Normalization Techniques for Microarray Data

2002-04-12 박윤주

1. Introduction

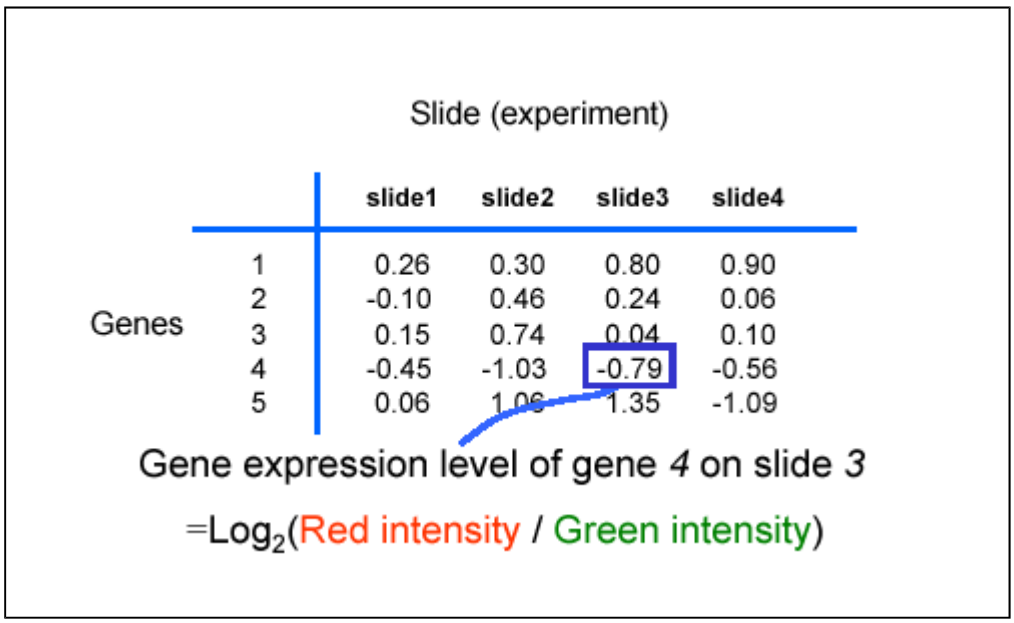
수 천 수 만개의 gene의 expression level을 동시에 monitoring 할 수 있는 DNA microarray 는 현재 biotechnology 중요한 한 분야로 자리 잡아가고 있다. 이 microarray는 normal 과 disease cell 안의 gene regulation이나 interaction 과 같은 현상을 이해하는데 도움을 줌으로서 약학이나 임상적인 조사에 많은 도움을 주고 있다.

그런데 이런 microarray 실험에서는 빈번히 발생하는 많은 systematic variation에 의한 bias 들이 있는데, 이런 bias들은 gene expression level 을 정확하게 측정하지 못하게 함으로서 잘못된 결과를 도출하게한다. Normalization은 microarray data에서 이런 systematic variation을 제거 하여줌으로써 좀더 정확한 측정을 할 수 있게 support하는 기법이다.

아래에서 이런 normalization 방법들에 대하여 좀 더 자세히 알아보고자 한다.

2. Gene Expression Data

Gene expression level 은 아래와 같이 microarray 의 raw data로 부터 얻어진 red intensity 와 green intensity의 비율의 log값인 log-ratio의 값으로 표현되어진다. 앞으로 수행할 normalization은 이 gene expression level 을 표현하는 log-ratio 값으로 한다.



3. Purpose of Normalization and Source of Bias

위에서 언급한 바와 같이 normalization의 가장 큰 목적은 microarray data에서 나타나는 많은 bias 중에 어떤 bias가 어떤 형태로 나타나는지를 발견해서 보정(adjustment)해 주는 데에 있는데 구체적인 목적은 아래와 같다.

첫째 실험 data 상에서 나타나는 두 dye (red-Cy5 dye 와 green-Cy3 dye) 의 형광강도(fluorescence intensity)의 balance를 보정해주기 위함이고 둘째로 다른 실험(또는 slide)간의 gene expression level 을 비교하기위해서 normalization 작업이 필요하기 때문이다.

이런 bias 가 나타나는 것에는 많은 이유가 있는데 몇 가지를 들어보자면, 첫번째로는 두 dye의 물질적인 특성((physical property)을 들 수 있다. 즉 열이나 빛 등에 대한 민감도 때문에 나타나는데, 종종 green dye 가 높은 fluorescence intensity를 보인다. 이러한 상황 때문에 평균적으로 같은 dye intensity를 고려한다는 것은 무리가 따른다. 두번째로는 dye혼합(incorporation)의 효율성(efficiency)문제이고 세번째로는 data를 모으고 scanning 하는 과정에서 나타난다. 이외에도 pin-group 간의 차이, slide heterogeneity의 문제 들이 실험상의 bias를 유발시킨다.

4. Which Genes to use

위에서 언급한 이유들로 array data의 normalization이 필요한데, 이러한 normalization에서 gene을 선택하는 3가지 접근방법이 있는데 아래와 같다.

첫번째로는, array 위의 모든 gene들을 고려한 방법이다. 이 방법은 우선 적은 비율의 gene 들만 다르게 발현된다는 가정 하에서 array 위의 거의 모든 gene들을 대상으로 normalization한다. 즉, 두개의 mRNA sample에서 의미 있게 변하는 gene들의 비율이 상대적으로 적어야 하고, 이렇게 up-down regulated 된 gene들의 발현도는 대칭(symmetry)적인 분포를 이루어야 한다.

이것은 앞 절에서 언급했던 log-ratio의 값들이 0을 중심으로 대칭을 이룬다는 말과 같다. 즉, 이를 위하여 모든 gene 들의 분포를 이런 형태로 adjustment 한다는 의미이다.

두 번째 방법으로는 항상 발현되는(constantly expressed) gene들을 기준으로 normalization 하는데, 이는 모든 gene을 대상으로 normalization 시키는 대신에 좀더 작은 gene의 subset을 대상으로 한다는 것이다. 종종 이런 gene 들을 우리는 housekeeping gene(e.g. Beta actin) 이라고 부른다. 그러나 이 방법의 문제점은 실험에서 특수한 한가지 상황 (e.g. temporary)을 고려했을 때 나타나는 housekeeping gene 을 찾는 것은 어렵지않으나, 어떠한 조건 하에서도 동일하게 발현 되는 housekeeping gene을 찾는다는 것은 매우 힘들 다는데 있다.

세번째로 housekeeping gene 사용 대신에 gene 들을 의도적으로 array 상에 배치해놓은 spiked controls 이나 titration series of control sequences를 사용하여 normalization 시키는 방법등이 있다.

요즘 많이 사용되고 있는 방법은 우선 첫번째 array 상의 모든 gene 들을 사용하여 normalization 시킨 후 나머지 방법들을 사용하여 검증하는 단계를 택하고있다.

5. Methods of Normalization

5.1 Single-slide data displays

single-slide data는 일반적으로 red dye 의 log-intensity $\log_2 R$ 과 green dye의 log-intensity $\log_2 G$의 2차원 plot으로 표현되어지는데, 이 plot에서 $\log_2 R=\log_2 G$ 의 직선상에서 벗어난 gene의 양을 관찰한다. 또한, 이 값들을 이용한 log-intensity ratio인 M값과 mean log-intensity인 A 값을 이용하여 plotting 하기도 한다.

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$
$$A = \log_2\sqrt{RG} = (\log_2 R + \log_2 G)/2$$

이 M과 A 값을 이용한 plotting 은 위의 $\log_2 R$ 과 $\log_2 G$를 이용한 plot을 45°회전시킨 plot으로 0값을 기준선으로 gene data를 관찰한다.

5.2 Within-slide normalization

Within-slide normalization은 각 slide의 red intensity 와 green intensity의 균형을 맞추기 위해 행해지는데, location adjustment는 일반적으로 log-ratio 값들의 mean 값이나 median 값을 보정하여 normalization하는 것이고, scale adjustment는 variance값을 보정하여 normalization 시켜주는 것이다.

5.2.1 Location

5.2.1.1 Global normalization

Global normalization은 red intensity 와 green intensity는 상수에 의해 연관되어있다고 가정하고(R = k·G), log-ratio 의 분포의 중심을 아래식과 같이 상수의 가감에 의해 0에 맞춘다.

$$\log_2 R/G \Rightarrow \log_2 R/G - c = \log_2 R/(k \cdot G)$$

윗 식에서 위치변수 $C= \log_2 K$ 는 gene set 의 log-intensity ratio의 mean 이나 median 값이다.

이러한 global normalization 은 여전히 많이 쓰이는 normalization 기법이지만 intensity에 영향을 받는 dye bias 나 pin-group에서 나타나는 왜곡된 현상들을 바로 잡아줄 수 없다는 문제점이 있다.

5.2.1.2 Intensity dependent normalization

많은 경우 dye bias 는 spot intensity에 의존적으로 나타나는데 위 4.1 절에서 언급 했었던 M-A plot 을 그려보면 알 수 있다. 이런 이유 때문에 intensity를 고려 하지 않은 global normalization 보다는 intensity

또는 A-dependent dye normalization이 더욱 바람직 하다.

Robust scatter-plot smoother 'lowess' 를 사용하여 아래 식과 같이 A-dependent 하게 normalization 시킬 수 있다.

$$\log_2 R/G \Rightarrow \log_2 R/G - c(A) = \log_2 R/(k(A) \cdot G)$$

윗 식에서 c(A) 는 M-A plot 에서의 lowess 적합 값이다. 이런 lowess( ) function 은 robust 하기 때문에 M-A plot 에서의 outlier 같은, 즉 다르게 표현된 gene의 small percentage에 의해 영향을 받지 않는다.

5.2.1.3 Within-print-tip-group normalization

Array 위의 spot 들은 몇 개의 sector로 구성되어 있는데, 이때 각각의 sector 들은 서로 다른 print-tip(또는 pin)을 사용하여 spot을 print 한다. 이때 print-tip 들 사이에는 tip 의 opening 이나 길이 또는 많은 시간이 흐른 후 printing의 변형 같은 systematic한 bias 가 존재 할 수 가 있다 (print-tip-group -> proxies for spatial effect). Within-print-tip-group normalization 은 간단하게 (print-tip+A)-dependent normalization 으로 표현할 수 있고 아래 식과 같이 나태내어 진다.

$$\log_2 R/G \Rightarrow \log_2 R/G - c_i(A) = \log_2 R/(k_i(A) \cdot G)$$

윗 식에서 $C_i(A)$ 는 i번째 print-tip-group 에서의 M-A plot 의 lowess 적합 값이다. 여기에서 i = 1, ... ,I . I 는 print-tip 의 수이다.

즉, 이 방법은 A-dependent normalization처럼 전체적으로 lowess fitting 을 하지않고 각 print-tip-group 마다 lowess fitting 을 한 후 normalization 을 시켜준다.

5.2.2 Scale

Within-print-tip group normalization 후 서로 다른 print-tip group 들의 분포 중심은 0으로 normalized 되지만, 그들 분포의 variance는 서로 같지 않을 경우가 종종 있다. 이 때문에 scale adjustment가 필요한데, 이 within-slide scale normalization 을 하기 위해서는 몇 가지 가정을 해야한다.

가정: i 번째 print –tip group 으로 부터의 모든 log-ratio는 평균이 0이고 분산이 $a^2_i\sigma^2$ 인 정규분포를 따른다.

- σ^2 : the variance of the true log-ratios
- a^2_i : the scale factor for the ith print-tip-group

위와 같이 가정하고, print-tip-group 간의 다른 분산을 갖게 하는 scale factor a_i 를 아래와 같은 제약조건에서 MLE(maximum likelihood estimator)방법을 사용하여 추정한다. 이렇게 추정된 scale factor a_i 를 각 print-tip-group 에서 제거한 후 normalization 을 수행 한다.

사용된 제약식과 MLE 방법을 사용하여 추정한 âi는 아래식과 같다.

$$\text{제약식} \ : \ \sum_{i=1}^{I} \log a_i^2 = 0$$

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\Pi_{k=1}^{I} \sum_{j=1}^{n_k} M_{kj}^2}}$$

$$M_{ij} : the\ j th\ \log - ratio\ in\ the\ i th\ print - tip - group\ ,\ j = 1, \cdots, n_i$$

a_i 를 추정하는 또 다른 식은 아래식과 같이 MAD(median absolute deviation) 를 사용하는데, 이 방법은 중위수(median)를 사용하기 때문에 평균을 사용한 윗 식보다는 좀더 robust 하다.

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\Pi_{i=1}^{I} MAD_i}}$$

$$MAD_i = median_j[\ |M_{ij} - median_j(M_{ij})|\ ]$$

이 robust 통계량 MAD 는 앞 절에서 언급하였던 robust lowess smoother 와 비슷하게 M-A plot 상에서 outlier에 영향을 받지않는다.

5.3 Paired-slides normalization (dye-swap)

Paired-slides normalization 은 두 개의 mRNA sample에 red dye와 green dye 를 바꾸어 할당하는 dye-swap 방법이다.

slide 1 : trt. -> red ctl. -> green

slide 2 : trt. -> green ctl. -> red

이 normalization방법은 두 slide에서 log-ratio(gene expression level)의 분포대한 variance는 대략적으로 같다고 가정하고 location adjustment 만 시행한다. 따라서 앞 절에서 언급했던 within-slide-location normalization방법에 의해 두 슬라이드를 normalization 시키는데 아래와 같다.

$\log_2 R/G - c$ 에 의해 첫번째 slide의 log-ratio 값들이 normalization 되고, 마찬가지로 두 번째 슬라이드도 $\log_2 R'/G' - c'$ 에 의해 normalization 된다.

c 와 c′ 가 대략적으로 같다고 보고 모든 gene을 사용하여 normalization 한다고 가정한 후 두 slide의 normalized 된 log-ratio들을 결합시키면 아래식과 같이 표현되어진다.

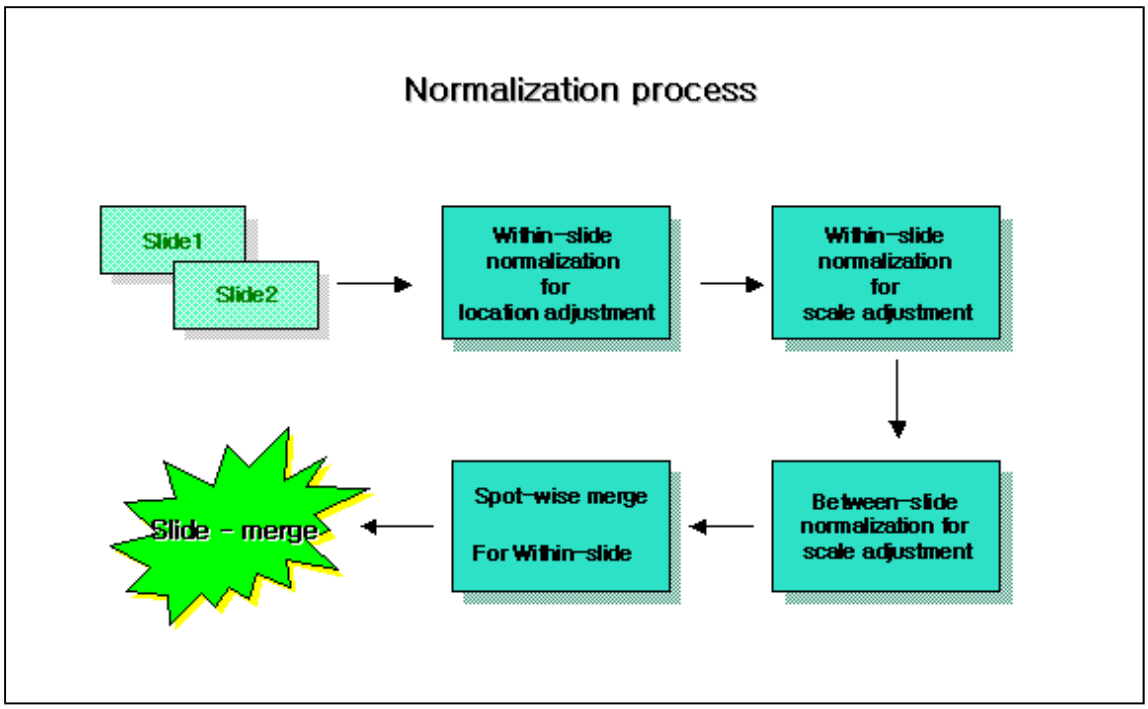
$$
\begin{aligned}
& [\,(\log_2(R/G) - c) - (\log_2(R'/G') - c')\,]/2 \\
\approx\ & [\,(\log_2(R/G)) - (\log_2(R'/G'))\,]/2 \\
\approx\ & [\log_2(RG'/GR')]/2 \\
=\ & (M - M')/2
\end{aligned}
$$

윗 식에서 gene 의 변화가 없다고 본다면 $[\log_2(RG'/GR')]/2 = (M - M')/2 = 0$ 이 된다.
결국 두 슬라이드에서 얻은 log-ratio의 평균 값을 통해 adjustment factor 인 c 와 c′ 가 제거되어 자연스럽게 normalization이 되어진다. 이러한 이유 때문에 이 방법을 self-normalization 이라고도 한다.

5.4 Multiple slide normalization

Within-slide-normalization 후 각 log-ratio 값들은 모두 중심이 0에 위치하는 분포를 가지게 된다. multiple slide normalization 방법은 slide (또는 experiment)간의 비교를 하기위한 기술인데, 위와 같이 Within-slide-normalization을 거쳤더라도 대체적으로 각 slide 간의 log-ratio의 scale이 같지않은 경우가 많다. 이 때문에 4.2.2절에서 사용한 within-slide scale normalization을 이용하여 각 slide 간의 multiple slide scale adjustment를 시행하여야 한다.

지금까지 보여왔던 array data에 대한 normalization의 전체적인 수행과정은 아래그림과 같이 도식화 할 수 있다.



REFERENCES

1. Y.H.Yang, S Dudoit, P Luu and T. P.Speed. , Normalization for cDNA Microarray Data. ,2000. (http://www.stat.berkeley.edu/users/terry/zarray/Html/normspie.html)
2. Harvard university., cDNA microarray experiments:pre-processing and experimental design., 2002. (http://biowww.dfci.harvard.edu/~bioconductor/workshops/ShortCourse012302/lectures/lect1b.pdf)
3. 최대우., Normalization Techniques. (http://www.digital-genomics.co.kr/pds/files/normalization.pdf)
4. Tsai Chen-An, Microarray : Normalization & Clustering. (http://binfo.ym.edu.tw/binfo/01s/Normalize.files/frame.htm)