cDNA Microarray Normalization에 대한 연구

김종영[1], 이재원[2]

<요약>

마이크로 어레이(microarray)실험에서 표준화(normalization)는 유전자의 발현수준에 영향을 미치는 여러 기술적인 변인을 제거하는 과정이다. cDNA microarray normalization에 있어 여러 방법이 제안되었지만, 이중 print-tip 효과가 존재할 때 사용되는 방법으로 print-tip lowess normalization이 대표적으로 사용된다. normalization에 사용되는 lowess 함수는 데이터의 특성에 따라 window width를 정해야만 연구의 목적에 맞는 결과를 도출할 수 있다. 본 논문에서는 각각의 tip에서 최적의 window width를 계산하는 절차를 논의하였다. 또한 이의 결과와 기존의 같은 window width를 사용하는 print-tip lowess normalization 결과와 비교 평가하여 normalization의 기본 원칙에 대한 타당성을 확인하였다.

주요용어 : Microarray, Normalization, Window width, Bootstrap, Crossvalidation

1. 서론

normalization이란 microarray 실험에서 유전자 발현수준에 영향을 미치는 기술의 변이를 찾아내어 제거하는 과정이다. cDNA microarry에 있어서 normalization의 주목적은 green Cy3와 red Cy5 dye간의 형광강도의 균형을 맞추고 각 실험(슬라이드)간 발현 수준을 비교할 수 있도록 하는데 있다. 이러한 상황에서 평균적으로 dye intensity들이 같게 되는 건 드문 일이고 intensity는 녹색염료에서 높게 나타나는 일이 흔하다. 이러한 편차는 dye의 물리적 성질(열이나 광 감도 등), dye 혼합의 효율성, 자료수집절차에서의 scanner setting 등의 다양한 요인에서 비롯된다.

최근에 다양한 normalization 방법들이 많이 제안되었는데 Yang et al.(2001)은 일반화 선형모형을 이용하여 normalization에 사용된 방법에 대한 모형을 평가하고, 두 인접한 spot의 background 강도에 대한 평가 방법을 제안하였다. 최근에는 그들이 제안한 intensity dependent 방법에 대한 실제적 응용 방법이 제안되고 있다. Tseng et al.(2001)은 quality filtering과정과 비교실험에서 normalization 곡선을 추정하기 위한 rank invariant방법을 제안하였는데, 한 슬라이드 내에서 여러 번 spot된 자료에 대한 두 형광광도의 비의 변동계수(CV)를 quality filtering의 한 측도로 삼았다. 통상적인 normalization 작업 후 특정 spot의 CV값에 대해서 threshold값을 지정하여 이를 만족하지 못하는 경우에는 한 슬라이드에서 반복된 spot중 하나의 특이치를 제거하고 다시 CV를 반복적으로 계산하여 일정기준을 만족하는지 확인하는 과정을 거치게 된다. Zien et. al(2001)은 각 반복 실험마다 획일적인 normalization 방법이 잘 적용될 수 없음을 지적하고, 반복 슬라이드에서 발현강도의 비에 대한 central tendency가 좋은 추정치임을 지적하였다.

1) 고려대학교 통계학과 석사과정, (130-701) 서울특별시 성북구 안암동 5가 1번지
2) 고려대학교 통계학과 교수, (136-701) 서울특별시 성북구 안암동 5가 1번지

microarray data의 특성중 하나는 spot intensity가 증가함에 따라 분산이 증가되는 현상을 확인할 수 있는데 통상적으로 logarithmic 변환을 함으로써 분산이 안정화됨을 볼 수 있다. 그러나 logarithmic 변환 후 high intensity와는 달리 low intensity spot의 경우에는 값의 변동의 차가 심함을 확인하고 여러 가지 분산안정화방법이 제안되었다. Huber et al(2002)는 low intensity 에서의 분산안정화 방법으로 arsinh transformation을 제안하였고 Cui et al.(2003)는 microarray data에 linear model을 이용하여 low intensity 에서는 linear 변환을 그리고 high intensity 에서는 logarithmic 변환을 하는 linlog transformation 방법을 제안하였다.

microarray 통계분석에서 normalization은 후의 분석에 큰 영향을 미치므로 조심스런 작업을 요한다. 본 논문은 Yang et al(2001)이 제시한 print-tip lowess normalization에 대해 논의 하고자 한다. Yang et al.(2001)이 제안한 normalization들은 전체 유전자로 시행하는 것으로 대다수의 유전자가 발현되지 않는다는 사실을 가정으로 한다. lowess의 특성상 window width를 정해줘야 함은 자명한 사실이고 가정에 충실히 하기 위해 window width를 가능한 작게 하여 normalization 하는 것이 옳은 방향이겠으나 분산이 줄어들고 상대적으로 bias가 커지게 되어 잘못된 결과를 도출할 수 있을 것이다. 또한 window width를 크게 하여 normalization을 실행한다면 가정에 안 맞을 수도 있을 뿐 아니라 lowess normalization의 초점인 intensity dependent effect를 보정하기 어려운 가능성을 배재할 수 없을 것이다. 본 연구는 각 tip마다 각각 다른 window width가 있다는 것을 확인하였으며 방법(cross validation, bootstrap)으로 찾은 window width를 normalization에 사용함으로서 동일한 window width를 사용하는 기존의 방법과 비교 평가하여 normalization의 기본 원칙에 대한 타당성을 확인하였다.

2. 방법

논문에서 사용된 lowess의 window width를 찾는 방법으로 다음과 같다.

- **Cross validation**

model selection에서 가장 직관적이고 흔히 쓰이는 방법은 cross validation일 것이다. N개의 유전자를 랜덤하게 K개의 그룹으로 분할한 다음 I번째 그룹을 제거한 I-1개의 그룹으로 lowess curve을 구하여 가장 작은 값을 가질 경우의 window width를 선택 한다

$$CV(f) = 1/N \sum_{i=1}^{N} L(M_i, f^{-h(i)}(A_i, f))$$

- **Bootstrap** - Jhun et al, 1993

window width 선택의 기준으로 MISE(Mean Integration Square Error)를 고려한다.

$$MISE(f) = E\int [g(\hat{x};f) - g(x)]^2 dx$$

MISE의 bootstrap추정치로 BMISE(Bootstrap Mean Integration Square Error)를 최소로 할 경우의 window width를 선택한다.

$$BMISE(f, f_0) = B^{-1} \sum_{i=1}^{B} \int [g_i^*(x;f) - \hat{g}(x;f_0)]^2 dx$$

여기서 $f^*(x;f)$는 bootstrap sample의 lowess curve, $f(x;f_0)$는 초기값을 갖을 때의 lowess curve이다. 초기값 f_0는 cross validation 선택이나 bootstrap 한 후 찾은 값을 다시 초기값으로 선택할 수 있다.

3. 결과

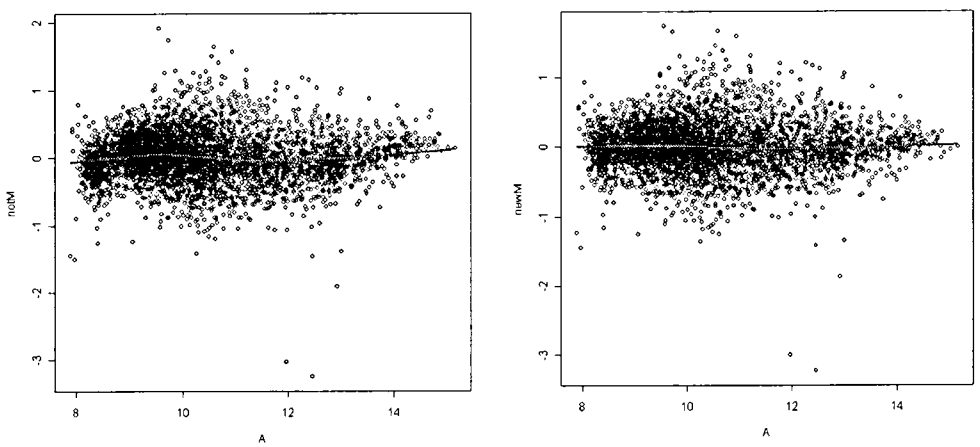
본 연구에서는 2가지의 방법(cross validation, bootstrap)을 HDL deficient mouse(M. J. Callow, 2001)데이터에 적용시켜 비교 분석을 실시하였다. 이 실험은 신진대사에 어떤 유전자가 영향을 미치는지 알아보는 것이 목적이다. 슬라이드는 16장, 4*4 print-tip, 각 tip마다 spot개수는 399개이다. 우선 background 보정을 한 후 2가지의 방법으로 구한 window width로 print-tip lowess normalization을 실시한 결과와 그리고 동일한 window width로 print-tip lowess normalization한 결과를 비교하였다. 공개 소프트웨어인 R을 사용하였고 동일한 window width를 이용한 분석에서는 lowess에서 제공하는 기본값을 사용하였다.

<표 1>은 한 슬라이드에서의 2가지 방법으로 찾은 각 tip에서의 window width이다. tip마다 다양한 window width가 발견됨을 볼 수 있다. 이 결과로 인하여 print-tip lowess normalization할 경우에 tip에 동일한 window width로 행하는 것보다는 tip에 맞는 window width로 normalization하는 것이 타당해 보인다. <그림 1>은 normalization결과를 MA plot으로 확인한 것으로 우측 그림이 bootstrap으로 찾은 window width로 normalization 한 결과이다. 좌측보다는 우측그림이 normalization 결과가 좋음을 알 수 있다.

bootstrap은 초기값에 민감하여 실행 후 window width가 조금씩 변하지만 어떤 초기값이나 넣어도 bootstrap후 값을 다시 초기값으로 bootstrap하는 작업을 반복하면 일정한 값에 수렴하는 과정을 보였다. <표 1>은 bootstrap을 이용한 window width의 수렴된 결과를 나타낸 것이다. Jhun et al.(1993)은 bootstrap의 초기값으로 cross validation 결과나 가능한 큰 window width를 사용하는 것을 제안하였지만 많은 경우에 cross validation 결과를 bootstrap에 이용하는 것이 상대적으로 수렴횟수를 적게 하는 결과를 보였다.

Tip	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
window width	0.38	0.16	0.5	0.4	0.34	0.2	0.3	0.76	0.38	0.34	0.5	0.38	0.6	0.32	0.5	0.5

<표 1> cross validation결과를 초기값으로 한 bootstrap 후의 window width



<그림 1>동일한 window width(좌)와 bootstrap(우)후의 window width로 실행된 normalization 결과

참고 문헌

1. Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. Bioinformatics 19(2):185-193
2. Hastie(2001), The Elements of Statistical Learning, The Springer. (NewYork)
3. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 1:1-9
4. M. J. Callow, S. Dudoit, E. L. Gong, T.P. Speed, and E. M. Rubin. Microarray expression proling identies genes with altered expression in hdl decient mice. *Genome Research*, Submitted.
5. M. K. Kerr, M. Martin, and G. A. Churchill. Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7:819-837, 2000.
6. George C. Tseng, Min-Kyu Oh, Lars Rohlin, James C. Liao, and Wing Hung Wong. (2001) Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research.29: 2549-2557
7. Myoungshic Jhun(1993), Bootstrap choice of smoothing parameter of locally weighted linear regression. : J.Japanese Soc, Comp, Statist, 5(1993), 25-32
8. S. Dudoit, Y. H. Yang, T. P. Speed, and M. J. Callow (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica, Vol. 12, No. 1, pp. 111-139.
9. T. B. Kepler, L. Crosby, and K. T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. Genome Biology, 3(7) : research 0037.1-12, 2002.
10. W. S. Cleveland and S. J. Devlin. Locally-weighted regression: An approach to regression analysis by local fitting. Journal of the American Statistical Association, 83 : 590 - 610, 1988.
11. Y.H.Yang, S. Dudoit, P. Luu, and T. P. Speed (2001). Normalization for cDNA microarray data. InM. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds), Microarrays: Optical Technologies and Informatics, Vol. 4266 of Proceedings of SPIE.
12. Y. H. Yang, S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research, 30(4):e15, 2002.
13. Y.H.Yang and N. Thorne (2003) Normalization for Two-color cDNA Microarray Data. Science and Statistics: A Festschrift for Terry Speed, D. Goldstein (eds.),IMS Lecture Notes, Monograph Series, Vol 40, pp. 403-418.
14. Xiangqin Cui, M. Kathleen Kerr, and Gary A. Churchill (2003) "Transformations for cDNA Microarray Data", Statistical Applications in Genetics and Molecular Biology.
15. Zien A, Aigner T, Zimmer R, Lengauer T (2001). Centralization: a new method for the normalization of gene expression data. Bioinformatics 17 Suppl.:S323-S331