Proceedings of the Autumn Conference in 2001
and the 30th Anniversary of the Korean Statistical Society

Microarray 자료분석에서 표준화

이성곤1) 박태성2) 최호식3)

요 약

본 논문은 microarray를 분석하기위한 표준화에 대한 여러 방법들을 소개하고 비교해보았다. Microarray 연구는 Human Genome Project에서 파생된 여러 생명공학 기술 중 가장 널리 사용되는 기술로 기존에는 하지 못했던 총체적인 유전자의 발현상황을 탐색할 수 있다는 장점을 지니고 있으나, 자료들에 일정한 패턴이 나타나거나 잡음이 첨가되어 정보의 추출이 용의하지 않다는 단점을 지니고 있다. 특히 자료에 일정한 패턴이 있는 경우에 올바르지 못한 결론을 이끌어낼 수도 있기에 이 패턴을 제거하는 표준화작업은 microarray 분석에 있어서 매우 중요한 처리과정이다. 본 논문에서는 표준화방법들을 소개하고 각각 가지고 있는 장단점을 실제 국내에서 얻어진 자료를 통해 비교하였고, 그 결과 LOWESS 적합을 통한 표준화방법이 타 방법에 비해 유용한 점이 많음을 확인할 수 있었다.

주요용어 : DNA chip, cDNA microarray, 표준화, LOWESS

1. 개요

Human Genome Project의 10여년간의 연구 결과로 우리는 인간이 지니고 있는 30억개의 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, 한두개의 유전자의 기능탐색이라는 종래의 한계를 벗어나 생명현상과 관련된 유전체수준의 연구가 가능해졌다라는 것을 뜻한다. 이러한 DNA chip 기술에는 cDNA chip 방식과 Affimatrix사의 oligochip방식이 있다. 이 중 Affimatrix사의 oligochip 방식은 반도체 집적기술을 접목시켜 높은 집적도와 응용성뿐만 아니라 신뢰성 높은 결과물을 제공하고 있어 주목받고 있는 기술이며 현재 여러 회사에서 개발에 성공했거나 추진중에 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 실험자료에 비해 잡음이 많이 포함되어 있으며 또한 자료에 일정한 패턴을 보이는 경우가 많다.

이중 첫번째 문제는 실험자의 숙련도가 낮아 제어를 제대로 하지 못했거나 실험에 쓰이는 여러 화학물질의 처리가 정밀하지 않은 경우 등으로 발생하는 것으로 잡음의 첨가가 일정하지 않

1 (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과, 박사과정

2 (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과, 교수

3 (151-742) 서울시 관악구 신림동 산 56-1, 서울대학교 통계학과, 석박사과정

결과적으로 microarray 분석에서 통계적 검정력을 약화시키는 결과를 가져온다. 두번째 문제는 발생원인은 비슷하다할 수 있으나, 그 효과가 판이하게 다른 경우이다. 슬라이드 수준에 전반적인 패턴을 보이거나 어떤 특정 경향성을 가진 경우가 이에 해당하며 이 경우는 자료에 전반적인 성향을 변화시키므로 이를 적절히 처리하지 않고 분석을 하게되면 편의가 발생하여 잘못된 결론을 도출할 수도 있다. 이러한 문제점들은 실험자의 숙련도가 높아지고 실험과정에 표준화가 이루어지면 줄어들 문제이기는 하나 현재 활성화되기 시작한 DNA chip실험수준에서는 중요하게 고려되야할 고민임에 분명하다.

이렇게 잡음을 제거하거나 일정한 패턴을 제거하는 등의 과정을 표준화 (normalization) 라고 하며 그간 microarray를 이용한 연구에서 여러 가지 표준화방법들이 제안되었다. 처음으로 cDNA microarray에서 통계적인 분석법을 제시한 Chen *et al.*(1997)은 Cy3의 intensity와 Cy5의 intensity가 일정한 비율을 하고 있을 것으로 가정한 후 표준화와 검정을 동시에 처리하였다. 또한 여러 image analysis program에서는 회귀분석을 통한 표준화를 많이 사용하고 있다. 그리고, Yang *et al.*(2001)은 2001년 1월 SPIE BiOE에서 발표한 글에서 여러 표준화 방법들을 소개하였고 거기에 자신이 제시한 LOWESS 적합을 이용한 표준화방법을 설명하였다. 또한 Dudoit *et al.*(2000)는 Yang이 제시한 LOWESS적합방법을 통해 표준화한 자료로 분석하였다. 이에 반해 사용된 microarray 데이터가 이미 표준화되었다고 가정하거나(Newton, 2001) 아예 이런 것에 대한 고려가 없었던 경우도 많다(Eisen, 1998). 그러나, 이러한 경우의 사용된 자료를 보면 일정한 패턴이 남아 있었음을 확인할 수 있었다.

본 본문에서는 cDNA의 전반적인 제작과정을 소개하고 선행 논문과 여러 image analysis program에서 제시된 여러 표준화 방법들을 소개하고 각 방법들을 비교하여 보도록 한다.

2. DNA chip의 제작과 Microarray 데이터의 생성

2.1 cDNA chip의 제작

cDNA chip은 비교하고자하는 두 종류의 세포의 핵에 들어있던 여러 mRNA의양의 비를 측정하는 방법을 이용한다. 특정 유전자를 transcription 하여 만들어진 mRNA는 추후 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 시에는 유전자가 비활성화되었다는 것을 뜻한다. 그러나, 세포에서 추출된 mRNA는 소량이므로 실험에서 요구되는 양을 얻기위해 cDNA로 변환한 후 RT-PCR 방법을 이용해 증폭한다. 이렇게 증폭된 cDNA에 하나는 Cy3(녹색)를 부착하고 다른 하나는 Cy5(적색)를 부착한다. 그런 다음 동일양을 섞은 후 미리 준비된 관련 cDNA와만 결합할 수 있는 probe를 화학반응(hybridization)하게한 후 슬라이드에 촘촘히 배열한다. 이런 식으로 슬라이드에 기록되는 cDNA는 수는 적게는 수백개에서 많게는 몇만개에 달한다. 즉 cDNA실험을 통해서 동시에 수백개에서 몇만개에 달하는 유전자의 발현양상을 한번에 살펴볼 수 있음을 뜻한다.

2.2 Oligochip의 제작

Oligochip은 우선 각 유전자의 고유한 20개정도 길이의 DNA sequence를 찾아낸 후 그 온전한 것과 이 sequence 중 가운데 하나의 염기만 다른 것을 반도체 제조공정과 비슷한 과정을 통해 합성한다. 그런 후, 세포에서 추출한 mRNA를 cDNA로 만든고 형광물질을 부착한 후 슬라

이드에 화학반응하게 한다.

이런 식으로 제작된 cDNA chip과 oligochip을 confocal microscope를 통해 각 형광물질의 intensity를 측정하여 cDNA chip의 경우 두장의, oligochip의 경우 한 장의 image 파일을 생성하고 image analysis program 등을 통해 각 유전자의 발현량을 수치화한다.

3. 표준화 기준

위에서 얻은 microarray 자료를 얻은 후 실제의 분석에 이용하기위해 표준화과정을 거쳐야한다. 이때 표준화는 두가지 방식의 표준화 기준으로 처리할 수 있는데, 첫번째로는 표준화만을 위해 일부로 넣은 몇개의 유전자만을 사용하여 표준화하는 방법과, 두번째로 전체 유전자를 사용하여 표준화하는 방법이 있다.

3.1 일부의 유전자만을 기준으로 삼는 방법

이 방법은 분석에 사용될 유전자 정보외에 표준화만을 위해 삽입된 유전자로 표준화하겠다는 전형적인 방법이다. 여기에 쓰이는 유전자로는 세포내에서생명활동을 위해 항상 일정한 양이 발현되고 있다고 생각되는 housekeeping gene을 이용하거나 아니면 전혀 발현되지 않으리라는 가정하에 spiked gene을 이용하여 표준화에 사용하고자 한다.

3.2 전체 유전자를 기준으로 삼는 방법

두번째로 제시된 방법은 위의 방법이 암과 같은 경우 같이 housekeeping gene이 제 역할을 못하는 경우나 또는 Cy3가 강하게 나타나는 것과 같이 슬라이드에서의 일정한 변이등이 나타나는 경우 등과 같이 일부의 유전자만으로는 보정하기 힘든 경우가 실제 실험에서 나타나기 시작하자 새로운 대안으로 제시된 방법이다. 이 방법은 수천, 수만개의 유전자를 실험하는 경우에 전체 유전자를 모두 사용하여 보정하는 방법이다. 이 방법은 이런 실험에서 대다수의 유전자는 비슷하게 발현되고 몇몇 특정의 유전자만 다르게 발현될 것이라는 가정을 전제로 한 것이다.

4. 표준화 방법

이러게 얻어진 표준화에 사용될 데이터에서 Cy3의 intensity를 G_j라고 Cy5의 intensity를 R_j이라고 하자. 이때 j는 각 유전자를 구분하는 id이다. 여기에 추가로

$$M_j = \log R_j/G_j = \log R_j - \log G_j$$
$$A_j = \log \sqrt{RG} = (\log R_j + \log G_j)/2$$

한 척도를 사용하도록 한다. 물론 여기에 사용되는 R_j와 G_j의 자료들은 위의 표준화 기준에

서 선택된 유전자 자료이다.

4.1 Global normalization

이 방법은 기본가정으로 전체 R_j와 G_j가 발현의 차이를 보이지 않는 이상 항상적으로 일정한 비를 이루고 있을 것으로 가정한 경우이다. 즉,

$$R = c \cdot G$$

한 관계를 가정한다. 여기에 log를 취해주면 단순한 상수의 차이로 표현될 수 있다. 다시 쓰면,

$$\log R = \log G + \log c$$

이다. 따라서, M_j^{Global}

$$M_j^{Global} = M_j - \hat{k} \quad (1)$$

으로 할 수 있고, 이때 표준화된 R_j^{Global}는 G_j를 기준으로 한다면 $e^{-\hat{k}} R_j$라고 할 수 있다. 그리고, 여기서 추정량 k는 보통 M_j의 중앙값이 쓰이며 그밖에 별도의 가정하에 MLE 추정량을 사용하기도 한다.(Chen, 1997) 그리고, 여기서는 G_j를 reference로 두어 값을 변화시키지 않는 것으로 한다.

4.2 Global regression normalization

이 방법은 confocal microscope에서 측정되는 강도가 보통 Cy3가 Cy5보다 강하게 측정된다는 것을 고려한 것이다. 그러나, chip이 여러 원인으로 인해 일정한 비율로만 되어있는 경우는 거의 없고 슬라이드상에 삐뚤어져 있는 경우가 많다. 이에,

$$\log R = \beta_{RG0} + \beta_{RG1} \log G$$

와 같이 가정하고 모수들을 추정한 b_{RG0}와 b_{RG1}이용하여 새로운 측도로

$$\log R_j^{RG} = \frac{\log R_j - b_{RG0}}{b_{RG1}}$$

를 이용한다. 따라서, M_j^{RG}는

$$M_j^{RG} = \frac{\log R_j - b_{RG0}}{b_{RG1}} - \log G_j \quad (2)$$

이 된다. 여기서 추정량 b_{RG0}와 b_{RG1}는 보통 회귀분석을 통해 얻어지며 영향을 많이 미치는 outlier를 제거한 후 다시 제 적합하는 반복작업을 통해 보다 정확성을 기할 수도 있다. 또한, 이 방법을 많은 image analysis program에서 사용하고 있는 표준화방법이기도 하다.

4.3 Intensity dependent regression normalization

cDNA microarray의 가장 큰 문제로 지적되는 점은 R_j와 G_j가 작은 값을 가지는 경우 매우 불안정하다는 것이다. 이 때문에 비율이 100, 1000과 같은 값이 손쉽게 나오게 되는 현상이 발생한다. 이러한 문제를 피하기 위해 microarray의 전반적인 intensity를 나타내는 새로운 척도 A_j를 사용하는 표준화하는 방법을 생각해볼 수 있다. 따라서, 가정하기를

$$M = \beta_{MA0} + \beta_{MA1} A$$

와 같이 하고 M_j^{MA}는

$$M_j^{MA} = M_j - \widehat{M}_j \tag{3}$$

가 된다. 즉, M_j^{MA}는 회귀분석의 잔차값이 된다. 또한 $R_j^{MA} = \exp(M_j^{MA} + \log G)$가 된다.

4.4 Intensity dependent LOWESS normalization

위의 방법들은 관계가 선형이라 가정하는 것이지만, 실제의 자료를 살펴보게 되면 자료가 선형인 경우외에 휘어지거나 굽어있는 경우를 볼 수 있다. 이러한 관계는 선형으로 볼 수 없기도 하거니와 산포가 비균등한 경우가 많아 최적의 비선형모형을 마련하는 일도 그리 쉬운 일은 아니다. 그렇기에 이러한 비선형모형 중 outlier에 robust한 LOWESS 적합을 이용하여 보정하는 방법을 고려해볼 수 있다(Cleveland, 1979; Yang, 2001). 즉,

$$M = k(A)$$

와 같은 비선형관계가 있음을 가정한다. 그리고, 이를 LOWESS 적합해서 얻은 M_j^{LOWESS}는

$$M_j^{LOWESS} = M_j - \widehat{kf}(A_j) \tag{4}$$

이다. 그리고 그때의 $R_j^{LOWESS} = \exp(M_j^{LOWESS} + \log G)$가 된다.

참고문헌

Chen *et al.* (1997), Ratio-Based Decisions and the Quantitative Analysis of cDNA Microarray Images, *Journal of Biomedical Optics*, 2(4):364-374

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

Dudoit *et al.* (2000), Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report 578, Department of Biochemistry, Stanford University School of Medicine

Eisen *et al.* (1998), Cluster analysis and display of genome-wie expression patterns. In *Proceedings of the National Academy of Sciences*, pages 14863-14868

Newton *et al.* (2001), On Differential Variability of Expression Ratios: Improving Statistical Inference about Gene Expression Changes from Microarray Data, *Journal of Computational Biology*, 8:37-52

Yang *et al.* (2001), Normalization for cDNA Microarray Data, In *SPIE BioE*