Bioinformatics and Biosystems 2006, Vol. 1, No. 1, pp. 8-16

Microarray 자료의 표준화 방법에 대한 고찰

이은경, 박태성

서울대학교 통계학과 생물정보통계연구실

초 록

DNA microarray 기술은 동시적으로 수천 개의 유전자의 발현상황을 탐색할 수 있다. 이 기술을 통해 얻어진 자료는 분석하기에 앞서 전처리 과정으로 배경보정 (background correction), 표준화 (normalization) 그리고 요약 (summarization)이 필요하다. 표준화란 microarray 실험에서 기술상의 문제로 첨가되는 일정한 잡음을 인식, 제거하기 위해 필요한 기법으로 그 동안 여러 방법들이 제시되어 왔다. 또한 마이크로어레이 자료의 분석을 위한 요약 방법으로도 많은 방법들이 연구되었다. 본 글에서는 표준화 방법들과 요약 방법들의 특성을 분석, 비교하고자 한다.

개 요

Human Genome Project의 10여 년간의 연구 결과로 우리는 인간이 지니고 있는 30억 개의 DNA 염기서열을 모두 해독하게 되었으며 이는 생명공학의 급속한 발전과 함께한 결과이다. 또한 이는 수많은 종류의 파생기술을 탄생시켰으며 이중에는 DNA chip 기술도 포함되어 있다.

DNA chip 기술은 기존 연구와 근본적인 차이를 보이는 획기적인 연구방법으로 다수 또는 전체 유전자 발현상황을 총체적으로 탐색할 수 있는 기반 기술을 제공하고 있다. 즉, 한 두 개의 유전자의 기능탐색이라는 종래의 한계를 벗어나 생명현상과 관련된 유전체수준의 연구가 가능해졌다라는 것을 뜻한다. 이러한 DNA chip 기술에는 cDNA chip 방식과 Affimatrix사의 oligochip 방식이 있다. 이 중 Affimatrix사의 oligochip 방식은 반도체 집적기술을 접목시켜 높은 집적도와 응용성뿐만 아니라 신뢰성 높은 결과물을 제공하고 있어 주목 받고 있는 기술이며 현재 여러 회사에서 개발에 성공했거나 추진 중에 있다. 그리고, cDNA chip은 비교적 적은 비용과 쉬운 제작방식으로 인해 현재 널리 사용되고 있다.

이러한 DNA chip에서 얻어진 자료를 DNA microarray 자료, 간단히 microarray 자료라고 한다. 이러한 자료는 보통의 실험 자료에 비해 잡음 (또는 비 생물학적 변이)이 많이 포함되어 있으며 또한 자료에 일정한 패턴을 보이는 경우가 많다. 잡음이 추가될수록 자료의 품질은 떨어지기 마련이며 실험자의 숙련도와 실험에 사용된 화학물질 등에 의해 그 정도가 달라질 수 있다. 특히 일정한 패턴을 지닌 잡음은 분석 결과에서 치명적인 효과를 발휘할 수도 있기에, DNA microarray 자료를 분석하는데 있어 전처리 과정을 마련하거나 분석에서 이러한 잡음을 분석모형에 고려하는 등으로 처리하여 잡음을 제거하는 과정을 거친다. 이를 표준화 (normalization)라 한다. 본 논문에서는 여러 가지 배경보정, 표준화, 그리고 요약의 방법들을 살펴보고 기존의 실험에서 얻어진 자료를 통해 방법들을 비교하도록 한다.

본 연구는 과학기술부지정 국가지정연구실인 생물정보통계연구실의 지원을 받았음

DNA microarray 자료

DNA microarray chip은 작은 고형체 기판 위의 특정 위치에 폴리뉴클레오티드 (polynucleotide)가 배열되어 있는 것으로 기판 위에 고정된 폴리뉴클레오티드를 probe라 부른다. 이 프로브가 cDNA 인 경우를 cDNA 칩, oligonucleotides 인 경우를 올리고 칩이라 부른다. 마이크로어레이를 만드는 방법에 따라서 크게 슬라이드 위에서 프로브를 합성 (synthetic)하는 형태의 마이크로어레이와 이미 외부에서 만들어진 프로브를 특정한 위치에 고정시키는 (spotted) 형태의 마이크로어레이로 나눌 수 있다.

cDNA 마이크로어레이 칩은 DNA를 특정한 위치에 고정시키는 형태의 마이크로어레이로 여러 개의 subarray로 나뉘며, subarray는 다시 같은 핀으로 찍히는 핀 그룹으로 구성된다. 이 핀 그룹은 각각 스팟으로 구성되어 있는데, 보통 이 스팟이 개개의 유전자를 의미한다. 그림 1은 cDNA 마이크로어레이 실험에 대한 그림이다. cDNA 마이크로어레이를 만들고 관심 있는 대상 (sample)에서 RNA를 추출한다. 추출한 RNA에 형광염료 (fluorescent dye)로 표지 (labeling)를 한다. 이렇게 표본을 준비하여 마이크로어레이에 혼성화 시켜준다. 결합이 되지 않은 유전자들을 씻어낸 다음에 스캐너를 통해 형광의 강도를 측정하여 수치로 나타내어 준다. 두 개의 대상 (sample)을 쓰는 경우 각각 적색 (Cy5)과 녹색 (Cy3)의 다른 형광염료로 표지 한 후 혼합하여 사용하고 하나의 대상을 쓴 경우는 한가지 형광염료로 표지 한다.

Affymetrix은 photo-lithography라는 기술로 만들어 지는 대

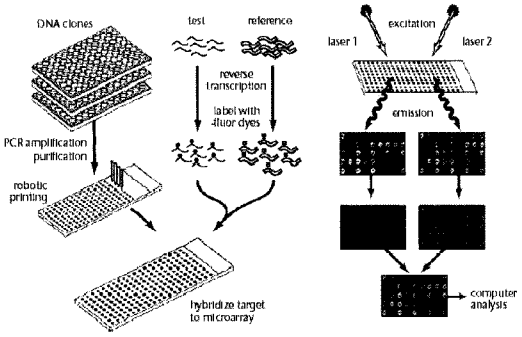

그림 1. cDNA 마이크로어레이 실험(Duggan *et al.*, 1999)

표적인 올리고 칩으로 유전자 서열 중에서 그 유전자를 대표할 만한 부분을 선별하여 25pb길이의 oligonucleotide를 이용한다. 하나의 유전자는 probe set이라고 불리는 16-20개의 probe pair로 이루어져 있다. 이들 pair는 perfect match (PM)와 mismatch (MM)로 구성되어 있다. 칩 위에 올려져 있는 프로브의 길이가 길면 정확히 일치되는 유전자가 아님에도 불구하고 비슷한 서열이 있는 부분이 프로브와 반응하여 발현 값을 보이는 경우가 있는데 이를 보완하기 위하여 mismatch (MM)을 사용한다. MM은 PM의 13번째 base를 바꾸어 만든 probe로 PM과 나란히 배열하여 불특정 결합을 측정하기 위한 것이다. 그러므로 하나의 스팟으로부터 하나의 유전자에 대한 발현정보를 얻을 수 있는 cDNA 마이크로어레이 칩과는 달리 Affymetrix 올리고 칩은 하나의 유전자에 대해 16-20개의 probe pair로부터 정보를 얻게 된다.

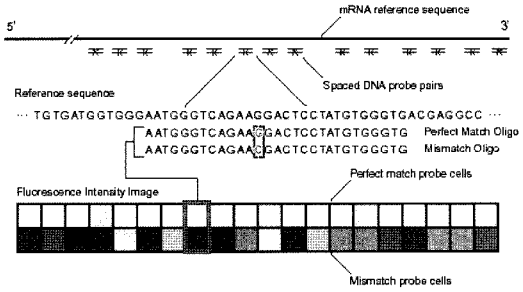

그림 2. Expression probe와 array design(Lipshutz *et al.*, 1999)

또 다른 종류의 올리고 칩으로는 spotted oligonucleotide microarray와 synthetic oligonucleotide microarray가 있다. 두 종류 모두 Affymetrix에서 사용하는 probe보다 긴 형태의 oligonucleotide를 이용하는 것으로 마이크로어레이를 만드는 방식이 다른 것이다. synthetic oligonucleotide는 긴 형태의 oligonucleotide를 슬라이드 위에 합성하는 형태로 만들어 지는 것으로 Affymetrix 마이크로어레이와는 사용하는 probe만 다른 형태가 된다. spotted oligonucleotide microarray는 Agilent 사에서 독점적으로 만드는 것으로 긴 형태의 oligonucleotide를 cDNA 마이크로어레이와 같이 염기 하나하나를 특정한 위치에 붙여 만드는 것이다.

이 두 가지 종류의 칩에서 나오는 자료는 크게 세가지로 분류할 수 있다. 첫 번째는 two channel array로 cDNA 마이크로어레이 실험에서 다른 형광염료로 표지 된 두 개의 sample을 혼합하여 사용한 실험에서 나오는 자료로 슬라이드를 적색 (Cy5)과 녹색 (Cy3) 채널에 대하여 스캔 한 후, 각각 gray 스케일의 이미지로 만들어 16-bit Tagged Image File Format (TIFF)형식의 파일로 저장한다. 그림 1의 cDNA 마이크로어레이 실험으로부터 얻어진 자료가 two channel 자료이다. 하나의 유전자에 대하여 적색과 녹색 채널로부터 각각 하나씩의 발현 량을 측정하게 된다. 두 번째는 one channel array로 cDNA 마이크로어레이 실험에서 하나의 sample을 하나의 형광염료로 표지 하여 사용한 실험이나 spotted oligonucleotide 마이크로어레이 또는 synthetic oligonucleotide 마이크로어레이를 사용한 실험에서 나오는 자료이다. 하나의 유전자에 대하여 하나의 발현 량을 얻게 된다. 그리고 마지막으로는 Affymetrix의 oligo chip으로부터 나오는 자료이다. 위에서 설명한 바와 같이 하나의 유전자에 대하여 PM, MM 각각 16-20개씩의 짧은 oligonucleotide를 프로브로 사용하여 32-40개의 발현 량을 얻게 된다.

이미지 분석 (Image analysis)

마이크로어레이 실험에서 스팟의 위치를 찾아내고 그 스팟의 정보를 얻어내는 과정을 통칭하여 이미지 분석이라 한다. Affymetrix사의 올리고 칩의 경우에는 이미지 분석 방법이 실험과정에 통합되어 있으므로 스팟을 이용한 cDNA칩을 중심으로 분석방법을 살펴보도록 한다. 적색과 녹색, 각 채널로부터 얻은 두 개의 이미지에서 픽셀이 높은 휘도값 (intensity value)을 갖게 되는 것은 적색/녹색 형광염료의 양이 각각 많은 부분인 것을 알 수 있고 이는 mRNA의 양이 많은 부분으로 생각할 수 있다. 이러한 픽셀들의 영역은 핀에 의하여 찍히는 슬라이드의 부분이며, 높은 휘도 값을 가질 것이다. 이 영역을 스팟 (spot) 또는 전경 (foreground)라 부르며 분석에 사용할 관심영역이 된다. 이 부분 이외의 휘도 값을 갖는 부분을 배경 (background)라하며 분석과정에서 보조적으로 사용된다.

cDNA칩 실험에서의 이미지 분석은 크게 격자화 (girding), 세분화 (segmentation)및 자료추출 (data ex□traction)단계로 구성되어 있다. 먼저 격자화는 실험정보를 이용하여 슬라이드 이미지에서 스팟이 있는 위치 (좌표 값)을 잡아주는 단계이다. 마이크로어레이 이미지의 기본적인 구조는 arrayer에 따라 결정된다. 실험 자는 실험에 사용된 핀 (pin)의 개수가 몇

개인지, pin-arrayer가 어떤 형태로 구성되어 있는지 그리고 부 격자 (subgrid)간의 거리와 각 부 격자 내의 스팟의 거리 등을 이용하여 격자화를 실시한다. 격자화 단계에서 얻게 되는 스팟 위치의 대략적 정보 (target mask)를 이용해서, 그 안의 픽셀들을 mRNA의 양을 나타내는 전경과 여러 가지 노이즈 (noise)에 의해 나타내는 배경으로 나누는 작업을 세분화라고 한다. 이와 같은 단계를 거쳐 최종적으로 생긴 스팟의 위치정보를 target patch라 한다. 마지막으로 자료추출단계에서는 스캔 되어 있는 이미지에서 target patch를 이용하여 발현정보를 나타내는 정보를 얻는 과정이다 (박태성 외, 2005).

전처리 과정(Preprocessing)

위에서 얻어진 마이크로어레이 자료는 많은 잡음을 포함하고 있는데 그 종류는 크게 두 가지로 나누게 되는데, 개개 자료에 랜덤 하게 되는 잡음과 자료 전체적 수준에서 일정한 패턴을 갖춘 체 첨가되는 잡음이다. 앞에 것은 개개 자료에 랜덤 한 만큼 제거하기가 어려우며 해당 자료의 산포를 증대시켜 통계적 검정력을 약화시킨다. 그에 반해 뒤의 것은 통계적 검정의 결과를 유도할 수 있는 위험성을 지닌 것으로 표준화의 목적이 이 잡음을 제거하는 것이라 할 수 있다.

전처리 (preprocess)란 실험에서 얻어진 자료를 실제로 분석하기 전에 거치는 단계로 마이크로어레이 자료에서는 크게 배경보정 (background correction), 표준화 (normalization), 그리고 요약 (summarization)의 세 단계로 나눌 수 있다. 배경보정 (background correction)은 이미지 분석에서 실제 신호(signal)가 검출되는 전경에서 함께 나타내어지는 배경의 효과에 대한 영향을 제거하기 위한 것으로 전경의 대표값에서 배경의 대표값을 빼주는 형태를 가장 많이 취한다.

표준화란 자료에 포함된 오차들 중 생물학적인 요인 이외에 다른 체계적인 요인으로 인해 발생되는 편의를 제거하는 과정이다. 마이크로어레이 자료는 실험자의 숙련도와 실험 설정에 따라 오차를 많이 줄일 수 있으나 실험 단위에서 발생하는 체계적인 편의가 관측되는 경우를 흔히 볼 수 있다. 이와 같이 필요한 정보자체를 왜곡시키는 체계적인 편의는 연구결과에 큰 영향을 미치게 된다. 그리고 마지막으로 요약 과정에서는 마이크로어레이 분석에서 각 유전자에 대하여 이를 대표하는 발현정보를 하나의 값으로 요약을 하여 차후의 분석에 이용하게 된다 (박태성 외, 2005).

이 장에서는 2장에서 설명한 세가지 다른 종류의 자료에 대하여 다양한 전처리 과정에 대하여 알아보도록 한다

Two channel data

우선 DNA microarray 실험에서 얻어진 자료에서 Cy3의 intensity를 G, Cy5의 intensity를 R이라 표기하도록 한다. 자료의 변화가 지수적인 생물학적인 자료의 특성상 두 세포의 발현의 차이를 나타내는 적절한 척도는 차이가 아니라 비율이 적절하므로 이를 로그변환을 한다. 즉, $M = \log\frac{R}{G} = \log R - \log G$가 된다. 여기서 로그변환을 하는 것은 자료의 형태가 전통적으로 다뤄왔던 선형 형태에 유사하다라는 점과 잡음의 효과가 줄어든다는 점 등의 장점을 얻을 수 있기 때문이다. 이에 추가로 M과 직교하는 척도를 마련한다. 이는 Yang et al. (2001)에서 intensity라 제안된 척도로 $A = \log\sqrt{GR} = \frac{1}{2}(\log G + \log R)$과 같다.

표준화 방법은 크게 global normalization과 intensity dependent normalization로 나눌 수 있다. 각각을 살펴보기로 하자

(1) Global normalization

가장 간단한 가정은 G와 R이 한 슬라이드 내에서 일정한 비를 이루고 있는 경우이다. 즉, R=c · G가 된다. 이 가정은 투입되는 Cy3, Cy5 형광물질의 특성상 이러한 잡음이 첨가되리라는 것에 기초하였다. 이를 다시 정리하면 log$R$= log$G$+log$c$이다. 따라서, M^{Gobal}은 각 유전자마다 표준화한 값으로 $M_j^{Global} = M_j - \hat{k}$와 같이 된다. K의 추정방법으로는 보통 이상치에 영향을 받지 않기 위해 $med\{M_1, \wedge, M_N\}$가 쓰이지만 (Yang *et al.*, 2002), 평균을 사용하거나 그밖에 일정 분포가정 하에 최대우도추정량 (maximum likelihood estimate, MLE)을 사용하기도 한다 (Chen, 1997). 이 방법을 여기서는 global median of log ratio 표준화 방법이라고 부른다.

그러나, 자료의 성질을 탐색해보면 이 정도의 가정으로는 부족한 점이 많음을 알 수 있다. 위에서 고려한 형광물질 차이뿐만 아닌 다양한 잡음이 첨가되기 마련이다. 그리하여 위의 가정을 보다 확장하여 $\log R = \beta_0^{RG} + \beta_1^{RG}\log G$ 와 같이 가정하고 모수들을 추정한 b_0^{RG}와 b_1^{RG}을 이용하여 표준화된 $\log R_j^{RG}$를 $\log R_j^{RG} = \frac{\log R_j - b_0^{RG}}{b_1^{RG}}$와 같이 하고, 이를 통해 표준화된 M_j^{RG}는 $M_j^{RG} = \log R_j^{RG} - \log G_j = \frac{\log R_j - b_0^{RG}}{b_1^{RG}} - \log G_j$와 같이 된다. 이 방법은 이미 개발되고 상용화된 여러 이미지 분석 소프트웨어에서 이용하는 표준화 방법이다. 이 또한 여기서는 global linear regression 표준화 방법이라 하겠다.

이를 여러 슬라이드의 경우로 확장시킨 방법이 Kerr et al. (2000)의 방법이다. 그러나, Kerr *et al.*은 별도의 표준화 과정을 마련하지 않고, 표준화에 사용되는 가정을 분석 모형 내에 포함한 ANOVA 모형을 제안하였으며, 여기엔 여러 슬라이드에 따른 변이를 반영하고자 하였다. 그 모형은 $y_{ijkl}=\mu+A+V_k+G_l+(VG)_{kl}+e_{ijkl}$과 같은 형태로, A는 슬라이드의 변이, V는 위에서 사용한 방식으로 해석하자면 Cy3 Cy5의 변이, G

는 유전자마다의 차이를 나타내는 인자이고, y는 Cy3 Cy5에서 나온 두 intensity를 하나로 모은 반응변수로 로그 변환한 값이다. 여기서 분석의 목적인 VG의 효과가 유의한 유전자가 어떤 것인지를 탐색하는 것이고 나머지는 표준화를 하기 위하여 추가된 것이다.

Wolfinger *et al.* (2001)은 Kerr et al. (2002)의 방법을 다시 표준화 과정과 분석과정으로 나누었고 표준화 과정에서 평균모형뿐만 아니라 분산모형도 고려하고 분석 또한 분산모형을 추가해 기존 통계이론으로 해석할 수 있도록 하였다. 표준화 모형은 $y_{ijk} = \mu + A_i + T_j + (AT)_{ij} + \varepsilon_{ijk}$이 되고 여기서 A와 AT가 랜덤 효과를 가진다고 가정하여 분산모형을 추가시켰다. 이렇게 추정한 값에서 잔차 $r_{ijk} = y_{ijk} - (\hat{\mu} + \hat{A}_i + \hat{T}_j + (\hat{AT})_{ij})$를 표준화된 값으로 한다.

그러나, 이러한 방법들은 선형적인 잡음만을 제거할 수 있고 슬라이드들을 살펴보았을 때 잡음이 비선형적으로 첨가된 자료가 상당수임을 볼 때 한계점을 지닌다 할 수 있다.

(2) Intensity dependent normalization

앞에서 소개한 방법들은 한가지 문제점을 지니고 있다고 할 수 있는데, 그것은 바로 G나 R값 중 하나의 값을 고정한 후에 표준화한다는 점이다. 비율의 특성상 자료의 형태는 기본적으로 R=G에 대칭적으로 분포하게 되는데 한쪽을 고정하여 고려하면 이러한 대칭성이 깨어지게 된다. 예를 들자면 생물학적인 정보를 담고 있지 않은 유전자 자료에서 (100, 120)과 (120, 100)은 잡음이 어디에 추가되었는가의 문제일 뿐 잡음이 추가된 정도에 있어서는 동일한 자료이다. 그러나, G를 기준으로 했을 때는 100과 120으로 별도의 성격을 지닌 자료로 분류되고 있는 것이다. 이러한 고려하에 Yang et al. (2001)은 intensity A를 제안하고 이것을 기준으로 표준화할 것을 제안하였다.

가장 간단한 표준화에 대한 가정은 $M = \beta_0^{MA} + \beta_1^{MA} A$와 같은 선형적인 형태가 될 것이다. 이에 표준화된 값 M_j^{MA}는 $M_j^{MA} = M_j - \hat{M}_j$가 된다. 여기서 $\hat{M}_j$은 회귀분석의 추정치를 이용할 수 있고 이때 M_j^{MA}는 잔차 값을 이용하게 된다.

그러나 슬라이드를 살펴보았을 때 잡음의 형태가 항상 선형적인 형태만이 아님을 알 수 있다. 그리고 이 경우 모수적으로 접근하기에는 낭비적인 면이 많아, Yang et al. 은 표준화 모형을 에 의존한 비선형적인 모형으로 확장을 제안하였다. 즉, 가정하기를 $M=f(A)$와 같은 A에 의존하는 일반적인 함수형태를 가정하고 이상점 (outlier)에 상대적으로 덜 민감한 LOWESS 함수 추정 법으로 추정하는 것이다 (Cleveland, 1979). 표준화된 M_j^{LOWESS}는 $M_j^{LOWESS} = M_j - \hat{f}(A)$와 같이 된다.

더 나아가 추가로 print-tip 별로 잡음의 효과가 다르게 나타날 수 있으므로 print-tip (PT) 별도 함수를 따로 추정할 수 있다. 즉, 가정하기를 $M = f_k(A)$이라 하고 표준화도 $M_{jk}^{LOWESS} = M_j - \hat{f}_k(A_{jk})$와 같이 될 것이다.

Affymetrix data

Affymetrix data는 two channel data와는 달리 하나의 프로브에 대하여 하나의 intensity 값만을 갖기 때문에 위에서 이용한 표준화 방법을 그대로 적용시키기 어려운 점들이 있다. 그리고 하나의 유전자에 대하여 perfect match (PM) 값과 mismatch (MM) 값을 계산하고 이 값으로부터 요약 값을 구하는 과정도 중요한 역할을 하게 된다.

먼저 배경보정으로는 가장 손쉽게 할 수 있는 $PM_j - MM_j$가 있다. Affymetrix사에서 제공한 초기의 MicroArray Suite (MAS 4)에서 제공한 방식으로 MM 값이 PM 보다 커지면 음수 값이 나오므로 로그변환을 취할 수 없어지고 요약 값 (expression value)또한 음수 값을 갖는 경우가 생긴다. 최근의 Microarray Suite (MAS 5)에서는 $PM_j - MM_j^*$ 를 이용한다. 여기서 MM_j^*은 짝이 되는 PM_j 값보다 크지 않은 MM_j로 $PM_j - MM_j$의 값이 0보다 크도록 보정을 해준 것이다. 이렇게 함으로써 로그변환으로부터는 자유로워 졌으나 PM 값이 MM 값보다 작은 경우에 대해서는 정보를 잃게 되는 것이다. Irizarry et al. (2003)은 불특정 결합 (non-specific binding)을 측정하기 위하여 제안된 MM 값이 대응되는 PM 보다 커지는 경우가 많아지는, 특히 유전자의 1/3이상이 되는 것을 발견하고 MM 값이 불특정 결합 (non-specific binding) 이라기 보다는 다른 유전자에 해당되는 특정결합 (specific binding)이라고 지적하고 PM 값만을 이용한 배경보정 방법으로 $PM = background + signal$ 형태의 모형을 제안하였다. 여기서 $background$는 정규분포, $signal$은 지수분포를 가정한 후 $E(Signal \mid PM)$의 추정치를 배경보정을 한 값으로 사용한다.

다음 단계인 자료를 표준화시키는 가장 손쉬운 방법은 Affymetrix사에서 제공하는 microarray suite (MAS)에서 사용하는 방법으로 각 칩 별로 전체적인 scale을 보정하는 방법을 이용한다. 이는 각 칩 내의 편의는 보정해 줄 수 있으나 칩간의 편의는 보정할 수 없다는 단점이 있다.

Li and Wang (2001)은 MAS에서 사용한 global scaling 표준화 방법으로는 해결할 수 없는 비선형표준화를 해야 한다고 주장하면서 이 위하여 먼저 각 칩과 베이스라인 칩에서 순위가 변하지 않는 유전자들의 집합 (invariant set)을 정의하였다. 그리고 이 집합을 표준화를 위한 기준 유전자로 사용하여 조각 별 선형함수 (piecewise linear function)를 적용하였다. 이는 MAS의 global scaling과는 달리 칩간의 편의까지 보정해준다는 장점이 있다.

Dudoit et al. (2002)는 cDNA 마이크로어레이의 표준화

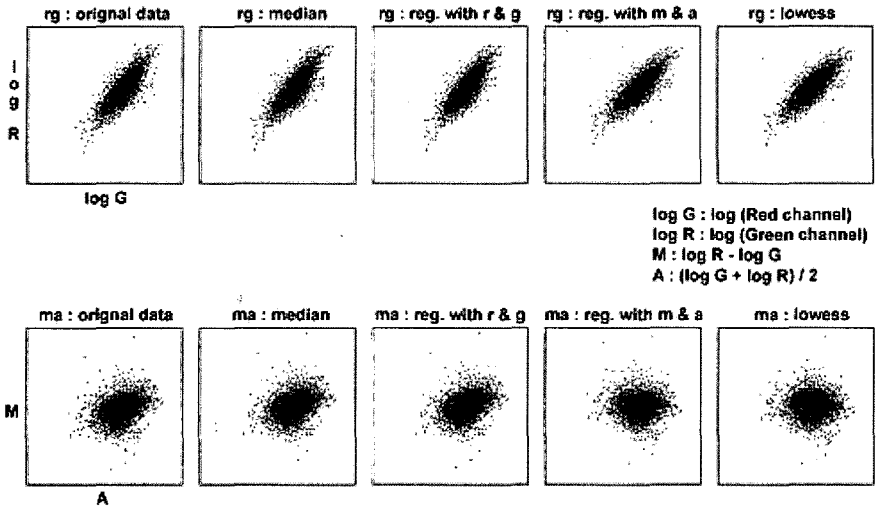

그림 3. 하나의 슬라이드에 4가지 표준화 방법을 적용한 그림. 왼쪽에서 오른쪽 순서로, 원자료, global median of log ratio, global linear regression, intensity dependent linear regression, intensity dependent LOWESS 표준화 방법을 적용한 결과이다 (이성곤 외, 2005)

방법 cyclic loess 를 제안하였다. 각 칩으로부터의 intensity를 하나의 channel로 간주하고 한 번에 두 개의 칩을 이용하여 MvA plot을 그리고 two channel 자료에서 사용한 LOWESS 방법을 적용한 것으로 둘 이상 $(i=1,2,\cdots,n)$의 칩이 있을 때에는 하나의 칩 k에 대해 모든 가능한 짝 $\{(k,1),(k,2),\cdots(k,k-1),(k,k+1),(k,n)\}$을 만들어 위의 방법을 시행한 후 칩 k를 보정한다. 이 과정을 표준화 될 때까지 반복한다. 이 방법은 표준화는 잘되는 반면 시간이 많이 걸리는 단점이 있다. Astrand (2003)는 cyclic loess의 변형형태인 대비를 이용한 표준화 방법 (contrast based method)을 제안하였다

Bolstad et al. (2003) 은 분위수 표준화 (quantile normalization)를 제안하였다. 이는 각 칩마다 probe intensity의 분포가 같도록 만들어 주는 방법으로 각 칩마다 계산된 분위수의 칩간 평균으로 표준화 분포를 삼는다. 즉, quantile-quantile plot이 직선이 되도록 맞추기 위하여 각 칩에서의 k번째 quantile의 probe intensity를 모아놓은 벡터 $Y_k=(y_{k1},\cdots,y_{kn})$를 $d=(\frac{1}{\sqrt{n}},\cdots,\frac{1}{\sqrt{n}})$로 사영 (projection)시켜 표준화를 시키는 방법이다.

마지막 단계인 요약 값 (summary value)를 정의하는 과정은 Affymetrix 자료에서 중요한 부분을 차지한다. 초기 Affymetrix 회사에서 제공한 값 (MAS4)은 $AvgDiff = \frac{1}{\#G}\sum_{j\in A}(PM_j - MM_j)$로 G는 각 유전자에서 $PM_j - MM_j$의 최소, 최대값을 제외한 평균에서 3SD안에 들어 있는 프로브들의 부분집합이다. 여기서 $PM_j - MM_j$는 j번째 프로브의 PM 값에서 MM 값을 뺀 것을 의미한다. 그 이후 MicroArray Suite (MAS 5)에서는 로그변환을 이용하여 새로운 요약 값인 $Signal = TukeyBiweight\{\log(PM_j - MM_j^*)\}$을 제공한다. 여기서 Li and Wong (2001)은 MAS4에서 제공하는 $AvgDiff$가 적절하지 않음을 발견하고 모델을 이용한 표현 값 (model based expression index : MBEI)을 정의하였다. MBEI는 하나의 유전자에 대한 요약 값 (summary value)를 구하기 위하여 모든 칩에 대한 해당 유전자의 모든 프로브 값들을 모아 프로브 집합을 만들고 각 프로브 집합에서 모델을 $PM_{ij} - MM_{ij} = \theta_i\phi_j + \varepsilon_{ij}$로 정의하고 ε_{ij}에 대하여 정규성 가정을 한 후 칩의 효과에 해당되는 θ_i의 MLE를 구한 것이다. 여기서 ε_{ij}는 i 번째 칩의 j번째 프로브에 해당되는 오차를 나타낸다.

Irizarry et al. (2003)은 PM만을 이용하여 robust multichip average (RMA)라는 방법을 제안하였다. 로그 변환된 PM 값을 Y_{ij}이라 할 때 선형 가법모형 (linear additive model) $Y_{ij} = \mu_i + \alpha_j + \varepsilon_{ij}$을 이용한다. 여기서 α_j는 프로브의 효과, μ_i는 i번째 칩의 효과, 그리고 ε_{ij}는 오차 항을 나타낸다. Affymetrix 기술은 평균적으로 유전자 발현을 잘 나타내는 intensity를 가지는 프로브만을 선택한다는 가정 하에서 μ_i를 median polish 방법 (Holder et al, 2001)을 이용하여 추정한 후 이를 i번째 칩에 대한 요약 값으로 사용한다.

Huber et al. (2002)과 Durbin et al. (2002)은 산포를 줄이기 위해 intensity에 따라 산포가 다를 때 산포를 일정하게 해주는 분산안정표준화 (variance stabilizing normalization :

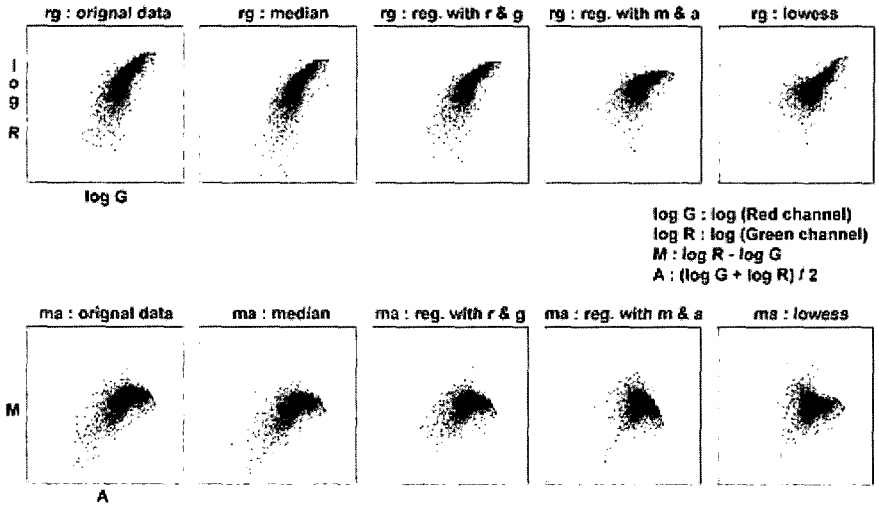

그림 4. 다른 슬라이드에 4가지 표준화 방법을 적용한 그림. 왼쪽에서 오른쪽 순서로, 원자료, global median of log ratio, global linear regression, intensity dependent linear regression, intensity dependent LOWESS 표준화 방법을 적용한 결과이다 (이성곤 외, 2005).

VSN) 을 이용하였다. 또한 Wu et al. (2004)는 probe sequence를 이용한 배경보정 방법인 GC-RMA를 제안하였다.

one channel

One channel 마이크로어레이 자료에 대해서는 위의 두 가지 마이크로어레이 자료에서 사용한 방법들을 조합한 표준화 방법을 이용할 수 있다. 먼저 two channel 마이크로어레이 자료에서 이용한 방법을 사용하여 슬라이드 내의 표준화를 시키고, 그 후 다음 단계로 분위수 표준화 (quantile normalization), 또는 다른 affymetrix 자료를 위한 표준화 방법을 이용하여 슬라이드간 표준화를 시킨다.

예 제

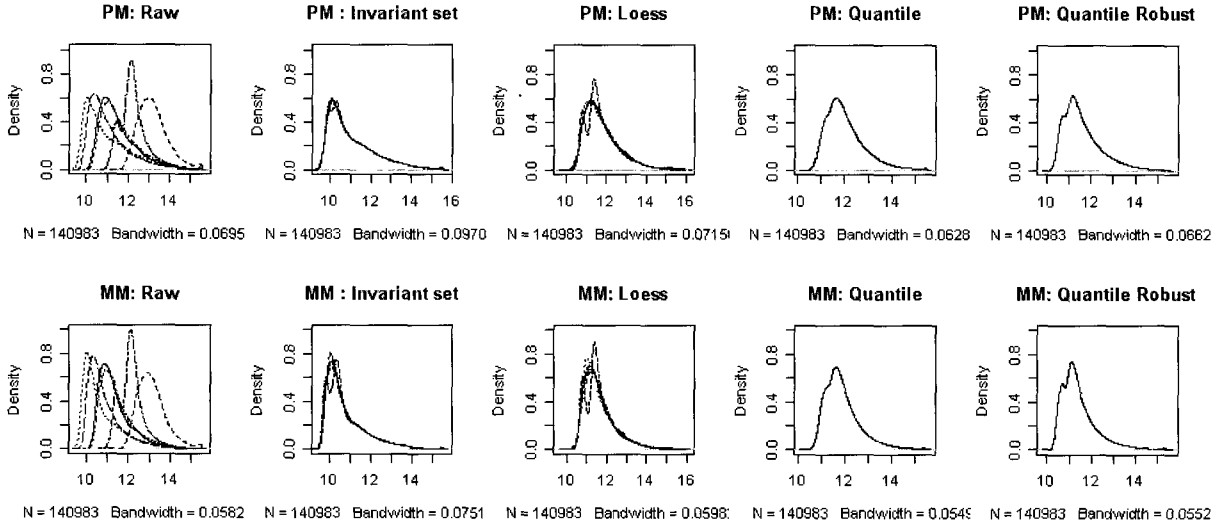

그림 5. 정상 brain sample 에 4가지의 다른 표준화 방법을 적용시키기 전과 후의 PM/MM 에 대한 각 슬라이드의 분포. 6개의 다른 색은 다른 슬라이드를 나타낸다.

위에서 설명한 여러 가지 표준화 방법들을 비교하기 위해 예제로 실험에서 얻어진 자료를 각 표준화 방법들로 표준화 한 결과이다.

Two channel cDNA data

흰쥐 embryo의 각 뇌신경조직 부위에서 분리된 신경기간세포 (neural stem cell)는 자가 증식 (self-renewal)과 가소성 (plasticity, mutipotency)이라는 잠재력을 가진다. 그리고 신경기간세포는 분화유도 시 신경세포 (neuron), 성상세포 (astrocyte), 및 oligodendrocyte로 분화되며, 이들 각 세포로의 분화를 촉진할 수 있는 cytokine (CNTF에 의해 astrocyte로의 분화, PDGF에 의해 신경세포로의 분화, T3에 의해 oligodendrocyte로의 분화 유도) 들의 소수가 밝혀져 있다 (Johe et al, 1996). 하지만, CNTF를 제외한 많은 다른 cytokine은 embryo의 나이와 뇌 부위에 따라서 운명 결정시의 선택이 변화하는 것으로 알려져 있다. 이 연구에서는 한 차례의 세포계대배양을 통해 완전한 신경기간세포군을 만들고, CNTF를 이용하여 신경기간세포를 성상세포로 분화 유도한 세포군과 단순히 mitogen 을 제거한 분화 유도를 통해 신경세포와 성상세포가 섞인 세포군을 확보하였다. 이러한 특징을 가진 세포군을 이용하여, 분화 기전에 관계하는 유전체의 조절 인자와 novel 한 세포 표지 유전체의 발현을 알아보는 cDNA chip 분석을 시행하였다 (이성곤 외, 2005).

실험 설계는 CNTF를 첨가한 그룹과 첨가하지 않은 그룹을 따로따로 마련하였고 각각 mRNA를 추출, Cy5로 염색한 다음, Cy3로 염색한 레퍼런스 cDNA와 교잡 (hybridization)하여 제작하였는데, 각 그룹을 분화 12시간, 1일, 2일, 3일, 4일, 5일이 지났을 때를 각각 3번씩 실험을 실시하였다. 즉, 두 개의 그룹에 세 번에 걸쳐, 여섯 시점에서 실험을 하여 총 36번의 cDNA microarray 슬라이드 얻었다. 이렇게 얻어진 스캐너로 스캔 하여 이미지 파일을 생성한 후에 ImaGene v.3으로 이미지 분석을 하여 자료를 얻었다.

그림 3과 그림 4는 위에서 소개한 실험에서 얻어진 자료 중 비교적 잘 된 실험과 잡음이 많은 실험에 대한 자료를 선택하여 몇 가지 cDNA 표준화 방법들을 적용, 비교해놓은 그래프이다. 단, 표준화된 $\log G$ 와 $\log R$ 은 $\log G$를 고정한 후 $\log R$의 값을 변화시킨 경우이다. 즉, $\log R^* = \log R - (\hat{M} - M)$ 이 된다.

그림 3의 자료는 다른 자료에 비해 비교적 실험이 잘된 경우인데 각 표준화 방법간에 큰 차이는 보이지 않음을 알 수 있다. 이에 반해 그림 4는 자료가 많은 잡음을 가지고 있으며 특히나 굽어지는 패턴이 있음을 볼 수 있다. 물론 이 패턴이 생물학적 정보를 담고 있으면 곤란한 면이 있지만 반복 실험된 다른 자료를 보았을 때 생물학적인 정보이기 보다는 잡음일 가능성이 다분하다. 즉, 표준화의 목적상 이 패턴을 인지하고 제거할 수 있어야 한다는 것이다. 이를 고려해보았을 때 intensity dependent LOWESS 표준화 방법이 이러한 문제점을 잘 해결함을 확인할 수 있었다. 이에 반해 다른 방법들은 선형적인 변화만을 보정했을 뿐 패턴 자체를 제거하지는 못했음을 볼 수 있다 (이성곤 외, 2005).

Affy data

신경 교아세포의 변종 (glioblastoma multiforme, GBM)은 뇌종양의 한 종류로 가장 악성인 종양 중 하나로 선행하는 조직, 기능상의 징후 없이 신경세포에 있는 약하던 악성 종양이 자라면서 강해져서 생기는 종양이다. 이에 대한 연구를 위하여 45개의 primary gliomas와 6개의 정상 brain sample에 대하여 Affy chip을 이용한 마이크로어레이 자료를 얻었다 (Rickman et al, 2001). 앞서 설명한 Affy 자료의 표준화 방법들을 비교하기 위하여 정상 brain sample 6개만을 이용하였다. 그림 5는 4가지의 다른 표준화 방법을 비교한 것이다. 첫 번째 행에 있는 그림은 표준화를 하기 이전의 perfect match (PM)과 miss match (MM)를 이용하여 각 슬라이드마다의 분포를 그린 것으로 6개의 sample이 모두 다른 분포를 갖는 것을 알 수 있다. 특히 하나의 sample이 다른 sample들과는 달리 이봉 분포 (bimodal distribution)을 갖고 있고 나머지 분포들은 모양은 같으나 위치가 다른 것을 볼 수 있다. Invariant set 과 loess 표준화 방법을 이용한 경우 PM, MM 모두 한 두 개의 sample을 제외한 나머지 sample들이 비슷한 분포를 갖는다. Quantile, Quantile robust 표준화 방법에서는 6개의 sample 모두 같은 분포를 갖는 것을 알 수 있다. 또한 Quantile Robust에서는 Quantile 방법에서보다 좀 더 확연하게 이봉 분포를 볼 수 있다.

그림 6과 7은 많이 사용되고 있는 3가지의 표준화, 요약방법인 MAS5, Li Wong, 그리고 RMA 방법을 비교한 그림이다. 상자그림과 분포그림 모두에서 MAS5는 sample마다 다른 분포를 갖는 것을 볼 수 있다. 그리고 Li and Wong과 RMA 방법에서는 하나의 sample을 제외하고 나머지 sample들은 서로 비슷한 분포를 갖는다. 이 결과에서도 알 수 있듯이 MAS5는 sample간 변동에 대한 보정을 고려하지 않는 반면 Li and Wong 이나 RAM 방법은 sample 간 변동까지 고려하여 표준화를 하게 된다.

그림 8은 산점도 행렬 그림을 이용하여 비교한 것이다. RMA 방법을 이용한 경우 요약 값들이 모두 y=x 부근에 모여있는 것을 볼 수 있으나 나머지 방법들은 그렇지 못하다. 특히 Li Wong 방법은 invariant set normalization을 이용하고 있으나 이 방법 이외에 비선형 보정을 해줄 수 있는 방법이 필요하다.

결과 및 토의

본 논문에서는 마이크로어레이 자료로부터 다양한 종류의 잡음을 제거하기 위한 표준화 방법에 대하여 살펴보았다. 마

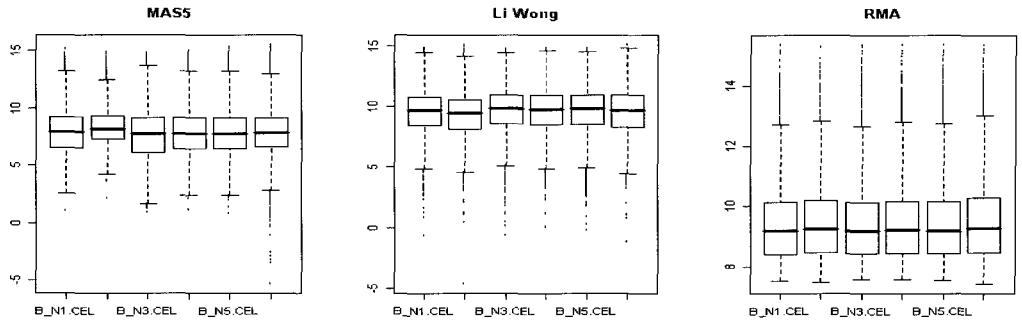

그림 6. 정상 brain sample 에 MAS5, Li and Wong, 그리고 RMA 방법을 적용하여 얻은 요약 값의 상자그림.

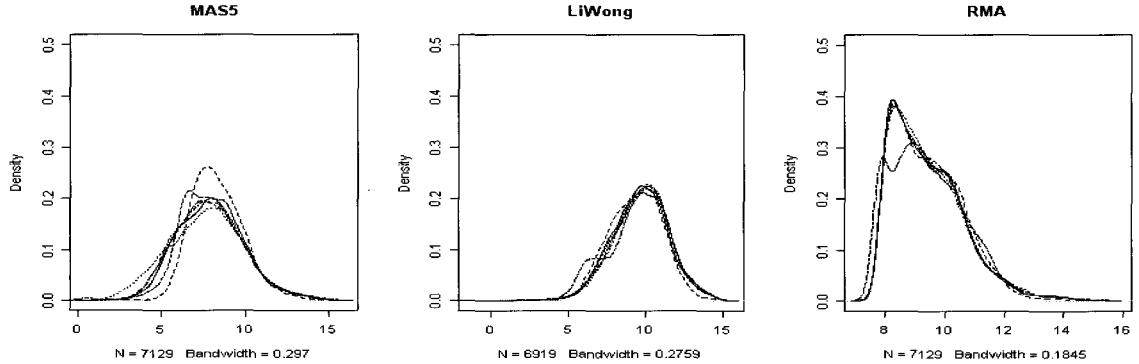

그림 7. 정상 brain sample 에 MAS5, Li and Wong, 그리고 RMA 방법을 적용하여 얻은 요약 값의 상자그림.

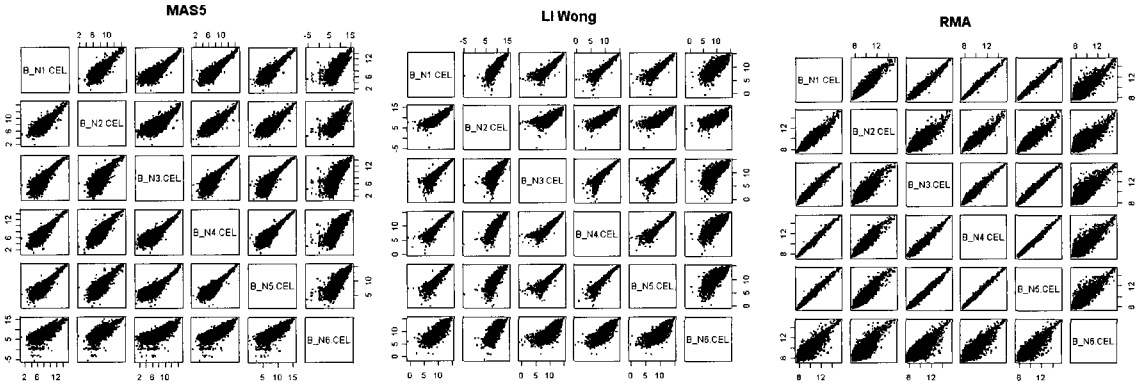

그림 8. 정상 brain sample 에 MAS5, Li and Wong, 그리고 RMA 방법을 적용하여 얻은 요약 값의 산점도 행렬 그림.

이크로어레이 자료에 대해 이러한 잡음을 제대로 제거하지 않으면 검정력의 약화를 초래할 수 있을 뿐만 아니라 심각한 편의(*bias*)를 갖는 잘못된 분석 결과를 초래할 수 있다 (이성곤 외, 2005). 따라서 가장 좋은 결과를 제공하는 표준화 방법을 선택하여 사용하는 것이 바람직하다.

본 논문에서 여러 가지 다양한 형태의 표준화 방법을 비교한 결과 cDNA 자료의 경우 원 자료에 첨가되는 잡음이 종종 비선형적인 형태로 나타날 수 있으므로 intensity dependent LOWESS 표준화 방법이 가장 효과적임을 예제의 분석을 통해서 확인할 수 있었다.

Affy 자료의 경우 RMA 방법인 perfect match만을 이용하여 quantile 표준화와 median polish 방법을 이용한 요약 값이

가장 좋은 결과를 보임을 알 수 있었다.

참고문헌

[1] 이성곤, 박태성, 강성현, 이승연, 이용성 (2005). DNA 마이크로어레이 자료의 PRINT-TIP 별 표준화 방법, 응용통계연구, 18 (1), 118-127.

[2] 박태성, 이승연, 김기웅, 이성곤, 최호식, 윤단규 (2005). 마이크로어레이 자료의 통계적 분석, 자유아카데미.

[3] Astrand, M. (2003) Contrast normalization of oligonucleotide arrays. *Journal of Computational Biology* 10 (1):95-102.

[4] Bolstad, B., Irizarry, R., Astrand, M., and Speed. T (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19 (2):185-193.

[5] 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*, 2:64-374.

[6] Cleveland (1979), Robust locally weighted regression and smoothing scatterplots. Journal of American Statistical Association, 74 (368):829-836

[7] Dudoit, S., Yang, Y. H., Callow, M. J., and Speed. T. P. (2002) Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments, *Statistica Sinica*, 12:111-139

[8] Duggan DJ, Bittner M, Chen Y, Meltzer P, Trend JM (1999) Expression profiling using cDNA microarrays *Nature Genetics*, 21 (Suppl. 1):10-14

[9] Durbin, B. P., Hardin, J. S., Hawkins, D.,M., and Rocke, D. M. (2002) A variance-stablizing transformation for gene-expression microarray data. *Bioinformatics* 10 (Suppl. 1), S105-S110

[10] Huber W., Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 1:1-9.

[11] Irizarry, R., Hobbs, F. C. B., Beaxer-Barcley, Y., Antonellis, K., Scherf, U., and Speed,. (2003) Exploration, normalization, and summaries of high densityh oligonucleotide array probe level data. *Biostatistics* 4, 249-264

[12] 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 (24):3129-3140

[13] 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.

[14] Li, C., and Wong, W. (2001) Model-based analysis of oligonucleotide arrays : Expression index computation and outlier detection. *Proceedings of the National Academy of Science USA* 98, 31-36

[15] Lipshutz RJ, Fodor SP, Gingeras TR, Lockhart DJ. (1999) High density synthetic oligonucleotide arrays. *Nature Genetics*, 21 (Suppl. 1):20-24.

[16] Richman, D. S., Bobek, M. P., Misek D. E., Kuick, R., Blaivas, M., Murnit, D. M., Taylor, J., and Hanash, S. M. (2001) Distinctive Molecular Profiles of High-Grade and Low-Grade Gliomas Based on Oligonucleotide Microarray Analysis. *Cancer Research* 61:6885-6891

[17] 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.

[18] Wu, Z., Irizarry, R., Gentleman, R., Martinex-Murillo, F., and Spender, F. (2004) A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* 99 (468):909-917

[19] Yang, Y. H, Dudoit, S. D., Luu, P., and Speed, T. P. (2001), Normalization for cDNA Microarray Data, In *SPIE BioE*

[20] 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.