Statistical Tests for Time Course Microarray Experiments

박태성[1], 이성곤[2], 최호식[3], 이승연[4], 이용성[5]

요 약

Microarray technology allows the monitoring of expression levels for thousands of genes simultaneously. In time-course experiments in which gene expression is monitored over time we are interested in testing gene expression profiles for different experimental groups. We propose a statistical test based on the ANOVA model to identify genes that have different gene expression profiles among experimental groups in time-course experiments. Using this test, we can detect genes that have different gene expression profiles among experimental groups. The proposed model is illustrated using cDNA microarrays of 3,840 genes obtained in an experiment to search for changes in gene expression profiles during neuronal differentiation of cortical stem cells.

주요용어 : ANOVA model, F-test, permutation test

1. 서론

Human Genome Project의 10여년간의 연구 결과로 우리는 인간이 지니고 있는 30억개의 DNA 염기서열을 모두 해독하게 되었으며 이는 생명공학의 급속한 발전과 함께한 결과이다. 그 결과로 최근 생명체들에 대한 유전체(genome) 연구가 활발하게 진행되고 있다. 이러한 연구는 유전체의 서열정보뿐만이 아니라 유전체들의 기능과 구조를 밝히는 것이 주 목적이다(Rashidi and Buehler, 2000). 그러나, 기존의 유전자 연구방법으로 한 연구자가 동시에 많은 수의 유전자에 대한 연구를 하는 것은 한계가 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 두개의 유전자의 기능탐색이라는 종래의 한계를 벗어나 생명현상과 관련된 유전체수준의 연구가 가능해졌다라는 것을 뜻한다(Schena등, 1995, Schena등, 1999; Schena, 2000). Eisen 등 (1998)은 microarry 자료의 분석을 통해서 발현패턴이 비슷한 유전자군을 찾기위한 군집분석 방법을 제안했으며, Golub 등(1999)은 microarray 실험이 새로운 종류의 암을 찾을 수 있는데 유용하게 사용될 수 있음을 입증했다. 앞으로 이러한 DNA microarray 관련 연구는 유전체 연구에서 보다 활발하게 응용될 것으로 기대된다.

이러한 DNA chip 기술에는 cDNA chip 방식과 Affimatrix사의 oligochip방식이 있다. 이 중

1) 서울대학교 통계학과 교수, 서울시 관악구 신림동산 56-1
2) 서울대학교 통계학과 박사과정, 서울시 관악구 신림동산 56-1
3) 서울대학교 통계학과 박사과정, 서울시 관악구 신림동산 56-1
4) 세종대학교 응용수학과 교수, 서울시 광진구 군자동 98
5) 한양대학교 생화학교실 교수, 서울시 성동구 행당동 17

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.

Microarray 자료를 이용한 연구의 목적 중에 하나가 여러 처리그룹들간에 발현패턴이 다르게 나타나는 유전자를 찾아내는 것이다. 예를 들어 위암조직과 정상조직을 사용하여 칩실험을 한 경우에 정상조직에 비해서 위암조직에서 강하게 발현되는 유전자를 찾을 수 있다. 이 유전자가 위암의 유발과 관련이 있는 유전자라는 것을 밝힐 수 있다. 칩 실험이 보편화되기 전에는 한 개의 칩 실험을 통해서 몇 배(예. 2배) 이상 발현이 되는 유전자를 찾는 단순한 방법이 사용되었으나 최근에 반복 실험이 보편화 되면서 보다 다양한 탐색 방법들이 소개되고 있다.

반복이 없는 실험에서는 Chen등(1997)에 의해 제안된 방법으로서, 각각의 발현정도가 정규분포(normal distribution)를 따른다고 가정한 후, 발현되는 정도에 대한 비율들의 평균에 대한 신뢰구간을 이용하는 방법이 사용될 수 있다. 또 다른 방법으로는 Newton등(2001)에 의해 제안된 방법이 있는데 이 방법은 각각의 발현정도의 분포가 감마분포(gamma distribution)를 따른다는 가정 하에서 비율에 대한 사후 오즈(posterior odds)를 이용하는 방법이다. Park등(2001)이 제안한 방법은 통계분석에 가장 보편적으로 사용되는 통계모형인 회귀모형을 사용하여 microarry 자료를 적합한 후에 잔차값을 이용하여 유의한 유전자를 찾는 방법이다. 이 방법은 Chen등과 Newton등의 방법에 비해 계산하기가 간단할뿐더러 기존의 통계분석프로그램을 이용하여 쉽게 결과를 얻을 수 있는 장점을 갖고 있다.

본 연구에서는 여러 처리 그룹간에 다르게 발현된 유의한 유전자를 탐색하는 실험에서 여러 다른 시간대에 걸쳐서 유전자의 발현패턴을 비교하기위한 탐색 방법을 제시하고자 한다. 먼저 여러 개의 칩 실험을 통해서 얻어진 다중(multiple) 슬라이드에 적용할 수 있는 탐색방법을 소개한 후에 이 방법을 시간별로 발현된 자료의 분석에 적용하고자 한다.

2. 처리그룹이 2개인 경우의 유의한 유전자의 탐색 방법

먼저 비교하고자하는 처리 그룹이 2개인 경우를 생각해 보자. 이표본 t-검정법을 이용한 방법을 Dudoit등(2000)이 제안하였다. 이 방법은 이표본 평균을 비교하기위한 t-통계량을 슬라이드에 있는 각 유전자별로 구한 후에 이 통계량값을 기초로 유의한 유전자를 찾는 방법이다. 이와 동일한 아이디어를 사용하여 Golub등(1999) 새로운 종류의 암을 발견한 바 있다. Kerr 등(2000)은 분산분석모형(ANOVA, analysis of variance)을 사용하고 Wolfinger등(2000)은 혼합모형(mixed model)을 사용하여 유의한 유전자를 찾는 방법을 제안하였다.

비교하고자 하는 처리그룹이 2개이고 한 개의 슬라이드에 N개의 유전자가 있다고 가정하고 첫 번째 그룹에서는 n_1 개의 슬라이드가 두 번째 그룹에서는 n_2 개의 반복실험된 슬라이드가 있다고 가정하자. 처리그룹을 $i(=1,2)$ 로 나타내고, 반복된 슬라이드를 $j(=1,\cdots,n_i)$ 로 나타내고, 유전자를 $l(=1,\cdots,N)$ 로 나타내자. x_{ijl} 을 로그변환된 적색과 녹색 발현강도의 비를 나타낸다고 가정하자. 그러면 각 유전자 별로 두 처리 그룹간에 평균값은 다음과 같다.

$$\bar{x}_{1.l} = \frac{1}{n_1}\sum_{j=1}^{n_1} x_{1jl} : \quad l \text{ 번째 유전자의 첫번째 처리그룹 } (i=1) \text{에서의 평균값}$$

$\bar{x}_{2.l} = \frac{1}{n_2}\sum_{j=1}^{n_2} x_{2jl}$: l 번째 유전자의 두번째 처리그룹 $(i=2)$ 에서의 평균값

이 평균값으로부터 다음과 같은 t-통계량을 유도할 수 있다.

$$t_l = \frac{\bar{x}_{1\cdot l} - \bar{x}_{1\cdot l}}{\sqrt{\frac{s_{1l}^2}{n_1} + \frac{s_{2l}^2}{n_2}}}, \quad l=1,\cdots,N$$

여기서 는 각 처리그룹의 표본분산을 나타내며 다음과 같이 정의된다.

$$s_{il}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ijl} - \bar{x}_{i\cdot l})^2, \quad i=1,2$$

이 t-통계량을 각각의 유전자마다 따로 정의된다. 이 통계량은 두 그룹의 평균을 비교하기위해 널리 사용되는 t-통계량과 동일한 형태를 갖는다. 만약 첫 번째 처리그룹과 두 번째 처리 그룹의 분산도 동일하다면 좀 더 간단한 다음의 식을 사용할 수도 있다.

$$t_l^* = \frac{\bar{x}_{1\cdot l} - \bar{x}_{1\cdot l}}{s_l\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad l=1,\cdots,N$$

여기서 s_l^2 은 두 처리그룹의 자료를 합하여 구한 공통분산에 대한 추정량이고 다음과 같이 정의된다.

$$s_l^2 = \frac{\sum_{j=1}^{n_1}(x_{1jl} - \bar{x}_{1\cdot l})^2 + \sum_{j=1}^{n_2}(x_{2jl} - \bar{x}_{2\cdot l})^2}{n_1 + n_2 - 2}$$

t_l 이나 t_l^* 같은 t-통계량값을 계산한 다음에는 t-분포를 이용하여 유의확률(p-값)을 계산할 수 있다. t-통계량의 절대값이 클수록 유의확률값은 작아지며 l 번째 유전자가 두 처리그룹에서 아주 다른 발현값을 갖고 있다는 것을 나타낸다. 따라서 t-통계량의 절대값을 기초로 큰 값에서부터 작은값 순서대로 의미가 있는 유전자를 찾아낼 수 있다.

이런 t-통계량을 사용하게 되면 특정 유전자가 여러 슬라이드에서 발현되는 분포를 고려하여 분산값을 추정하게 되므로 이 분산값에 비해 실제로 관찰된 발현값이 유의하게 차이가 나는지를 객관적으로 평가하게 된다. 따라서 분포이론에 근거한 통계적 추론 방법이 그냥 단순하게 2배 혹은 3배 차이에 근거한 탐색 방법에 비해 훨씬 객관적이고 정확한 결과를 제공한다고 할 수 있다.

그 다음으로 고려할 문제는 각 유전자마다 t-통계량을 구한후에 이 값을 기초로 결론을 구해야하는데 전체 N 개의 검정을 동시에 시행해야 하므로 제1종오류(type one error)를 조절할 필요가 있게된다. 흔히 사용되는 방법은 Bonferroni방법이 있는데 이 방법은 보통의 유의수준

α 대신에 α/N 을 사용하는 방법이다.

앞에서 소개한 t-검정은 자료의 분포가 정규분포를 따른다는 가정하에서 사용하는 방법이다. 그러나 많은 경우에 microarray에서 구한 발현값이 정규분포를 따르지 않는다. Dudoit등은 순열검정(permutation test)에 기초한 비모수적 검정법을 소개하였다. 이 순열검정법은 자료를 계속 순열변환시켜 새로운 종류의 자료를 생성시킨 후에 이 자료를 이용하여 다시 t-통계량값을 계산하는 작업을 수만번 반복하여 t-통계량의 분포를 경험적(empirical)으로 구하는 방법이다. 이 검정법은 t-분포를 사용하지 않으므로 정규분포에 대한 가정없이 검정결과를 구할 수 있는 방법이다. 이 과정을 수행하기위해 많은 계산이 요구되나 컴퓨터의 계산속도가 워낙 향상되어서 별 문제없이 쉽게 구할 수 있다. 이 순열검정법에서도 역시 N 개의 검정을 동시에 시행해야 하므로 제1종오류(type one error)를 조절할 필요가 있게된다. 이 경우에는 Westfall and Young (1993)의 step-down 방법이 널리 사용된다.

3 처리그룹이 3개 이상인 경우의 유의한 유전자 탐색 방법

비교하고자 하는 처리그룹이 $I(\geq 3)$ 개이고 한 개의 슬라이드에 N 개의 유전자가 있다고 가정하고 i 번째 그룹에서는 n_i 개의 슬라이드가 있다고 가정하자. 앞에서와 마찬가지로 처리그룹을 $i(=1,\cdots,I)$ 로 나타내고, 반복된 슬라이드를 $j(=1,\cdots,n_i)$ 로 나타내고, 유전자를 $l(=1,\cdots,N)$ 로 나타내자. x_{ijl} 을 로그변환된 적색과 녹색 발현강도의 비를 나타낸다고 가정하자. 그러면 각 유전자 별로 두 처리 그룹간에 평균값은 다음과 같다.

$$\bar{x}_{i.l} = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{1jl}, i=1,\cdots,I : \ l \text{ 번째 유전자의 } i\text{번째 처리그룹에서의 평균값}$$

이 평균값으로부터 각 처리그룹의 평균이 동일한지를 검정하기위해 다음과 분산분석표(analysis of variance table; ANOVA table)를 만들 수 있다.

분산분석표

요 인	제곱합	자유도	평균제곱	F-값	유의확률
처 리 잔 차	SStr SSE	I-1 n-I	MStr MSE	MStr/MSE	p-값
계	SST	n-1			

분산분석표에서 정의되는 제곱합(sum of squares)들과 평균제곱합(mean sum of squares)들은 다음과 같다.

$$SST = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ijl} - \bar{x}_{..l})^2$$

$$SStr = \sum_{i=1}^{I}(\bar{x}_{i\cdot l} - \bar{x}_{..l})^2, \quad MS_{tr} = \frac{SS_{tr}}{I-1}, \quad MSE = \frac{SSE}{n-I}$$

$$SSE = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(x_{ijl} - \bar{x}_{i\cdot l})^2$$

평균제곱합을 이용하여 다음과 같은 F-통계량을 구할 수 있다.

$$F = \frac{MS_{tr}}{MSE}$$

이 F-통계량은 각 처리그룹의 평균이 동일하다는 가정하에서 자유도가 $(I-1, n-I)$인 F-분포를 따르게 된다.

t-통계량과 마찬가지로 각 유전자별로 F-통계량을 계산한 다음에는 F-분포를 이용하여 유의확률(p-값)을 계산할 수 있다. F-통계량의 값이 클수록 유의확률값은 작아지며 l 번째 유전자가 I개의 처리그룹에서 아주 다른 발현값을 갖고 있다는 것을 나타낸다. 따라서 F-통계량의 값을 기초로 큰 값에서부터 작은값 순서대로 의미가 있는 유전자를 찾아낼 수 있다.

다음으로 또 고려할 문제는 각 유전자마다 N 개의 F-통계량을 구한후에 이 값을 기초로 결론을 구해야하므로 역시 제1종오류를 조절할 필요가 있게된다. 이 경우에는 흔히 사용되는 Bonferroni방법을 이용하면 된다.

이 F-검정도 역시 각 자료의 값들이 정규분포를 따른다는 가정하에서 사용하는 방법이다. 그러나 많은 경우에 microarray에서 구한 발현값이 정규분포를 따르지 않는다. Park등(2001)은 ANOVA 모형에서 얻어지는 잔차(residual)을 이용한 순열검정(permutation test)에 기초한 비모수적 검정법을 소개하였다. 이 순열검정법은 자료를 계속 순열변환시켜 새로운 종류의 자료를 생성시킨 후에 이 자료를 이용하여 다시 F-통계량값을 계산하는 작업을 수만번 반복하여 F-통계량의 분포를 경험적(empirical)으로 구하는 방법이다.

참고문헌

Chen, Y., Dougherty, E.R. and Bittner, M.L. (1997). Ration-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* 2(4), 364-374

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2000). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray expriments. Submitted.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci., 95:14863-14868.

Golub, T.R. Slonim, T.K. Tamayo, P. Huard, C., Gaasenbeek, M. Mesirov, J.P., Coller, H. Loh, M.H. Downing, J.R., Caligiuri, M.A., Bloomeld, C.D., and Lander, E.S.(1999).

Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, 286:531-537.

Myers, R.H. (1990). Classical and Modern Regression with Applications, PWS-KENT Publishing company

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarry data. *Journal of Computational Biology*, 8 37-52

Rashidi, H.H., Buehler, L.K. (2000). Bioinformatics Basic: Applications in Biological Science and Medicine, CRC Press

Schena, M. (1999) editor. DNA Microarrays : A Practical Approach. Oxford University Press.

Schena, M., Shalon, D. Davis, R.W., and Brown, P.O.(1995). Quantitative monitoring of gene expression patterns with a complementary dna microarray. Science, 270:467-470.

Schena, M. Shalon, D., Heller, R. Chai, A., Brown, P.O., and Davis, R.D. Parallel human genome analysis: microarray-based expression monotoring of 1000 genes (1996). Proc. Natl. Acad. Sci., 93:10614-10619.

Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C., and Wong, W. H. (2001). Issues in cDNA microarray analysis: qualit filtering, channel normalization, models of variations and assessment of gene effects. Nucleic Acids Research, 29, 12, 2549--2557.

Westfall, P. H. and Young, S. S. (1993) Resampling-based multiple testing: examples and methods for p-value adjustment. Wiley series in probability and mathematical statistics. Wiley.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology, 8, 6, 625--637.