혼합 모형을 이용한 Microarray자료의 특이 발현변이 유전자의 추정방법에 관한 연구

연세대학교 대학원
의학전산통계학과
이 명 희

혼합 모형을 이용한 Microarray자료의 특이 발현변이 유전자의 추정방법에 관한 연구

지도 김 동 기 교수

이 논문을 석사 학위논문으로 제출함

2002년 6월 일

연세대학교 대학원

의학전산통계학과

이 명 희

이명희의 석사 학위논문을 인준함

심사위원 _____ 인

심사위원 _____ 인

심사위원 _____ 인

연세대학교 대학원

2002년 6월 일

감사의 글

지금까지 저의 모든 것을 주관하시고 앞에서 올바르게 이끌어 주신 전지 전능하신 하나님께 감사드리며 영광을 돌립니다.

제가 무엇을 하든지 믿어주시고 항상 열심히 하라고 희망을 주셨던 아버님 이존식님과 어머님 정정례님을 비롯한 저희 가족 모두에게 감사드리며, 또한 곁에서 칭찬과 책망을 아끼지 않고 저의 버팀목과도 같았던 동준오빠에게 감사드립니다.

2002년 6월

이 명 희 올림

제 목 차 례

표 차 례

그 림 차 례

국 문 요 약

혼합모형을 이용한 Microarray 자료의 특이 발현변이 유전자에 대한 추정방법에 관한 연구

본 논문은 정규 혼합모형을 이용하여 Microarray 자료에서 특이하게 발현된 유전자를 추정하기 위한 알고리즘을 제안하고 또다른 통계적 방법과 비교하였다.

Microarray 기술은 짧은 시간에 엄청난 양의 데이터를 생성할 수 있는 가능성을 지니고 있어 생명공학 분야의 연구 도구로 각광을 받고 있으며, 수천개의 유전자들의 발현 수치들을 측정하는 큰 장점을 가지고 있지만, 잡음이 큰 자료들로부터 실제로 발현변이 유전자를 추정하는데에는 큰 어려움이 있다. 따라서 이 논문은 두 조건하에서 측정된 발현 수치를 이용하여 얻어진 t-통계량으로 EM 알고리즘을 이용하여 정규 혼합모형에 적합시키고 , 오즈비를 적용하여 특이하게 발현된 유전자를 확인할 수 있는 방법을 제안하였다.

중이염을 가진 쥐들과 그렇지 않은 쥐들의 1176개의 유전자들에 대한 발현 수치를 포함하고 있는 자료를 적용시킨 결과 15개의 유전자들의 발현변화가 확인되었다. 또한 정규 혼합모형을 모의 실험에 적용시킨 결과 평균이 2.0이상이거나 평균이 작더라도, 잡음이 아주 적을수록 원래 발현변이 유전자들을 잘 식별할 수 있었다

핵심되는 말 : cDNA Microarray, DNA chip, AIC, BIC, EM 알고리즘, 정규 혼합모형, Differential gene expression, 오즈비

제 1 장 서 론

인간의 모든 세포는 발생의 초기 단계에서부터 공통적으로 똑같은 유전적 정보(genetic information)를 가지고 어떤 임의의 결정된 역할들을 행하며 그 역할들에 의하여 특정 유전자들만 발현함으로 나머지 다른 세포들과는 구별이 되어진다. 평상시에 각 유전자들은 서로 긴밀한 관계를 유지하고 있으나 만약 이들에게 돌연변이(mutation)가 생기거나 발현이 변화(change of expression) 된다면 인간에게 질병(disease)이 발생하게 되는 것이다. 그러므로 동시에 수천개의 유전자에 대한 정보를 얻을 수 있는 생물공학(Biotechnology) 기술이 요구되었고, 주어진 세포(cell) 또는 조직(tissue)에서 수 만개의 유전자들에 대한 수천개의 발현 정도를 동시에 측정하기 위해서 만들어진 microarray의 기술이 넓게 확산되고 있다. 조건이 다른 두 상황에서의 유전자의 발현 측정치들을 비교함으로 생물학적 기능(biological process, function)이 제공되게 된다.

microarray를 사용한 여러 다양한 연구들이 진행되어 오면서 , 연구의 주관심사(focuss)는 잡음(noise)이 많은 자료로부터 유전자의 진짜 발현변이(expression change)들을 어떻게 감지 할 수 있을까 하는 문제였다. 이 문제에 대하여 여러 연구에서 제안된 방법들이 몇 가지 있다. 복합적인 microarray실험을 이용하여 다양한 유전자들의 변이성(가변성, variability)에 접근해 유전자의 발현 정도의 변동과 평균적인 발현 정도(Mean expression levels)와의 관련성을 제안하였으며(Chen et al 1997 , Ideker 2000 , Newton 2001), 유전자의 발현과 연관되어 있는 임의 오차(random error)의 분포적 특성을 비모수적인 방법으로 추정하여 변동(variability)이 큰 잡음(noise)으로부터 진짜로 발현의 정도가 변화된 유전자를 구별했다(Efron 2000, Tusher 2001, Pan 2001). 또한 유전자 발현정도(expression of gene)의 로그정규분포

(Log-normal distribution)와 감마분포(Gamma distribution)라는 강한 모형의 가정을 설정해서 추정하는 모수적 통계적인 방법을 사용하여 발현 변이 유전자들을 발견했으며(Black and doerge 2001) , 정규 혼합 분포 모형(Normal mixture model)에 접근하는 비모수적인 방법을 이용하기도 했다(Pan et al 2001).

본 논문에서는 질병이 있는 실험군(Case)과 질병이 없는 대조군(control)으로 나뉘어진 microarray 자료를 가지고 이표본 검정(two-sample test)에 대한 문제를 다루었다. 이 두 실험조건 하에서의 유전자의 발현을 측정한 자료로부터 유전자의 발현변이(expression change)가 있는가를 확인하기 위해서 두 표본의 t-통계량을 이용해서 정규 혼합모형을 이용하여 추정하였다. 이 논문에서 전체적으로 다루어질 내용은 2장에서 microarray 실험과 활용분야에 관해서 언급하고 3장은 특이하게 발현된 유전자의 추정을 하기 위해서 정규혼합모형을 바탕으로 한 비모수적인 접근방법과 이를 위해서 이용하는 EM 알고리즘을 소개하고 있으며, 4장에서는 실제 microarray 자료에서 통계 패키지중의 하나인 S-Plus를 이용하여 EM 알고리즘을 구현한 프로그램으로 직접 정규 혼합 모형을 추정하여 발현 변이 유전자를 탐색하고, 5장은 위의 EM 알고리즘으로 예측한 정규 혼합 모형의 효과를 평가하기 위해서 1000개의 모의 자료를 가지고 시뮬레이션 하였으며, 마지막으로 EM 알고리즘 프로그램의 성능을 비교하기 위해서 이론의 접근 방법이 다른 SAM(Significance Analysis of Microarray)을 이용하여 EM 알고리즘을 이용한 정규 혼합 모형 방법의 타당성을 보여 주고 , 그 최종 결과를 요약하는 것이다.

제 2 장 Microarray 실험

2.1 Microarray의 소개

1990년대 후반 생물정보학과 DNA 염기서열 분석기술의 발달로 인간을 비롯한 동식물, 미생물의 지놈(Genome)에 대한 방대한 양의 데이터들이 쏟아지기 시작하였다. 1997년에는 미생물의 일종인 효모의 지놈이, 2000년에는 초파리와 선충, 식물 연구의 모델로 쓰이는 애기장대의 지놈이 완성되었으며 2001년 2월에는 인간 지놈의 분석 결과가 발표되기에 이르렀다. 그리고 현재에도 각종 미생물, 쥐 등 많은 생물의 지놈 프로젝트가 진행 중에 있다. 이러한 과정을 통해 나온 지놈 정보는 유전자 발견(Gene Discovery), 의약 개발, 질병 진단 등 여러 영역에 새로운 가능성을 열어 놓고 있으며, 현재 이 지놈을 구성하고 있는 유전자들의 기능과 이들의 네트워크를 밝히는 작업이 필수적이다. 한편 기존의 전통적인 유전적 변화를 해독하기 위한 연구 방법은 단일 유전자(simple gene) 또는 단일 단백질(simple protein)에 대한 실험에 근거하여 진행되었기 때문에 그 경로가 또한 매우 제한적이고 유전자의 전체적인 움직임을 관찰하기에는 한계가 있었다. 그러나, 수천 혹은 수만 개의 생물데이터를 한꺼번에 만들거나 실험실에서 사용되는 여러 분석 과정을 하나의 칩에 집적한 바이오칩(Biochip) 기술의 발달로 인해 등장하게 된 DNA microarray(또는 microchip)란 새로운 개념의 기술이 도입되어 이제는 마이크로 단위로 실험하기에 이르렀다. Microarray는 수만 개 이상의 DNA나 단백질 등을 고밀도로 일정 간격으로 배열하여 붙이고, 분석 대상 물질을 처리하여 그 결합 양상을 분석할 수 있는 바이오칩을 말하는데

DNA칩, 단백질칩(Protein chip) 등이 그 예 이다. 이러한 microarray 기술을 통하여 하나의 칩(chip)상에서 전체 유전체(genome)의 발현양상을 탐색할 수 있게 되었고, 동시에 수천 개의 유전자들 간의 상호작용도 관찰이 가능하게 되었다. Microarray 기술은 짧은 시간에 엄청난 양의 데이터를 생성할 수 있는 가능성을 지니고 있어 생명과학 분야의 새로운 연구 도구로 각광을 받고 있다.

2.2 DNA chip(Microarray)의 제작

DNA microarray의 가장 기본적인 원리는 염기 결합(즉, DNA의 경우 A-T, G-C ; RNA의 경우 A-U, G-C)이다. array란 함은 표본이 순서대로 잘 정열된 것을 말한다. Array는 염기결합 법칙에 근거하여 알려지거나 알려지지 않은 유전자들을 결합시키는 매개체이고 , 알려지지 않은 유전자를 밝혀내는 과정을 자동화하는 매개체를 제공한다. Array 실험은 sample을 손으로, 혹은 robotic machine 을 이용하여 심음으로써 만들 수 있다.

DNA microarray는 array에 배열되는 유전자의 특성에 따라 ,즉 유전 물질의 크기에 따라서 cDNA microarray chip과 Oligonucleotide chip, 이 두 가지 형태로 나눌 수 있다.

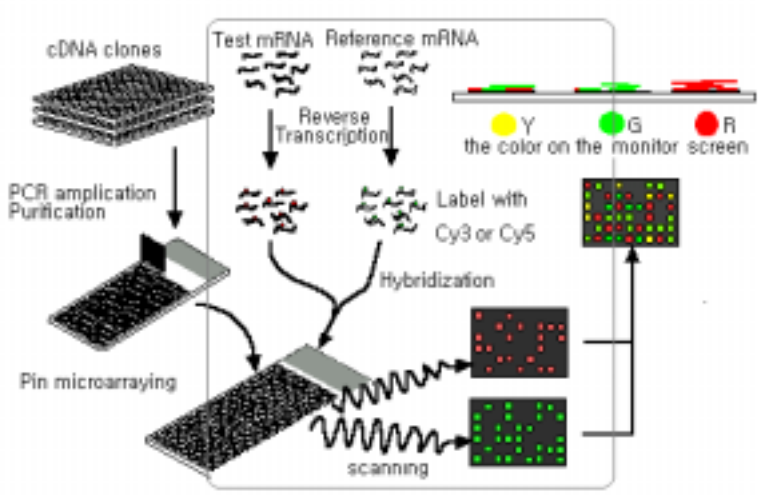

그림 2.1 cDNA microarray chip 생산과 검색 과정
(Duggan et al, 변형)

2.2.1 cDNA microarray

1995년 미국 Stanford 대학의 생화학과에서 처음 개발되었으며 약 3~4천개의 유전자들을 1cm2 안에 붙일 수 있다. 처음에 유전자 발현 측정을 목적으로 cDNA를 붙여 놓은 chip을 만들었기 때문에 cDNA microarray chip이라고 불린다. 또한 이 cDNA microarray chip은 두가지 다른 환경에서 발현되는 독특한 유전자들을 분석하는데 많은 도움이 되며 수천개 이상의 유전자 발현변이를 단 한번의 실험으로 탐색할 수 있는 것이다. 또한 이 chip은 최소한 500bp 이상의 유전자(full-length open leading frame 또는 EST)가 붙여져 있다. 실험과정을 살펴보면 다음과 같으며 아래의 그림은 위의 그림 중에서 네모친 부분의 과정을 더 자세히 그려 놓은 것이다.

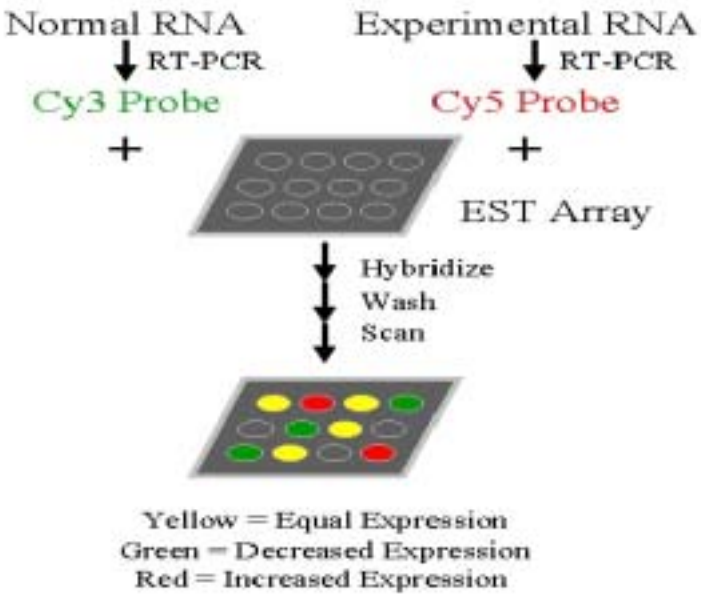

그림 2.2 cDNA Microarray

1단계 : DNA 칩을 제작하는 단계로 cDNA microarray인 경우 spotting 할 DNA clone(알려진 유전자 , EST)들을 PCR에 의해 증폭하고 분리정제한 후 96-well이나 384-well plate에 담게 된다. 고밀도로 이들 유전자를 유리 슬라이드(glass slide) 위에 robot을 이용하여 찍어서 칩을 제작하는데, 이부분을 조금 더 자세히 기술하면 미세하게 제작된 pin이 DNA를 plate로부터 담아서 유리 슬라이드로 컴퓨터가 지정한 똑같은 장소에 옮기는 것이다.

2단계 : 시료준비 단계로 두 개의 다른 환경에서 얻은 세포들인 test 시료와 reference 시료에서 total RNA를 각각 분리한 후, RT-PCR(reverse transcription)방법을 통하여 mRNA를 역전사시키는 과정에서 두가지 형광색소인 빨간색(Cy5)이나 녹색(Cy3)으로 각각 표지(labeling)하여 시료의 cDNA를 결합시킨다. 그러나 시료로부터 RNA를 추출하는 방법은 비교적

간단하지만 DNA 칩의 탐식자(probe)로 사용되는 cDNA 제조에 필요한 현재의 형광물질 표지방법은 비교적 많은 양의 RNA를 요구하기 때문에 시료의 양이 적을 경우 이를 이용하기 어려운 점이 있다. 이를 극복하기 위해 PCR을 이용한 total cDNA의 증폭, 또는 방사선 동위원소를 이용한 표지법의 개발 등의 방법이 제시되고 있다.

3단계 : hybridization 과정으로 합성한 두 개의 시료 cDNA를 같은 양으로 섞어서 제작해 놓은 하나의 microarray 칩에 결합시키고, 결합이 안된 cDNA들을 씻어낸다.

4단계 : 스캐닝 과정으로 레이져 형광 스캐너를 이용하여 각 spot의 형광정도를 읽어 들인다. 각각 유전자의 형광 정도는 그 유전자의 발현정도를 알려주는 것으로 이들 정보는 컴퓨터에 의하여 분석되어 진다.

5단계 : 각 유전자의 발현 수치의 데이터를 이용하여 분석하는 단계이다.

2.2.2 Oligonucleotide microarray

oligonucleotide(20-25mer) 혹은 peptide nucleic acid(PNA) probe가 chip 상에서(in situ) 합성되거나 합성된 후에 chip상에 고정화 된 다음, 표지(labeling)된 sample DNA에 의해 hybridization되면 상보적인 유전자의 염기서열이나 발현정도가 결정된다. 이러한 방법은 역사적으로 "oligoDNA chip 이라고 불리며 affymetrix 사에 의해 개발되었고, GeneChip이라는 이름으로 상품화하고 있다.

위의 실험 단계인 1단계에서 슬라이드 칩 위에 디자인된 oligonucleotide를 각각 합성하면서 칩을 제작하는 방법이 위의 방법과 차이점이다. 또한 수백 bp이상의 DNA로 제작된 chip과는 다르게 15 ~ 25 bp의 비교적 짧은 DNA로 제작된 DNA chip 이다. cDNA chip은 염기 서열이 긴 DNA를 사

용하기 때문에 유전자의 돌연변이 여부를 알 수 없지만 oligonucleotide chip은 유전자에 돌연변이가 있는지, 염기 차이 이상에 따른 질병 검사 , 진단에 이용할 수가 있다.

2.3 DNA chip의 활용분야

유전자 발현을 검색하는 방법들 중에서 cDNA chip이나 oligonucleotide chip이 기존의 방법들보다 뛰어나며, 일단 많은 수의 유전자들을 한번에 탐색하는 데에 그 의미가 있다고 할 수 있다. 아래의 표는 이 두 DNA chip의 사용 가능한 분야들을 요약한 것이다.

표 2.1 Microarray의 활용

cDNA chip	Oligonucleotide chip
·인체 유전자 기능분석 연구	·암관련 유전자 돌연변이 검색진단
·암 및 질병관련 유전자 진단	·유전병관련 유전자 돌연변이 검색 진단
·신약개발	·DNA 염기서열 분석
·유전자 치료	·유전자 변이 가계도 작성
·식품 안전성 검사	·장기 이식가능 조직 검사
·임상 병리학	·법의학(용의자확인, 친자 확인 등..)
·산업용 유전자 재조합 동식물 및 미생물 연구	

인간의 모든 세포는 똑같은 유전 정보를 가지고 있지만 발생 초기 단계에서 자신의 결정된 역할들에 의하여 특정 유전자들만 발현함으로써 다른

세포들과 구별되어지는데, 인간의 몸을 구성하고 있는 각각의 세포들에서는 그들만의 독특한 유전자들이 발현되는 것이다. 이들 유전자들은 서로 아주 긴밀한 관계를 유지하고 있으며 이들 유전자 중에서 하나라도 돌연변이가 생기거나 변화가 일어나면 질병이 발생하게 된다. 따라서 Microarray를 이용하여 게놈차원에서의 유전자 기능과 변화를 알아내는 것은 인류의 건강과 생명의 신비를 해석하는데 아주 중요하며 이러한 지식은 신약 개발이나 유전자 치료 등에도 많은 기여를 할 것이다.

제 3 장 정규혼합모형(Normal Mixture model)

3.1 자료와 통계적 모형의 가정

3.1.1 자료의 형태

	Case(condition 1)				Control(condition 2)			
	1	2	……	m	1	2	……	n
sample 1	x_{11}			x_{m1}	y_{11}			y_{n1}
sample 2								
sample 3								
sample 4								
⋮								
sample N	x_{1N}			x_{mN}	y_{1N}			y_{nN}

그림 3.1 Microarray data의 형태

위의 그림과 같이 표본이 되는 각 유전자 i에 대해서 , $i = 1, 2, 3 \cdots, N$ 개의 유전자가 있을 때, condition 1(실험군)하에서 m개의 microarray로부터 유전자의 발현수치(expression level)인 $X_{1i,} \cdots\cdots, X_{mi}$을 얻을 수 있고 ,condition 2(대조군)하에서는 n개의 microarray로부터 $Y_{1\,i,} \cdots\cdots, Y_{n\,i}$을 얻을 수 있다. 각각의 x_i와 y_i는 형광염료로 표지(labeling)된 cDNA microarray에서 Green 채널에 대한 Red의 상대적인 비(ratio)로 나타내어지는 발현 수치(expression level)들로 표현되어 진다.

여기에서 제안되어지는 방법은 어떤 특정 microarray 기술에 제한적이지 않으며, 일반적으로 전체 유전자인 표본(= N)의 수가 1000이상이고, 반면에 복제된 microarray의 수 m과 n이 5이하일 때 이용한다. 여기에서의 최종 목적은 이표본에 대한 평균의 비교(Two sample comparison)인데, $\{ X_{1i,} \cdots\cdots, X_{mi} \}$의 평균 $\overline{X}$와 $\{ Y_{1i,} \cdots\cdots, Y_{ni} \}$의 평균 $\overline{Y}$가 같은 유전자인지 아니면 다른 평균들을 가지고 있는 유전자인지를 확인하는 것이며 귀무가설과 대립가설을 다음과 같이 표현할 수 있다.

$$H_0 : \overline{X} = \overline{Y} \quad , \quad H_1 : \overline{X} \neq \overline{Y}.$$

그러나 아주 큰 표본개수를 가지고 있으면서도 아주 작은 replication의 수, m과 n을 가지고 있는 microarray의 독특한 특성 때문에 $t-test$ 또는 비모수적 순위 검정 방법(rank-based)등의 전통적인 통계적 검정을 하는 것은 효율적이지 않다(Thomas et al 2001). 따라서 다음에 설명하게 될 비모수적 모형(nonparametric model)에 기초한 정규 혼합모형을 언급하고자 한다.

3.1.2 통계적 모형에 대한 가정

지금부터 제안하게 되는 모형의 가정은 Efron et al(2000), Efron et al(2001), Tusher et al(2001), Wei Pan et al(2001)에서 사용된 가정과 본질적으로 같다. 첫째, 유전자의 발현 자료(expression data of genes)에 대해서 다음과 같은 비모수적인 모형을 가정하는데, 두 조건하에서의 유전자(gene) i에 대한 평균 발현수치(mean expression level)를 $\mu_{(1),i}$, $\mu_{(2),i}$라 정의할 때, 두 조건하에서의 모형은 다음과 같다.

$$X_{ji} = \mu_{(1),i} + \varepsilon_{ji} \quad \sim \quad X_{ji} = X_{1i}, \cdots, X_{mi} \quad \cdots\cdots\cdots \text{ Case,}$$

$$Y_{ki} = \mu_{(2),i} + \varepsilon_{ki} \quad \sim \quad Y_{ki} = Y_{1i}, \cdots, Y_{n_i} \quad \cdots\cdots\cdots \text{ Control,}$$

$$i = 1,2,\cdots,N(\text{\# of 표본}),$$

$$j = 1,2,\cdots,m(\text{\# of Case}),$$

$$k = 1,2,\cdots,n(\text{\# of Control}).$$

또한 실험군과 대조군의 독립 임의오차(independent random error)인 ε_{ji}와 ε_{ki}는 다음과 같은 평균과 분산을 가진다고 가정한다.

$$E(\varepsilon_{ji}) = E(\varepsilon_{ki}) = 0 \quad , Var(\varepsilon_{ji}) = \sigma^2_{(1),i} \, , Var(\varepsilon_{ki}) = \sigma^2_{(2),i}.$$

단, 전체 유전자들의 발현수치(expression level)는 같은 분산을 가진다고 가정하지 않는데, 이유는 이미 기존의 연구들에서(Newton et al 2001, Wei Pan et al 2001) 유전자 발현수치(expression level)에 대한 분산 $\sigma^2_{(c),i}$가 평균 발현수치(mean expression level)인 $\mu_{(c),i}$에 의존한다라고 알려져 있기 때문이다. 그러므로 모든 유전자의 발현수치(expression level)에 대한 분산이 서로 같다라고 가정할 필요도 없다.

3.2 비모수적 정규 혼합모형

3.2.1 이표본 $t-$통계량의 구성(Construct of two sample t-statistic)

앞에서 언급했던 실험군 그룹과 대조군 그룹의 유전자의 발현수치(expression level)의 관측 개체들을 이용하여 정규 혼합모형을 구축할 때 사

용할 data point를 얻기 위하여 기존의 연구(Wei Pan et al 2001)에서 이용하였던 이표본 $t-$통계량(Two sample $t-$statistic)인 값들을 y_i라 정의 하고, y_i를 계산하는 식은 다음과 같다.

$$y_i = \frac{z_{i1} - z_{i0}}{\sqrt{\frac{\sum_{j=1}^{m}(x_{ij} - z_{i1})^2}{m(m-1)} + \frac{\sum_{k=1}^{n}(y_{ik} - z_{i0})^2}{n(n-1)}}}$$

, $z_{i1} = \sum_{j=1}^{m} \frac{x_{ij}}{m}$ (실험군에서의 각 표본의 평균)

, $z_{i0} = \sum_{k=1}^{n} \frac{y_{ik}}{n}$ (대조군에서의 각 표본의 평균)

, $i = 1, 2, \cdots\cdots, N$(# of 표본)

, $j = 1, 2, \cdots, m$(# of Case)

, $k = 1, 2, \cdots, n$(# of Control).

y_i의 분자는 실험군과 대조군의 2조건하에서 평균적인 유전자 발현 수치(average gene expression level)의 차이를 의미하고, 반면에 분모는 분자의 표본 표준 오차이며 큰 변동(variation)을 가지는 유전자의 발현 수치에 보정을 해줌으로써 관찰된 차이를 표준화시키는 역할을 하게 된다.

$t-$통계량 값을 얻었음에도 불구하고 $t-$검정($t-test$)을 시행할 수가 없는데, $t-$검정을 하기 위해서 필수적으로 요구되는 의심스러운 정규성 가정을 지지해줄 만한 어떤 증거도 없기 때문이다. 또한 작은 표본 크기(m+n , replication의 수) 때문에 순열 검정(permutation test)이나 비모수 검정(nonparametric test)을 시행할 수도 없다. 따라서 다음장에서는 강한 분포적

인 가정이 없이도 사용할 수 있는 비모수적인 방법에 기반을 둔 정규 혼합 모형(Normal Mixture Model)에 접근하는 방법을 알아보기로 한다.

3.2.2 정규 혼합 모형

Finite mixture model을 사용해서 특이한 발현현상(differential expression)이 있는 유전자를 탐색하는 것은 정밀한 접근일 뿐만이 아니라 유연성 있는 추정방법을 제공한다. 유전자의 발현자료와 같은 연속형인 자료에 대해 혼합분포(mixture distribution)에서 Normal component를 사용하는 것은 자연스러운 일이다. 특이 발현변이 유전자를 알아내기 위한 정규 혼합모형을 기반으로 한 접근 방법(Normal mixture model-based approach)으로부터 몇 개의 집단으로 나누어지게 될 자료는 다양한 혼합 비율을 가지고 2개 또는 그 이상의 정규 분포(Normal distribution)를 가지는 몇몇의 하위 집단(subclass)으로 나누어지게 된다는 것을 가정한다. 확률밀도 함수를 가지는 정규 혼합분포(Normal mixture distribution)의 모형은 다음과 같다.

$$f(y;\Phi_g) = \sum_{i=1}^{g} \pi_i \phi(y;\mu_i, V_i) ,$$

$$\sum_{i=1}^{g} \pi_i = 1, \qquad \pi_i \geq 0,$$

- Φ_g — g($i = 1, \cdots, g$)개의 성분혼합모형(component mixture model)에서 모르는 모수들 (π_i, μ_i, V_i),
- $\phi(y;\mu_i, V_i)$: i-component에서의 평균 μ_i와 분산 V_i를 가지는 정규 분포,

- π_i — 각 성분(component)에서의 혼합 비율(mixing proportion).

위에서 성분(component)의 수는 자료를 기초로 하여 선택될 것이며 정규 혼합모형(Normal mixture model)에서 실제로 모수를 어떻게 추정하는지, 자료를 어떻게 적합시키는지 다음 장에서 기술하겠다.

3.2.3 정규 혼합 모형의 적합(Fit of Normal Mixture Model)

정규 혼합모형에서 모수(parameter)를 추정하기 위한 일반적인 접근 방법은 Expectation - Maximization(EM) 알고리즘을 적용하는 것이다(Dempster et al, 1977). 이 알고리즘은 혼합모형에서 모수들의 최대우도 추정을 수치적으로 근사하게 하기 위해 반복적으로 계산하는 방법(iterative method)이다. 여기에서는 f 를 추정하기 위해서 정규 혼합모형(Normal mixture model)에 적합시키는 방법은 다음과 같다(McLachlan and Basford 1988 ; Tittering et al 1985). 다른 2조건하에서의 실험군과 대조군의 표본의 개수를 N이라고 하자. N개의 자료에서 t-통계량인 $y_{1,} \cdots\cdots, y_N$이 있을 때, 최대 우도추정량 $\hat{\Phi}_g$를 얻기 위해서 다음과 같은 로그 우도함수(log-likelihood)를 최대화(maximization) 시켜야 한다.

$$\log L(\Phi_g) = \sum_{j=1}^{N} \log f(y_j ; \Phi_g).$$

EM 알고리즘은 다음에 따라오는 E-step과 M-step의 절차를 반복함으로써 $\hat{\Phi}_g$를 계산한다. k번의 반복을 한다라고 가정했을 때 모수의 추정값들은 $\pi_i^{(k)}$, $\mu_i^{(k)}$, $V_i^{(k)}$들이다.

E-step :

$$\tau_{ij}^{(k)} = \tau_i(y_j \mid \phi(y_j; \mu_i^{(k)}, V_i^{(k)})) = \frac{\pi_i^{(k)}\phi(y_j; \mu_i^{(k)}, V_i^{(k)})}{f(y_j; \Phi_g^{(k)})},$$

$$i = 1, \cdots\cdots, g(\#\text{ of component}),$$

$$j = 1, \cdots\cdots, N(\#\text{ of 표본}).$$

이 단계에서는 처음으로 각 표본(sample)이 각각의 집단 분포에 속하는지를 추정하는 단계인데, $\tau_{ij}^{(k)}$는 표본인 y_j가 집단 i에 속하는 사후 확률(posterior probability)이며 현재 모수들은 Φ_g에 대해서 $\Phi_g^{(k)}$를 추정한다.

M-step :

$$\pi_i^{(k+1)} = \sum_{j=1}^{N} \pi_{ij}^{(k)}/N,$$

$$\mu_i^{(k+1)} = \sum_{j=1}^{N} \tau_{ij}^{(k)} \times y_j / \sum_{j=1}^{N} \tau_{ij}^{(k)},$$

$$V_i^{(k+1)} = \sum_{j=1}^{N} \tau_{ij}^{(k)} (y_j - \mu_i^{(k+1)})^2 / \sum_{j=1}^{N} \tau_{ij}^{(k)}.$$

이 단계에서는 사후확률을 가중치(weight)로 이용하여 모수 추정을 하는 단계이다. 위와 같이 k번의 반복으로 수렴에서, 최대우도추정으로 $\hat{\Phi}_g^k = \Phi_g^{(\infty)}$를 추정할 수 있다. 그리고 EM 알고리즘으로 모수를 추정하는 과정에서 국부최대값(local maxima)이 발견될 수 있으므로 다양한 초기값들을 가지고 EM 알고리즘을 여러 번 계산을 해야 하며, 가장 큰 로그 우도

함수(log-likelihood)를 초래하는 결과를 최종 추정값으로 선택하는 것이 바람직하다.

다음으로 생각할 것은 몇 개의 집단으로 나누어야 하는지, 즉 몇 개의 정규 분포 성분들로 이루어져 있는지를 결정해야 하는데 다양한 모형 선택 기준 중에서 가장 잘 알려진 Akaike Information Criterion(AIC) 와 Bayesian Information Criterion(BIC)을 이용할 수 있다.(Schwartz 1978) :

$$AIC = -2\log L(\hat{\Phi}_g) + 2\nu_g \text{ (Akaike 1973)},$$

$$BIC = -2\log L(\hat{\Phi}_g) + \nu_g \log(N) \text{ (Schwartz 1978)},$$

ν_g : Φ_g에서 모수들의 수.

집단의 수인 g의 값을 1부터 다양하게 주어 정규 혼합모형에 적합시켜서 AIC, BIC중에서 첫 번째 국부 최소값(local minimum)에 해당하는 모형을 선택한다(Fraley and Raftery 1998). 이 논문에서는 g(=component)의 수를 정하기 위해서 BIC의 값을 사용하기로 한다.

또한 이와 다른 접근방법이 있는데, 가설검정을 통해서 g의 값을 결정하는 것이다. 이 방법은 다음과 같은 가설을 검정하기 위해서 우도비 검정(likelihood ratio test)을 시행한다.

$$H_0 : g = g_0,$$

$$H_1 : g = g_0 + 1,$$

$$LRT = \frac{2\log L(\hat{\Phi}_{go+1})}{2\log L(\hat{\Phi}_{go})}.$$

Mclachlan(1987)은 위의 귀무가설 하에서 LRT 통계량의 분포에 접근하기 위해서 붓스트랩(bootstrap)을 이용하는 것을 제안하였다. 이 결과로 추정되어진 $P-value$를 기초로 하여 귀무가설을 기각하는지의 여부를 결정할 수 있다.

마지막으로 각 자료들이 어떤 집단에 속하는지를 결정해야 한다. Bayes rule을 이용하여 각 자료들이 모형에서 어느 정규 분포모형에 속할 사후 확률(Posterior probability)을 추정하여 가장 높은 사후 확률을 가진 집단으로 분류하는 것이며 다음과 같다.

$$\hat{\tau}_{ij} = \frac{\hat{\pi}_i \phi(y_j; \hat{\mu}_i, \hat{V}_i)}{f(y_j; \hat{\Phi}_g)} \ ,$$

$$i = \ \# \ of \ component \quad , \quad j = \ \# \ of \text{ 표본} \ .$$

또 한가지 방법은 오즈비(Odds ratio)의 개념을 적용해서 각 집단으로 분리하는 것이다. 유전자들이 2개 이상의 정규분포 모형으로 이루어진다고 해도 크게는 발현이 변화된 유전자와 발현이 변하지 않은 유전자로 다음과 같이 분리할 수가 있으며 다음의 식으로 표현할 수가 있다.

$$f = p_0 f_0 + p_1 f_1 \ ,$$

p_0 : 유전자의 발현이 변화되지 않은 집단의 비율(unchanged class) ,

p_1 : 유전자의 발현이 변화된 집단의 비율(changed class) ,

f_0 : 유전자의 발현이 변화되지 않은 집단의 정규분포 모형 ,

f_1 : 유전자의 발현이 변화된 집단의 정규분포 모형 .

그러므로 위의 추정된 정규 혼합물의 모형을 이용해서 실험군의 유전자의 발현이 대조군에 비해서 그 비율이 얼마나 변하는지를 오즈비를 이용하는데 다음과 같다.

$$odds_1 = \frac{\text{각 자료가 } f_1 \text{에 속하는 } probability}{\text{각 자료가 } f_1 \text{에 속하지 않을 } probability} = \frac{\Lambda_1}{1-\Lambda_1},$$

$$odds_2 = \frac{\text{각 자료가 } f_0 \text{에 속하는 } probability}{\text{각 자료가 } f_0 \text{에 속하지 않을 } probability} = \frac{\Lambda_2}{1-\Lambda_2},$$

$$odds\ ratio = \frac{odds_1}{odds_2}.$$

따라서 오즈비는 각 자료가 유전자의 발현이 변화되지 않은 집단에 속하는 오즈에 대한 유전자의 발현이 변화되는 집단에 속할 오즈의 비(ratio)라고 말할 수 있다. 각 자료를 다음과 같은 기준으로 분리할 수 있다.

$odds\ ratio > 1$: 유전자의 발현이 변한 집단 ,

$odds\ ratio < 1$: 유전자의 발현이 변하지 않은 집단 .

따라서 이 논문에서는 g(=component)의 수를 정하기 위해서 BIC의 값을 사용하고, 정규 혼합모형에 적합시킨 다음에 오즈비를 이용해서 각각의 집단으로 분리하기로 하겠다.

제 4 장 Microarray자료를 이용한 특이 발현 유전자의 탐색

4.1 자료(Data)

폐렴 구균성 중이염(Pneumococcal Otitis media)는 고막 안쪽에 위치한 중이강이라는 공간에 염증이 생기는 것을 말하며 특히 소아의 경우 병원을 찾아오는 환자 중 가장 많음 비율을 차지하는 질환이다. 중이염의 발생률은 출생 후 3세까지의 유소아 중에 1/3이 세 번 이상 급성 중이염에 걸린다고 하고 전체 소아의 2/3에서 3세가 될 때까지 한 번 이상 중이염에 걸릴 정도로 흔한 질환으로 알려져 있다. 따라서 이 중이염의 발병학을 이해하기 위해서 폐렴 구균 중이염에 반응하는 것과 관련된 유전자를 식별하고 중이염에서 그 유전자들의 역할을 연구하는 것이 아주 중요하다고 할 수 있다.

여기에서 쓰인 자료는 미네소타 대학(University of Minnesota)에서 2001년도에 수행한 연구의 자료를 이용하였으며 아래의 웹페이지에 가면 그 자료들을 다운받을 수 있으며, 여기에서는 이 자료들을 부분적으로 사용하였다.

http://www.biostat.umn.edu/~weip/paper/ratdata.html .

여기서 이용한 microarray 자료의 구성은 쥐의 중이점막층(midle ear mucosa)에 아급성 폐렴 구균성 중이염(Subacute Pneumococcal otitis infection)을 가진 5개의 실험군(Case)과 이 중이염을 가지고 있지 않은 2개의 대조군(Control)으로 이루어져 있다. 자료의 구성은 다음과 같다.

표 4.1 중이염 자료의 형태

	Control		Case				
	rat 1	rat 2	rat 1	rat 2	rat 3	rat 4	rat 5
gene 1	16161	17550	19640	18375	25415	26446	20764
gene 2	15993	16993	15898	15637	18791	18814	17245
gene 3	15636	17146	15315	15079	19702	19069	16381
gene 4	16339	17829	19951	18329	25747	24818	19079
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
gene 1176	22202	19899	18661	20150	27136	32597	34639

4.2 자료의 처리와 표준화

원자료의 수치는 관찰된 유전자의 발현 수치들을 측정한 것이며, 다음과 같이 자료를 처리한다. 자료의 1,2열($j = 1, 2$)은 대조군에 속하는 microarray이고, 나머지 열들($j = 3, 4, 5, 6, 7$)은 실험군에 속하는 microarray이다. 첫째, 유전자의 발현수치들이 정규 분포에 더 근사하도록 하기 위해서 전체 자료에 대해서 자연 로그(natural logarithm)로 변환시킨다. 실험군의 3열($j = 3$)의 자료를 변환시키기 전과 후의 그림은 다음과 같다.

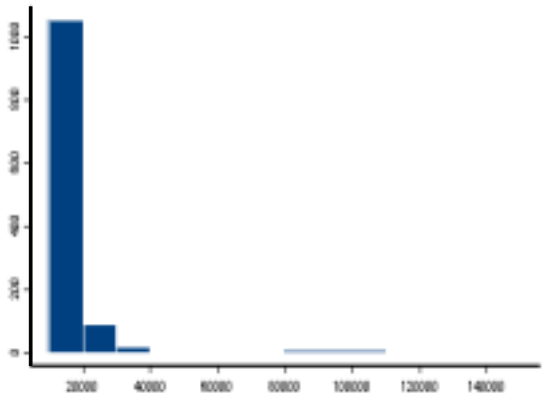

그림 4.1 intensity

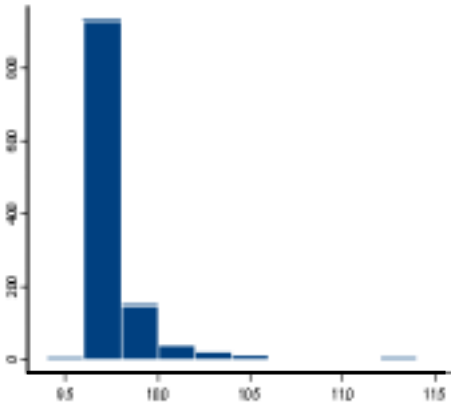

그림 4.2 log(intensity)

둘째, 각각의 7개의 실험에 대하여 로그로 변환(log-transformation)된 유전자의 발현 수치에서 각각의 중앙값(median value)을 빼줌으로써 표준화 시켜준다. 이 표준화는 대부분의 유전자들이 적어도 $\frac{1}{2}$은 발현이 되지 않았을 것이라는 것에 바탕을 두고 있으며, 평균값(mean)을 빼주는 것보다는 이상치에 대해서 더 강하기(robust) 때문에 중앙값을 이용한다.

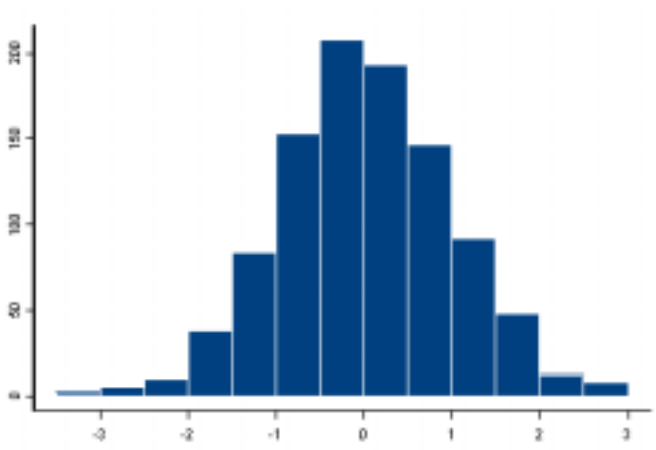

그림 4.3 Normalization(intensity)

마지막으로, 위에서 표준화된 microarray의 자료를 기초로 각 유전자들에 대하여 이표본 t-통계량을 다음과 같이 계산한다.

$$y_i = \frac{z_{i1} - z_{i0}}{\sqrt{\frac{\sum_{j=1}^{m}(x_{ij} - z_{i1})^2}{m(m-1)} + \frac{\sum_{k=1}^{n}(y_{ik} - z_{i0})^2}{n(n-1)}}},$$

$$z_{i1} = \sum_{j=1}^{m}\frac{x_{ij}}{m}, \qquad z_{i0} = \sum_{k=1}^{n}\frac{y_{ik}}{n},$$

$i = 1, 2, \cdots\cdots, N$(표본의 개수) ,

$j = 1, 2, \cdots\cdots, m$(실험군의 복제(replication)의 수) ,

$k = 1, 2, \cdots\cdots, n$(대조군의 복제(replication)의 수) .

여기에서 y_i의 분자는 중이염에 감염된 실험군과 감염되지 않은 대조군에서의 유전자들의 평균적인 발현 수치에 대한 차이를 의미한다. 또한 y_i의 분모 부분은 실험군과 대조군에서 관찰된 발현 수치의 차이들이 변동이 아주 크므로 이 변동을 보정해 주는 역할을 하게 된다.

4.3 정규 혼합모형의 모수 추정

본 논문에서는 유전자의 발현 수치 자료에 대해서 정규 혼합모형(Normal mixture model)을 이용해서 EM 알고리즘을 통해 각 집단(class)에서의 혼합비율(Mixing proportion)과 평균(μ_i), 분산(V_i)을 추정하였다. 또한 이 알고리즘에서 이용한 초기값(initial value)에 대해서, EM 알고리즘의 가중치로 사용되는 혼합비율은 π, 각 집단의 수를 g라고 가정한다면 다음과 같다.

$$\pi = \frac{1}{g}, \qquad \sum_{i=1}^{g} \pi_i = 1, \qquad i = 1, \cdots, g.$$

초기값으로 사용되는 평균과 분산은 정규분포의 난수(random number)를 발생시켜서 이용하였으며, 집단의 수는 $1 \sim 5$까지 미리 정해 놓고 각 집단의 수에 따라 여러번 다양한 초기값을 주어 반복하였으며, 아래의 모수들을 구하는 EM 알고리즘의 추정 방법에 따라서 수렴이 될 때까지 반복시키고 허용한계는 10^{-5}으로 두고 최대 반복수는 5000회로 설정하였다. 이 결과들 중에서 제일 큰 로그 우도함수(log - likelihood)의 값이 나온 결과를 선택하였다.

표 4.2 정규 혼합모형의 적합 결과(g = 1~5)

g	$\hat{\pi}$	$\hat{\mu}$	$\hat{\sigma}$
1	1.0	0.468	2.289
2	0.247	1.783	3.249
	0.753	0.037	1.655
3	0.335	-0.265	1.091
	0.657	0.724	2.433
	0.008	9.710	1.306
4	0.008	9.918	1.166
	0.285	-0.201	1.028
	0.481	1.132	2.444
	0.226	-0.415	1.957
5	0.007	9.932	1.157
	0.008	-0.846	0.005
	0.247	-0.457	1.965
	0.458	1.213	2.434
	0.280	-0.153	1.044

각각의 모수들을 1~5까지의 집단의 범위에 따라서 5개의 혼합모형에 적합한 결과는 앞의 표에 제시한 것과 같다.

4.4 집단(component, class)의 수 결정

EM 알고리즘을 통해 각 성분(component)의 수에 따라서 우리가 알지 못했던 모수들을 추정한 후에 , BIC(Bayesian Information Criterion) 또는 AIC(Akaike Information Criterion)의 기준에 맞추어 과연 몇 개의 정규 혼합모형으로 이루어져 있는지를 선택해야만 한다. 그 판단의 기준이 되는 것은 BIC 또는 AIC의 값이 최초에 최소(Local minimum)가 되는 값을 선택하면 된다. 다음은 각 집단의 정규 혼합분포를 이용해서 구한 BIC와 AIC의 추정치를 나타낸 것이다.

표 4.3 g의 변화에 따른 BIC , AIC의 결과

g	BIC	AIC	$\log L$
1	5299.0	5288.9	-2642.4
2	5204.5	5184.2	-2588.1
3	5203.1	5172.7	-2580.4
4	5216.5	5175.9	-2580.0
5	5217.7	5167.0	-2573.5

위의 표를 살펴보면, g=1에서 g=2로 증가할 때 로그 우도함수 값이 아주 급격한 증가를 보이며 g=3이 되면서부터는 거의 완만한 로그 우도값의 형태를 보인다. 또한 BIC의 값이 g= 3일 때 처음으로 최소가 되며 AIC의

값도 첫 번째 국부 최소값(local minimum)이 된다. 따라서 이 모형에서는 g=3인 경우를 선택한다. 집단의 수 g를 결정하는 것도 아주 중요한 문제이기는 하지만 이 논문에서는 정규 혼합모형의 모수를 추정해서 모형에 적합시키는 것이 목적이므로 일단 집단의 수를 1~5까지 미리 정해놓고 BIC의 국부 최소값(Local minimum)이 되는 g를 결정한다.

따라서 위의 결과에 따라서, 이 microarray 자료의 정규 혼합 모형의 분포가 3개의 집단으로 이루어져 있음을 알 수 있으며 g=3일 때 적합한 모형은 다음과 같은데,

$$f(y;\hat{\Phi}) = 0.335 * N(-0.265, 1.190) + 0.657 * N(0.724, 5.919) + 0.008 * N(9.710, 1.706).$$

약 99%이상의 자료들이 평균이 0에 가까운 첫 번째 정규 분포의 집단과 2번째 정규 분포의 집단에 속한다. 평균이 0에 가까운 집단들은 대부분의 유전자들에 대해서 유전자의 발현의 변화가 거의 없거나 약간의 변화가 있음을 말한다. 반면에 10개의 유전자들은 유전자 발현 정도가 특이하게 변했다고 말할 수 있는 3번째 집단에 속한다고 할 수 있는데, 평균이 0에서 많이 동떨어져 있기 때문이다.

제 5 장 모의 실험

5.1 자료

본 논문에서는 표본크기(Sample size)를 1000으로 가정하고, 제일 간단한 모형으로 좀 더 쉽게 보여주기 위해서 혼합 정규 분포의 성분이 2($g = 2$)인 것으로 가정한다. 모의 실험 자료의 구성은 혼합 정규 분포는 각 집단에 대한 표본의 혼합 비율이 0.2(Group=A), 0.8(Group=B)로 나누었으며, 5개의 대조군과 5개의 실험군으로 가정하였다. 대조군의 분포는 평균이 0이고 분산이 1.0인 정규분포이며, 실험군중에서 800개의 평균도 평균이 0이고 분산이 1.0인 정규분포로 A집단과 B집단사이에 평균의 차이가 없는 집단, 즉 유전자의 발현의 변화가 없는 집단으로 가정하였다. 반면에 실험군중에서 200개의 표본은 분산을 1.0으로 일정하게 두고 평균은 각각 0.2, 0.4, 0.6, 0.8, 1.0, 2.0, 3.0, 4.0, 5.0으로 설정하여 대조군과의 평균 차이가 있는 집단, 즉 유전자의 발현의 변화가 있는 집단으로 9가지의 경우를 다루었으며 , noise가 같을 때 평균의 차이가 날수록 발현이 있는 유전자를 얼마나 잘 찾는지를 알아보았다. 또한 noise가 증가함에 따라서 발현이 변화된 유전자를 얼마나 잘 예측하는지를 알아보기 위해서 각각의 평균은 1.0으로 일정하게 두었으며, 표준 편차(Standard deviation, SD)는 각각 0.2, 0.6, 1.0, 1.5, 2.0, 2.5으로 가정하였다.

이 15가지의 경우에 대해 seed를 다르게 정해 놓고 난수 발생을 시킨 후에 각각의 경우에 대해서 EM 알고리즘을 통해서 5000번의 반복으로 로그 우도함수($\log L(\Phi_g)$)를 최대화(maximization) 시켜서 각 정규 분포에 대한 혼합 비율, 평균 , 분산 등의 모수 추정(parameter estimation)을 하였다. 여기에서 쓰인 초기값들은 10번정도 여러 다른 값을 주어서 로그 우도함수

(log-likelihood)의 값이 제일 최대인 것으로 추정한 모수를 선택하였고 최대화(maximization) 단계에서 5000번 이전에 수렴한 값들을 사용하였다.

여기에서 나온 결과를 토대로 민감도(sensitivity), 특이도(specificity) , 정분류율(accuracy rate), 오분류율(overall error rate), 가양성율(false discovery rate), 가음성율(false negative rate) 등을 구하여 이 알고리즘이 얼마나 정확하고 효율적으로 모형을 잘 예측하는지를 살펴보았다.

5.2 EM 알고리즘을 이용한 정규 혼합모형의 추정 (분산 일정)

5.2.1 예측모형의 모수 추정

각각의 표본에 대한 모수 추정은 다음 표와 같다.

표 5.1 정규혼합 모형 모수 추정

평균차Δ	$\hat{\pi}$		$\hat{\mu}$		$\hat{\sigma}$	
	A	B	A	B	A	B
0.2(z_1)	0.971	0.029	-0.031	0.110	1.105	2.124
0.4(z_2)	0.871	0.129	0.009	-0.111	1.001	1.978
0.6(z_3)	0.054	0.946	0.921	-0.067	2.591	1.071
0.8(z_4)	0.071	0.929	0.931	-0.053	2.450	1.117
1.0(z_5)	0.024	0.976	0.910	-0.005	3.326	1.293
2.0(z_6)	0.501	0.499	0.954	-0.535	2.188	0.766
3.0(z_7)	0.361	0.639	2.429	-0.557	3.216	0.991
4.0(z_8)	0.300	0.700	4.106	-0.577	3.805	0.987
5.0(z_9)	0.203	0.797	7.915	-0.515	2.264	1.151

평균의 차가 2.0 이하일 때에는 A, B 분포중의 한 분포가 다른 한 분포에 강력하게 영향을 주지 못하므로 두 정규 분포 모형에서 200개의 평균 차이가 나는 표본들을 정확히 찾아 내지 못하는 경향을 보이고, 평균 차이가 2.0 보다 크게 되면서부터는 급격하게 A, B 표본 분포의 특성을 잘 가려내는 것을 볼 수 있었다. 위에서 예측된 모형을 통해서 각각의 평균 차이별로 다음과 같이 오즈비(Odds ratio)를 이용하여 A와 B 두 그룹으로 나누었다.

$$odds_A = \frac{\text{각 자료가 } A \text{ 에 속하는 확률}}{\text{각 자료가 } A \text{ 에 속하지 않을 확률}} ,$$

$$odds_B = \frac{\text{각 자료가 } B \text{ 에 속하는 확률}}{\text{각 자료가 } B \text{ 에 속하지 않을 확률}} ,$$

$$odds\ ratio = \frac{odds_A}{odds_B},$$

$odds\ ratio > 1$: A 집단,

$odds\ ratio < 1$: B 집단 .

5.2.2 예측 모형의 평가

다음은 EM 알고리즘을 통해서 예측된 모형이 얼마나 우수한 예측력을 보유하고 있는지 , 모형의 평가를 하기 위해 오분류표(Confusion Matrix)를 이용하였다. 오분류표(Confusion Matrix)는 원래 난수(random number) 발생시에 분산은 같고 평균 차이가 나도록 설정해 놓았던 두 개의 집단 A, B를 예측모형이 정확하게 원래 집단으로 잘 분류한 빈도($A \Rightarrow A$, $B \Rightarrow B$)와 그렇지 못한 빈도($A \Rightarrow B$, $B \Rightarrow A$)를 함께 제시한 표로써 실제의 범주와 예측모형에서의 범주 사이의 관계를 나타낸 표라고 정의할 수 있다. 이

오분류표를 이용하여 민감도와 특이도를 그림으로 나타내었다. 민감도와 특이도는 다음과 같이 계산했다.

- 민감도 $= \frac{(\text{실제 } A,\ \text{예측 } A)\text{인 관찰치의 빈도}}{\text{실제 } A\text{인 관찰치의빈도}}$,

- 특이도 $= \frac{(\text{실제 } B,\ \text{예측 } B)\text{인 관찰치의 빈도}}{\text{실제 } B\text{인 관찰치의빈도}}$.

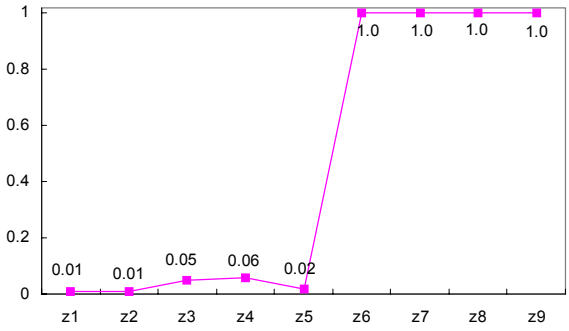

그림 5.1 민감도(분산일정 , 평균 차이)

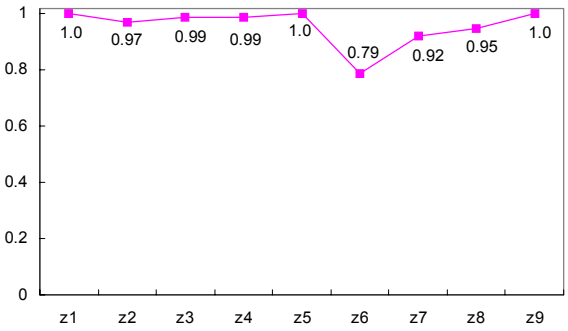

그림 5.2 특이도(분산일정 , 평균 차이)

민감도에서는 평균차이가 1.0인 표본까지는 200개의 A집단에서 최대 6%(11개)만을 제대로 식별함으로 상당히 판별능력이 부족하다고 말 할 수 있으나, 표본차이 2.0부터 99.5%(199개)로 급격하게 증가함으로 실제 A집단을 A집단이라고 판별하였고, 3.0이후로는 100%로 A집단을 정확하게 잘 예측하는 모형임을 알 수 있다. 특이도를 살펴보면 민감도가 급격하게 증가하는 z_6(평균차=2.0)에서 특이도가 79%(630개)로 그 이전의 구간과 급격하게 감소해서 B집단을 제일 낮게 예측했으며, 그 이외의 구간에서는 평균차이에 상관없이 최소 92% ~ 100%로 원래의 B집단을 B집단으로 아주 잘 예측하는 것을 볼 수 있다. 따라서 이 자료에서 위의 민감도와 특이도의 결과를 이용하여 다음과 같이 EM 알고리즘으로 유전자의 발현정도에 대한 정규 혼합모형의 예측 능력을 평가해 볼 수 있다. 발현의 변화가 있는 유전자를 A집단이라 할 수 있고, 발현의 변화가 없는 유전자를 B집단에 속한다고 할 수 있는데 평균의 차이가 크면 클수록 발현의 변화가 있는 유전자를 잘 식별할 수 있었으며, 민감도가 급격하게 증가하는 구간을 제외한 모든 구간에서는 평균의 차이에는 상관없이 발현이 없는 유전자를 잘 가려낼 수 있다는 것을 알 수 있었다.

다음 그림들은 정분류율(accuracy rate), 오분류율(overall error rate), 가양성율(false discovery rate), 가음성율(false negative rate)의 모의 실험 결과를 나타낸 것이다.

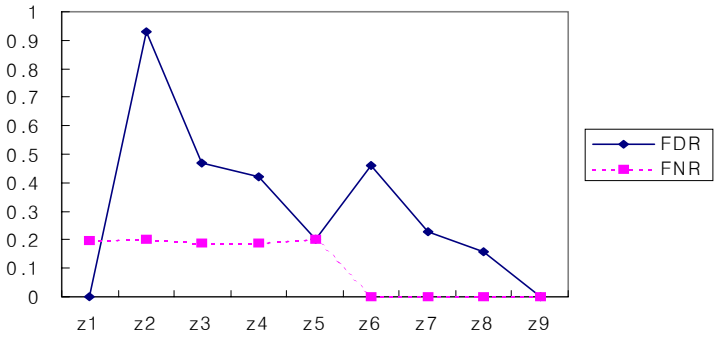

그림 5.3 가양성율과 가음성율(분산일정 , 평균 차이)

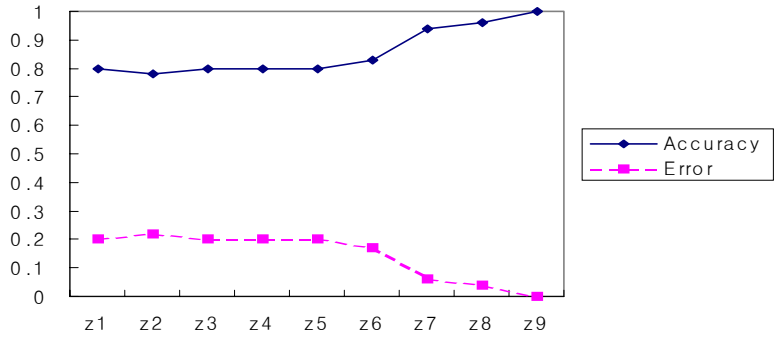

그림 5.4 정분류율과 오분류율(분산일정 , 평균 차이)

위에서 얻어진 값들을 구하는 식은 다음과 같다.

$$\text{가양성율} = \frac{(\text{실제 } B,\ \text{예측 } A)\text{의 빈도}}{\text{예측 } A\text{빈도}} \ ,$$

$$\text{가음성율} = \frac{(\text{실제 } A,\ \text{예측 } B)\text{의 빈도}}{\text{예측 } B\text{빈도}} \ ,$$

$$\text{정분류율} = \frac{(\text{실제 } A,\ \text{예측 } A)\text{의 빈도} + (\text{실제 } B,\ \text{예측 } B)\text{의 빈도}}{A+B\text{의 총 빈도}} ,$$

$$\text{오분류율} = \frac{(\text{실제 } A,\ \text{ 예측 } B)\text{의 빈도} + (\text{실제 } B,\ \text{ 예측 } A)\text{의 빈도}}{A+B\text{의 총 빈도}}.$$

위의 결과에서 가양성율(False Discovery Rate)은 평균 차이 2.0 이하의 구간에서 최소 20%에서 최대 93%사이를 불규칙하게 증가하고 감소하다가 3.0이상부터는 규칙적으로 최소 0%까지 감소함을 알 수 있으며, 가음성률(False Negative Rate)의 경우를 살펴보면 평균차이 2.0 이전까지는 평균의 차이에 관계없이 20%정도의 실험군을 대조군으로 잘못 식별하였으며 그 이후부터는 거의 0%에 근사함을 알 수 있다. 또한 정분류율(accuracy rate)은 평균 차이 2.0이상부터 증가하기 시작하여 그 이후로 최대 100%까지 예측을 잘 하였으며, 오분류율(overall error rate)은 두 표본의 평균 차이 2.0 ~3.0 사이에서 급격하게 감소해서 3.0이상부터는 거의 최소 0%에 근사해짐을 알 수 있었다. 따라서 이 모의 실험 자료에서, 위의 6가지의 결과로 미루어 볼 때 EM 알고리즘으로 혼합 정규 분포 모형을 예측하는 것은 평균차이가 2.0 ~ 3.0 이상만 되면 모형을 예측하는데 있어서 아주 좋은 추정 방법이 된다는 것을 알 수 있다.

5.3 EM 알고리즘을 이용한 정규 혼합모형 추정 (평균 일정)

5.3.1 예측 모형의 모수 추정

앞장에서는 9개의 표본들을 분산이 동일하고 평균 차이만 나도록 난수 발생을 시켰고, 지금부터 사용하는 표본들 5개는 평균은 동일하고 표준편차만 차이가 나도록 하여서 noise가 더 커질수록 모형이 어떻게 예측되는지를 살펴보겠다. 아래의 정규 혼합모형의 모수 추정 결과를 분석해 보면 noise

의 차가 클수록 혼합비율이 더 커지는 것을 볼 수가 있는데, noise를 최소 0.2로 두었을 때에는 거의 정확하게 0.2(A 집단) : 0.8(B집단)의 혼합비율을 잘 맞추었고, noise가 크게 증가할수록 혼합비율의 정확도도 조금씩 감소하는 경향을 보임을 볼 수 있으며 noise 크기가 평균과 같은 1.0이상이 되면서부터 모형을 정확하게 예측하는 능력이 급격하게 떨어짐을 알 수 있다.

표 5.2 정규 혼합모형의 모수 추정

SD차 Δ	$\hat{\pi}$		$\hat{\mu}$		$\hat{\sigma}$	
	A	B	A	B	A	B
0.2(z_1)	0.209	0.791	8.184	-0.623	3.257	1.086
0.6(z_2)	0.491	0.509	0.772	-0.486	2.059	0.926
1.0(z_3)	0.024	0.976	0.910	-0.005	3.326	1.293
1.5(z_4)	0.058	0.942	0.975	-0.067	2.594	1.083
2.0(z_5)	0.032	0.968	1.174	-0.012	2.874	1.095
2.5(z_6)	0.071	0.929	0.378	-0.025	2.348	1.060

5.3.2 예측 모형의 평가

모의 실험에서 noise를 조금씩 증가하며 변경해 주었을 때, 예측된 모형이 어떻게 평가되는지를 알아보기 위해서 앞의 5.2.2와 같은 방법으로 모형을 평가하였다.

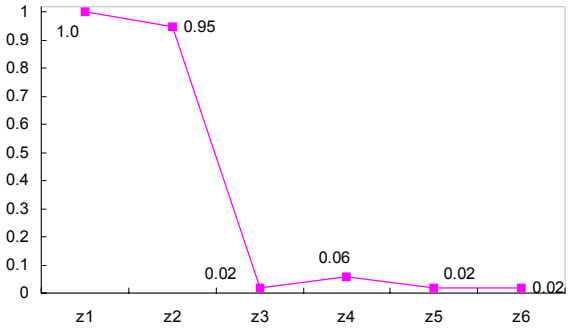

그림 5.5 민감도(평균일정 , SD 차이)

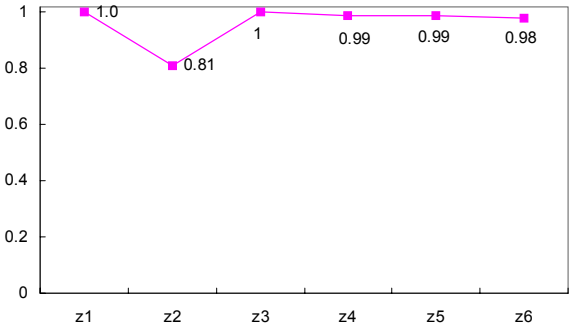

그림 5.6 특이도(평균일정 , SD 차이)

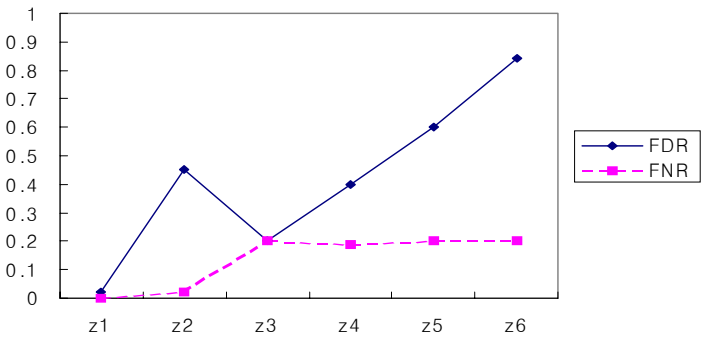

그림 5.7 가양성율과 가음성율(평균일정 , SD 차이)

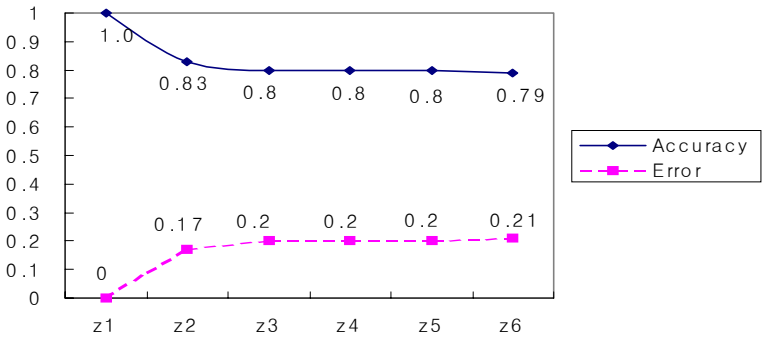

그림 5.8 정분류율과 오분류율(평균일정 , SD 차이)

앞의 그림 결과를 보면 민감도는 noise가 0.6일때 아주 좋은 예측율을 보이다가 0.6 ~ 1.0사이에 급격하게 예측율이 2%로 감소하고 1.0이후로는 거의 비슷한 양상을 보임을 알 수 있다. 또한 특이도는 noise가 0.2 ~ 0.6일 때 급격하게 감소하였을뿐 그 이외의 구간에서는 noise의 차이에 관계없이 유전자의 발현이 없는 집단을 잘 식별하였다. 특히 가양성율(False Discovery rate)의 경우 noise 0.6 ~ 1.0사이에 감소하기는 하지만 전체적으로 noise가 증가할수록 계속 증가하는 경향을 나타내고, 가음성율(False Negative rate)은 noise가 1.0일 때까지 증가하다가 그 이후로 더 이상 증가하지 않는 경향을 보임을 알 수 있다. 또한 정분류율과 오분류율은 noise가 0.6이상이 나면서부터 모형을 예측하는 능력이 떨어짐을 볼 수 있다.

지금까지의 결과를 정리해 보면, 이 모의실험 자료에서 EM 알고리즘으로 정규 혼합모형을 추정할 때에 평균의 차이가 적을 때에는 정규 혼합 모형을 잘 예측하지 못하지만 평균 차이가 2.0 ~ 3.0 이상만 나면 모형을 예측하는 데에 아주 좋은 추정 방법이 된다는 것을 알 수 있었고, 평균이 같고 noise를 증가시킨 경우에는 noise가 1.0 이하일 때 모형의 예측율이 좋다가 1.0이상의 차이가 나면서부터는 가양성율(False Discovery Rate)이 크게

증가하는 것으로 보아 모형을 예측하는데 좋지 않은 결과를 보임을 알 수 있었다. 따라서 이 자료의 모의실험 결과에서 평균의 차이가 2.0이상으로 되거나, 평균의 차이가 그리 크지 않더라도 noise가 0에 가까울수록 모형을 예측하는 능력이 좋음을 알 수 있다.

제 6 장 SAM과의 비교(Significance Analysis of Microarray)

2001년에 스탠포드 대학의 Tusher, Tibshirani and Chu가 Microarray 실험의 집합에서 강한 모형에 대한 가정을 하지 않으며, 유의한 유전자를 밝혀내기 위한 통계적인 방법을 제안했는데 바로 SAM(Significance Analysis of Microarray)이다. 이 프로그램의 가장 큰 장점은 사용자가 직접 기준점(cut-off point)을 정해 놓고 가양성율(False Discovery Rate)을 조정할 수 있다는 것이다. 앞의 4장에서 모의 실험을 했던 것과 같이 SAM을 이용하여 발현의 변화가 있는 유전자를 얼마나 잘 가려내는지, 즉 통계적으로 유의한 유전자를 EM 알고리즘으로 추정한 것과 비교해서 얼마나 잘 식별하는지를 알아보기로 한다.

6.1 SAM을 이용한 유의하게 발현변이한 유전자의 추정

먼저 두 조건(대조군, 실험군)에서의 유전자 발현의 상대적인 차이(relative difference)를 의미하며 t-통계량이라 정의하는 SAM score, d_i를 추정한다.

$$d_i = \frac{\bar{x}(i) - \bar{y}(i)}{s(i) + s_0}, \quad i = 1, \cdots, N(\text{유전자의 수}),$$

$\bar{x}_i$: 실험군에서의 유전자의 평균 발현 수치 ,

$\bar{y}_i$: 대조군에서의 유전자의 평균 발현 수치 ,

$s(i)$: 반복 발현 측정치(repeated expression measurements)

의 표준 편차(standard deviation) ,

s_0 : fudge factor(분산의 계수를 최소화시키기 위한 상수) .

다음으로 유전자 발현이 유의하게 변했는지를 알아보기 위해서 유전자들을 추정된 d_i의 순위대로 정렬시키는데,

$$d_{(1)} \leq d_{(2)} \leq \cdots \leq d_{(N)} .$$

B개의 순열(Permutation)을 바탕으로 각 순열 b에 대해서 통계량 $d_{(i)}{}^{*b}$를 계산하는데 다음과 같은 순서 통계량에 일치한다.

$$d^{*b}_{(1)} \leq d^{*b}_{(2)} \leq \cdots \leq d^{*b}_{(N)} , \qquad b = 1, 2, \cdots, B .$$

그리고 다음과 같은 기대 순서통계량(expected order statistics)을 추정한다.

$$\bar{d}_{(i)} = \frac{1}{B} \sum_{b=1}^{B} d^{*b}_{(i)} .$$

마지막으로, 유전자의 발현이 잠재적으로 유의하게 변화했는지를 알아보기 위해서 d_i(관찰된 score) : $\bar{d}_{(i)}$(기대 순서통계량)의 산점도(scatter plot)를 사용한다. 가양성율(False Discovery Rate)을 조절하기 위해서 고정되어 있는 적당한 상수 △(delta)를 결정할 수 있는데, 만약에 $d_i - \bar{d}_{(i)} > \triangle$ 이면 , 유의하게 변한 (significant positive) 유전자들이며, $\bar{d}_i - d_{(i)} > \triangle$ 이면 유의하게 변하지 않은(significant negative) 유전자들이라고 말할 수 있

다. 상수 △는 유의하게 변한 유전자들 사이에서 가장 작은 score d_i를 상위 기준점(upper cut-point , $cut_{up}(\triangle)$)이라 하고, 유의하게 변하지 않은 유전자들 사이에서 가장 작은 음수의 score d_i를 하위 기준점(lower cut-point , $cut_{low}(\triangle)$)이라고 한다. 이 기준점들을 기초로 하여 B개의 순열(permutation)로부터 가짜로 유의한 유전자(flase significant genes) 들의 평균적인 수를 추정하게 되는 것이다.

6.2 중이염 자료의 적용

1176개의 유전자를 가지고 있으면서 폐렴 구균성 중이염을 가지고 있는 5개의 실험군과 가지고 있지 않은 2개의 대조군의 자료를 이용하여 SAM을 이용하여 유의하게 발현된 유전자를 탐색하였다. 적용시킨 결과는 479개의 유전자가 발현 변화에 있어 유의하고 그 중에서 220개의 유전자가 유의하지 않은데 유의한 결과를 나타냈다.

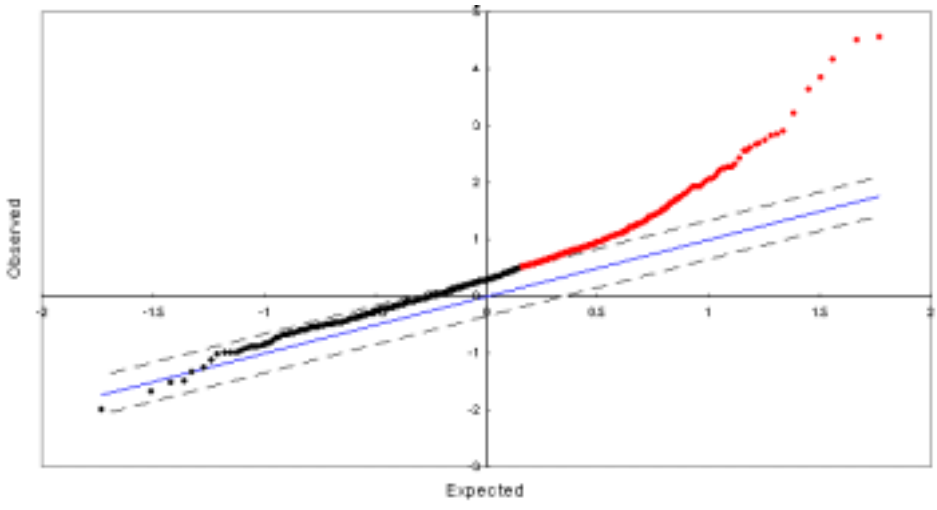

그림 6.1 중이염 자료에 대한 SAM의 결과

앞에서 EM 알고리즘을 통해 정규 혼합 모형에 적합한 결과에서는 단지 10개의 유전자만이 유의한 것으로 나왔는데, SAM과의 결과와는 너무나도 다른 차이를 보인다는 것이 흥미롭다. 따라서 다음 부분에서는 앞에서 모의 실험하였던 자료를 가지고 SAM으로 다시 모의 실험을 한 결과를 언급하겠다.

6.3 모의 실험 자료의 적용

앞장에서 사용하였던 자료와 동일하고 그 중에서 5개의 경우만을 다루기로 하겠다. 다음의 그림들은 SAM을 이용하여 위의 자료들중에서 5개의 경우, 즉 분산이 1.0으로 동일하고 평균이 0.2, 1.0, 2.0, 5.0인 경우에 대한 자료들의 결과를 나타낸 것이며, 위의 자료들에서 100개의 순열(permutation)로부터 가짜로 유의한 유전자(flase significant genes) 들의 평균적인 수를 추정한 결과는 다음과 같다.

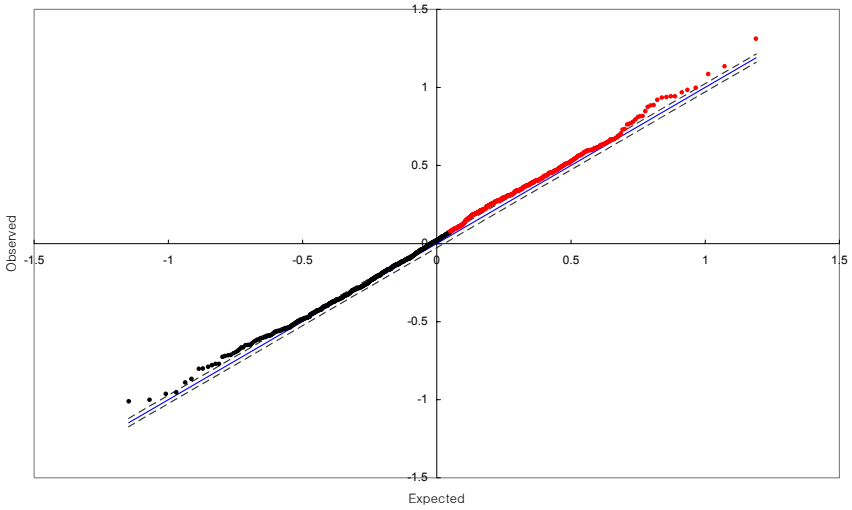

그림 6.2 N(0.2, 1.0) & N(0.0, 1.0)

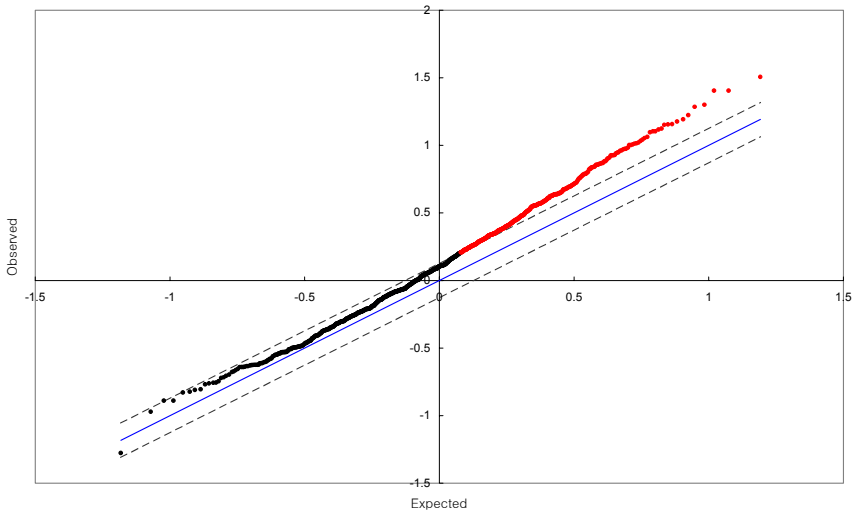

그림 6.3 N(1.0, 1.0) & N(0.0, 1.0)

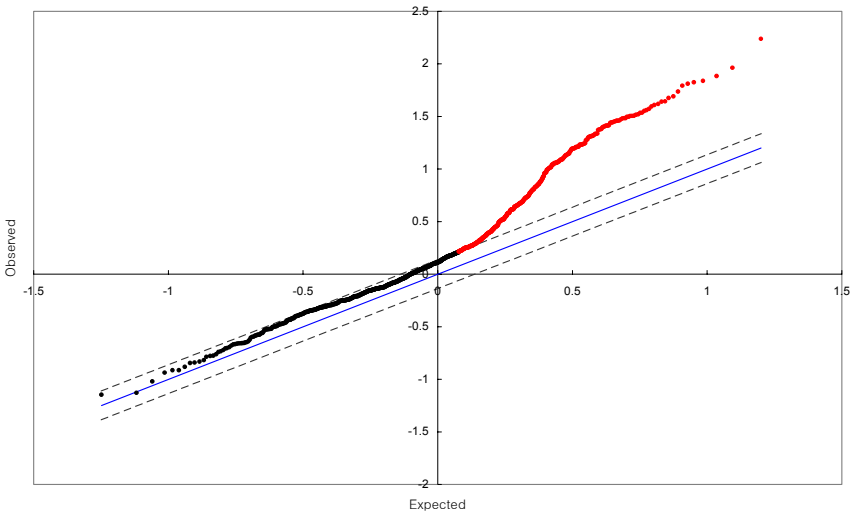

그림 6.4 N(2.0, 1.0) & N(0.0, 1.0)

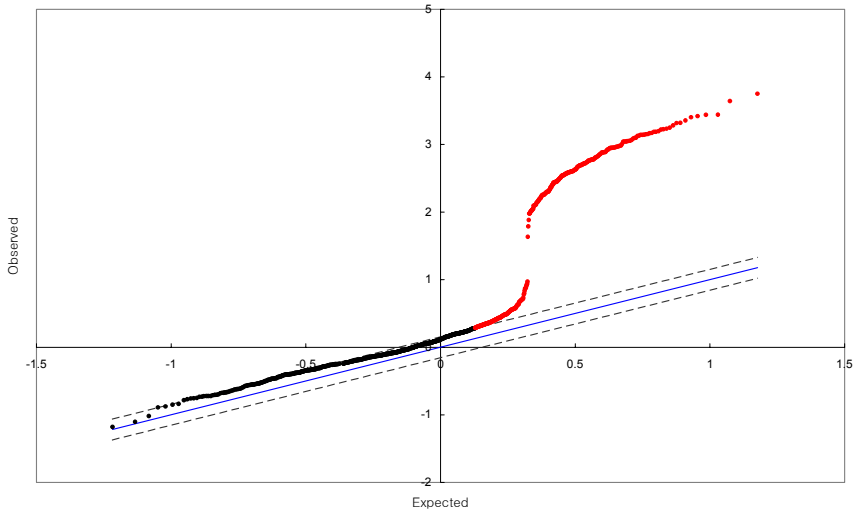

그림 6.5 N(5.0, 1.0) & N(0.0, 1.0)

실험군과 대조군의 평균의 차이가 0.2일 때에는 442개의 유의한 유전자중에서 Median 379.8개의 유전자가 가짜로 유의하다고 나와 가양성율이 86% 정도로 아주 높았고 그림을 통해서도 알 수가 있다. 또한 평균의 차이가 커질수록 유의한 유전자들 중에서 가짜로 발현 변화가 유의하다고 하는 유전자들의 수가 꾸준히 감소함을 알 수 있다. 또한 EM 알고리즘으로 발현 변화의 유의한 유전자들을 추정하는 것보다 과추정(over-estimated) 하는 경향을 보이고 있다. 실제로 모의실험 자료에서는 200개의 유전자만이 발현의 변화가 있는데, 이 결과들에서는 평균차이가 5.0에서 조차도 351개의 유전자가 유의한 것으로 결과가 나와 실제의 유의한 유전자보다 최소 1.75배정도 높게 추정되었음을 알 수 있다. 다음의 표는 실제로 유의한 유전자(# of significant gene)의 개수, 가짜로 유의한 유전자(# of false significant gene)의 개수, 가양성율(False discovery Rate)의 결과를 제시한 것이다.

표 6.1 SAM의 결과

		#Significant	#False pos (MEDIAN)	FDR (MEDIAN)
(A)	N(0.2, 1.0)	442	379.8	0.859
(B)	N(0.6, 1.0)	490	415.3	0.848
(C)	N(1.0, 1.0)	405	224.4	0.554
(D)	N(2.0, 1.0)	403	211.2	0.526
(F)	N(5.0, 1.0)	351	151.5	0.432

제 7 장 토의 및 결론

지금까지 microarray의 자료에 대해서 정규 혼합모형(normal mixture model)을 이용하여 실험군(Case)과 대조군(Control)으로 나누어진 이표본의 t-통계량에 대한 분포를 추정해서 실제로 특이하게 발현 변이된 유전자를 탐색하는 것을 제안하였다. 강한 모형의 가정이 확실치 않기 때문에 비모수적인 방법에 접근해서 추정을 하였는데, 먼저 정규 혼합모형(normal mixture model)의 정의를 설명하자면, 정규 혼합모형(normal mixture model)이란 하나의 원자료(Raw data)의 분포가 여러개의 다른 혼합 비율로 구성되어 있는 여러 부집단의 혼합모형을 의미하고, 이 논문에서는 이 혼합비율로 이루어진 각각의 혼합분포에 포함된 여러 부집단(subclass)의 모수와 여러 부집단의 혼합비율을 추정하였다. 이 추정된 분포의 결과로 어떤 유전자들이 질병이 없는 대조군의 유전자 발현 정도(expression level)에 비해 질병이 있는 실험군의 유전자 발현정도가 특이하게 변하였는지 변화가 없었는지를 알 수가 있었다.

또한 정규 혼합모형에서 모수를 추정할 때 최대 우도 추정량(maximum likelihood estimator)을 구하는 과정에서 로그우도함수(Log likelihood function)를 최대화시키는데 Expectation-Maximization(EM) 알고리즘을 이용하고 임의로(randomly) 여러번 다양한 초기값을 주어서 가장 큰 로그우도(Log-likelihood)값을 초래하는 추정된 모수들을 선택하였다. 실제 microarray 자료를 가지고 각 분포들의 모수를 추정하였고, 10개의 유전자의 발현에 변화가 있는 것으로 나타났다. 위의 결과에 대해서 정규 혼합모형을 EM 알고리즘으로 추정한 모형의 적합성을 검토하였는데, 1000개의 난수(random number)를 발생한 자료의 모의실험(simulation) 결과 두 집단의 분산이 같은 상태에서 평균의 차이를 증가시켰을 때에 평균차이가 1.0일 때에

는 민감도(sensitivity)가 2%로 원래의 발현변이(expression changed gene) 유전자를 식별하는 능력이 많이 부족하였다. 평균차이 2.0이상이 되면서 99.5%의 급격한 증가율을 나타내었고 3.0 이상부터는 100%의 아주 높은 발현변이 예측율(high prediction rate)을 나타내는 것으로 보아, 이 모의 실험 자료에서 EM 알고리즘을 이용한 정규 혼합모형이 원래의 발현변이 유전자를 식별하는 능력이 아주 좋음을 알 수 있었다. 또한 특이도(specificity)의 경우에는 민감도가 갑자기 증가하는 구간인 평균이 2.0일 때를 제외하고는 평균차이에 거의 상관없이 최소 92% ~ 100%로 원래 발현이 안된 유전자의 (expression unchanged gene) 식별을 잘 해내었고, 가양성율(False Discovery Rate)은 평균이 3.0일 때 거의 23%, 4.0일 때 16%, 5.0일 때에는 거의 0%에 가까운 결과를 나타내었으며 평균차이가 증가할수록 가양성율이 감소하는 양상을 나타냈다. 따라서 이 모의 실험자료에서는 질병을 가지는 실험군에서의 평균과 질병을 가지지 않은 대조군에서의 평균에 대한 차이가 2.0 ~ 3.0이상만 나면 정규 혼합모형으로 발현변이 유전자를 추정하는 것이 적합한 방법인 것으로 나타났다.

마지막으로 유전자 발현변이를 추정하는 또 다른 접근 방법인 SAM(Significance Analysis of Microarray)을 이용하여 위와 같은 모의 실험을 하였는데 원래 유전자 발현 변이 200개에 대해서 평균 차이 1.0에서 405개, 5.0에서조차도 351개가 발현 변화에 있어 유의한 유전자라는 결과를 보임으로 과추정(over-estimation)하였다. 또한 가양성율도 평균 차이 5.0에서 median 43%로 정규 혼합 모형과 비교했을 때 아주 높은 수치를 나타내었으며, 평균의 차이가 적어짐으로 최대 median 85%의 높은 가양성율을 나타내었다. 따라서 이 microarray 자료와 모의 실험 자료에서는, 발현변이 유전자를 탐색하기 위한 추정 방법으로 SAM에 비교해서 정규 혼합모형의 추정 방법이 더 적합하다는 것을 알 수 있었다.

본 논문에서 미처 다루지 못한 부분은 첫째로, 가양성율(False Discovery Rate)과 관련된 부분인데, 실제 microarray 자료를 가지고 정규 혼합 분포의 모형을 적합시킬 때 가양성율을 조절하면서 추정할 수 있는 어떤 기준점(cut-off point)을 마련한다면 더 신뢰적이고 타당한 추정 방법이 될 것이다.

둘째로, 평균 '0'을 중심으로 거리의 멀고 가까움에 따라서 발현 변이 유전자와 변화되지 않은 유전자의 집단으로 나누었는데, 앞에서 모의 실험하였을 때 두 분포의 평균이 많은 차이가 나지 않는 경우를 보았다. 이럴 경우에 발현이 변화된 집단과 발현이 변화되지 않은 집단으로 어떻게 구분하는지, 그 기준이 다소 모호하므로, 이것에 대한 절대적인 기준이 필요하다고 생각된다. 따라서 위의 두가지 문제는 앞으로 더 연구되어야 할 부분이다.

참고 문헌

김정숙 ,나종화 , S-Plus 사용법 및 프로그래밍, 자유아카데미, 2000

유종영, 이승천, 차경준 , 허문열 , S-Plus를 이용한 통계 계산, 박영사, 1997

Akaike, H. Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory(eds. B. N. Petrov and F. Csaki), 267-281 ,Akademiai Kiado, Budapest, 1973

Alexander Kamb, and Mani Ramaswami . Simple method for statistical analysis of intensity differences in microarray-derived gene expression data. *BMC Biotechnology. 1-8, 2001*

Bradley Efron, John D. Storey, and Robert Tibshirani. Microarrays, Empirical Bayes Methods, and False Discovery Rates. Stanford Technical report., 2001

Botstein, D. and Brown, P. Exploring the new world of the genome with DNA microarrays. *Nature Genetics(Suppl.)*,21 , 33-37, 1999

Dempster, A. P., Laird, N. M. and Rubin, D. B. Maximum likelihood estimation from incomplete data via the EM algorithm(with discussion). *J. R. Statist. Soc.* B, 39, 1-38 ,1977

Dudoit S, Yang YH, Callow MJ and Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Tech Rept, Stat Dept, UC-Berkely, 2000

Efron, B. , Tibshirani, R. , Storey, J. D. and Tusher, V. Empirical Bayes analysis of a microarray experiment. Manuscript, 2001

Eisen, M., Spellman, P., Brown, P. and Botstein, D. Cluster analysis

and display of genome wide expression patterns. *Proc. Nat. Acad. Sci.*, 95, 14863-14868, 1998

Gil Chu, Balasubramanian Narasimhan, Robert Tibshirani, Virginia Tusher. SAM "Significance Analysis of Microarrays" Users guide and technical document, 2001

Ideker, T., Thorsson, V., Siehel, A. F. and Hood. L. E. Testing for differentially expressed genes by maximum likelihood analysis of microarray data. *Journal of Computational Biology*, 7, 805-817, 2000

Kerr, M. K., Martin, M. and Churchill, G. A. Analysis of variance for gene expression microarraydata. *Journal of Computational Biology*,7 , 819-837, 2000

Lander, E. S. Array of hope. *Nature Genetics(Suppl)*, 21, 3-4, 1999

Lee, M-L T., Kuo, F. C., Whitmore, G. A. and Sklar, J. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Nat. Acad. Sci*, 97, 9834-9839, 2000

McLachlan, G. L. and Basford, K. E. Mixture models Inference and Applications to Clustering. Marcel Dekker, New York, 1988

McLachlan, G. L., Peel D, Basford KE, Adams P. Fitting of mixtures of normal and t-components. Journal of Statistical Software , 1994

Marina Sapir and Gary A. Churchill. Estimating the Posterior Probability of Differential Gene Expression from Microarray Data, 2001

Newton, M. A., Kendziorski, C. M. Richmond, C. S. Blattner, F. R. and Tsui, K. W. On differential variability of expression ratios improving statistical inference about gene expression changes from microarray data.

Journal of Computational Biology, 8, 37-52, 2001

Pan, W, Lin, J. and Le, C. A mixture model approach to detecting differentially expressed genes with microarray data. Technical report 2001-011, Division of Biostatistics, University of Minnesota, 2001

Pan, W., Lin, J. and Le, C. How Many Replicates of Arrays Are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach. *GenomeBiology, 2002*

Pan, W., Lin, J. and Le, C. Model-Based Cluster Analysis of Microarray Gene Expression Data. *GenomeBiology, 2002*

Steve selvin, Modern applied biostatistical method using S-PLUS, Oxford University Press, 1998

Titterington, D. M., Smith, A. F. M. and Makov, U. E. Statistical Analysis of Finite Mixture Distributions. Wiley, NewYork, 1985

Virginia Goss Tusher, Robert Tibshirani, and Gilbert Chu. Significance analysis of microarray applied to the ionizing radiation response. *PNAS,* , 98, 5116-5121, 2001

W.N.Venables, B.D. Ripley, Modern applied statistics with S-PLUS, Springer, 1999

ABSTRACT

Method for identifying of differentially expressed gene in microarray data using mixture model

Lee, Myung Hee
Dept. of Biostatistics and Computing
The Graduate School
Yonsei University

In this thesis, an algorithm is proposed for identifying differentially expressed genes , by using Normal mixture model in a microarray data.

Microarray technology is in the limelight of research tool for biotechnology because this makes it possible to generate the large amount of data at a short time and includes a merit to measure thousands of gene expression levels at once, However it is difficult to identify differentially expressed genes from the data with large noise, in practice.

In this thesis, the method is suggested to fit the normal mixture model using EM algorithm. And also suggested method identifies differentially expressed gene using odds ratio.

This method is applied to the data set containing expression levels of 1176 genes of rats with and without pneumococcal middle ear infection. Three components were founded, Two of the three components contain 99% genes whose expression levels are almost not changed, whereas the rest contains 15 genes with changed expression levels.

The simulation study shows that change of expression is identified well when the mean value of data is more than 2.0. However, eventhough the mean value is less than 2.0, then change of expression is also identified well when the noise is small.

Key words : cDNA Microarray, DNA chip, AIC, BIC, EM algorithm, Normal Mixtures, Differential gene expression, Odds ratio