Journal of Korean
Data & Information Science Society
2007, Vol. 18, No. 4, pp. 1005 ~ 1021

Monitoring of Gene Regulations Using Average Rank in DNA Microarray: Implementation of R[1)]

Changsoon Park[2)]

Abstract

Traditional procedures for DNA microarray data analysis are to preprocess and normalize the gene expression data, and then to analyze the normalized data using statistical tests. Drawbacks of the traditional methods are: genuine biological signal may be unwillingly eliminated together with artifacts, the limited number of arrays per gene make statistical tests difficult to use the normality assumption or nonparametric method, and genes are tested independently without consideration of interrelationships among genes. A novel method using average rank in each array is proposed to eliminate such drawbacks. This average rank method monitors differentially regulated genes among genetically different groups and the selected genes are somewhat different from those selected by traditional P-value method. Addition of genes selected by the average rank method to the traditional method will provide better understanding of genetic differences of groups.

Keywords : Average Rank Method, Between Array, DNA Microarray, Gene Expression Data, LOWESS Regression, Mock Array, P-value Method, Quantile Normalization, Within Array

1. 서론

DNA 마이크로어레이 실험의 주목적은 둘 또는 그 이상의 그룹(specimen, group, class)간의 유전자 발현수준을 비교하여 특정그룹에서 나머지 그룹과 비교할 때 발현수준이 차별적으로 나타나는 유전자를 확인하는 것이다. 마이크로어레이는 수천, 수만 유전자의 발현수준을 동시에 측정할 수 있게 하는 대용량처리(high throughput) 실험으로서 어떤 특정요인이 서로 상관성이 있는 유전자군에 미치는 영향을 전체 유전체(genome)의 윤곽(profile)을 통해 알아볼 수 있게 함으로써, 분자생물학적 연관성을 총

1) This research was supported by the Chung-Ang University research grants in 2006.
2) Professor, Department of Statistics, Chung-Ang University, Seoul 156-756, Korea. E-mail : cspark@cau.ac.kr

체적으로 이해할 수 있게 하였다. 이것은 전통적 통계적 방법에서 서로 상관성이 있는 여러 개의 변수를 연구할 때 다변량기법을 사용하는 것과 유사한데, 마이크로어레이 데이터에서는 변수(유전자)의 수가 반복치(어레이)의 수보다 훨씬 더 많다는 점이 주된 다른 이유가 된다. 마이크로어레이 데이터의 분석방법은 지금까지 많은 연구가 되어오고 있으며 이에 관한 문헌으로는 Draghici (2003), Krane and Raymer (2003), Amaratunga and Cabrera (2004), 박태성 등(2005), Augen (2005) 등을 들 수 있다.

유전자 발현수준을 나타내는 마이크로어레이 데이터에는 과도한 인위적 편차(artificial bias)가 포함되어 있어 생물학적 특성을 제대로 표현하지 못하는 현상이 발생한다. 이러한 오류는 실험의 각 단계, 즉 마이크로어레이 제작, mRNA준비, 혼성화(hybridization), 스캐닝(scanning), 이미징(imaging) 등에서 발생하는 것으로서 일정치의 편차를 유발하여 발현수준의 추정을 정밀하지 못하게 하여 어레이간 비교를 어렵게 한다. 이러한 문제점을 해결하기 위한 대표적 방법은 preprocessing 과 표준화(normalization)가 있다. Preprocessing과 표준화는 관측된 발현수준에서 마이크로어레이 제조과정에서 발생하는 비생물적 변동이나 편차를 제거하고자 하는 활동으로서 우연편차(random bias)는 제거할 수 없으나 구조적편차(systematic bias)는 몇가지 가정하에서 제거가 가능하다 (Knudsen, 2004).

Preprocessing은 크게 배경휘도(background intensity)조정과 척도변환(scale transform)으로 구분한다. 배경휘도 조정은 유전자만에 의한 발현수준을 측정하려는 것으로서, cDNA 칩에서는 전경휘도(foreground intensity)에서 배경휘도를 빼고, 올리고 칩에서는 PM(perfect match)휘도에서 MM(mismatch)휘도를 빼주게 된다. 마이크로어레이 데이터는 극단적 우측편향(positive skewness)분포를 나타내어 대부분의 데이터가 0에 아주 가까이 밀집된 형태를 보이고 있어 데이터분포의 분석에 어려움이 있다. 이를 해결하기 위해 척도변환으로 로그변환(밑이 2)을 사용하여 변환된 발현수준이 근사적으로 정규분포(Chen et al, 1997), 또는 감마분포(Newton et al, 2001)를 따르도록 한다.

표준화는 어레이내(within array) 표준화와 어레이간(between array) 표준화로 구분한다. 마이크로어레이에 혼성화되는 mRNA 표본이 한 종류인 경우를 일 채널(single channel), 두 종류인 경우를 이 채널(two channel)이라 한다. 이 채널 마이크로 어레이에서는 두 종류의 형광염료(Red, Green)를 사용하여 두 종류의 mRNA 표본을 동시에 혼성화하게 된다. 일 채널 어레이인 경우에는 어레이간 표준화를 사용하여 표본에 따른 유전자의 발현 수준을 비교하게 된다. 반면에 이 채널 어레이인 경우에는 어레이내 표준화를 사용하여 동일한 어레이 내에서 형광염료에 차이에 따른 오차를 먼저 수정한 다음, 어레이간 표준화를 사용하게 된다. 어레이간 표준화에는 분위수(quantile) 표준화가 널리 사용되고 있다. X_{ij}를 i번째($i=1,2,\cdots,l$) 어레이에서 j번째($j=1,2,\cdots,G$) 유전자의 로그변환된 발현수준이라 하고, $X_{i(j)}$는 i번째 어레이 내에서 j번째 순서통계량, R_{ij}는 X_{ij}의 순위라 하자. 이 때 대표 어레이(mock array)는 $Q_j = \sum_{i=1}^{l} X_{i(j)}/l$ 로 표현되고 X_{ij}의 표준화된 발현수준은 $Q_{R_{ij}}$ 가 된다. 분위수 표준화를 사용하게 되면 주어진 어레이 내에서 G개 유전자의 순위는 동일하게 유지되지만, 각 어레이 내에 있는 유전자의 발현수준들은 다른 어레이의 발현수준들과 모두 같아짐을 알 수 있다. 이 채널인 경우에는 두 형광 염료 R(red) 과 G(green)가 하나의 어

레이에 동시에 혼성화되므로 두 채널 사이의 휘도에는 염료편차(dye bias)가 발생하여 참 휘도를 오도하게 된다. 따라서 어레이 내 표준화가 필요하게 되고 주로 사용하는 표준화 방법은 MA plot을 사용하는 방법이다 (Deshmukh and Puroit, 2006).

MA plot에서는 LOWESS(locally weighted regression) 회귀분석[Cleveland(1979), Cleveland and Devlin(1988)]으로 추정된 $\hat{M}$을 M에서 뺀 값으로 M을 대체하게 된다. 이 때 표준화된 로그변환 발현수준은

$$logR = logR - \hat{M}/2$$
$$logG = logG + \hat{M}/2$$

로 표현된다.

Preprocessing과 표준화가 끝난 마이크로어레이 데이터를 유전자 발현행렬이라 하고 이 행렬을 사용하여 생물학적 특성을 관찰하기 위해 여러 가지 통계적 분석을 실시한다. 이에 사용되는 주된 통계적 방법은 개개의 유전자에 대해서 통계적 가설검정을 실시하고 P-value의 크기순으로(작은 값부터) 유의한 정도를 순서화하여 일정수의 유전자를 선택하게 된다. 이렇게 선택된 유전자에서는 그룹 간 발현수준이 서로 다르게 나타나는 것으로 판단하여 해당 유전자의 역할을 연구하게 된다. 이러한 다중가설검정(multiple hypothesis test) 문제에서는 검정의 제1종 오류(Type I error)가 증가하는 문제점을 조정하기 위해 false discovery rate (FDR)을 고려하여 보정하게 된다.

마이크로어레이에서 사용되는 다중가설검정은 동시에 수천, 또는 수만 개의 유전자에 대한 검정이 실시되지만, 각 검정은 유전자 각각에 대한 별도의 검정이 이루어진 것이다. 이 각각의 검정에서 데이터의 정규성을 검토하고, 정규성이 인정되면 모수적 방법(t-test, F-test)을 사용하고 그렇지 않으면 비모수적 방법을 사용하게 된다. 마이크로어레이 데이터는 근본적으로 유전자 상호간의 연관성을 가진 대용량의 다변량 데이터임을 감안하면, 유전자 간의 상관성을 고려하여 유의한 유전자를 선별하는 것이 중요함을 알 수 있다. 이 논문에서는 유의한 유전자를 선별할 때 P-value에 의한 개별적 유의성으로 판단하는 대신 어레이 내에서의 순위를 사용하는 방법에 대해 연구하였다. 즉, 각각의 어레이 내에서 각 유전자의 상대적 발현수준을 순위로 표현하고, 다른 어레이에서 해당 유전자의 순위와 비교하는 방법을 통해 유전자를 선별하는 방법을 제시하고 있다. 이 방법은 기존의 P-value에 의한 유전자의 선별방식 하에서는 선별되지 않았던 유전자중 상호간의 연관성에 의해 중요한 역할을 하는 유전자를 추가로 선별할 수 있어 유전자 비교에 도움을 줄 수 있다고 판단된다. 평균순위 방법과 P-value 방법을 적용하는 과정은 R-언어를 사용하여 구현하였으며, 특히 R-graphic은 그림을 이용한 분석에 도움이 되었다. R을 이용한 통계적 분석과 그래픽 방법은 Everitt and Hothorn (2006), Murrell (2006)등에 잘 설명되어 있다.

2. 평균순위(average rank)

한 유전자가 여러 어레이내에서 해당하는 순위들은 서로 다른 값을 가질 수 있다. 이러한 서로 다른 순위를 동일 그룹내의 여러 어레이에 대해서 취한 평균값을 평균순

위라 한다. 이 논문에서는 편리상 순위는 가장 큰값부터 그 다음 작은 순으로 부여하는 것으로 가정한다. 또한 발현치는 preprocessing과 표준화가 되어진 값으로 간주한다.

특정 순위에 해당하는 유전자의 번호를 나타내는 함수를 다음과 같이 정의한다. 유전자 번호 i에 대한 발현치를 X_i, 해당 어레이 내에서 X_i의 순위를 R_i라 할 때, 함수 V(R_i)는 순위 R_i에 해당하는 유전자 번호는 i 를 나타내도록 정의한다. 동일한 어레이 내에서 발현치 X_i의 순서통계량을 $X_{(i)}$라 하면 $X_{(i)} = X_{V(i)}$임을 알 수 있다. 아래 표처럼 G개의 유전자의 발현치 X가 어떤 순위를 가진다고 하자.

유전자 번호:i	1	2	3		G
발현치: X_i	X_1	X_2	X_3		X_G
순위: R_i	4	7	1		49

위 표로부터

$$V(1)=3,\ V(4)=1,\ V(7)=2,,,,V(49)=G$$

임을 알 수 있고, 따라서 발현치의 순서통계량은 다음과 같이 표현된다.

$X_{(1)} = X_{V(1)} = X_3,\ X_{(4)} = X_{V(4)} = X_1,\ X_{(7)} = X_{V(7)} = X_2,\ X_{(49)} = X_{V(49)} = X_G$

만일 특정 유전자가 그룹 간 발현정도가 다르다면 동일 그룹내에 있는 어레이에서는 발현정도가 유사하지만, 다른 그룹에 있는 어레이에서는 발현정도가 다르게 나타날 것이다. 이 때 상대적 크기인 순위도 유사한 양상을 보이게 된다.

3. 일채널 이그룹 문제

유전자의 수를 G, 그룹의 수를 2[첫째 그룹은 대조(control, wild)그룹, 둘째 그룹은 처리(treatment, mutant)], j 번째 그룹에서 반복어레이의 수를 l_j 라 하면, 각 유전자의 발현수준은 $\{X_{ijk},\ i=1,2,\cdots,G,\ j=1,2,\ k=1,2,\cdots,l_j\}$로 표현한다. 이 때 주어진 j,k 에 대해 $X_{.jk} = (X_{1jk}, X_{2jk}, \cdots, X_{Gjk})$는 하나의 어레이를 형성한다. 어레이 $X_{.jk}$의 순위는 $\{R_{ijk},\ i=1,2,\cdots,G\}$로 나타낸다. 그룹 j에서 i번째 유전자의 평균순위는

$$\overline{R}_{ij.} = \frac{\sum_{k=1}^{l_j} R_{ijk}}{l_j}$$

가 되고, 두 그룹에 대한 평균순위의 차는 $D_i = \overline{R}_{i1.} - \overline{R}_{i2.}$ 로 정의한다. D_i의 절대값을 $A_i = |D_i|$라 하고 이를 크기 순으로 나열하여 차별적으로 발현되는 유전자를 탐지한다.

$\{A_i, i=1,2,\cdots,G\}$의 순위를 $\{T_i, i=1,2,\cdots,G\}$라 하면 $V(T_i)=i$가 되고 순서통계량은 $A_{(i)}=A_{v(i)}$임을 알 수 있다. 전체 유전자중 그룹간 평균순위가 가장 차별적으로 발현된 100p%의 유전자를 탐지하고자 하면 선택된 유전자 번호는, Gp에 가장 가까운 정수값 g에 대해,

$$V(1), V(2), \cdots, V(g)$$

가 되고, 이에 해당하는 유전자의 두 그룹간 평균순위의 차는

$$D_{V(1)}, D_{V(2)}, \cdots, D_{V(g)}$$

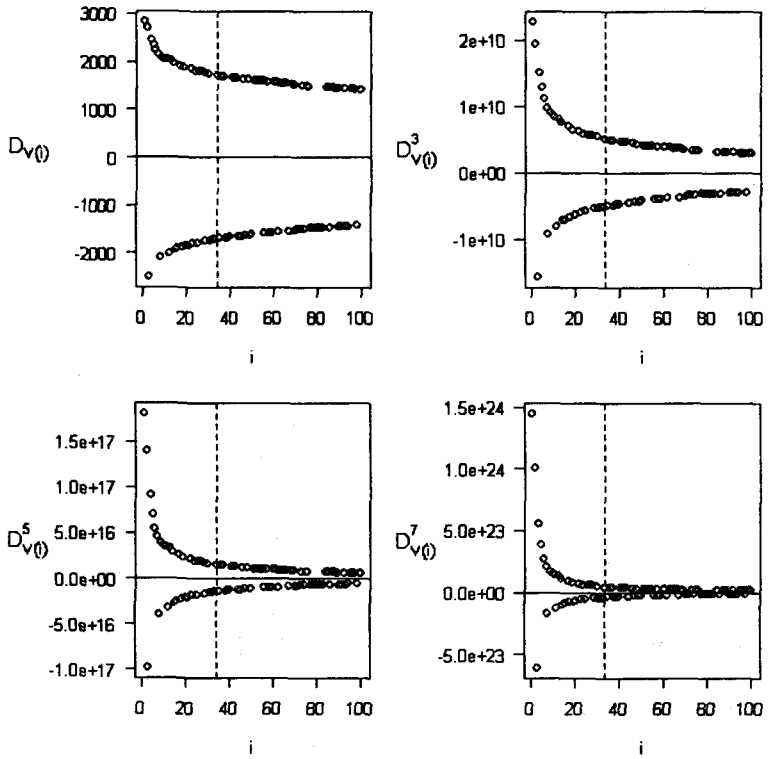
으로서 절대값이 가장 큰 g개의 발현치가 됨을 알 수 있다. 어떤 유전자에 대한 평균순위의 차, D_i, 가 양의 값을 가지면 이 유전자가 처리그룹에서 대조그룹보다 발현양이 상대적으로 감소하는 것을 알 수 있고, 반대로 음의 값을 가지면 발현량이 상대적으로 증가하는 것을 알 수 있다.

평균순위의 차에 대한 통계적 유의성에 대한 검정문제는 평균순위의 분포를 통해 근사적으로 알아볼 수 있으나, 차별적 유전자를 탐지하는 문제에서는 통계적 유의성보다는 차별성의 정도에 따라 몇 개를 선택하느냐 하는 문제가 실질적으로 더 중요한 문제가 된다. 평면좌표 상에 점 $\{(i, D_{V(i)}), i=1,2,\cdots,G\}$를 찍어보면 i=1일 때 y축 상의 수평선(y=0)의 위 또는 아래의 멀리 떨어진 곳에서 시작하여 i 값이 증가함에 따라 수평선(y=0)에 급격히 가까워지다가 더 이상은 y값에 거의 변화가 없는 수평선을 유지하게 되는 점의 형태를 볼 수 있다. 이와 같은 변화가 명확히 구분하기 어려운 경우에는 $[D_{V(i)}]^3$, $[D_{V(i)}]^5$, 또는 $[D_{V(i)}]^7$ 과 같이 홀수제곱(3,5,7,...)을 점찍어보면 명확히 구분할 수 있다. 차별적으로 발현되는 유전자의 수, g,를 결정하는 한 가지 방법은 이 평면좌표 상에서 점들이 급격히 선 y=0에 가까워지다가 변화가 거의 없게 되는 x 좌표에서 수직선을 긋고 이 선의 왼쪽에 있는 점의 수를 g로 선택할 수 있다 (<그림 1> 참조). 이것은 주성분분석에서 고유근을 크기순으로 그릴 때, 급격히 줄다가 거의 변화가 없는 점까지의 개수만큼 주성분변수를 선택하는 scree plot과 유사한 기준을 사용하게 된다.

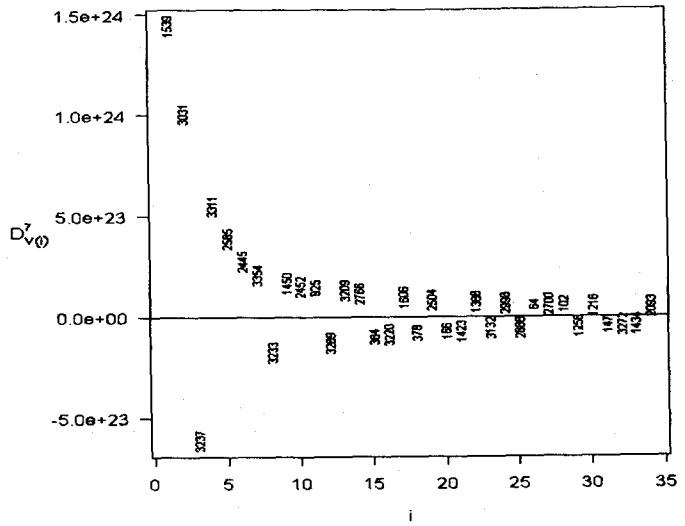
Mice2 데이터는 두 그룹(wild, mutant)에서 각각 4개씩의 표본을 얻어 총 3434개의 유전자에 대해 발현치를 얻은 데이터이다. 즉, G=3434, $l_1=l_2=4$ 인 일채널 이그룹 데이터이다. 평균순위를 이용한 데이터의 분석을 위해 R언어를 사용한 프로그램과 데이터는 http://cau.ac.kr/~cspark 에 있는 <프로그램 1>과 <Mice 2 data>와 같고 그 결과는 <그림 1>, <그림 2>와 <그림 3>에 나타나 있다.

<그림 1>에서는 100개까지의 평균순위차를 나타내고 있는데 수평선 y=0 에 점점 접근함을 알 수 있다. 평균순위차의 3제곱, 5제곱, 7제곱에 대한 점들을 보면 제곱의 차수가 커질수록 변화의 차를 구분하기 쉽게 나타난다. 수직점선은 i=34 이며 이는 전체유전자의 1%에 해당한다. <그림 2>는 그림 1의 결과를 전체 유전자수의 1%(i=1~34)에 해당하는 차별적 발현 유전자의 평균순위차를 나타내고 각 점에 해당하는 유전자번호를 대신 표시하여 해당 유전자를 쉽게 알 수 있도록 하였다. <그림 3>

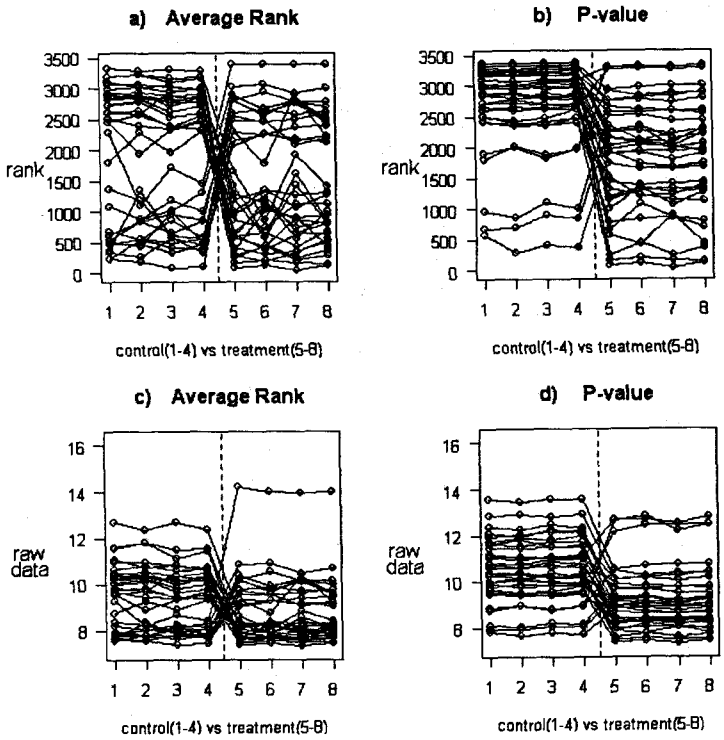
은 평균순위를 사용한 경우와 P-value 방법을 사용한 경우에 탐지된 유전자에 해당하는 평균순위와 원 발현치를 대조그룹(1~4)과 처리그룹(5~8)간에 비교할 수 있도록 나타내었다. 그림 a)는 평균순위를 사용한 경우 두 그룹의 순위를, b)는 P-value를 사용한 경우에 두 그룹의 순위를, c)는 평균순위를 사용한 경우 두 그룹의 발현치를, b)는 P-value를 사용한 경우에 두 그룹의 발현치를 나타내었다. 그림 a)에서는 평균순위를 사용하여 선택된 유전자들이 대조와 처리 그룹간의 순위가 서로



<그림 1> 차별적으로 발현된 유전자의 평균순위차 $D_{V(i)}$와 3제곱,5제곱,7제곱값: Mice2 데이터



<그림 2> 전체유전자의 1%에 해당하는 차별적 발현유전자의 평균순위차 $D^7_{V(i)}$와 해당 유전자 번호:Mice2 데이터



<그림 3> 대조-처리 그룹간 평균순위 방법과 P-value 방법에 의해 선택된 차별적 유전자의 순위와 원 데이터: Mice2 데이터

상반되게 나타나 상대적으로 다르게 발현됨을 쉽게 알 수 있다. 반면에 b)에서는 P-value 방법에 의해 선택된 대부분의 유전자에서 대조그룹에서는 큰 순위에 집중되어 있으나 처리그룹에서는 넓은 범위의 순위에 퍼져있어, P-value 방법이 순위에 있어서는 그다지 차별적이지 못함을 알 수 있다. 그림 c)에서는 원 발현치의 값이 대조와 처리 그룹간에 상반되게 나타나지만 그 범위가 전체 취하는 값의 아래 쪽 반에 밀집되어있다. 그림 d)에서는 원 발현치의 값이 역시 대조와 처리 그룹간에 상반되게 나타나지만 처리그룹에서 작은 값을 나타내는 것이 대부분임을 알 수 있다.

<표 1>에서는 P-value방법과 평균순위법을 사용할 때 선택된 유전자의 번호와 함께 각 유전자의 상대방법에서 해당하는 순위를 나타내고 있다. 두 방법에서 공통적으로 선택되어지는 유전자는 유전자 번호 앞에 * 표시가 되어 있다. 예를 들면 P-value 방법에서 유전자 1586이 가장 차별적인 유전자로 나타났으나 평균순위 방법에서 이 유전자는 1005번째로 전혀 차별적이지 못함을 보여주고 있다. 또한 평균순위 방법에서 유전자 1539는 가장 차별적이지만 P-value 방법에서 이 유전자는 47번째로 차별적이다. 이 표를 통해 보면 총 34개의 유전자중 7개만이 두 방법에서 공통적으로 선택되고 나머지 27개는 서로 다른 유전자가 선택되어짐을 알 수 있다. 또한 공통적으로 선택된 유전자들도 각 방법에서의 순서가 불규칙적으로 나열됨을 알 수 있다. 여기서 P-value방법에 의해 선택된 유전자와 더불어 평균순위에 의해 선택된 유전자를 추가적으로 고려하면 유전자 발현현상을 더 효율적으로 연구할 수 있다고 판단한다.

<표 1> P-value와 평균순위 방법에 의해 선택된 유전자의 순위와 상대방법에서 나타난 해당 유전자의 순위: Mice 2 데이터

a. P-value 방법에 의해 선택된 순서화된 유전자 번호
b. 열 a에 해당하는 유전자의 평균순위 방법에서 해당하는 순위
c. 평균순위 방법에 의해 선택된 순서화된 유전자 번호
d. 열 c에 해당하는 유전자의 P-value 방법에서 해당하는 순위

순위	a	b	c	d	순위	a	b	c	d
1	1586	1005	1539	47	18	*2585	5	378	278
2	1915	1070	*3031	7	19	547	41	2504	103
3	2518	211	3237	162	20	*3311	4	166	354
4	976	174	*3311	20	21	1838	1086	1423	342
5	1778	254	*2585	18	22	*2445	6	1388	77
6	1446	252	*2445	22	23	987	287	3132	50
7	*3031	2	3354	84	24	2714	289	2998	445
8	*2766	14	3233	238	25	1536	1230	2888	1171
9	3063	48	1450	59	26	3195	837	64	143
10	1546	99	2452	140	27	2472	272	2700	577
11	* 102	28	925	95	28	3358	72	* 102	11
12	1458	1024	3289	324	29	2110	611	1256	55
13	3012	199	*3209	14	30	996	146	1216	256
14	*3209	13	*2766	8	31	1758	1958	147	62
15	1671	333	384	61	32	603	111	3272	58
16	2810	74	3220	111	33	2784	543	1434	106
17	1454	277	1606	217	34	1409	70	2093	470

4. 일채널 다그룹 문제

유전자의 수를 G, 그룹의 수를 K(>2), j 번째 그룹에서 반복어레이의 수를 l_j 라 하면, 각 유전자의 발현수준은 $\{X_{ijk}, i=1,2,\cdots,G, j=1,2,\cdots,K, k=1,2,\cdots,l_j\}$로 표현한다. 이 때 주어진 j,k 에 대해 $X_{.jk}=(X_{1jk},X_{2jk},\cdots,X_{Gjk})$는 하나의 어레이를 형성한다. 어레이 $X_{.jk}$의 순위는 $\{R_{ijk}, i=1,2,\cdots,G\}$로 나타낸다. 그룹 j에서 i번째 유전자의 평균순위는

$$\overline{R}_{ij.}=\frac{\sum_{k=1}^{l_j} R_{ijk}}{l_j}$$

가 되고, i 번째 유전자의 총평균순위는

$$\overline{\overline{R}}_{i..}=\frac{\sum_{j=1}^{K} \overline{R}_{ij.}}{K}$$

가 된다. 이 때 평균순위차는 총평균순위에 대해

$$W_{ij}=\overline{R}_{ij.}-\overline{\overline{R}}_{i..}$$

로 정의한다. 이 때 유전자 i에 대해 K개의 그룹에서 발생하는 평균순위차의 제곱합은

$$W_i^2=\sum_{j=1}^{K} W_{ij}^2$$

이 되고 이를 크기순으로 나열하여 차별적으로 발현되는 유전자를 탐지하게 된다. W_i^2의값 크기에 따라 선택된 g개의 유전자 번호를 $V(1), V(2),\cdots, V(g)$라 하고, 이에 해당하는 유전자의 평균순위차의 제곱합은 $W_{V(1)}^2, W_{V(2)}^2,\cdots, W_{V(g)}^2$ 이라 하자. 이제 통계량 $W_{V(i)}^2$를 통해 제 3절에서와 유사한 방법으로 차별적으로 발현되는 유전자를 탐지할 수 있다.

Khan 데이터는 4개의 그룹에서 각각 23, 8, 12, 20개씩의 표본을 얻어 총 2308개의 유전자에 대해 발현치를 얻은 데이터이다. 즉, G=2308, l_1=23, l_2=8, l_3=12, l_4=20 인 일채널 다그룹(K=4) 데이터이다. 평균순위를 이용한 데이터의 분석을 위해 R언어를 사용한 프로그램과 데이터는 http://cau.ac.kr/~cspark 에 있는 <프로그램 2>와 <Khan

data>와 같고 그 결과는 <그림 4>, <그림 5>와 <그림 6>에 나타나 있다.

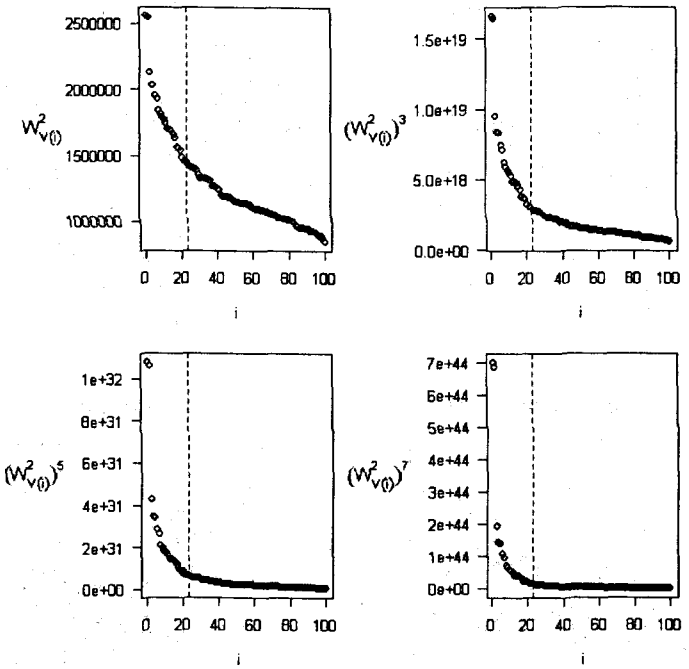
<그림 4> 에서는 평균순위차의 제곱합으로 나타내어 제 3절의 <그림 1>과는 달리 항상 양수값만을 가지게 된다. <그림 4>에서는 100개까지의 평균순위차의 제곱합을 나타내고 있는데 i가 커질수록 수평선 y=0 에 점점 접근함을 알 수 있다. 평균순위차 제곱합의 3제곱, 5제곱, 7제곱에 대한 점들을 보면 제곱의 차수가 커질수록 변화의 차를 구분하기 쉽게 나타난다. 수직점선은 i=23 이며 이는 전체유전자의 1%에 해당한다. <그림 5>에서는 유전자 187과 107이 다른 유전자에 비해 특히 차별적으로 나타남을 알 수 있다. 그림 6에서는 그룹당 어레이의 수가 둘 이상이면서 서로 다르므로 개개의 순위나 원 데이터 대신 그룹의 평균순위와 원 데이터의 평균값을 표시하였다. <그림 6>에서 4개 그룹에 대한 차별성을 보면 그룹별 평균순위 뿐만 아니라 원 데이터의 평균값도 평균순위 방법에 의해 선택된 유전자들이 P-value 방법에 의해 선택된 유전자들 보다 조금 더 차별적임을 알 수 있다. 또한 <표 2>에서는 23개의 유전자중 9개는 두 방법에서 공통적으로 선택되었으며 나머지 14개는 서로 다르게 선택되었다. <표 1>에서와 마찬가지로 공통으로 선택된 유전자들도 두 방법에서의 순서가 불규칙적으로 뒤섞여있다.

5. 이채널 이그룹 문제

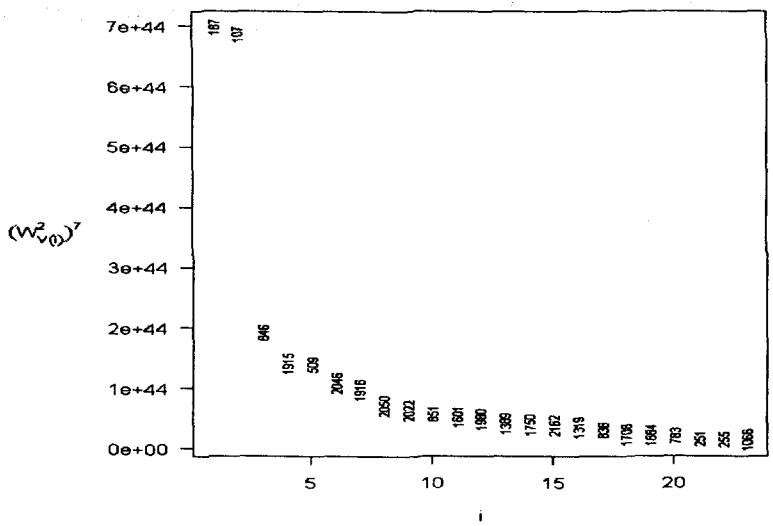
유전자의 수를 G, 첫째 그룹은 red dye, 둘째 그룹은 green dye를 사용하여 동일한 어레이에 같이 혼성화(co-hybridize)한 경우를 나타낼 때, 각 그룹에서 반복어레이의 수를 l 이라 하면, 각 유전자의 발현수준은 $\{(X_{ik}, Y_{ik}), i=1,2,\cdots,G,\ k=1,2,\cdots,l\}$로 표현한다. 이 때 주어진 k에 대해 $(X_{\cdot k}, Y_{\cdot k})=\{(X_{1k}, Y_{1k}),(X_{2k}, Y_{2k}),\cdots,(X_{Gk}, Y_{Gk})\}$는 하나의 어레이쌍을 형성한다. 주어진 어레이쌍 k에서 발현치 $\{X_{ik}, i=1,2,\cdots,G\}$ 간의 순위를 $\{R_{ik}, i=1,2,\cdots,G\}$라 하고, 발현치 $\{Y_{ik}, i=1,2,\cdots,G\}$ 간의 순위를 $\{S_{ik}, i=1,2,\cdots,G\}$라 하자. 어레이 k에서 i번째 유전자의 순위차는 $D_{ik}=R_{ik}-S_{ik}$ 로서 한 쌍의 순위 데이터끼리의 차를 나타내며, 평균순위차는

$$D_{i\cdot}=\frac{\sum_{k=1}^{l} D_{ik}}{l}$$

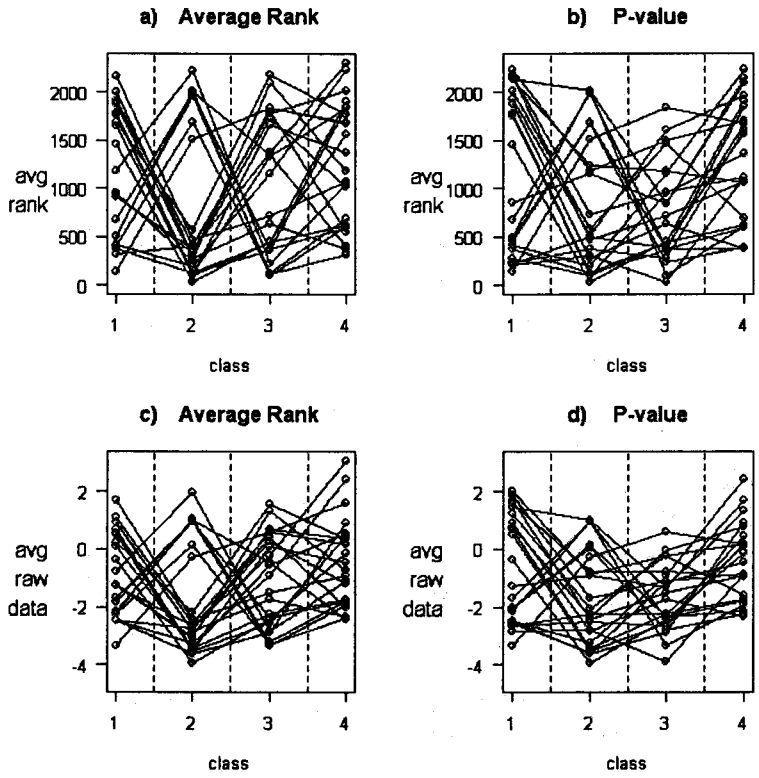
로 정의한다. 통계량 $D_{i\cdot}$을 이용하여 제 3절에서 적용한 방법을 사용하면 차별적으로 발현되는 유전자를 탐지할 수 있다.



<그림 4> 차별적으로 발현된 유전자의 평균순위차 제곱 $W^2_{V(i)}$ 와 3제곱,5제곱,7제곱값: Khan 데이터



<그림 5> 전체유전자의 1%에 해당하는 차별적 발현유전자의 평균순위차 제곱 $(W^2_{V(i)})^7$와 해당 유전자 번호:Khan 데이터



<그림 6> 그룹간 평균순위 방법과 P-value 방법에 의해 선택된 차별적 유전자의 그룹별 평균순위와 원 데이터의 평균: Khan 데이터

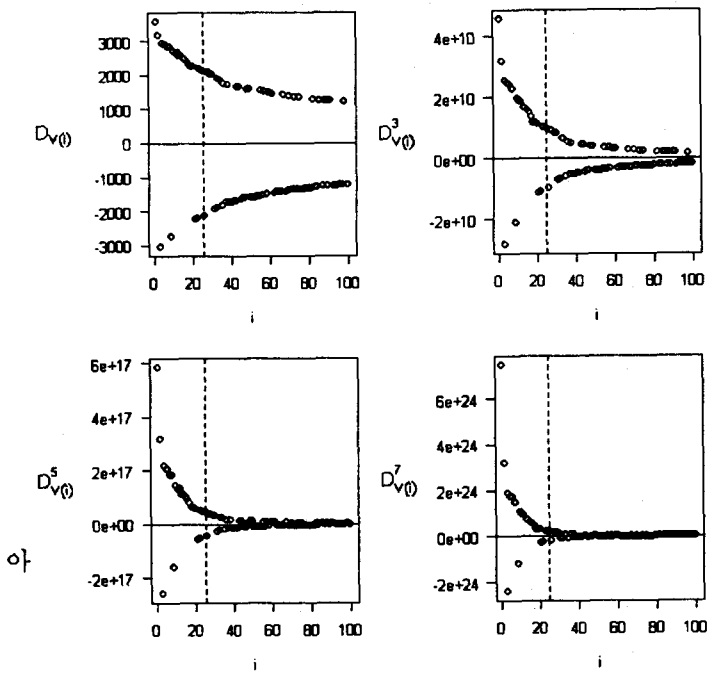
<표 2> P-value와 평균순위 방법에 의해 선택된 유전자의 순위와 상대방법에서 나타난 해당 유전자의 순위: Khan 데이터

a. P-value 방법에 의해 선택된 순서화된 유전자 번호
b. 열 a에 해당하는 유전자의 평균순위 방법에서 해당하는 순위
c. 평균순위 방법에 의해 선택된 순서화된 유전자 번호
d. 열 c에 해당하는 유전자의 P-value 방법에서 해당하는 순위

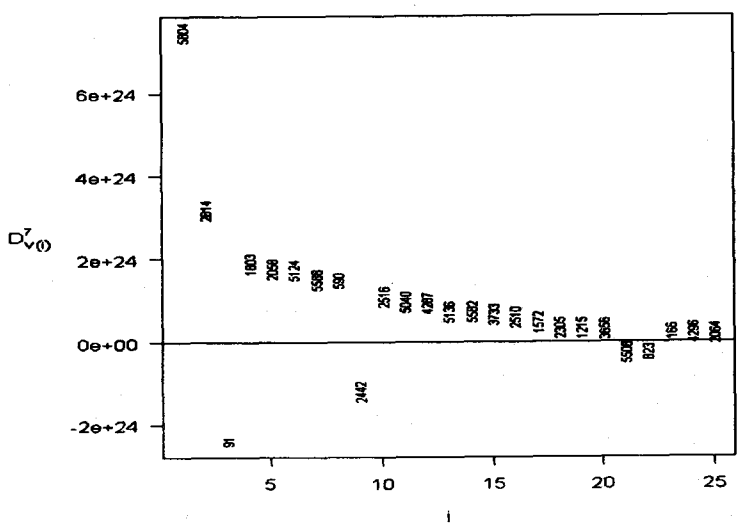
순위	a	b	c	d	순위	a	b	c	d
1	1955	56	* 187	12	13	842	44	*1389	2
2	*1389	13	* 107	11	14	1387	41	1750	187
3	1003	37	* 846	21	15	*2046	6	*2162	19
4	*2050	8	1915	89	16	1645	28	*1319	10
5	246	40	509	25	17	123	33	* 836	18
6	1954	121	*2046	15	18	* 836	17	1708	20
7	1194	70	1916	62	19	*2162	15	1884	55
8	545	174	*2050	4	20	*1708	18	783	26
9	174	135	2022	27	21	* 846	3	251	84
10	1319	16	851	28	22	1	26	255	48
11	* 107	2	1601	68	23	129	36	1066	34
12	* 187	1	1980	46					

E-coli 데이터는 wild(red dye)와 mutant(green dye)그룹에서 각각 6개씩의 표본쌍을 얻어 총 5128개의 유전자에 대해 발현치를 얻은 데이터이다. 즉 E-coli 데이터는 G=5128, l=6인 이채널 이그룹 데이터이다. 이 데이터 분석을 위한 프로그램과 데이터는 http://cau.ac.kr/~cspark 에 있는 <프로그램 3>과 <E-coli data>와 같고 그 결과는 <그림 7>, <그림 8>과 <그림 9>에 나타나 있다. 이 프로그램에서는 LOWESS 회귀분석을 사용하여 어레이내 표준화를 실시하고 있다. 표준화된 데이터에서는 정규성을 가정에 문제점을 발견하고 비모수적 방법으로 어레이쌍에 대해 Wilcoxon의 부호순위검정(signed rank test)을 실시한 결과 총 5128개의 유전자에 대한 검정중 1770(약 35%)의 검정에서 부호순위가 취할 수 있는 최대값(6(6+1)/2=21) 또는 최소값(0)이 나타나는 문제점이 발견되었다. 즉 모수적 방법은 정규성의 문제로 인해 결과를 신뢰할 수 없고 비모수적 방법으로는 차별적 유전자로 판명되어도 유의성에 차이가 나지 않는 너무 많은 유전자가 선택되는 현상이 발생하였다. 평균순위 방법에서는 이러한 문제점이 발생하지 않아 특히 유용하게 사용될 수 있음을 알 수 있다.

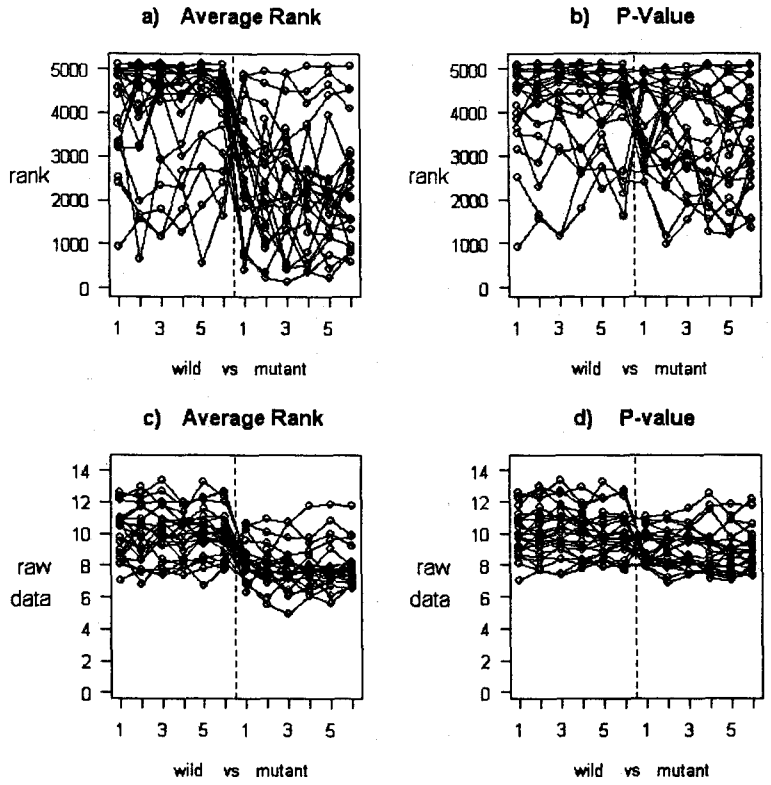
<그림 7>과 <그림 8>은 제 3절의 <그림 1>, <그림 2>와 유사한 형태를 보이고 있으며 보는 방법은 동일하다. 유전자 5804가 가장 차별적으로 나타남을 알 수 있고 대부분의 차별적 유전자에서 발현치가 크게 나타남(up-regulated)도 알 수 있다. <그림 9>에서는 두 방법에 대한 wild와 mutant 그룹의 순위와 원 데이터를 표시하고 있는데, 순위와 원 데이터 모두에서 평균순위를 사용한 방법이 P-value 방법보다 두 (wild, mutant) 그룹에서 더 차별적임을 알 수 있다.



<그림 7> 차별적으로 발현된 유전자의 평균순위차 $D_{V(i)}$와 3제곱,5제곱,7제곱값: E-coli 데이터



<그림 8> 전체유전자의 0.5%에 해당하는 차별적 발현유전자의 평균순위차 $D^7_{V(i)}$와 해당 유전자 번호: E-coli 데이터



<그림 9> 대조-처리 그룹간 평균순위 방법과 P-value 방법에 의해 선택된 차별적 유전자의 순위와 원 데이터: E-coli 데이터

<표 3> P-value와 평균순위 방법에 의해 선택된 유전자의 순위와 상대방법에서 나타난 해당 유전자의 순위: E-coli 데이터

a. P-value 방법에 의해 선택된 순서화된 유전자 번호
b. 열 a에 해당하는 유전자의 평균순위 방법에서 해당하는 순위
c. 평균순위 방법에 의해 선택된 순서화된 유전자 번호
d. 열 c에 해당하는 유전자의 P-value 방법에서 해당하는 순위

순위	a	b	c	d	순위	a	b	c	d
1	* 85	3	4867	219	14	*1932	18	4692	144
2	3701	701	2389	93	15	4581	78	*3156	20
3	3001	735	* 85	1	16	1728	3562	2130	26
4	*4234	11	1536	121	17	3015	2940	1320	115
5	3903	28	1747	30	18	1776	198	*1932	14
6	3311	27	4316	163	19	*3627	12	1043	293
7	1717	409	4698	269	20	*3156	15	3080	221
8	3258	1401	521	145	21	3346	189	4628	736
9	1424	244	*2066	13	22	*4328	13	702	146
10	*1753	25	2136	123	23	114	1049	158	40
11	3714	112	*4234	4	24	4005	256	3636	33
12	3266	128	*3627	19	25	96	2047	*1753	10
13	*2066	9	*4328	22					

6. 결론

마이크로 어레이 데이터에 대한 통계적 분석은 여러 그룹의 발현치를 그룹별로 유의한 차가 있는가를 개개의 유전자에 대해서 P-value 방법을 통해 검정하는 것이 주를 이루고 있다. 이러한 방법에서의 문제점은 유전자간의 상관성을 검토할 수 없고, 경우에 따라서는 관측값의 개수(어레이의 수)가 유의한 통계적 결론을 낼 만큼 충분하지 못하기도 할 뿐만 아니라 제 5절의 예처럼 정규성이 위배되는 경우에 사용하는 비모수적 방법에서 너무 많은 유전자가 동일한 정도의 통계적 유의성을 나타내어 검정의 의미를 상실하는 경우도 발생하게 된다. 또한 표준화 과정에서 생물적 특성이 소실될 가능성도 배제할 수 없기도 하다. 이 논문에서는 어레이 내에서 유전자들의 상대적 크기(순위)를 구한 다음 다른 어레이 내의 해당 유전자들의 순위와 비교함으로써(평균순위차) 유의한 유전자를 식별하고자 하였다. 이와 같이 평균순위를 사용하는 경우에는, 상대적 크기를 비교함으로써 유전자간의 상관성에 대한 검토를 할 수 있고, 적은 관측값의 개수에도 영향을 받지 않을 뿐만 아니라 표준화를 사용하지 않아도 되기 때문에 생물적 특성을 보존하는 통계적 분석을 할 수 있다.

제 3, 4, 5절의 예에서 나타난 결과처럼 평균순위 방법에 의한 분석의 결과는 P-value 방법의 결과와는 사뭇 다른 차별적 유전자를 선택하게 되는 것이 일반적이고, 따라서 전통적 P-value 방법에서 선택된 유전자와 함께 평균순위 방법에서 선택된 유전자를 추가하면 생물적 특성을 판단하는데 유용하게 사용될 수 있다고 판단된

다.

참고문헌

1. 박태성 등 (2005), 마이크로어레이 자료의 통계적 분석, 자유아카데미.
2. Amaratunga, D. and Cabrera, J. (2004), *Exploration and Analysis of DNA Microarray and Protein Array Data*, Wiley.
3. Augen, J. (2005), *Bioinformatics* in the Post-Genome Era, Addison Wesley.
4. Chen Y, Dougherty E.R., Bittner M.L. (1997), Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics*, Vol. 2, 364-374.
5. Cleveland, W.S. (1979), Robust Locally Weighted Regression and Smoothing Scatterplots, *Journal of the American Statistical Association*, Vol. 74, 829-836.
6. Cleveland, W.S. and Devlin, S.J. (1988), Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting, *Journal of the American Statistical Association*, Vol. 83, 596-610.
7. Deshmukh, S.R. and Puroit, S.G. (2006), *Statistical Methods in Microarray Data Analysis with R*, under publication.
8. Draghici, S. (2003), *Data Analysis Tools For DNA Microarrays*, Chapman & Hall/CRC.
9. Everitt, B.S. and Hothorn, T. (2006), *A Handbook of Statistical Analyses Using R*, Chapman & Hall/CRC.
10. Knudsen, S. (2004), *Guide to Analysis of DNA Microarray Data, 2nd edition*, Wiley-Liss.
11. Krane, D.E. and Raymer, M.L. (2003), *Fundamental Concepts of Bioinformatics*, Benjamin Cummings.
12. Murrell, P. (2006), *R graphics*, Computer Science and Data Analysis Series.
13. Newton M.A., Kendziorski C.M., Richmond C.S., Blattner F.R., Tsui K.W. (2001), On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* Vol. 8, 37-52.

[2007년 11월 접수, 2007년 11월 채택]