 유전자라고 판단한다.  $d(i)$ 와  $d_E(i)$ 에 대하여 고정된 임계치  $\Delta$ 값을 사용하여  $d(i) - d_E(i) > \Delta$ 를 만족하는 유전자들을 양의 유의한 유전자(Significant positive gene)라 하고,  $d_E(i) - d(i) > \Delta$ 를 만족하는 유전자들(그림에서 녹색점)을 음의 유의한 유전자

표 2. 유의한 유전자의 비율  $p(\alpha)$  과 오류율(FDR)의 관계

(1)	유의하지 않은 유전자 (음성,Negative)	유의한 유전자 (양성,Positive)
정확한 판단(True)	TN	TP
부정확한 판단(False)	FN	FP
$\Sigma$	$1-p(\alpha)$	$p(\alpha)$

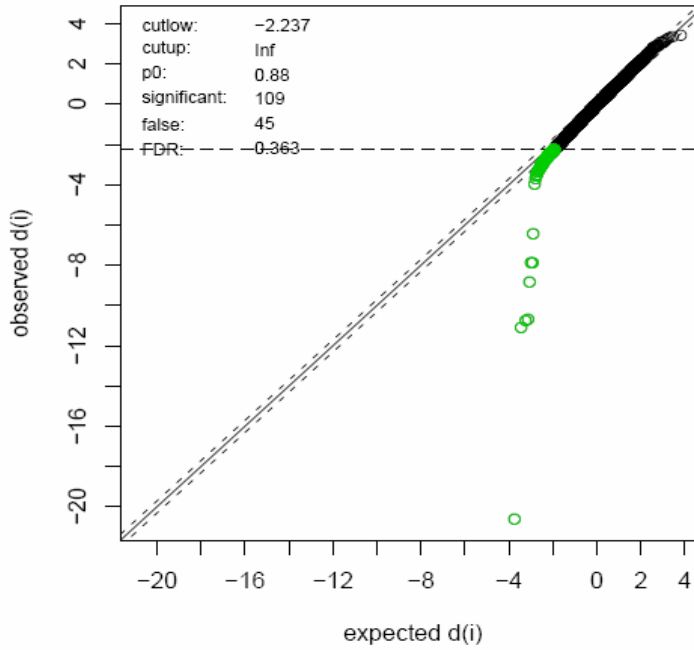


그림 5. ApoAI 실험 자료의  $d(i)$ 와  $d_E(i)$ 의 산점도  
( $\Delta=0.3$ 일 때, 오류율 36.3%)

(1) Broberg (2003)

(Significant negative gene)라 한다. 양의 유의한 유전자들 중에서 가장 작은  $d(i)$  값을 상위 절단점(upper cut-point, 그림에서 cutup)으로, 음의 유의한 유전자들 중에서 가장 큰  $d(i)$  값을 하위 절단점(lower cut-point, 그림에서 cutlow)으로 둔다. 각 치환자료에서 유의하다고 판단한 유전자(falsely significant gene) 수의 평균을 유의한 유전자들의 수로 나눈 값의 백분율을 구할 수 있는데, 이를 오류율(FDR)이라 한다(Tusher et al., 2001; Broberg, 2003).

이 방법을 이용하였을 때, [그림 5]과 같이 사용자가 정한 고정된 값을  $\Delta$ 로 하여 유의한 유전자를 검색할 수 있다. ApoAI 실험 자료에서는 109개의 유의한 유전자를 식별하였으며 이 중 45개의 유의하지 않은 유전자를 유의하다고 판단하여, 이때 오류율은 36.3%가 된다.

### 3.3 점수를 이용한 유전자 선택방법

비교하고자 하는 집단 간에서 유의한 차이를 보이는 유전자는 추가분석을 위해 자료의 차원을 감소시키며 집단간의 차이에 대해 연구하거나 분류분석을 할 때 중요한 역할을 한다. 유의한 유전자들은 실제로 집단 간의 차이를 유발하는 유전자일 확률이 높아 특정 생물학적 현상과 크게 관련(Biological relevance)되어있는데, 이 관련성 정보를 이용하여 유의한 유전자(또는 정보적 유전자)를 선택할 수 있다(Ben-Dor et al., 2000). 이 절에서는 각 유전자에 대하여 적절한 관련성 점수를 주어, 그 점수를 이용하여 유전자를 선택하는 방법들에 대해 설명한다.

$n$ 개( $i = 1 \dots n$ )의 반복어레이(Replicated array slide)가 있으며, 각 어레이마다  $N$ 개( $g = 1 \dots N$ )의 유전자가 있다고 가정한다. 그리고 각 어레이  $x_i$  마다 실험군과 대조군을 식별하게 해주는 표시자(label) 변수  $l_i$  가 있는데, '+1'와 '-1'으로 표시한다.

### 3.3.1 TNoM 점수

Ben-Dor et al(2000)에서 제시한 TNoM 점수(Threshold Number of Misclassification score)는 유전자의 발현수준 임계치(Threshold)를 이용하여 분류 분석을 하였을 때 비교하고자 하는 두 집단을 얼마나 성공적으로 구별하였는지를 측정하는 방법이다. 여기서, 각 유전자의 TNoM 점수는 그 유전자의 발현수준에 대한 임계치를 이용하여 분류분석 시에 오분류(Misclassification)의 개수로 정의된다. 예를 들어, 한 유전자의 발현수준 임계치로 두 집단을 정확하게 구별할 수 있다면 그 유전자의 TNoM 점수는 0이 된다.

우선 각 유전자에 대하여 최적의 Decision Stump을 구한 후, 그 결정규칙을 이용하였을 때 오분류의 개수를 계산한다. 여기서 사용하는 Decision Stump 규칙은  $g$ 번째 유전자에 대하여 그 집단(unknown)을 예측하기 위해서 유전자의 발현수준 임계치를 이용하는 방법으로 다음과 같은 식으로 나타낼 수 있다.

$$g(x_i : g, t, d) = \begin{cases} d & M_{i(g)} > t \\ -d & M_{i(g)} < t \end{cases}$$

여기서,  $M_{i(g)}$ 는  $i$ 번째 어레이에서  $g$ 번째 유전자의 발현수준을 말한다. 이 규칙은 2가지 모수( $t, d$ )에 의해 정의되는데,  $t$ 는  $g$ 번째 유전자에 대한 임계치(Threshold)이며,  $d$ 는 방향모수(direction parameter)로  $\{+1, -1\}$ 의 값을 가진다. 이 방법은  $sign(d(x-t))$ 으로 집단을 예측하는데, 이를 이용하여 오류의 개수를 다음과 같이 계산할 수 있다.

$$Err(d, t|g) = \sum_i 1[l_i \neq sign(d \cdot (M_{i(g)} - t))]$$

이 오류의 개수를 최소화하는  $t, d$ 를 구할 수 있는데, 이 때 오류의 개수를 각 유전자의 TNoM 점수라고 하며 다음과 같은 식으로 정의된다.

$$TNoM(g) = \min_{d,t} Err(d, t|g)$$

이렇게 구한 TNoM 점수의 통계적 유의성을 검정하기 위하여, 임의로 표시되어진(labeled) 자료에서 한 유전자가 주어진 점수  $s$ 보다 더 좋은 점수를 가질 확률을 추정할 수 있는데, 이 값을 TNoM 점수  $s$ 의  $p$ -value라고 한다. 비교하고자 하는 실험군( $n_T$ ), 대조군( $n_C$ )으로 이루어진 자료와 점수  $s$ 가 주어졌을 때,  $p$ -value는 다음과 같이 정의된다.

$$p-val(s, n_T, n_C) = Prob(TNoM(U) \leq s)$$

여기서,  $U$ 는  $n_T$ 개의 '양성(+)'과  $n_C$ 개의 '음성(-)'들이 균등하게 이루어진 순위 벡터를 말한다.

TNoM 점수는, 최적의 규칙(Best decision stump, best rule)에 의한 예측 품질(Quality of the Prediction)에 대하여 부분정보만을 제공해 준다는 단점이 있다. 예를 들어  $k$ 개의 단측 오류를 정하는 규칙과,  $k$ 개의 오류이더라도 한 종류의  $k/2$ 개 오류와 다른 종류의  $k/2$ 개 오류로 정해주는 규칙을 구별하지 않는다는 것이다. 그러나 단측 오류만을 정하는 규칙과 다른 규칙들을 구분하여 고려하는 것은 중요한 일이며, 이를 위해 고안된 방법이 Info 점수이다. 여기서, 단측 오류(one-sided error)는 비교하고자 하는 두 집단 중 하나(+1)를 '-1'로 잘 못 예측하여 발생하는 오류를 말한다(Ben-Dor et al., 2000).



### 3.3.2 INFO 점수

Ben-Dor et al(2000)에서 제시한 Info 점수(Mutual Information score)는 분류 분석 시 임계치의 양 쪽에서 발생하는 정보손실(information lost, entropy)을 이용하여, 비교하고자 하는 두 집단을 얼마나 성공적으로 구별하였는지를 측정한다.

비교하고자 하는 두 집단 중 실험군( $n_T$ )을 ‘양성(+)'으로 표시하고 대조군( $n_C$ )을 ‘음성(-)'으로 표시한다고 가정하였을 때, 실험군에서 과대발현(over-expressed)된 유전자는 양성분할(positive partition)과 관련되었으며, 대조군에서 과대 발현된 유전자는 음성분할과 관련되었다고 판단한다. 여기서 분할에 관련된 모든 유전자들은 둘 중 한 가지 경우에 속한다고 가정한다.  $g$ 번째 유전자에 대하여 어레이들( $n = n_T + n_C$ )을 그 발현수준에 따라 가장 작은 것부터 크기 순으로 나열한다. 여기서, 각 어레이의 양성, 음성의 표시{-, +}들로 이루어진 벡터를  $g$ 번째 유전자의 순위벡터(Rank Vector)  $v$  라고 한다.

예를 들어,  $g$ 번째 유전자가 실험군(+)에서 과소발현(under-expressed)되었다면 왼쪽에 ‘+'가 집중되어 있고 오른쪽에 ‘-'가 집중되어 있는 순위벡터  $v$ 를 가지게 된다. 그 반대로, 대조군(-)에서 과소발현(under-expressed)되었다면  $g$ 번째 유전자는 왼쪽에 ‘-'가 집중되어 있고 오른쪽에 ‘+'가 집중되어 있는 순위벡터  $v$ 를 갖는다. 이 때 순위 벡터  $v$ 를  $x, y$ 의 두 부분으로 나눌 수 있는데, 왼쪽에 ‘-'가 집중된 부분을 앞부분(Prefix)  $x$ 로 부르며, 벡터의 오른쪽에 ‘+'가 집중된 부분을 뒷부분(suffix)  $y$ 라고 한다. 만약  $g$ 번째 유전자의 순위 벡터  $v$ 가 동질적인  $x, y$ 로 이루어져 있다면 유의한 차이를 보이는 유전자로 판단할 수 있다.

Info 점수는 상호정보(Mutual Information), 조건부 엔트로피(Conditional entropy) 등의 정보이론을 이용한다.  $x$ 가 {-, +}들로 이루어진 벡터이고  $p$ 는  $x$  벡터 내에서 ‘+'의 비율(fraction)을 나타낸다고 가정한다면,  $x$ 의 엔트로피는 다음과 같이 정의된다.

$$H(x) = -p \cdot \log(p) - (1-p)\log(1-p)$$

엔트로피는 벡터  $x$ 의 정보(information)를 측정하는데,  $p$ 가 0 또는 1일 경우를 제외하고는 항상 양(positive)의 값을 가진다.  $x$ 가 동질적이어서  $p$ 가 0 또는 1인 경우에는 엔트로피는 최소치인 0이 되고,  $x$ 가 동일한 수의 '+'와 '-'로 구성되어 있는 경우에는 최대치인 1이 된다. 따라서  $H(x)$ 를 이용하여 정보이론의 동질성(또는 비동질성)을 측정할 수 있다.

순위벡터  $v$ 의 앞부분  $x$ 의 엔트로피와 뒷부분  $y$ 의 엔트로피의 최소가중합 (Minimal weighted sum)을  $v$ 의 INFO점수라고 하며 다음과 같이 정의된다.

$$INFO(v) = \min_{x,y=v} \left( \frac{|x|}{|v|} \cdot H(x) + \frac{|y|}{|v|} \cdot H(y) \right)$$

여기서,  $|\cdot|$ 는 벡터의 길이를 말하며 이 점수는 반복 어레이들을 비교하고자 하는 두 집단으로 분류하였을 경우에 순위벡터의 조건부 엔트로피이며, 0이거나 양의 값을 가진다. 두 집단의 순위벡터  $v$ 가 완전하게 분할되어서,  $x$ 는 '+'만을 가지고  $y$ 는 '-'로만 이루어져 있거나 그 반대로  $x$ 는 '-'만을 가지고  $y$ 는 '+'로만 이루어져 있다면 INFO점수는 0의 값을 가진다.

각 유전자의 INFO점수는 그 유전자의 발현수준에 대한 임계치의 양 쪽에서 발생하는 엔트로피를 최소로 하는 임계치를 이용하여 분류분석을 하였을 때, 오분류의 개수로 정의된다. 따라서,  $g$ 번째 유전자의 INFO 점수가 낮을수록, 관련성 정보를 많이 가지고 있는 정보적 유전자라고 판단한다(Ben-Dor, A., Friedman, N., Yakhini, Z., 2002).

이렇게 구한 Info 점수의 통계적 유의성을 검정하기 위하여, 임의로 표시되어진 (labeled) 자료에서 한 유전자가 주어진 점수  $s$ 보다 더 좋은 점수를 가질 확률을 추정할 수 있는데, 이 값을 Info 점수  $s$ 의  $p$ -value라고 한다.

비교하고자 하는 실험군( $n_T$ ), 대조군( $n_C$ )으로 이루어진 자료와 점수  $s$  가 주어졌을 때,  $p$ -value는 다음과 같이 정의된다.

$$p\text{-val}(s, n_T, n_C) = \text{Prob}(\text{Info}(U) \leq s)$$

여기서,  $U$ 는  $n_T$ 개의 ‘양성(+)'과  $n_C$ 개의 ‘음성(-)'들이 균등하게 이루어진 순위 벡터를 말한다.

[그림 6]과 같은 유의한 유전자들의 과잉정도그림(Overabundance graph)을 통하여, ApoAI 실험 자료에서 주어진 Info 점수보다 더 좋은 점수를 가질 것이라 기대되어지는 유전자들의 수(Expected)와 실제 자료에서 주어진 Info 점수에 의해 식

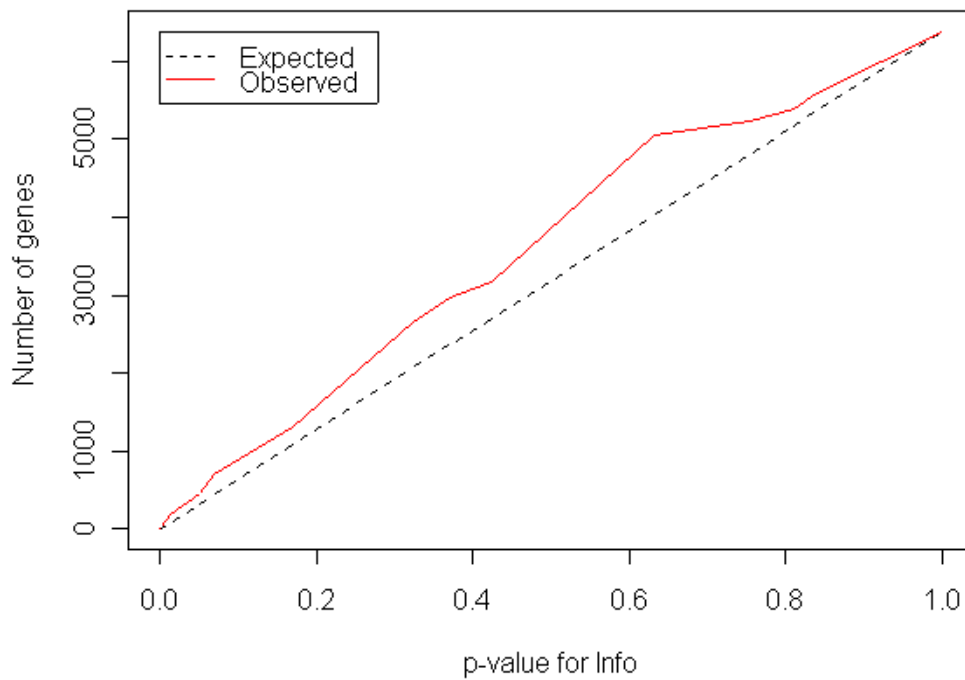


그림 6. ApoAI 실험자료의 Info 방법에 의해 관측된 유의한 유전자들의 수와 임의로 표시되어진 자료에서의 기대유전자수의 비교

별된 유전자들의 수(Observed)를 비교할 수 있다. 이 때 기대유전자들의 수는  $p - val(s, n_T, n_C) \times (n_T + n_C)$  로 구할 수 있다. 기대되어지는 유전자들의 수와 식별된 유전자들의 수의 차이는, 분석하고자 하는 자료가 얼마나 많은 정보적 유전자들을 가지고 있는지를 알려준다(Overabundance analysis).

### 3.3.3 Separation 점수

Slonim et al(2000)에서 제시한 Separation 점수는 비교하고자 하는 두 집단의 유전자 발현수준 분포 간의 상대적인 집단간격 정도를 측정하며, 정규성 가정에 근거한다. 각 유전자는 길이  $n$ 인 집단벡터  $l$ 과 유전자발현벡터  $g$ 을 가진다고 가정하였을 때, 비교하고자 하는 두 집단이 얼마나 떨어져 있는지를 알기 위하여 상대적인 집단간격(relative class separation)을 측정한다. 이를 Separation 점수(또는 Correlation metric)라 하며 다음과 같이 정의된다.

$$Sep(g) = \frac{|\mu_{+(g)} - \mu_{-(g)}|}{\sigma_{+(g)} + \sigma_{-(g)}}$$

여기서,  $g$ 번째 유전자에 대하여 실험군(+)에서의 평균 발현 수준을  $\mu_{+(g)}$ 이라 하며 표준편차를  $\sigma_{+(g)}$ 이라 한다. 마찬가지로  $\mu_{-(g)}$ 와  $\sigma_{-(g)}$ 는 각각  $g$ 번째 유전자에 대해 대조군(-)에서의 평균 발현 수준과 표준편차를 말한다. 비교하고자 하는 두 집단 간의 간격(separation)은 그 평균 간의 거리와 비례하는데, 이 거리는 표준편차로 정규화 하여야 한다. 표준편차가 큰 경우에 집단 평균으로부터 멀리 떨어져 있어서 집단 간의 간격이 크지 않을 수 있다.

$g$ 번째 유전자에 대하여 비교하고자 하는 두 집단의 분포가 떨어져있는 정도가 클수록 두 집단간에 유의한 차이를 보이는 유전자이다.  $g$ 번째 유전자에 대하여 두

집단 평균의 중간값과의 거리와 Separation 점수를 이용하여 집단을 예측하는데, 이때 오류의 개수를 유전자의 Gaussian error (Gaussian score)라고 하는데, 두 집단 분포 간의 중복(Overlap)을 의미하며 다음과 같은 식으로 나타낼 수 있다.

$$Err(g) = \sum_i 1 \left[ l_i \neq \text{sign} \left( M_{i(g)} - \frac{\mu_{+(g)} - \mu_{-(g)}}{\sigma_{+(g)} - \sigma_{-(g)}} \right) \cdot \frac{\mu_{+(g)} - \mu_{-(g)}}{2} \right]$$

각 집단의 유전자 발현수준들이 정규 분포한다면, Separation 점수는 두 집단이 얼마나 떨어져 있는지를 잘 측정할 수 있다. 그러나 정규 분포하지 않을 경우에는 잘못된 측정을 할 수 있는데, 예를 들어 유전자 발현수준들이 한 집단에서 비대칭적으로 분포한다면 분산의 추정에 편의가 있을 수 있으며 이는 잘못된 점수를 도출해 낸다(Slonim et al., 2000).

### 3.4 요약

지금까지 여러 유전자선택방법들에 대하여 살펴보았다. 이를 요약하면 [표 3]과 같다.

표 3. 유전자 선택방법들의 요약

유전자선택방법		유의한 유전자 선택기준	유의성검정	비고
통계량을 이용하는 방법	T-통계량	$ t $ 가 큰 유전자들	$p - value$	FDR 계산
	B-통계량	$ b $ 가 큰 유전자들	.	
	D-통계량	$ d_{(i)} - d_{E(i)}  > \Delta$ 인 유전자들	$q - value$	
점수를 이용하는 방법	TNoM 점수	TNoM점수의 $p - value$ 가 작은 유전자들	$p - value$	과잉정보 추정
	INFO 점수	INFO점수의 $p - value$ 가 작은 유전자들	$p - value$	과잉정보 추정
	Separation 점수	Gaussian error가 작은 유전자들	.	

## 제4장 실제자료에 적용

### 4.1 실험 배경

유전자 발현자료를 분석할 경우에는 관측치의 개수보다 변수의 개수가 더 많은 특수한 자료구조와 많은 오차요인들로 인하여, 한 가지 방법이 아니라 여러 통계적 방법을 이용하는 것이 일반적이다. 본 논문에서는 실제 자료에 각 유전자선택 방법들을 적용한 결과들을 여러 가지 방법으로 비교하고, 이 결과들을 종합하여 자료에 적합한 선택방법에 대하여 알아보려고 한다. 각 방법들의 수행능력은 기존에 알려진 유의한 유전자들과 분류분석의 예측정확도를 이용하여 평가하였다.

유의한 유전자들을 이용하여, 각 방법의 유의한 유전자 식별능력과 유의한 유전자들의 평균 순위를 살펴보았다. 여기서 평균 순위는, 각 방법들에 의해 유의한 유전자일 것 같은 가능성에 따라 유전자들의 순위가 정해지는 것을 이용하는 방법으로, 유의한 유전자들의 평균 순위가 높은 방법일수록 이들을 검출하는 능력이 좋은 방법이라 평가내릴 수 있다(Broberg, 2003).

각 방법들에 의해 선택되어진 유전자들로 분류분석을 하였을 때 그 예측 정확도로 유전자 선택 방법들을 비교하였다. 만약 비교하고자 하는 두 집단 간의 유의한 차이를 보이는 유전자들이 선택되어졌다면, 분류분석을 하였을 때 그 예측정확도가 높을 것이다. 본 논문에서는 예측정확도를 추정하기 위하여 Leave-One-Out Cross-Validation(LOOCV)을 이용하였다. LOOCV는  $n$ 개의 표본들이 있다고 가정하였을 때,  $n - 1$ 개의 표본(training sample)들을 대상으로 분류기(classifier)를 훈련시킨 후 나머지 1개의 표본(test sample) 집단을 정확하게 예측하는지를 검사하는 방법이다. 이 과정을  $n$ 개의 표본들이 1번씩 남겨지도록  $n$ 번 반복하였을 때 정확하게 예측하지 못한 정도를 분류분석의 오류율(error rate)로 사용하였다. 분류분석방법은 잡음이 많거나 희박자료에 로버스트한(robust) SVMs(Support Vector

Machines)을 이용하였으며. 원형기준 핵함수(Radial Basis Function(RBF) kernel function)를 사용하였다(Jaeger et al., 2003).

## 4.2 실제 자료 (ApoAI 실험자료)

Callow et al (2000)에서 사용된 ApoAI자료는 단일 슬라이드 cDNA 마이크로 어레이 실험에서 얻어진 쥐의 유전자 발현자료이다. ApoAI(apolipoprotein AI) 유전자는 고밀도 리포 단백질(high density lipoprotein, HDL)의 신진대사(작용, metabolism)에 중요한 역할을 한다고 알려져 있다. ApoAI유전자가 넉아웃(knocked out)된 쥐는 HDL 콜레스테롤 수위가 낮다. 이 실험의 목적은 ApoAI유전자의 부족이 간(liver)의 다른 유전자의 활동에 어떻게 작용하는지를 밝히는 것이다. 이 실험에서는 8마리의 ApoAI 넉아웃 쥐들과 8마리의 정상 C57BL/6 쥐들을 비교하며, 이들의 간조직 mRNA를 추출하여, 빨강색(Cy5)으로 염색된 cDNA를 합성하여 얻었다. 준거설계 실험 자료로, 모든 어레이에 공통으로 사용되는 준거(reference) RNA는 8마리의 정상 쥐들로부터 추출하여, 녹색(Cy3)으로 염색된 cDNA를 합성하였다. 각 어레이는 6,384개의 유전자들(spot)로 이루어져 있으며, 4 × 4 핀을 이용하였다(spotting).

통계적인 분석에 들어가기 전에 표준화 과정을 거쳐야 한다. 이 자료의 M-A 그림을 살펴보면 수평축(0) 중심에서 벗어나는 정도가 유전자 발현강도에 따라 다를 수 있는데, 이를 발현강도 의존 염료 편이(intensity-dependent dye bias)가 있다고 한다. 또한 각 핀(pin, print-tip)별 발현강도의 차이가 분명하여 핀에 대한 표준화도 필요함을 알 수 있다. 본 논문에서는 비선형적인 추세와 핀 효과에 의한 편이를 최소화하기 위하여 Yang et al(2002)에서 제시한 print-tip group lowess 표준화 방법을 이용하였다. 이는 자료를 핀 별로 구분한 후 각각에 대하여 lowess 함수를 이용하여 지역적으로 적합(robust locally linear fit)하는 방법이다.



### 4.3 실험 결과

본문에서 설명한 6가지 유전자 선택 방법들 중 T-통계량, B-통계량, D-통계량, Separation 점수를 이용한 방법은 유전자자료가 각 이론적 배경에서 제시하는 분포를 따른다고 가정하는 모수적인 방법들로, 이들은 실제 자료의 분포에 따라 결과에 영향을 많이 받을 수 있음을 염두에 두었다.

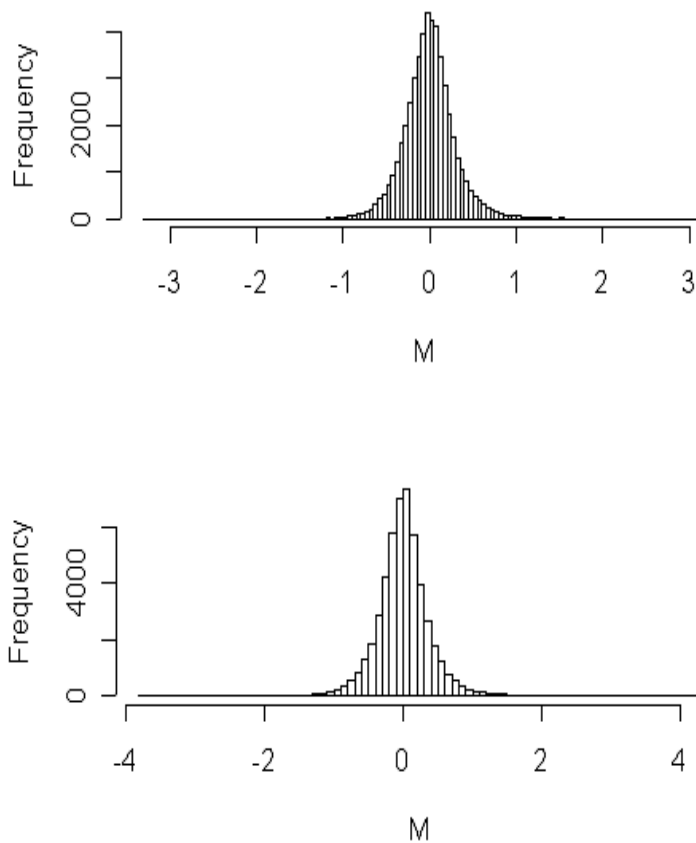


그림 7. 정상 쥐들(위)과 ApoAI 녹아웃 쥐들(아래)의 M값들의 히스토그램

실제적인 유전자 선택과정에 들어가기 앞서, 비교하고자 하는 두 집단, 정상 쥐들(control)의 집단과 ApoAI 녹아웃 쥐들(KO)의 집단별 분포에 대하여 살펴보았다. [그림 7]을 살펴보면 실제 자료가 정규분포에 근사한 분포를 따르고 있음을 알 수 있으므로, 정규성 가정에 근거하는 T-통계량, Separation 점수를 이용하는 방법을 적용하였을 때, 나온 결과를 도출해 낼 수 있을 것이라 예상하였다.

기존의 알려진 10개의 유의한 유전자들을 포함하고 있는 ApoAI 실험 자료에 6가지 방법들을 적용하고 각 방법마다 유전자들의 순위를 정하였다. TNoM, INFO 점수를 이용하는 방법은 그 점수의  $p$ -value를 이용하였고 Separation 점수는 Gaussian error rate을 이용하였다.

각 방법들에 의해 계산되어진 통계량과 점수들 중 유의한 유전자들에 해당하는 값들은 [표 3]과 같았다. T-통계량, D-통계량을 적용하였을 때 가장 좋은 결과를 보이고 있으며, 그 다음으로 Separation 점수, TNoM, INFO 점수 순이나, 그 차이는 근소하였으며 B-통계량의 평균 순위가 가장 낮았다.

[그림 8]-[그림 10]을 통하여 유의한 유전자들을 이용한 평가결과를 검토하였다. [그림 8]에서 유의한 유전자들을 나타내는 점들이 직선에서 벗어나 있음을 확인할 수 있었다. 낮은 수행능력을 보여주었던 B-통계량의 경우를 살펴보면 유의한 유전자들은 무리에서 벗어나 있었으나 이를 구별시켜 줄 만한 적절한 경계값(cut-off)을 B-통계량을 이용하여 정하기가 어렵다는 것을 알 수 있었다[그림 9]. 또한 실제 자료에서 주어진 TNoM 점수에 의해 식별된 유전자들의 수(Observed)와 임의로 표시되어진 자료에서 그 점수보다 더 좋은 점수를 가질 것이라 기대되어지는 유전자들의 수(Expected)를 비교하였다[그림 10]. 기대 유전자수와 자료를 통하여 구한 유전자수의 차이가 클수록, TNoM 점수를 이용하여 두 집단을 분류 분석하였을 때의 결과를 신뢰할 수 있게 되는데(Ben-Dor et al., 2002), 이 자료의 경우, TNoM 점수에서 유의한 유전자수와 기대유전자수 간에 확연한 차이가 있다는 것을 확인할 수 있었다.

표 3-1. 유의한 유전자들에 해당하는 통계량과 유전자 순위

Name	동일분산 T-통계량		이분산 T-통계량		B-통계량		D-통계량	
	통계량 (순위)	p-value	통계량 (순위)	p-value	통계량 (순위)	통계량 (순위)	p-value	
Apo AI, lipid-Img	-23.1044 (6384)	0.01	-23.1287 (6394)	0.01	0.1059 (55)	-20.6117 (6384)	0.000	
EST, ID:439353	-11.7625 (6382)	0.01	-11.7703 (6382)	0.01	-0.6102 (307)	-11.0562 (6383)	0.000	
CATECHOLO-METHYLTRAN	-11.7591 (6381)	0.01	-11.7389 (6381)	0.01	0.2783 (130)	-10.6115 (6381)	0.000	
EST, ID:374370	-12.9824 (6383)	0.01	-13.0296 (6383)	0.01	4.6860 (3225)	-10.7492 (6382)	0.000	
ApoCIII, lipid-Img	-10.4301 (6380)	0.01	-10.4280 (6380)	0.00	2.5953 (1489)	-8.7855 (6380)	0.000	
EST, ID:483614	-9.0186 (6378)	0.01	-9.0286 (6378)	0.01	1.6782 (901)	-7.8533 (6378)	0.000	
est, ID:484183	-9.0874 (6379)	0.01	-9.0846 (6379)	0.01	3.1819 (1875)	-7.8621 (6379)	0.000	
similar to yeast sterol	-7.2089 (6377)	0.01	-7.2006 (6377)	0.00	6.3317 (6282)	-6.3937 (6377)	0.000	
EST, ID:353292	-4.4343 (6375)	0.01	-4.4321 (6375)	0.01	-1.0171 (532)	-3.9226 (6376)	0.089	
NA, ID:317638	-4.2509 (6374)	0.01	-4.2519 (6374)	0.01	6.0130 (6265)	-3.5609 (6374)	0.198	
평균순위	6379		6379		2106.1		6379	

표 3-2. 유의한 유전자들에 해당하는 점수와 유전자 순위

Name	TNoM 점수		Info 점수		Separation 점수	
	점수 (순위)	<i>p</i> - value	점수 (순위)	<i>p</i> - value	점수 (순위)	
Apo AI, lipid-Img	0 (6380)	0.000155	0 (6380)	0.000155	0.00113058	(6384)
EST, ID:439353	0 (6380)	0.000155	0 (6380)	0.000155	0.000155481	(6383)
CATECHOLO-METHYLTRAN	0 (6380)	0.000155	0 (6380)	0.000155	0.000836859	(6381)
EST, ID:374370	0 (6380)	0.000155	0 (6380)	0.000155	0.000231192	(6382)
ApoCIII, lipid-Img	0 (6380)	0.000155	0 (6380)	0.000155	0.00113058	(6380)
EST, ID:483614	0 (6380)	0.000155	0 (6380)	0.000155	0.0042474	(6379)
est, ID:484183	0 (6380)	0.000155	0 (6380)	0.000155	0.0049238	(6378)
similar to yeast sterol	0 (6380)	0.000155	0 (6380)	0.000155	0.023173	(6377)
EST, ID:353292	1 (6363)	0.002486	1 (6363)	0.002486	0.110154	(6368)
NA, ID:317638	2 (6272.5)	0.018648	2 (6272.5)	0.013054	0.123729	(6358)
평균순위	6368		6368		6377	

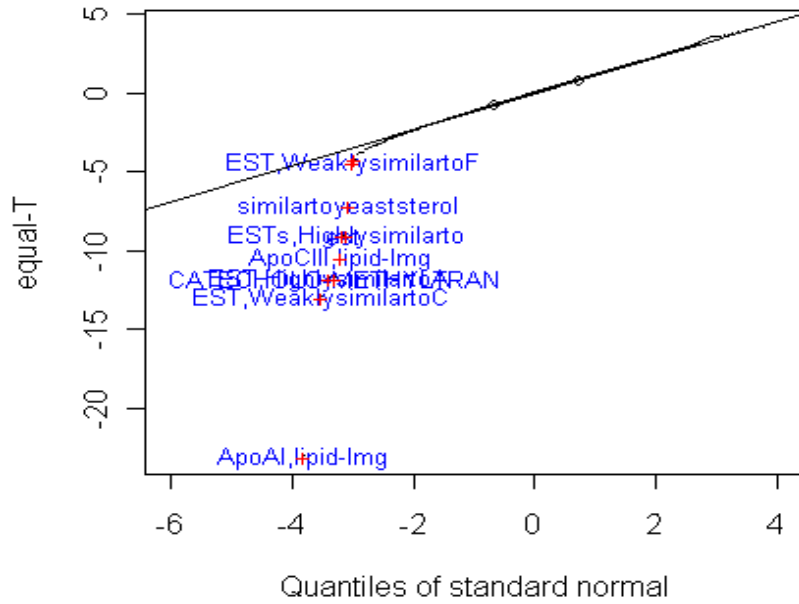


그림 8. 2-표본 동일분산 T-통계량의 정규 Q-Q 그림에서 유의한 유전자들의 위치

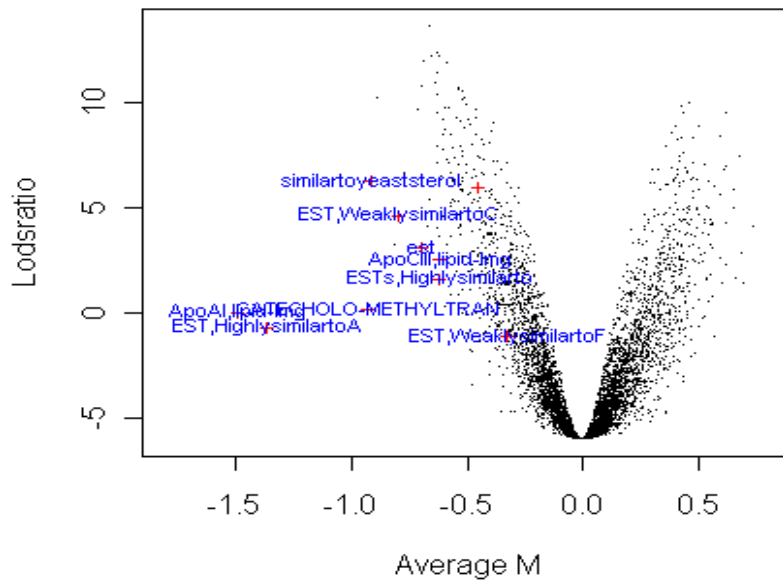


그림 9. B-통계량과 M값 그림에서 유의한 유전자들의 위치

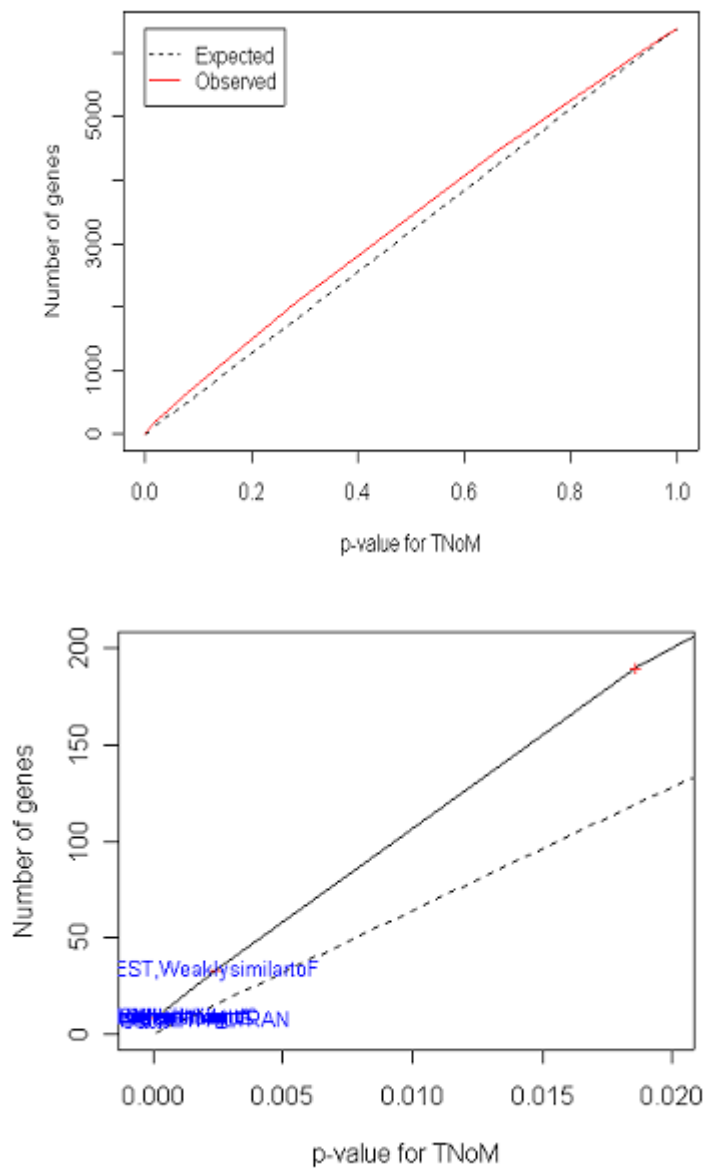


그림 10. TNoM 방법에서 유의한 유전자수와 임의로 표시되어진 자료에서의 기대유전자수의 비교(위). 유의한 유전자들의 위치(아래, 부분확대)

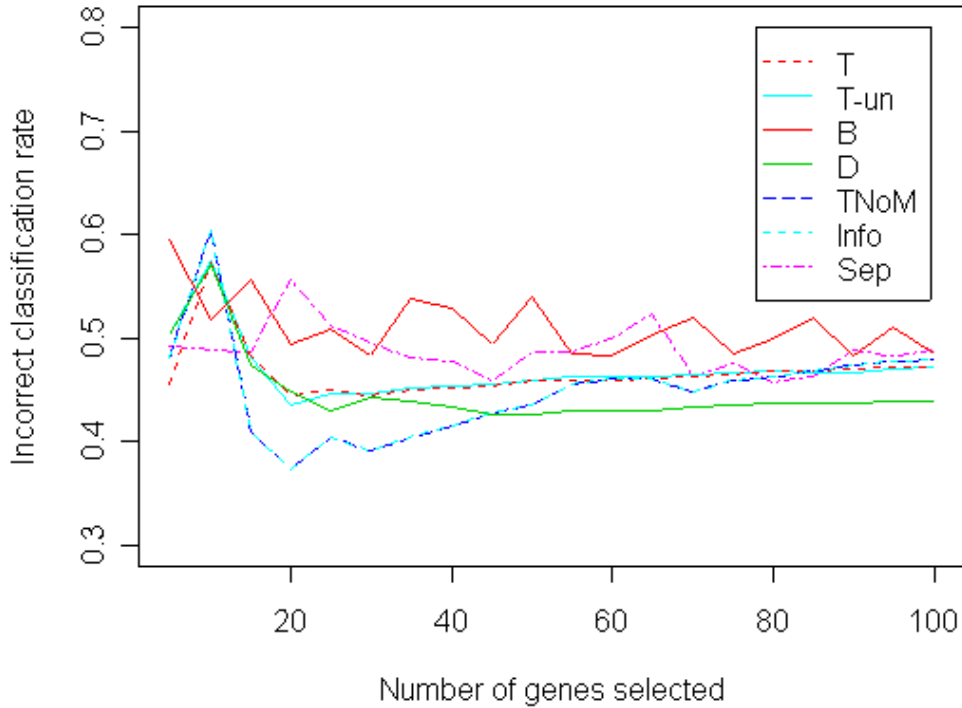


그림 11. 각 방법에 의해 선택된 유전자들을 이용한 분류분석의 정확성 비교

각 방법들에 의해 선택되어진 유전자들로 분류분석을 하였을 때 그 예측 정확도로 유전자 선택 방법들을 비교하였다. 이 때 예측정확도는 LOOCV를 이용하여 추정하였다. ApoAI 실험 자료는 총 16개의 표본들로 이루어져 있으므로, 15개의 훈련표본(training sample)들을 대상으로 분류기(classifier)를 훈련시킨 후, 나머지 1개의 표본(test sample)이 실험군에 속하는 표본인지 또는 대조군에 속하는 표본인지를 예측하였다. 이 과정을 16개의 표본들이 1번씩 남겨지도록 16번 반복하여, 정확하게 예측하지 못한 정도(오류율, Error rate)를 구하였다. 분류기를 훈련시키는 과정에서, 그 대상이 되는 훈련표본은 전체 유전자들(6384개의 유전자들)이 아니라, 각 방법에 의해 선택되어진 유의한 유전자들(5개-100개의 유의한 유전자들)이다. 만약 비교하고자 하는 두 집단 간의 유의한 차이를 보이는 유전자들이 선택

되어졌다면, 분류분석을 하였을 때 그 예측정확도가 높을 것이다. 분류분석방법은 SVMs을 이용하였으며. 원형기준 핵함수를 사용하였다(Jaeger et al., 2003).

[그림 11]에서 세로축은 각 방법들에 의해 선택되어진 유의한 유전자들로 분류 분석을 하였을 때 정확하게 예측하지 못한 정도(오류율, Error rate)를 나타내며, 가로축은 훈련표본으로 사용된 유의한 유전자들의 수를 나타낸다.

TNoM, Info 점수를 이용하였을 때 가장 낮은 오류율(그림에서 유전자 20개일 때)을 보여주고 있으며 유전자의 개수가 40개 이상에서는 직선의 기울기가 완만히 상승하는 형태를 띠고 있음을 확인할 수 있었다. 이 두 가지 방법의 수행능력은 유사하였는데, 이는 직선의 기울기가 거의 일치하고 있는 것으로 확인할 수 있다.

T-통계량과 D-통계량을 이용한 경우도 유전자의 개수를 증가시킬수록 직선의 기울기가 완만해지는 형태를 보여주었다. 그림에서 알 수 있듯이 분석에 이용되는 유전자의 수가 많아질 때 직선의 기울기가 감소하지 않는다는 것을 확인할 수 있었는데, 이러한 결과는 분류분석 전 단계로 유전자 선택과정을 수행할 필요성을 보여주는 것이다. 반면에, B-통계량과 Separation-점수의 경우 직선의 기울기가 불안정하여 이들 방법을 이용하였을 때 분류분석 수행능력의 상승효과는 기대할 수 없었다.



## 제5장 모의 실험을 통한 비교

### 5.1 실험 배경과 실험 자료

모의 실험 자료에 각 유전자선택 방법들을 적용해 보아, 앞서 실제 자료를 통한 비교 결과들을 검증하였다. Baldi et al(2001)에서 사용한 E.coli 실제자료를 모형으로 이용하여 정규분포 자료를 생성하였다. 총 100개의 모의자료를 생성하였으며 각 모의자료(10,000×10)는 5개의 대조군과 5개의 실험군으로 구성되었다. 각 어레이는 10,000개의 가상 유전자 발현 수준자료를 포함하고 있다. 10,000개의 유전자들은 1%의 유의한 유전자들을 포함하기 위하여, 99%의 동일한 분포의 집단과 1%의 상이한 분포의 집단으로 생성하였다. 좀 더 현실적인 상황을 만들기 위하여, 1%의 상이한 분포는[표 4]에서 아래 3줄의 분포들 중에서, 99%의 동일한 분포는 위 3줄의 분포들 중에서 임의로 선택하였다(Broberg, 2003).

표 4. 모의실험 자료(정규분포)의 평균과 표준편차들

어레이	대조군		실험군	
	평균	표준편차	평균	표준편차
99%의 유전자	-8	0.2	-8	0.2
	-10	0.4	-10	0.4
	-12	1.0	-12	1.0
유의한	-6	0.1	-6.1	0.1
1%의 유전자	-8	0.2	-8.5	0.2
	-10	0.4	-11	0.7

모의실험 자료에 6가지 유전자선택 방법을 적용한 결과는 ROC곡선과 100개의 유의한 유전자들의 평균 순위를 이용하여 비교하였다. ROC 곡선은 실제 유의한 유전자들(100개) 중 검출되지 않은 유전자들의 비율(FN, falsely negatives)과 유의하다고 잘 못 판단하여 검출된 유의하지 않은 유전자들의 비율(FP, falsely positives)에 대한 그림으로, 좋은 ROC 곡선을 가지는 방법일수록 이를 이용하여 유전자들을 선택하였을 때, 더 많은 유의한 유전자들과 더 적은 유의하지 않은 유전자들이 검출되어지며 발견하지 못하는 유의한 유전자들의 수가 적을 것이다.

## 5.2 실험 결과

[그림 12]의 ROC 곡선을 살펴보았을 때, T-통계량이 가장 좋은 방법이었으며 그 다음으로 Separation 점수, TNoM(Info) 점수 순으로 좋은 결과를 보여주었다. 이 때 TNoM, Info 점수의 ROC 곡선은 거의 일치하여 실제자료의 결과와 동일하였다. 그리고 B-통계량의 경우 좋지 않은 ROC 곡선을 보여주고 있어, 이 자료에 적절치 않아 보인다.

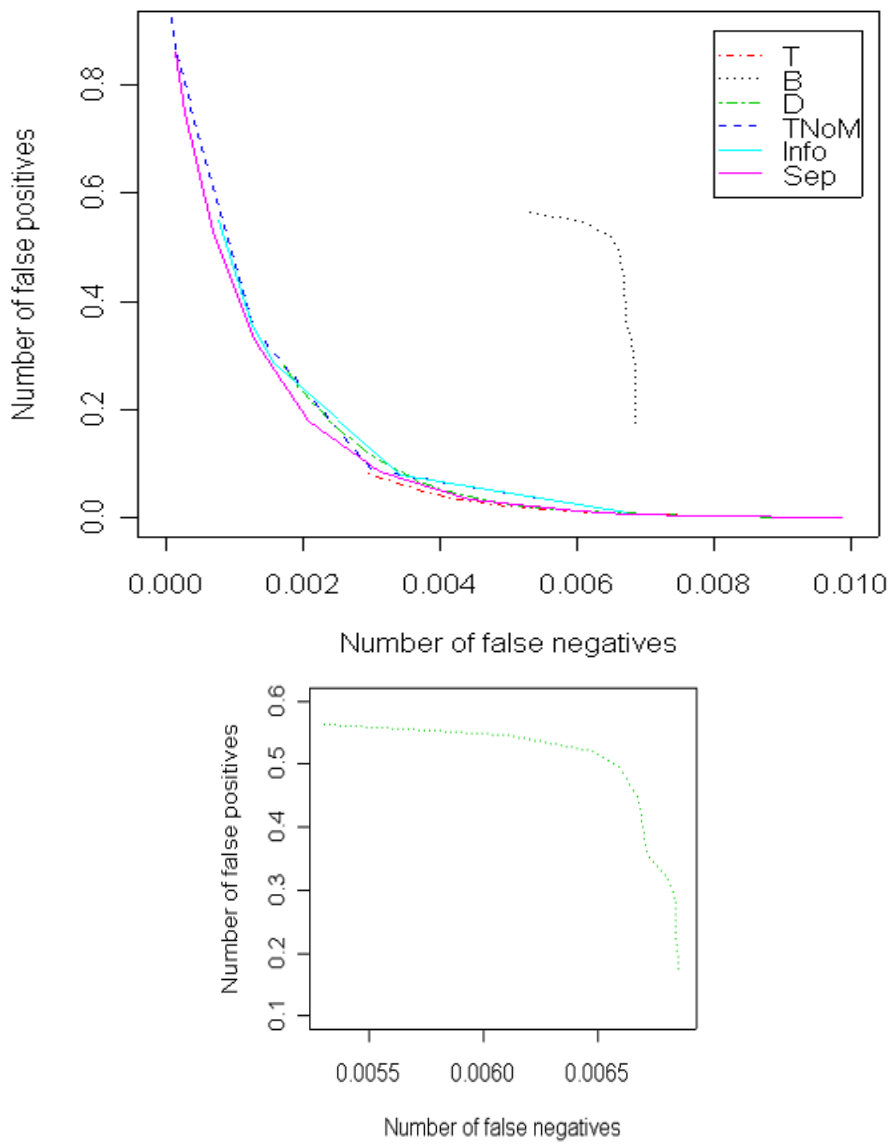


그림 12. 유전자 선택방법들의 ROC 곡선(위)과 B-통계량방법의 ROC 곡선(아래)

표 5. 각 방법의 유의한 유전자 평균 순위 비교

통계량 이용	유의한 유전자의 평균 순위	점수 이용	유의한 유전자의 평균 순위
T-통계량	8870	TNoM 점수	8509
B-통계량	4634	Info 점수	8573
D-통계량	8532	Separation 점수	8749

각 모의 실험 자료에서 100개의 유의한 유전자들의 평균 순위를 계산되어지며 이들 평균 순위들의 평균은 [표 5]와 같았다. T-통계량과 Separation 점수의 평균 순위가 가장 높았고 그 다음으로 Info, D-통계량, TNoM 점수의 순으로 그 차이는 크지 않았다. 그리고 B-통계량의 평균 순위는 가장 낮아 실제자료의 결과와 다르지 않는 결과를 보여주었다.

## 제6장 결론 및 향후과제

지금까지 유의한 유전자선택방법들에 대해 알아보고 각 방법들의 수행능력을 유의한 유전자들의 평균 순위와 분류분석 예측정확도를 이용하여 비교하였다.

실제 자료에 6가지 방법들을 적용한 결과, T-통계량과 TNoM, Info 점수 방법의 수행능력이 가장 우수하였고 B-통계량을 이용하는 방법은 낮은 수행능력을 보여주었다. B-통계량의 낮은 수행능력은 준거설계 자료가 Lönnstedt et al(2002)에서 가정하고 있는 반복어레이구조(Replicated array structure)에 적합하지 않기 때문이라고 할 수 있다(Smyth et al., 2004). 분류분석 예측정확도 비교에서 TNoM, Info점수의 낮은 오류율은, 이들 방법의 비모수적인 접근이 모수적인 방법들에 비하여 낮은 상관유전자(correlated gene)들을 검출할 가능성을 높게 해주기 때문이다(Jaeger et al., 2003).

모의 실험을 통한 비교에서도 T-통계량과 TNoM, Info 점수방법은 좋은 ROC 곡선을 보여 주고 있어, 자료에 적합한 방법이라는 결론을 내릴 수 있었으며, B-통계량의 좋지 않은 ROC 곡선은 정규분포 자료에 적용하는 것 또한 적절치 않음을 보여주고 있다.

유전자선택방법들은 분석하고자 하는 자료의 분포에 영향을 받을 수 있으며, 실험계획에 따라서도 다른 결과를 도출할 수 있어서, 본 논문의 결과를 일반화하기는 어려우므로 다양한 분포의 모의자료와 여러 실험설계 환경자료를 이용한 비교연구가 앞으로의 과제로 여겨진다. 또한 유전자발현자료의 특성상 다른 방법들을 병행하여 여러 방법들에 의해 공통적으로 검출되는 유전자들을 선택하는 것이 신뢰할 수 있는 결과를 도출하는 데 도움을 줄 것이다. 이러한 여러 유전자 선택방법들의 결과를 동시에 비교할 수 있는 프로그램의 개발에 대하여 연구되어야 할 것이다.

## 참 고 문 헌

- Baldi, P., Long, A.D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics* 17: 509 - 519
- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M., Yakhini, Z. (2000) Tissue classification with gene expression profiles. *J Comput Biol* 7: 559-583
- Ben-Dor, A., Friedman, N., Yakhini, Z. (2002) Overabundance Analysis and Class Discovery in Gene Expression Data. Technical Report 2002-50. School of Computer Science & Engineering, Hebrew University
- Broberg, P. (2003) Statistical methods for ranking differentially expressed genes. *Genome Biol* 4: R41
- Callow, M.J., Dudoit, S., Gong, E.L., Speed, T.P., Rubin, E.M. (2000) Microarray expression profiling identifies genes with altered expression in HDL-deficient mice. *Genome Res* 10: 2022-2029
- Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays. *Nat Genet Supp* 32: 490-495
- Draghici S (2002) Statistical intelligence: effective analysis of high-density microarray data. *Drug Discovery Today* 7 11: S55-S63
- Dudoit, S., Yang, Y.H., Speed, T.P., Callow, M.J. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12: 111-139
- Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M. (1999) Expression profiling using cDNA microarrays. *Nat Genet* 21: 10-14
- Efron, B., Tibshirani, R., Goss, V., Chu, G. (2001) Microarrays and their use in

- a comparative experiment. Tech. report, Stanford University
- Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.G. (2001) Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc* 96: 1151-1160
- Jaeger, J., Seugupta, R., Ruzzo, W.L. (2003) Improved gene selection for classification of microarrays. *Pac Symp Biocomput* 8:53-64
- Kaminski, N., Friedman, N. (2002) Practical Approaches to Analyzing Results of Microarray Experiments. *Am J Respir Cell Mol Biol* 27: 125-132
- Kerr, M.K., Churchill, G.A. (2001) Experimental design for gene expression microarrays, *Biostatistics* 2: 183-201
- Lönnstedt, I., Speed, T.P. (2002) Replicated microarray data. *Stat Sinica* 12: 31-46
- Pan, W. (2002) A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics* 18: 546-554.
- Slonim, D.K., Tamayo, P., Mesirov, J.P., Golub, T.R., Lander, E.S. (2000) Class prediction and discovery using gene expression data. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology (RECOMB)*
- Smyth, G.K., Yang, Y.H., Speed, T.P. (2003) Statistical issues in microarray data analysis. *Functional Genomics: Methods and Protocols* 224: 111-136
- Smyth, G.K. (2004) Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. *Statist Applications Genet Mol Biology* 3
- Thomas, J.G., Olson, J.M., Tapscott, S.J., Zhao, L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Research* 11:1227-1236

- Tusher, V.G., Tibshirani, R., Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci USA* 98: 5116-5121
- Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J., Speed, T.P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30: e15



## ABSTRACT

### A Comparison of Statistical Methods for Selecting Significant Genes in cDNA Microarray Experiments

Bang, Jeong Sook

Dept.of Biostatistics and Computing

The Graduate School

Yonsei University

Significant genes are defined as genes in which the expression level characterizes a specific experimental condition. Such genes in which the expression levels differ significantly between different groups are highly informative relevant to the studied phenomenon. Also, For the analysis of gene expression data it is need to reduce the dimension using the statistical methods of selecting significant genes due to systemic variations of cDNA microarray experiments and special data structure. The aim of this paper is to compare different methods, the T-statistics, log posterior likelihood ratio B-statistics, D-statistics using SAM, TNoM score, Info score, and Separation score. Using real and simulated data, we suggest a proper method to select significant genes in each data.

---

Key words : Gene expression data, Significant genes, T-statistics, B-statistics, D-statistics, TNoM score, Info score, Separation score