응용통계연구 제 18권 1호, 2005년, pp. 115-127

DNA 마이크로어레이 자료의 PRINT-TIP별 표준화(NORMALIZATION) 방법*

이성곤[1] 박태성[2] 강성현[3] 이승연[4] 이용성[5]

요 약

DNA 마이크로어레이 기술은 수천 개 또는 수만 개의 유전자의 발현을 동시에 탐색할 수 있는 새로운 과학 기술이다. 표준화(normalization)는 마이크로어레이 실험에서 다양한 원인에 의해 발생하는 잡음(noise)을 줄이거나 제거하는 과정을 나타낸다. print-tip의 변동은 잡음의 주요 요인으로 인지되어왔다. 본 논문에서는 잡음의 주요 발생 요인이 되는 print-tip의 변동을 조절하기위한 print-tip 표준화 작업에 대한 객관적인 비교 및 그 타당성 평가를 하였다. 먼저 그동안 제안된 여러 표준화 방법들 중에서 가장 널리 사용되고 있는 방법들을 정리해서 소개한 후에, 잡음이 많이 포함된 실제 cDNA 실험자료를 이용하여 각 표준화 방법의 특성을 비교해 보았다. 또한 실험자료와 유사한 모의분포를 생성한 후에 print-tip 표준화 작업에 대한 체계적인 비교를 해 보았다.

주요용어: DNA 마이크로어레이, print-tip, 표준화.

1. 개요

Human Genome Project (HGP)의 10여년간의 연구 결과로 우리는 인간이 지니고 있는 30억개의 DNA 염기서열을 모두 해독하게 되었다. HGP는 생명공학의 급속한 발전과 함께 수많은 종류의 새로운 첨단 기술을 탄생시켰으며 이중에는 DNA chip 기술도 포함되어 있다.

DNA chip 기술은 다수의 유전자 발현상황을 총체적으로 탐색할 수 있는 기반 기술을 제공하고 있다. 즉, 한 두개의 유전자의 기능탐색이라는 종래의 한계를 벗어나 생명현상과 관련된 유전체수준의 연구가 가능해졌다는 것을 뜻한다. 이러한 DNA chip 기술에는

* 이 연구는 과학기술부의 학술연구비에 의하여 지원되었음 (Korean Systems and Biology Research Grant, M1030970000-03B5007).

1) (151-747) 서울시 관악구 신림9동, 서울대학교 자연과학대학 통계학과, 박사과정
E-mail: skon@biostats.snu.ac.kr
2) (151-747) 서울시 관악구 신림9동, 서울대학교 자연과학대학 통계학과, 교수
E-mail: tspark@stats.snu.ac.kr
3) (151-747) 서울시 관악구 신림9동, 서울대학교 자연과학대학 통계학과, 박사과정
E-mail: gadin@biostats.snu.ac.kr
4) (143-747) 서울시 광진구 군자동 98, 세종대학교 응용수학과, 교수
E-mail:leesy@sejong.ac.kr
5) (133-791) 서울시 성동구 행당동 17, 한양대학교 생화학교실, 교수
E-mail:yongsung@hanyang.ac.kr

Affimetrix사의 oligochip방식과 cDNA chip 방식이 있다. Affimetrix사에서 개발한 oligochip 방식은 반도체 집적기술을 접목시킨 방식으로 높은 집적도를 제공할 뿐만 아니라 신뢰성 높은 결과물을 제공하고 있다. 그리고, cDNA chip은 비교적 적은 비용과 쉬운 제작방식으로 인해 소규모 실험실에서 널리 사용되고 있다.

DNA chip에서 얻어진 자료를 간단히 마이크로어레이 자료라고 한다. 이러한 자료는 잡음(noise)이 많이 포함되어 있으며 또한 자료에 일정한 패턴을 보이는 경우가 많다. 특히 실험자의 숙련도와 실험에 사용된 화학물질 등에 따라 잡음의 양이 달라질 수 있다. 잡음이 추가될수록 자료의 품질은 떨어지기 마련이며 특히 일정한 패턴을 지닌 잡음은 분석 결과에서 큰 영향을 미칠 수 있다. 따라서 마이크로어레이 자료를 분석하는 초기 단계에서 잡음을 간단한 통계 모형을 고려하여 잡음을 제거하는 과정을 거친다. 이런 과정을 표준화(normalization)라고 한다.

초기 마이크로어레이 연구에서는 이러한 표준화 과정을 거치지 않고 실험에서 얻어진 결과를 그대로 분석했으나 점차 표준화의 중요성이 인식되면서 다양한 표준화 방법들이 제시되었다. DNA 마이크로어레이 분석에서 최초의 통계적인 분석법을 제시한 Chen *et al.*(1997)은 정규분포에 근거한 ML 추정법을 사용하여 표준화 방법을 제안하였다. Kerr *et al.*(2000, 2001)은 ANOVA 모형을 사용한 표준화 방법을 제시하여 표준화와 통계적 분석을 동시에 실시하는 방법을 제안하였다. Wolfinger *et al.*(2001)은 혼합모형(mixed model)에 기초하여 표준화과정을 위한 모형과 통계적 분석을 위한 모형을 각각 제안하였다. Yang *et al.*(2002)은 기존의 로그변환된 발현값의 비 M외에 이와 직교하는 A라는 새로운 측도를 제안하고 M을 A의 함수로 가정한 후에 이 함수를 Locally weighted scatterplot smoothing(LOWESS, Cleveland, 1979) 방법으로 추정하는 표준화 방법을 제안하였다. Tseng *et al.*(2001)은 Yang *et al.*(2002)의 방법을 모든 유전자에 대해 적용하지 않고 rank invariant 유전자들을 뽑은 후에 이들을 이용하여 표준화하는 방법을 제안하였다.

Quackenbush (2001, 2002)은 표준화 과정에 대한 중요성을 지적하였고 변수변환의 필요성을 강조하였다. Wang *et al.*(2002) 과 Chen *et al.*(2003)은 기존에 제안된 표준화 방법을 향상시키기 위한 새로운 방법을 제시하였다. 또한 Kernel이나 local polynomial regression을 이용한 비선형 표준화 방법들도 새로 제안되었다 (Workman *et al.*, 2002). Park *et al.*(2003)은 반복실험된 마이크로어레이 자료들로부터 각 유전자들 간의 분산값들을 추정한 후에 이 변동값들을 기준으로 어떤 표준화방법이 더 좋은지를 비교하였다. 또한 간단한 모의실험을 통하여 자료를 생성한 후에 역시 분산값들을 이용하여 각 표준화 방법들을 비교하였다.

마이크로어레이 실험에서 print-tip이 잡음의 주요 요인으로 지적되고 있다. 실제로 많은 마이크로어레이 자료에서 print-tip 별로 다양한 패턴이 나타난다 (Yang *et al.*, 2002). 이러한 print-tip간의 변동을 조절하기위해 print-tip 별로 표준화를 실시하는 print-tip 표준화가 널리 사용되고 있으나 이에 대한 필요성 및 타당성에 대한 평가는 거의 없다. Park *et al.*(2003)에서도 print-tip에 대한 연구 결과는 포함되어있지 않다. 본 논문에서는 그동안 제안된 여러 표준화방법들 중에서 가장 널리 사용되고 있는 몇 가지 표준화 방법들을 개괄적으로 소개하고, print-tip 표준화 방법에 대한 비교평가를 실시하고자한다. 먼저 실제 cDNA 실험에서 얻어진 자료를 통해 표준화 방법들을 소개하고 실험 자료와 유사한 형태의 모의

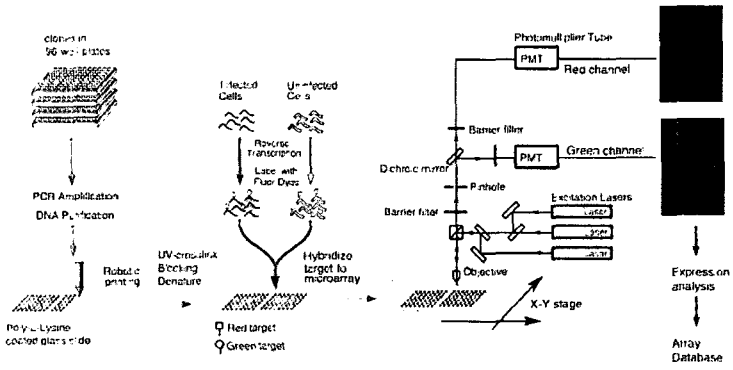

그림 2.1: cDNA 마이크로어레이의 제작 공정

실험 자료를 생성하여 print-tip 표준화 방법에 대한 비교 결과를 제시하고자한다.

2. DNA 마이크로어레이 실험과정

DNA 마이크로어레이 실험을 하는 방법에는 Oligochip을 사용하는 방법과 cDNA chip을 사용하는 방법으로 크게 나눌 수 있다. Affymetrix사에서 개발한 oligochip은 우선 각 유전자의 고유한 20개 bp정도 길이의 DNA sequence를 찾아낸 후 이 모든 sequence를 갖고 있는 온전한 것과 이 sequence 중 가운데 하나의 염기만 다른 sequence를 반도체 제조공정과 비슷한 과정을 통해 합성한다. 그 다음에 세포에서 추출한 mRNA를 cDNA로 만들어 형광물질을 부착한 후 미리 준비한 슬라이드에 화학반응을 시킨 후에 이미지 분석을 통해 발현값을 얻게 한다.

cDNA chip은 비교하고자하는 두 종류의 세포에 들어있던 mRNA의 상대적인 양을 측정한다. 여기 존재하는 mRNA의 양이 그대로 단백질의 양과 동일한 것은 아니지만, 앞으로 만들어질 단백질의 양을 나타내고 있다고 볼 수 있다. 특정 유전자를 전사(transcription)하여 만들어진 mRNA는 번역(translation)과정을 통해 단백질을 생성하는데 사용되므로 mRNA의 양은 해당 유전자의 발현정도를 나타내는 척도로 삼을 수 있다. 즉, 측정된 mRNA의 양이 많을 때에는 해당 유전자가 활성화되었다는 것을 뜻하며 적을 때에는 유전자가 비활성화되었다는 것을 뜻한다.

그림 2.1은 cDNA 마이크로어레이 실험에 대한 그림이다. 먼저 알려진 유전자를 대량으로 증폭하여, 미세하게 제작된 침을 이용하여 유리 슬라이드 위에 유전자를 찍는다. 다음에는 그림 2.1에서 보는 것과 같이 비교하기 위한 두 개의 다른 조직이나 세포들로부터 mRNA를 추출한다. 이 mRNA를 역전사(reverse transcription)시킬 때 두 개의 다른 색의 형광물질인 적색(Cy5)과 녹색(Cy3)으로 염색시킨 cDNA를 합성한다. 이렇게 합성된 두 개의 cDNA를 같은 양으로 섞은 후에 먼저 찍어 놓은 슬라이드 위의 유전자와 합성시킨다(hybridize). 그 다음 과정으로 합성이 안 된 유전자들을 씻어낸 후에 스캐너를 이용하여 합성된 각 유전

자의 발현강도를 이미지 분석을 통해 측정한다. 이런 식으로 슬라이드에 기록되는 cDNA는 수는 적게는 수백개에서 많게는 수만개에 달한다. 즉 cDNA실험을 통해서 동시에 수많은 유전자의 발현양상을 살펴볼 수 있다.

3. 표준화

마이크로어레이 자료에는 많은 잡음이 포함되어 있다. 예를 들어, cDNA 마이크로어레이 실험에서는 녹색 Cy3와 적색 Cy5 염료간의 형광 물리적 차이에 의해서 잡음이 발생할 수 있으며 형광염료의 혼합비율의 차이에 의해서도 잡음이 발생할 수 있다. 또한 이미지 분석에서 스캐너의 레이저 강도 등의 다양한 요인에 의해서도 역시 잡음이 발생할 수 있다. 표준화는 유전자 발현값에 영향을 미치는 다양한 형태의 잡음을 찾아내어 제거하는 과정이라고 할 수 있다.

3.1. 표준화하는 데 사용하는 유전자

표준화의 첫단계는 우선적으로 어떠한 유전자 자료를 가지고 표준화 할 것인가를 결정하는 단계이다. 표준화하는 자료를 선정하는 방법에서 크게 두가지로 나눌 수가 있다.

첫 번째 방법은 기준 유전자를 이용하는 방법으로 실험전에 미리 표준화에 적절하다고 생각되는 기준 유전자를 선택하여 이를 일부러 마이크로어레이에 첨가하는 방법이다. 기준 유전자로는 세포내 생명활동을 위해 항상 일정량이 발현된다고 생각되는 housekeeping gene 또는 실험에서 발현되지 않도록 제작된 spiked gene등이 사용된다. 그러나 실제 실험을 통해서 기준 유전자들도 일정한 발현값을 보이지 않고 실험에 따라 다양한 발현값을 보이고 있는 것으로 나타나 적절치 못한 방법으로 인식되고 있다.

두 번째 방법은 실험 실시 후에 표준화에 사용할 유전자를 사후에 선택하는 방법으로 유전자들 중에서 일정한 발현양상을 보이는 rank invariant gene만을 선별한 후 표준화하는 방법 (Tseng *et al.*, 2001)과 전체 유전자 자료를 다 사용하는 방법이 있다.

3.2. 표준화 방법

DNA 마이크로어레이 실험에서 얻어진 자료에서 Cy3의 발현값을 G, Cy5의 발현값을 R이라고 하자. 실험 대상의 전체 유전자 수를 p라고 하고 각각의 유전자를 j로 나타내자. 발현값의 비(ratio) M과 intensity A는 다음과 같이 정의된다 (Yang *et al.*, 2002).

$$M = \log \frac{R}{G} = \log R - \log G, A = \log \sqrt{GR} = \frac{1}{2}(\log G + \log R)$$

표준화 방법은 global(G) 표준화방법과 A를 고려하는 intensity dependent(ID) 표준화 방법으로 구분한다. G 표준화방법은 각 유전자별로 M을 다음과 같이 표준화한다.

$$M_j^{Global} = M_j - \hat{c}$$

여기서 $\hat{c}$는 M의 중앙값을 이용하여 추정할 수 있다. ID 표준화 방법에는 선형관계를 가정하는 경우와 비선형관계를 가정하여 표준화하는 방법으로 나눌 수 있다. ID 비선형 표준화 경우에는 LOWESS와 같은 비선형 모형을 이용하여 다음과 같이 표준화한다.

$$M_j^{LOWESS} = M_j - \hat{c}(A_j)$$

여기서 $\hat{c}$가 적합한 비선형함수이다.

마이크로어레이 실험에서는 여러 개의 print-tip을 사용하여 실험하므로 print-tip이 잡음 발생의 큰 원인이 될 수 있다. print-tip 표준화 방법은 각 표준화 방법들을 print-tip별로 실시한다. ID 비선형 표준화에 적용해보면

$$M_{jk}^{LOWESS} = M_j - \hat{c}(A_{jk}) \quad k = 1, \cdots, K$$

과 같이 된다. 여기서 j는 유전자를 나타내고 k는 print-tip을 나타낸다. 표준화 방법에 대한 더 자세한 설명은 Yang *et al.*(2002)과 Park *et al.*(2003)을 참조하기 바란다.

4. 예제

4.1. 실험 자료

이 실험의 목적은 흰쥐 embryo의 각 뇌신경조직 부위에서 분리된 신경기간세포(neural stem cell)를 이용하여 분화 기전에 관계하는 유전체의 조절 인자와 novel 한 세포 표지 유전체의 발현을 알아보는 것이다. 이를 위해 태생 14일 된 흰쥐 태아의 대뇌피질 신경조직(cortex) 에서 추출한 신경기간세포를 신경세포 (neuron), 성상세포 (astrocyte), oliodendrocyte의 세포로 분화유도하였다. 또한 신경기간세포를 CNTF를 매일 10ng/ml 로 첨가하여 성상 세포만으로 분화 유도하였다. 이 두 비교그룹을 편의상 No CNTF 군과 CNTF 군으로 부르자. 이 두 세포군의 분화 1일부터 5일 동안의 유전체 발현 변화를 살펴 보기 위해서 2400개 이상의 알려진 유전체와 1700 여개의 novel 유전체가 찍힌 유리칩을 이용하여 cDNA 마이크로어레이 실험을 수행하여 유전체 발현 양상을 관찰하였다.

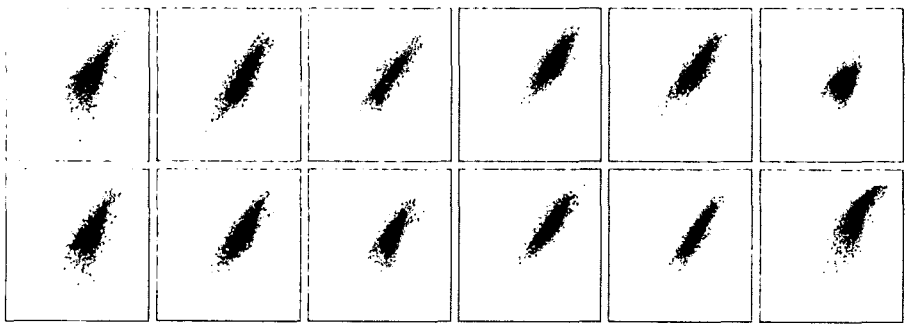

그림 4.1: 원자료의 산점도 (일부)

마이크로어레이 실험 설계는 CNTF를 첨가하지 않은 No CNTF 그룹과 첨가한 CNTF 그룹 각각에 대하여 mRNA를 추출하여, Cy5로 염색한 다음, Cy3로 염색한 reference cDNA와 hybridization하여 제작하였다. 각 처리 그룹을 분화 12시간, 1일, 2일, 3일, 4일, 5일이 지났을 때에 각각 3번씩 반복 실험을 실시하였다. 즉, 두개의 그룹에 세 번에 걸쳐, 여섯 시점에서 실험을 하여 총 36번의 cDNA 마이크로어레이 슬라이드를 얻었다. 이렇게 얻어진 자료를 스캔하여 이미지 파일을 생성한 후에 ImaGene v.3으로 이미지 분석을 하여 자료를 얻었다. 그림 2은 이렇게 얻어진 36개의 슬라이드의 자료 중 일부 슬라이드의 산점도를 그린 그림인데, 가로축은 Cy3로 염색된 reference 유전자의 발현량을 log 변환한 값이고 세로축에는 Cy5로 염색된 각 그룹 세포에서의 유전자 발현량을 log 변환한 값을 나타낸다.

그림 4.1의 산점도에서 가로로 3개의 슬라이드는 반복 실험된 슬라이드이다. 이 그림으로부터 이 실험의 재현성(reproducibility)이 많이 떨어지는 것을 알 수 있다. 이는 슬라이드마다 잡음이 많이 포함되고 있다는 것을 나타내는 것이며 따라서 표준화가 필요함을 보여주고 있다.

4.2. 표준화 방법들의 비교

각 표준화방법들을 구체적으로 비교하기 위해 실험에서 얻어진 자료를 각 표준화 방법들로 표준화하여 보았다. 그림 4.2과 그림 4.3는 36개의 슬라이드 중에서 선택한 두 개의 슬라이드에 대하여 앞에서 소개한 표준화 방법들을 적용시켜 구한 결과들이다. x-축은 $\log G$값을 y-축은 $\log R$값을 나타낸다. 표준화된 $\log R^*$값은 표준화된 $\log G^*$값이 표준화하기전의 $\log G$값과 동일하도록 고정한 후에 $\log R$값을 아래와 같이 변환시켜 구한 값이다. 즉,

$$\log R^* = \log R - (\hat{M} - M)$$

이 된다. 그림 4.2의 슬라이드는 비교적 실험이 잘된 경우를 보여준다. 이 슬라이드에 대해서는 각 표준화 방법 간에 큰 차이가 나타나지 않았다. 이에 반해 그림 4.3는 잡음이 많이

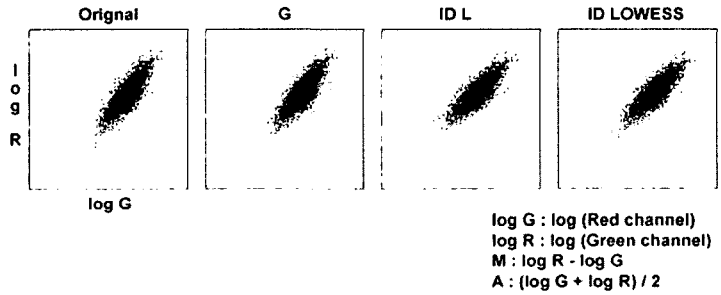

그림 4.2: 하나의 슬라이드에 3가지 표준화 방법을 적용한 그림. 원쪽에서 오른쪽 순서로, 원자료, global median of log ratio, ID linear regression, ID LOWESS 표준화 방법을 적용한 결과이다.

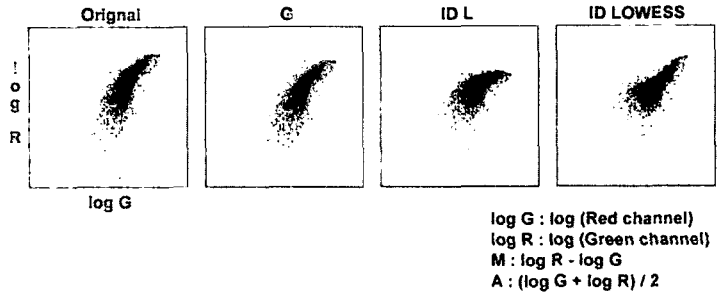

그림 4.3: 다른 슬라이드에 적용한 예

포함된 슬라이드를 보여준다. 원자료에는 굽어지는 비선형 패턴이 있음을 볼 수 있다. 물론 이 패턴이 생물학적 패턴일 수 있으나 반복 실험된 다른 자료를 비교해보면 잡음임을 쉽게 알 수 있다. 즉, 표준화 작업을 통해서 이런 패턴을 제거시키는 것이 바람직하다. 이 자료에 대해 표준화 방법을 적용해 보면 ID LOWESS 표준화 방법만이 이러한 패턴을 없앨 수 있음을 볼 수 있다. 이에 반해 다른 방법들은 선형적인 변화만을 보정했을 뿐 패턴 자체를 제거하지는 못함을 볼 수 있다. 그러한 어떠한 표준화 방법도 발현값이 작은 부분에 일반적으로 나타나는 과산포(over-dispersion) 문제에 대해서는 큰 효과를 보여주지 못했다.

그림 4.4는 그림 4.1에 있는 여러 슬라이드에 대해 ID LOWESS 표준화 방법을 적용하여 표준화시킨 후의 그림이다. 역시 비선형의 패턴이 모두 없어졌음을 확인할 수 있다.

5. 모의실험을 통한 print-tip 표준화 방법의 비교

본 연구에서는 앞에서 소개한 다양한 표준화 방법들 중에서 어떤 표준화 방법이 가장 효과적인지를 알아보기 위해서는 참값을 알고 있는 상태에서 실제 실험 자료와 유사한 형

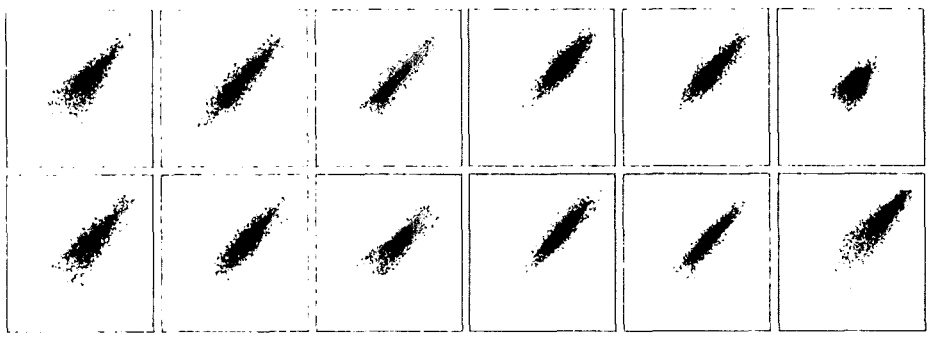

그림 4.4: 그림 4.1에서 소개된 원자료를 ID LOWESS 표준화 방법으로 표준화한 후의 산점도 ($\log G$ 값을 고정하여 그려진 산점도)

태가 나오는 잡음들을 첨가하여 가상의 실험 자료를 생성하고 표준화한 후에 그 결과와 참값을 비교해 보고자 한다. 특히 Park *et al.*의 모의실험을 확장하여 print-tip 별 표준화 방법을 평균오차제곱합(mean square error, MSE)를 이용하여 비교해 보고자 한다.

5.1. 자료의 생성

실험자료를 살펴보면 크게 몇가지 유형으로 나눌 수가 있다. 아무런 패턴이 없으며 산포도 고른 경우 (Type I), 산포는 고른 편이나 패턴이 존재하는 경우 (Type II), 패턴은 없지만 산포가 고르지 않은 경우 (Type III), 패턴이 있으면서 산포도 고르지 않은 경우 (Type IV)로 나눌 수 있다. 각 경우를 고려하여 유사한 분포를 나타내는 모양 4가지를 생성해보았다. 자세한 모형 생성과정은 Park *et al.*(2003)에 설명되어 있다. 그림 5.1이 각 유형별로 생성된 마이크로어레이 자료의 전형적인 모습을 보여준다. Type I은 산포가 고르며 패턴이 들어가지 않은 이상적인 분포로서, 이상적인 실험결과에서 기대되는 형태일 것이다. 여기서는 어떤 유전자도 다르게 발현되고 있지 않다고 가정하고 있다. 이 자료는 다른 자료들과 비교할 원자료가 된다.

Type II는 산포는 고른편이나 특정 패턴이 들어간 분포를 모사한 분포이다. Type III는 발현값이 작은 부분에서 산포가 증가하는 경우를 모사한 분포이고 Type IV는 발현값이 작은 부분에서는 산포가 증가하면서 또한 동시에 특정한 패턴을 갖고 있는 경우를 나타낸다.

그림 5.2은 4종류의 Type들을 두 개씩 묶어서 새로 생성한 6종류의 새로운 Type들을 보여준다. 이 그림은 서로 다른 효과를 갖고 있는 두 종류의 print-tip이 존재하는 가상적인 슬라이드의 자료를 생성한 것이다. 각 표준화 방법별로 print-tip 효과를 고려한 경우와 고려하지 않은 경우로 구분하여 표준화 방법을 적용하여 print-tip의 효과를 추정해 보았다.

5.2. 평균제곱합(Mean square errrors)의 비교

편의상 모의실험 자료를 생성할 때에 $\log G$에는 잡음이 첨가되지 않는다는 가정을 하

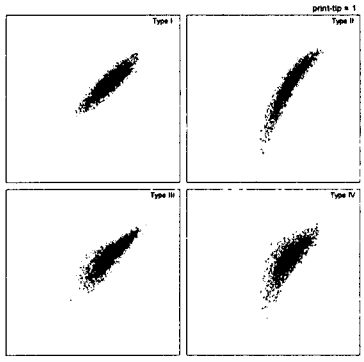

그림 5.1: 4종류의 모의 모사분포: Print-tip이 1개인 경우

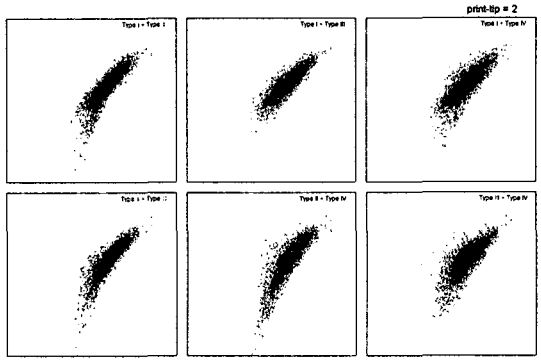

그림 5.2: Print-tip이 두 개로 두 종류의 모사분포 Type이 섞여있는 경우

였고 표준화된 $\log G$값도 변하지 않게 설정하였다. 따라서 표준화된 $\log R$이 $\log R^N$이라 할 때, 각 유전자마다의 변화량은 $|\log R_j^N - \log R_j|$이고 전체 유전자의 평균적인 변화량은 아래와 같은 표본 평균오차제곱합 MSE(sample mean squares error)를 이용하여 구할 수 있다.

$$\frac{1}{n}\sum_{j=1}^{n}(\log R_j^N - \log R_k)^2$$

이 표본 MSE 값이 작으면 작을수록 잡음이 잘 제거된 것이라 할 수 있다.

표준화방법들을 비교하기 위해 5000개의 유전자를 지닌 유사분포를 500번씩 생성하여 표본 MSE를 구하고 그의 평균값을 구한 값을 표 5.1과 그림 5.3에 제시하였다. 여기서 참

표 5.1: Type 별로 유사분포 생성후 각 표준화 방법을 적용하여 얻은 MSE의 평균값

Case	Types	G	L	N	G-pt	L-pt	N-pt
1	Type I+I	8.8e-05	1.2e-04	5.2e-04	1.8e-04	2.3e-04	6.2e-04
2	Type II+II	1.7033	0.6230	0.5146	1.7031	0.6227	0.5167
3	Type III+III	0.7849	0.7507	0.7216	0.7848	0.7505	0.7227
4	Type IV+IV	1.0234	0.6978	0.5613	1.0234	0.6974	0.5644
5	Type I+II	1.1791	0.7470	0.6305	1.1259	0.5852	0.5321
6	Type I+III	0.6648	0.6503	0.6426	0.6651	0.6480	0.6339
7	Type I+IV	0.7967	0.6657	0.5865	0.7821	0.6208	0.5550
8	Type II+III	1.2863	0.8148	0.6893	1.2408	0.6858	0.6187
9	Type II+IV	1.3654	0.6924	0.5624	1.3585	0.6594	0.5402
10	Type III+IV	0.9153	0.7515	0.6536	0.9024	0.7222	0.6421

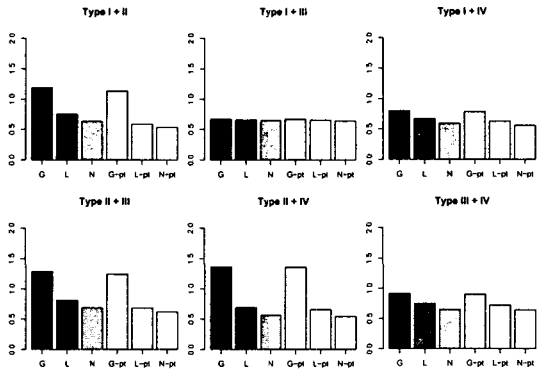

그림 5.3: 2종류의 print-tip이 있고 Type이 섞여 있는 경우에 유사 자료 생성후 각 표준화 방법을 적용하여 얻은 MSE의 평균값

값은 생성된 유사분포의 Type I의 $\log R$이다. 여기서 G는 global 표준화 방법을, L은 ID linear 표준화 방법을 N은 ID nonlinear LOWESS 표준화 방법을 나타낸다. 각 방법 뒤에 붙은 "-pt"은 print-tip 별로 표준화 방법을 적용한 결과를 나타낸다.

표 5.1에서 Case 1부터 4까지는 2개의 핀을 사용했지만 핀이 동일한 Type을 갖는 경우를 나타낸다. Type I의 경우는 표준화를 하더라도 아무런 효과가 없으므로 MSE값이 다 0으로 나온다. 즉 이 자료는 표준화가 필요 없는 형태의 자료라는 의미이다. Type II, III, IV의 경우는 정도의 차이는 있지만 N이 L보다 더 좋고, 또 L이 G보다 더 좋은 결과를 보여준다. 그리고 print-tip을 사용한 경우와 사용하지 않은 경우가 거의 차이가 없이 나오는 것을 알 수 있다. 즉, print-tip 효과가 없는 경우에 print-tip을 이용한 표준화를 했을 때에 과추정(over-parameterization) 문제가 거의 발생하지 않는 것을 확인할 수 있다.

Case 5부터 10까지는 2개의 pin을 사용하고 pin이 서로 다른 Type을 갖고 있는 혼합 형태를 갖는 경우를 나타낸다. 이 경우도 MSE값을 살펴보면 Case 1에서 4까지의 결과와 유사한 결과를 보여준다. 즉, 6가지 혼합 형태에 대해서도 N이 L보다 더 좋고, 또 L이 G보다 더 좋은 결과를 보여준다. 그리고 print-tip을 사용한 경우와 사용하지 않은 경우가 조금 차이가 나기는 하지만 G와 L의 차이나 G와 N의 차이만큼 크게 나타나지는 않음을 알 수 있다. 즉 print-tip 표준화 방법을 적용할지 여부를 결정하는 것보다 어떤 표준화 방법을 선택하는 것이 더 중요한 문제임을 알 수 있다.

5.3. 결과 및 토의

마이크로어레이 실험에서 얻어진 원자료에는 다양한 종류의 잡음이 포함되어 있다. 표준화 과정은 본격적인 마이크로어레이 자료의 통계적 분석 이전에 실시되는 가장 중요한 전처리 과정(pre-processing) 분석 방법이다. 초기에는 주로 간단한 G 표준화방법이 많이 사용이 되었지만 점차 비선형 LOWESS 방법 같은 복잡한 방법이 많이 사용되고 있다. 최근

에는 다양한 Kernel 이나 smoothing 기법에 근거한 표준화 방법들이 제시되었다. 처음 이 표준화 방법에 제안되었을 때에는 이 방법의 사용 여부에 대하여 실험연구자들과 분석자들 간에 많은 논란이 있었으나 이제는 마이크로어레이 자료의 분석 과정에 꼭 필요한 단계라는 것에 동의하고 있다.

본 논문에서는 대표적으로 널리 사용되는 G 표준화 방법과 ID 표준화 방법을 실제 cDNA 마이크로어레이 실험자료를 통해 비교해보고 모의실험 자료를 이용하여 print-tip별 표준화 방법을 비교해 보았다. 이 비교 연구를 통해서 다음과 같은 결론을 얻을 수 있었다. 실제로 print-tip 효과가 없는 경우에 print-tip 별로 표준화를 실시하더라도 심각한 과추정의 문제를 발생하지 않는 것을 알 수 있었다. 따라서 가능하면 모든 경우에 대하여 print-tip 별로 표준화를 실시하는 것이 바람직하다고 생각된다. 그러나 print-tip 효과가 있는 경우에 print-tip 표준화를 실시하더라고 효과는 생각보다 그리 크게 나타나지 않았다. 비록 그 효과가 작다할지라도 이후의 분석에 크게 영향을 미칠 수 있다. 즉, 유의한 유전자를 탐색하는 분석이나 유전자 군을 찾는 군집분석 등에 크게 영향을 미칠 수 있다. 따라서 그 효과를 무시해서 print-tip 표준화를 하지 않고 그 다음 단계의 분석으로 직접 넘어가는 것은 바람직하지 않을 것으로 판단된다. 또한 본 연구에서 ID 선형 표준화 방법도 LOWESS 방법과 유사한 결과를 보여 준 것도 주목해야할 점이라고 생각된다. 따라서 시간이 많이 걸리는 LOWESS 방법에 비해 훨씬 간단하게 표준화를 수행할 수 있는 ID linear 표준화 방법도 고려하여 print-tip 별로 표준화를 실시하는 것이 가장 바람직할 것으로 생각된다.

참고문헌

Chen, Y., Dougherty, E.R., and Bittner, M.L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images, *Journal of Biomedical Optics*, **2**, 364-374.

Chen Y.J., Kodell, R., Sistare, F., Thompson, K.L, Morris, S., and Chen, J.J. (2003). Normalization methods for analysis of microarray gene-expression data, *Journal of Biopharmceutical Statistics*, **13**, 57-74.

Cleveland (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association*, **74**, 829-836.

Johe, K.K., Hazel, T.G., Muller, T., Dugich-Djordjevic, M.M., and McKay, R.D. (1996). Single factors direct the differentiation of stem cells from the fetal and adult central nervous system, *Genes & Development*, **10**, 3129-3140.

Kepler T.B., Crosby L., and Morgan K.T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression, *Genome Biology*, **3**, research 0037.1-0037.12.

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

Kerr, M.K., Afshari, C.A., Bennett, L., Bushel, P., Martinez, J., Walker, N., and Churchill, G.A. (2001). Statistical analysis of a gene expression microarray experiment with replication, *Statistica Sinica*, **12**, 203-217.

Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., and Simon, R. (2003). Evaluation of normalization methods for microarray data, *BMC Bioinformatics*, **4**, 33.

Quackenbush J. (2001). Computational analysis of microarray data, *Nature Review Genetics*, **2**, 418-427.

Quackenbush J. (2002). Microarray data normalization and transformation, *Nature Genetics*, **32**, Suppl:496-501.

Tseng, G.C, Oh, M.K., Rohlin, L., Liao, J.C., and Wong, W.H. (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects, *Nucleic Acids Research*, **29**, 2549-2557.

Yang, Y.H., Dudoit, S., Luu, D.M, Peng, V., Ngai, J., and Speed, T.P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide sysytematic variation, *Nucleic Acids Research*, **30**, e15.

Wang, Y., Lu, J., Lee, R., Gu, Z., and Clarke, R. (2002). Iterative normalization of cDNA microarray data, *IEEE Transactions on Information Technology Biomedicine*, **6**, 29-37.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R.S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models, *Journal of Computational Biology*, **8**, 625-637.

Workman C., Jensen L. J., Jarmer H., Berka R., Gautier L., Nielser H. B., Saxild H. H., Nielsen C., Brunak S., and Knudsen S. (2002). A new non-linear normalization method for reducing variability in DNA microarrayexperiments, *Genome Biology*, **3**, research0048.1-research0048.16.

[2003년 5월 접수, 2004년 10월 채택]

Print-tip Normalization for DNA Microarray Data*

Sung-Gon Yi[1] Taesung Park[2] Sung Hyun Kang[3] Seung-Yeoun Lee[4] Yong Sung Lee[5]

ABSTRACT

DNA microarray experiments allow us to study expression of thousands of genes simultaneously. Normalization is a process for removing noises occurred during the microarray experiment. Print-tip is regarded as one main sources of noises. In this paper, we review normalization methods most commonly used in the microarray experiments. Especially, we investigate the effects of print-tips through simulated data sets.

Keywords: DNA microarray; Print-tip; Normalization.

* This work was supported by a grant from the Korean Ministry of Science and Technology (Korean Systems and Biology Research Grant, M1030970000-03B5007).

1) Graduate Student, Department of Statistics, Seoul National University, Shinrim-dong, Kwanak-gu, Seoul, 151-747, Korea
E-mail:skon@biostats.snu.ac.kr

2) Professor, Department of Statistics, Seoul National University, Shinrim-dong, Kwanak-gu, Seoul, 151-747, Korea
E-mail:tspark@stats.snu.ac.kr

3) Graduate Student, Department of Statistics, Seoul National University, Shinrim-dong, Kwanak-gu, Seoul, 151-747, Korea
E-mail:gadin@biostats.snu.ac.kr

4) Professor, Department of Applied Mathematics, Sejong University, Kunja-dong 98, Kwanjin-gu, Seoul, 143-747, Korea
E-mail:leesy@sejong.ac.kr

5) Professor, Department of Biochemistry, Hanyang University, Hengdang-dong 17, Sungdong-gu, Seoul, 133-791, Korea
E-mail:yongsung@hanyang.ac.kr