Normalization of DNA Microarray data

권 성우, 한 종훈

(sw74@postech.ac.kr)

포항 공과 대학교 화학 공학과

1. 서론

최근 급격히 증가하는 유전체 데이터를 분석하기 위해서는 새로운 방식의 분석 기술이 필요하다. 이러한 기술들 중에서 대표적인 것이 바로 Microarray 기술이다. Microarray 기술에는 DNA chip, protein chip, lab-on a chip 등이 있다. DNA Microarray는 고형체에 고정 시킨 DNA를 mRNA나 다른 DNA와의 잡종 형성을 통해 유전자의 발현 양상을 알 수가 있게 된다. 이를 통해 특정 상태의 유전자 발현 양상을 연구할 때 많이 사용한다.

한편, Microarray 실험은 실험 과정에서 편차가 높기 때문에 재현 성이 떨어지는 단점이 있으며, 이로 인해 유전자들의 발현 양상을 정확하게 측정하지 못하는 경우가 많고 잘못된 결과를 도출 하기도 한다. 따라서 재현 성을 높이기 위해서는 실험 과정의 핵심 요소들의 편차를 보정 시켜야만 하며 이를 위해 Normalization을 수행 한다. 즉, Normalization은 microarray data를 보정 하여줌으로써 좀더 정확한 측정을 할 수 있게 해준다. 본 글에서는 DNA microarray 데이터에 편차가 생기는 이유와 최근 개발된 Normalization의 방법에 대해서 알아보고자 한다.

2. 로그 비

빨간색으로 염색한 샘플과 녹색으로 염색한 샘플을 비교하기 위해서 Normalization 에서는 이미지 분석을 통해 정량화된 데이터들 중, 빨간색 강도와 녹색 강도의 비(ratio)를 이용한다. 비를 로그 변환을 시킨 후에 normalization을 많이 한다. 로그 변

환을 많이 하는 이유는 첫째로, 이미지 분석 시에 16 비트의 범위로 전환하기 때문에 0에서부터 65535라는 넓은 구간을 가지게 되지만, 대부분의 DNA microarray의 정량화 값들은 1000이하의 값을 가지게 된다. 따라서 1000이하의 값을 가지는 구간에서 각 점들간의 차이가 나지 않는 문제가 생긴다. 따라서 범위를 일정하게 해 주기 위해서 로그 변환을 취한다. 둘째로는 평균 시그널의 강도에 따라서 random variation은 선형적으로 증가하는데 로그 변환을 시키면 일정한 값을 가지게 되는 장점이 있다. 셋째로는 로그 변환 값은 빨간색 강도와 녹색 강도의 비 (R/G)를 차 ($M = \log R - \log G$)로 바꾸어 나타낼 수가 있게 되어 의미 분석이 쉬워지는 장점이 있다. 또한 각 점의 밝기 정도를 나타내는 파라메터로 $A = (\log R + \log G)/2$ 를 사용한다.

3. 편차 발생 원인과 normalization 기준

Normalization은 편차를 보정하는 것이 목적인데, 편차가 발생하는 주 원인은, 첫 번째로는 두 염색 약의 물질적인 특성을 들 수 있다. 즉 열이나 빛 등에 대한 민감도 때문에 나타나는데, 종종 녹색 염색 약이 높은 형광 강도를 보인다. 이러한 상황 때문에 평균적으로 같은 염색 강도를 고려한다는 것은 무리가 따른다. 두 번째는 각 염색 약 별로 혼합 효율성의 차이로 인한 문제이고 세 번째로는 DNA microarray 로봇의 물리적인 조건 차이에 의해서이다.

한편, 유전자를 선택하는 방법에 따라서 normalization의 기준이 나누어 지는데, 많이 사용하는 방법에는 두 가지가 있다. 첫 번째로는, DNA microarray의 모든 유전자들을 고려한 방법이다. 이 방법은 우선 적은 비율의 유전자들만 다르게 발현된다는 가정 하에서 DNA microarray위의 모든 유전자들을 대상으로 normalization 한다. 즉, 두 개의 표본에서 의미 있게 변하는 유전자들의 비율을 상대적으로 감소 시키도록 하고, 과다 발현하거나 억제 발현 하는 유전자들이 대칭적인 분포를 이루게 하도록 보정하는 방법이다. 한편 두 번째 방법은 모든 유전자를 대상으로 normalization 시키는 대신에 housekeeping 유전자와 같이 항상 발현되는 유전자들을 기준으로 normalization 하는 것이다. 이 방법의 경우 모든 실험 조건에서 동일하게 발현 되는 housekeeping 유전자를 찾기가 어려운 단점이 있다.

4. 위치 보정 방법

쉽고 가장 많이 사용하는 normalization 방법은 global normalization으로서 로그 비 값과 상수 값 (C)의 차로써 나타낸다. 이때 상수는 유전자의 로그 비 평균값이나 중위 값을 사용한다. 이 방법에서는 강도에 영향을 받는 염색 약 편차나 microarray의 핀으로 인한 위치 편차를 고려하지 못하는 단점이 있다.

$$\log_2 R/G - C = \log_2 R/(k \times G), \quad \text{k는 상수}$$

염색 약 편차를 가지는 global normalization의 단점을 보완한 방법으로 intensity dependent normalization가 있다. 보통 A-dependent 하게 normalization 시키는데 A란 로그 비의 평균 값이며 이 값으로 빼준 후 global normalization을 수행한다. 이 방법은 강건성이 높기 때문에 이상치 (outlier)에 영향을 받지 않는 장점을 가진다.

$$\log_2 R/G - C(A) = \log_2 R/(k(A) \times G)$$

Microarray 위의 점들은 몇 개의 부분으로 구성되어 있는데, 이때 각각의 부분들은 서로 다른 핀을 사용하여 점을 프린트 한다. 이때 핀들 사이에는 팁의 개 패나 길이들 등 여러 물리적 요소가 변화할 수 있다. 이와 같은 이유로 인해 위치 편차가 발생하는데 (그림 1), 이러한 점을 intensity dependent normalization은 고려 하지 못하는 단점이 있다. 이를 보완하기 위해 개발된 방법이 print-tip-group normalization이다. 이 방법은 intensity dependent normalization에서 C(A)항을 C_i(A)으로 바꾸어 준 것으로써 C_i(A)는 i번째 print-tip-group에서의 로그 비의 평균값이고 i는 print-tip의 수이다 (그림 2).

$$\log_2 R/G - C_i(A) = \log_2 R/(k_i(A) \times G)$$

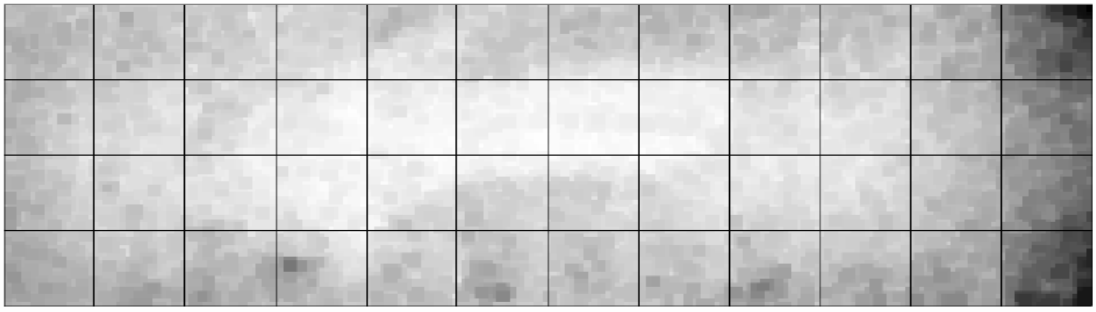

그림 1. DNA microarray에서 이미지 분석시 나타난 background 이미지의 강도 분포. 회색은 background가 낮은 것을 의미 하고 검은색은 높은 것을 의미한다. DNA microarray의 왼쪽 모서리는 오른쪽 모서리에 비해서 background 이미지가 상대적으로 낮다. 이러한 원인은 각

print-tip에 의해서 나타나는 편차이다. 그림의 DNA microarray는 4개의 print-tip을 가진 로봇으로 만든 것으로 총 19200점이 있다.

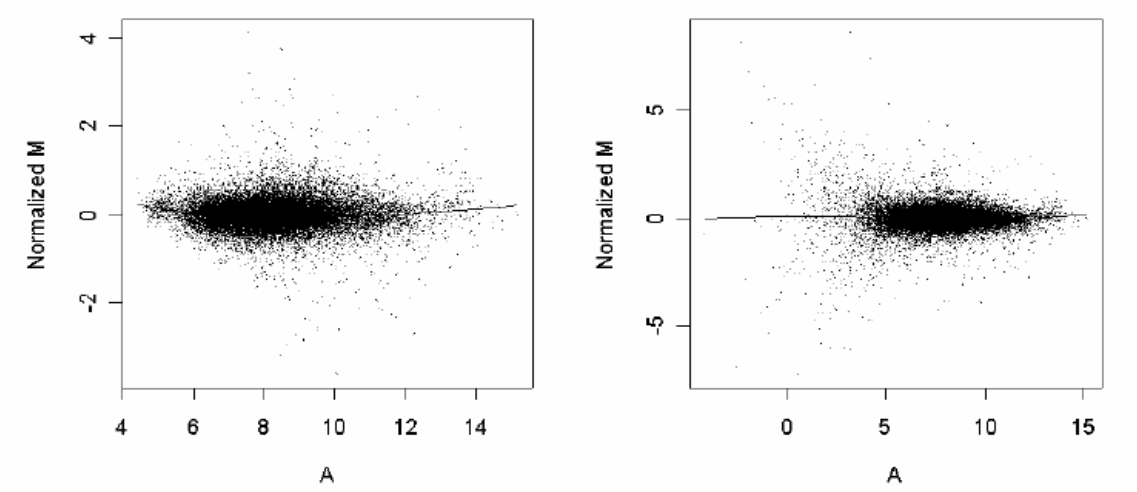

그림 2. print-tip group normalization한 후의 MA-plot으로 왼쪽은 global normalization의 경우이며 오른쪽 그림은 housekeeping 유전자를 이용하여 수행한 경우이다.

5. 스케일 보정 방법

위치 보정 후 분산이 같지 않을 경우가 있다 (그림 3). 이 때문에 스케일 보정이 필요한데, 스케일 보정 방법에는 MLE (maximum likelihood estimator) 방법과 MAD (median absolute deviation) 방법이 많이 사용 된다. MLE 방법은 기존의 편차를 a_i라고 하고 위치 보정으로부터 얻어 지는 모든 로그 비 값의 평균은 0이고 분산은 $a_i\sigma_i$인 정규 분포를 따른다고 한 후에 MLE 방법을 통해서 a_i를 구한다.

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\left(\prod_{k=1}^{I}\sum_{j=1}^{n_i} M_{ij}^2\right)^{1/I}}, j = 1,2,3...,n_i$$

단 M_{ij}는 i번째 print-tip 그룹의 j번째 log-ratio

MAD의 경우 평균값을 사용하는 MLE와는 달리 중위 수를 사용하여 a_i를 구한다. 이 방법은 중위 수를 사용하기 때문에 MLE방법 보다는 좀더 강건성이 높은 장점이

있다.

$$\hat{a}_i = \frac{MAD_i}{\left(\prod_{k=1}^{I} MAD_i\right)^{1/I}}$$

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

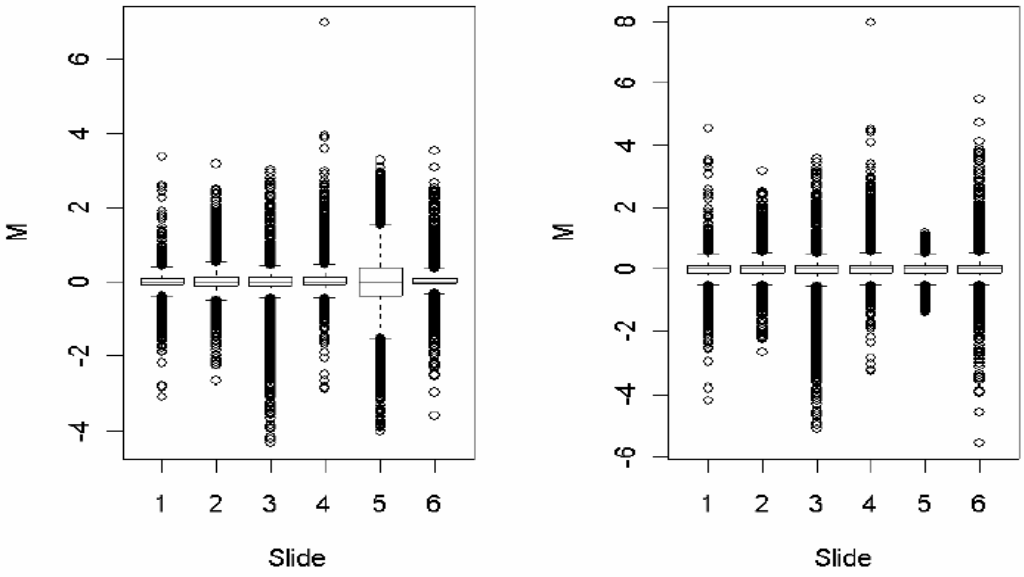

그림 3. 위치 보정후 각 DNA microarray별 정량화 시킨 데이터의 범위를 박스 그림으로 표시한 것이다. 왼쪽 그림에서 5번째 microarray의 데이터는 다른 것에 비해 넓게 분포하는 것을 알 수가 있다. 오른쪽 그림은 스케일을 MAD 방법으로 보정한 데이터를 박스 그림으로 표시한 것이다.

참고 문헌

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.

Cleveland, W. S., and Loader, C. L. (1996). Smoothing by Local Regression: Principles and Methods. In W. Haerdle and M. K. G. Schimek, editors, *Statistical Theory and Computational Aspects of Smoothing*, Springer, New York, pages 10-49.

Gordon. K. S., Yang, H. Y., and Speed, T. (2002). Statistical issue in cDNA Microarray data analysis. Research Report, Walter and Eliza Hall Institute of Medical Research, Melbourne, Australia.

Sandrine. D., and Robert. G. (2002). cDNA Microarray experiments: pre-processing and experimental design. Statistics and Genomics Short Course. Department of Biostatistics. Harvard school of public health.

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.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds.), *Microarrays: Optical Technologies and Informatics*, 4266. the international society for optical engineering.