응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 65–80*

소규모 경시적 마이크로어레이 실험의 통계적 분석*

이근영[1] 양상화[2] 김병수[3]

요 약

소규모 경시적 마이크로어레이 실험이란 시점의 개수가 적은 경시적 마이크로어레이 실험으로서 현재까지 보고된 경시적 마이크로어레이 실험의 약 80%를 차지한다. 최근 들어 소규모 경시적 마이크로어레이 실험을 대상으로 하는 통계적 분석 방법이 몇 가지 제안되었다. 최근에 제안된 세 가지 방법들을 실제 소규모 경시적 마이크로어레이 실험자료에 적용하여 분석하고 모의실험 자료를 생성하여 각 방법들의 검정력과 위양성율을 비교해 보았다. 그 결과 낮은 위양성율을 보이는 STEM방법이 다른 방법에 비해서 우위에 있음이 드러났다.

주요용어: 소규모 경시적 마이크로어레이, QR 방법, maSigPro, STEM, 위발견율.

1. 서론

생물의 성장, 호흡, 소화, 질병 등의 수많은 생명현상은 생물체의 세포내에 존재하는 유전자들의 유기적인 상호작용의 표현이라 할 수 있다. 따라서 이러한 생물학적 현상에 대한 연구에 있어서 유전자 연관성을 연구하는 것은 필수적이라고 할 수 있다. 본 연구의 중심인 올리고뉴클레오타이드(Oligo-nucleotide) 마이크로어레이는 수천, 수만의 유전자가 점적된 작은 유리판을 이용하여 서로 다른 조건 하에서 특이적으로 발현하는 유전자를 검색할 수 있는 생물학 실험이다. 마이크로어레이 위에 우리가 관찰하고자 하는 생물 표본의 mRNA를 염료로 표지하고 점적된 올리고뉴클레오타이드와 보합시킨 후 레이저 스캐너를 통해 스캔(scan)하면 각각의 유전자마다 보합된 mRNA의 양에 따라 형광 강도(fluorescent intensity)를 나타내게 되는데 그것으로써 수많은 유전자의 발현양상을 한꺼번에 관찰할 수 있게 되었다.

일반적으로 한 시점에서의 마이크로어레이 실험보다는 시간이 경과함에 따른 유전자들의 발현양상을 연구하는 것이 동적인 유전자 발현 양상을 규정하는 데에 더욱 많은 정보를 제공한다. 마이크로어레이 실험이 본 궤도에 오르면서 실험자들은 경시적 마이크로어레이

* 본 연구는 2004년도 과학기술부의 재원으로 한국과학재단의 지원을 받아 수행된 연구임 (R01-2004-000 -10057-0).

1) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 석사 후 연구원.
E-mail: dung90@yonsei.ac.kr
2) (120-749) 서울시 서대문구 신촌동 134, 연세대학교 의과대학 암전이 연구소, 연구교수.
E-mail: ysh@yumc.yonsei.ac.kr
3) (120-749) 교신저자. 서울시 서대문구 신촌동 134, 연세대학교 응용통계학과, 교수.
E-mail: bskim@yonsei.ac.kr

실험(time-course microarray experiment)에 더욱 많은 관심을 가지게 되었다. 본 논문에서는 소규모 경시적 마이크로어레이 실험의 분석에 관해서 다루고 있다. 소규모 경시적 마이크로어레이 실험이란 경시적 마이크로어레이 실험에서 시점의 개수가 적은 경우를 말한다. 경시적 마이크로어레이 실험에서 표본을 채취하는 시점의 개수는 간혹 충분치 못한 경우가 있을 수 있다. 예를 들어 인간의 유전자를 대상으로 한 실험인 경우에 윤리적인 문제로 인해 여러 시점에서 유전자를 채취하기가 힘들고 또한 경제적인 제약으로 인해 여러 시점에서 마이크로어레이를 제작하기가 쉽지 않은 경우도 있다. Ernst 등 (2005)이 Stanford Microarray Database의 마이크로어레이 실험들을 조사한 결과 80%의 실험들이 8개 이하의 시점을 가지고 있는 것으로 나타났다. 이와 관련 본 연구에서는 소규모 경시적 마이크로어레이 실험을 시점의 개수가 8개 이하인 실험을 의미하기로 한다. 그런데 지금까지 제시된 경시적 마이크로어레이 분석 방법은 많은 수의 시점을 전제로 하고 있다. 따라서 소규모 경시적 마이크로어레이자료를 분석하는 방법은 많이 발표되지 않았고 또한 지금까지 제안된 방법들 중 어떠한 방법이 소규모 경시적 마이크로어레이 자료에 적합한지 아직 연구된 바 없다. 따라서 본 연구는 시점이 많지 않은 경우에 제안된 여러 방법들 중 어떠한 방법이 가장 좋은 성능을 보이는지에 대해서 비교 검토를 해 보고자 한다. 본 논문에서는 Liu 등 (2005), Conesa 등 (2006), 그리고 Ernst 등 (2005)이 제안한 방법들을 실제 소규모 경시적 마이크로어레이 실험자료에 적용하여 분석하고 모의실험 자료를 생성하여 각 방법들의 검정력과 위양성율을 비교해 보았다.

2. 분석방법

2.1. QR 방법 (Quadratic Regression method - QR method)

Liu 등 (2005) 은 유전자 프로파일의 양상을 인식하기 위해서 다음과 같은 이차회귀모형을 사용하는 것을 제안하였다.

$$y_{ij} = \beta_{0j} + \beta_{1j}x + \beta_{2j}x^2 + \epsilon_{ij}, \tag{2.1}$$

단, x : 시점 (연속형 변수), y_{ij} : i번째 반복표집(replicate)의 j번째 유전자의 발현치이며, 각 유전자별 반복표집의 개수는 일치하지 않을 수도 있음.

이 이차 회귀모형은 시점을 연속형 독립변수로 가정하고 시간의 흐름에 따라 다르게 발현하는 유전자 양상을 검정하기 위해 유의수준 α_0와 α_1을 사용한다. 먼저 j번째 유전자 프로파일이 특이 발현 하는지 검정하기 위해서 α_0를 사용한다. 이때의 귀무가설은 $H_0 : \beta_{1j} = \beta_{2j} = 0$이다. α_0는 다양한 다중검정에서의 p값 조정방법에 따라 선택하게 되는데 본 논문에서는 Benjamini와 Hochberg (1995)가 제안한 위발견율(FDR) 조절 절차를 이용하여 α_0를 선택하였다. 모형 (2.1)을 이용한 유전자 프로파일의 양상은 다음의 과정에 따라 결정된다.

(가) 만약 j번째 유전자 프로파일에서 모형 (2.1)의 전체 p값이 α_0보다 크다면 귀무가설이 기각되지 않는다. 즉, j번째 유전자 프로파일은 시간에 따른 특이발현을 하지 않는다고 결정한다. 이 경우 유전자 프로파일의 양상은 "수평(flat)"이다. 여기서 전체 p값은 다음과 같은 분산분석표를 구성하여 F통계량을 통해 계산된다.

표 2.1: QR 방법에서의 분산분석표

제곱합(sum of square)	해석	F 통계량
$SS(linear)$	선형효과에 의한 변동량	$\frac{SS(linear)/df1}{SS(residual)/df3}$
$SS(quadratic\|linear)$	선형효과에 의한 변동량을 제외한 이차효과에 의한 변동량	$\frac{SS(quadratic`\|`linear)/df2}{SS(residual)/df3}$
$SS(residual)$	잔차의 변동량	

표 2.1에서 $df1, df2, df3$는 각각 $SS(linear), SS(quadratic|linear), SS(residual)$의 자유도이고 이때 전체 p값은 다음의 F통계량을 통해 계산된다.

$$F = \frac{SS(linear) - SS(quadratic|linear)/(df1 + df2)}{SS(residual)/df3} \sim F_{df1+df2,df3}.$$

(나) 만약 j번째 유전자 프로파일에서 모형 (2.1)의 전체 p값이 α_0보다 작다면 귀무가설이 기각된다. 이 경우 유전자 프로파일의 양상을 분류하기 위해 다음과 같은 과정을 거친다.

a. 이차효과(β_{2j})에 대한 p값과 선형효과(β_{1j})에 대한 p값이 둘 다 α_1보다 작을 경우 j번째 유전자 프로파일은 이차항과 선형항을 모두 가지고 있는 것으로 간주한다. 이 경우 발현 양상은 "2차-1차(quadratic-linear)"이다.

b. 이차효과에 대한 p값은 α_1보다 작고 선형효과의 p값은 α_1보다 클 경우 j번째 유전자는 이차효과만 유의한 것으로 간주한다. 이 경우 발현 양상은 "2차(quadratic)"이다.

c. 이차효과에 대한 p값이 α_1보다 클 경우 j번째 유전자 프로파일은 모형(2.1)에서 이차항을 뺀 다음과 같은 모형으로 다시 재적합시킨다.

$$y_{ij} = \beta_{0j} + \beta_{1j}x + \epsilon_{ij}. \tag{2.2}$$

재적합 후의 선형효과에 대한 p값이 α_1보다 작을 경우 이 유전자 프로파일은 선형항을 가지고 있는 것으로 판단한다. 이때의 발현 양상은 "1차(linear)"이다. 만약 모형 (2.2)의 선형효과에 대한 p값이 α_1보다 클 경우 발현 양상은 "수평"이다.

앞서 분류된 양상들인 "수평", "2차–1차", "2차", "1차"를 다시 유전자 프로파일에서 처음 시점과 마지막 시점의 발현값의 차이에 따라 다시 "선형상향(inear up; LU)", "선형하향(linear down; LD)", "2차 볼록(quadratic concave; QC)", "2차 오목(quadratic convex; QV)", "2차선형 볼록상향(quadratic-linear concave up; QLCU)", "2차선형 볼록하향(quadratic-linear concave down; QLCD)", "2차선형 오목상향(quadratic-linear convex up-QLVU)", "2차선형 오목하향(quadratic-linear convex down; QLVD)", "수평"의 9가지 양상으로 분류된다. 그림 2.1은 이러한 양상의 예를 나타낸다.

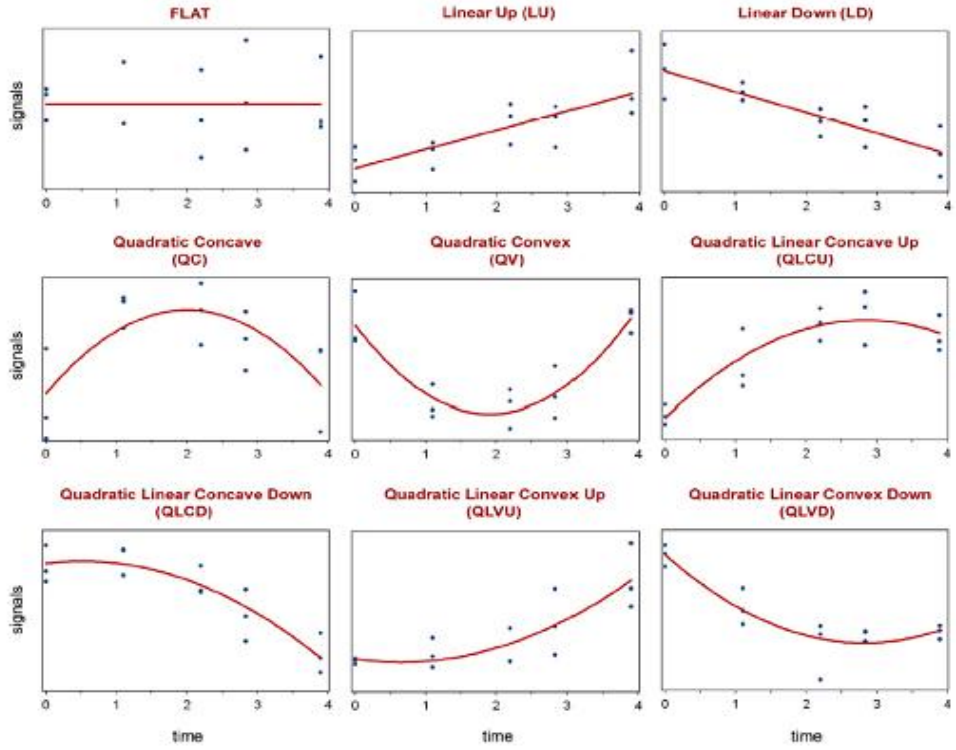

그림 2.1: QR 방법의 9가지 패턴의 예 (Liu 등, 2005, Fig. 4)

2.2. maSigPro 방법 (microarray Significant Profiles)

Conesa 등 (2006) 이 제안한 이 방법은 유전자 프로파일의 양상을 인식하기 위하여 다항 회귀모형을 사용한다. 이 모형에서는 유전자 발현에 영향을 끼치는 요소로 두 가지 변수를 고려한다. 하나는 양적 변수인 시간이고 다른 하나는 질적 변수로서 서로 다른 실험군을 나타내는 가변수이다. 우선 다음과 같은 표기를 정의하도록 하겠다.

I : 실험군의 개수,

J : 시점의 개수,

R_{ij} : i번째 실험군의 j시점에서의 반복표집의 개수,

y_{ijr} : i번째 실험군의 j번째 시점에서 r번째 반복표집의 발현치 $(r = 1, \ldots, R_{ij})$.

이때 maSigPro 방법에서는 유전자 프로파일의 양상을 인식하기 위해 다음과 같은 $J-1$차 다항 회귀모형을 사용한다.

$$\begin{aligned} y_{ijr} = {} & \beta_0 + \beta_1 D_{1ijr} + \cdots + \beta_{(I-1)} D_{(I-1)ijr} \\ & + \delta_0 T_{ijr} + \delta_1 T_{ijr} D_{1ijr} + \cdots + \delta_{(I-1)} T_{ijr} D_{(I-1)ijr} \end{aligned}$$

$$+\gamma_0 T_{ijr}^2 + \gamma_1 T_{ijr}^2 D_{1ijr} + \cdots + \gamma_{(I-1)} T_{ijr}^2 D_{(I-1)ijr}$$
$$\vdots$$
$$+\lambda_o T_{ijr}^{J-1} + \lambda_1 T_{ijr}^{J-1} D_{1ijr} + \cdots + \lambda_{(I-1)} T_{ijr}^{J-1} D_{(I-1)ijr} + \epsilon_{ijr}, \quad (2.3)$$

단, $D_{(i')ijr}$: $i'+1$번째 실험군을 나타내는 가변수 $(i' = 1, \ldots, I-1), D_{(i')ijr} = \begin{cases} 1, & i' = i, \\ 0, & i' \neq i, \end{cases}$
T : 시간 (연속형 변수), $\beta, \delta, \gamma, \ldots, \lambda$: 회귀계수, ϵ_{ijr} : 모형에 포함된 효과를 고려한 후 남은 오차항으로서 $N(0, \sigma^2)$를 따름.

대조군을 제외한 실험군이 한 個인 경시적 마이크로어레이 실험의 경우 모든 가변수 $D_{i'(ijr)}$의 값이 0이므로 다음의 모형이 적용된다.

$$y_{jr} = \beta_0 + \delta_0 T_{jr} + \gamma_0 T_{jr}^2 + \cdots + \lambda_o T_{jr}^{J-1} + \epsilon_{jr}. \quad (2.4)$$

maSigPro 방법의 실행은 다음의 절차에 따라 진행된다.

(가) 유전자 선택 단계

위의 모형 (2.3)이나 모형 (2.4)에 유전자 프로파일을 적합시켜서 p값을 계산하여 특이발현 유전자로 추측되는 후보 유전자를 찾는다. p값을 계산하는 방법은 QR 방법과 마찬가지로 표 2.2와 같은 분산분석표를 구성하여 F통계량을 이용한다. 표 2.2에서 계산된 F통계량을 기초로 각각의 유전자 프로파일마다 p값을 계산한다. 각각의 유전자 프로파일에서 계산된 p값이 유의수준보다 낮다면 해당 유전자 프로파일을 선택한다. 이때 유의수준의 설정은 QR 방법과 마찬가지로 위발견율을 사용한다.

(나) 변수 선택 단계

유전자 선택 단계에서 선택된 유전자 프로파일들의 각각에 적합시킨 모형의 회귀계수는 해당 유전자 프로파일이 시간에 따라 혹은 실험조건에 따라 발현에 차이를 보이는지를 확인하는데 사용된다. 하지만 모형 (2.3)에서는 사용된 변수들의 수가 많아

표 2.2: QR 방법에서의 분산분석표

변동요인	제곱합(SS)	자유도	F통계량
회귀	$SSR = \sum_{ijr} (\hat{y}_{ijr} - \bar{y})^2$	$df1^*$	$\frac{(SSR/p)}{SSE/[\sum_{i,j} R_{ij} - (p+1)]}$
잔차	$SSE = \sum_{ijr} (y_{ijr} - \hat{y}_{ijr})^2$	$df2^{**}$	
전체	$SST = \sum_{ijr} (y_{ijr} - \bar{y})^2$	$df3^{***}$	

$^* df1 = p$,
$^{**} df2 = \sum_{i,j} R_{i,j} - (p+1)$,
$^{***} df3 = \sum_{i,j} R_{i,j} - 1$

서 (가)단계에서 과적합의 문제가 발생하였을 우려가 있으므로 회귀모형의 변수선택 단계를 거쳐야 한다. 변수선택 방법은 전진선택법, 후진제거법, 단계별선택법 중의 하나를 선택하여 사용한다. 각 유전자 프로파일마다 변수선택 단계를 거친 후 결정계수를 이용하여 적합도 검정을 실시한다. 이러한 단계를 거쳐 최종적으로 특이발현 유전자 프로파일을 선택한다.

2.3. STEM 방법 (Short Time-series Expression Miner)

Ernst 등 (2005) 이 제안한 이 방법은 앞서 두 방법의 회귀모형을 사용하여 유전자 프로파일을 적합하여 특이발현 유전자를 검색했던 것과 달리 순열 검정에 기반을 둔 유전자 검색방법이다. 이 방법의 대략적인 실행 절차는 다음과 같다.

(가) 모형 프로파일 선택

우선 모든 유전자 프로파일을 대표할 수 있는 모형 프로파일 집합을 선택한다. 여기서 모형 프로파일은 경시적 마이크로어레이 자료의 시점의 개수와 같은 개수를 가지고 실험자가 정한 변동범위 내에서 변하는 整數로 이루어진 수열이다. 예를 들면, 시점의 개수가 6이고 변동범위가 ± 2라면 $(0, 1, -1, -1, -2, -2)$와 같은 수열이 모형 프로파일의 한 예이다. 이때 경시적 마이크로어레이 실험에서의 준거는 첫 번째 시점의 실험이라고 가정하고 모형 프로파일의 첫 번째 값은 항상 0으로 설정한다.

앞서 설정한 모형 프로파일의 정의에 따르면 시점의 개수를 n이라 하고 변동범위를 c라고 했을 때 생성가능한 모든 모형 프로파일의 개수는 $(2c+1)^{n-1}$이다. 그런데 만약 시점이 6개이고 변동범위를 2로 설정했을 경우 모든 가능한 모형 프로파일의 개수는 $5^5 = 3125$가 되어 유전자 프로파일의 개수에 비해 너무 많아질 우려가 있고 계산상의 어려움도 따른다. 따라서 모든 가능한 모형 프로파일 집합을 최대한 대표할 수 있는 m개의 집합을 선택하는 문제를 고려해야 한다.

P를 전체 모형 프로파일 집합이라고 하고 R을 P의 부분집합으로서 전체를 대표하는 m개의 모형 프로파일이라고 하자. 직관적으로 생각할 때 R에 속하는 원소 중에서 서로 가장 가까운 거리에 있는 두 모형 프로파일의 거리가 커질수록 R에 속하는 원소들이 서로 중복되지 않고 모든 유전자 프로파일을 대표할 수 있을 것이다. 이러한 개념은 다음과 같이 표현할 수 있다.

$$\arg \max_{R \subset P, |R|=m} \min_{p_1, p_2 \in R} d(p_1, p_2), \tag{2.5}$$

단, $|R|$: R에 포함된 유전자 프로파일의 개수, $d(p_1, p_2)$: P에 속하는 두 모형 프로파일 p_1, p_2간의 거리를 나타낸다.

(2.5)의 규칙을 통해서 선택된 집합을 R'라고 하고 R'에 속하는 최소의 거리를 가지는 두 모형 프로파일간의 거리를 $b(R')$라고 하자. $b(R')$는 우리가 찾아낼 수 있는 가장 이상적인 결과이다. 하지만 R'을 찾아내는 문제는 非決定難解(NP-hard)임이 증명되었다. (2.5)의 규칙에 따르면 $b(R)$은 클수록 좋은 결과임을 의미한다. 그런데 어떠한

알고리즘도 $b(R)$값의 하한(lower-bound)을 $b(R')/2$보다 크게 할 수는 없다. 즉 $b(R)$의 하한이 $b(R')/2$보다 크게 되는 R을 찾아내는 문제 또한 非決定難解이다. 결국 모형 프로파일 집합 R을 찾을 때 얻어낼 수 있는 최선의 알고리즘은 $b(R) \geq b(R')/2$을 보증하는 방법이다. 다음에 설명하고 있는 알고리즘은 $b(R) \geq b(R')/2$을 보증하는 R을 찾는 방법을 보여준다. 먼저 P에 속한 모든 모형 프로파일 중에서 가장 거리가 먼 두 개를 선택한 다음 그것을 우선 R에 포함시킨 후 다음에 포함될 모형 프로파일은 이미 포함된 것들과 가장 거리가 먼 것을 포함시킨다. p를 R에 포함시킬 모형 프로파일의 후보라고 할 때 이러한 규칙은 다음과 같이 표현할 수 있다.

$$\max_{p \in (P-R)} \min_{p_1 \in R} d(p, p_1). \tag{2.6}$$

(2.6)의 규칙에 따라 p를 포함하는 과정을 반복적으로 수행하여 최종적으로 m개의 모형 프로파일을 포함하는 집합이 선택된다.

(나) 유의한 모형 프로파일선택

(가)에서 선택한 모형 프로파일 집합을 M이라고 하고 전체의 유전자 프로파일의 집합을 G라고 하자. 우선 각각의 유전자 $g \in G$를 어느 한 모형 프로파일, 예를 들어 $m_i \in M (i = 1, \ldots, |M|)$에 배정해야 한다. e_g를 유전자 g의 프로파일이라고 할 때, e_g와 거리 $d(e_g, m_i)$이 최소가 되는 m_i에 배정한다. 여기서 $d(e_g, m_i)$는 상관계수를 이용한 거리로서 $1 - \rho(e_g, m_i)$로 계산된다. 그리고 이때 각각의 m_i에 배정된 유전자 프로파일의 개수를 $t(m_i)$라고 정의하도록 하겠다.

이제 위에서 유전자 프로파일이 배정된 m_i중에서 확률변동으로 인한 우발적 효과가 아닌 실제 특이발현 유전자들이 배정된 모형 프로파일 m_i를 찾아야 한다. 이 실험의 귀무가설은 유전자 프로파일의 어느 시점에서도 같은 값을 가진다는 것, 즉 유전자 프로파일의 발현이 시점의 변동과 무관하다는 것이다. 따라서 유전자 프로파일의 시점을 바꾸어 다시 $m_i \in M$에 배정한다고 해도 귀무가설을 따르는, 즉 특이발현 유전자가 배정되지 않은 m_i는 $t(m_i)$에 비해서 배정된 유전자 프로파일의 수가 많은 차이를 보이지 않을 것이다. 반면에 특이발현 유전자들이 배정된 m_i는 시점을 바꾸어 다시 배정했을 경우 원래 시점에서 배정했을 때의 $t(m_i)$와 많은 차이를 보일 것이다.

위에서 언급한 기본논리를 구현하기 위해 순열 검정을 사용한다. 시점의 개수를 n이라고 할 때 반복 순열의 횟수는 $n!$이다. 소규모 경시적 마이크로어레이 실험의 경우 n이 작으므로 계산시간은 큰 문제가 되지 않는다. s_i^j를 j번째 반복에서 m_i에 배정된 유전자 프로파일의 개수라고 정의하자. 그리고 $S_i = \sum_j s_i^j$라 하고 $E_i = S_i/(n!)$라고 하자. 즉, E_i는 $n!$번의 반복 순열에서 나온 s_i^j의 평균이다. 앞서 설명한 이 검정의 기본논리에 따르면 귀무가설을 따르지 않는 m_i는 $t(m_i)$가 E_i에 비해서 유의한 차이를 보이는 큰 수로 나타날 것이다. 따라서 m_i가 유의한 모형 프로파일인지에 대한 검정의 p값은 다음과 같이 계산할 수 있다. X를 $Binomial(|G|, E_i/|G|)$를 따르는 확률변수라고 정의했을 때 m_i에 대한 p값은 $P(X \geq t(m_i))$로 계산된다. 즉, $t(m_i)$가 E_i에 비해서 크면 클수록 p값이 작아질 것이다. 각각의 m_i에서 계산된 p값이 유의수준보

다 낮다면 해당 유전자 프로파일을 선택한다. 이때 유의수준의 설정은 앞선 두 가지 방법과 마찬가지로 위발견율을 사용한다.

(다) 선택된 모형 프로파일의 그룹화

위에서 설명한 과정을 거쳐 유의한 모형 프로파일이 선택이 되었고 각각의 모형 프로파일에 유전자가 배정되었다. 하지만 (가)의 과정에서 모형 프로파일을 선택할 때 서로 간의 거리가 최대한 멀어지도록 선택되었다 하더라도 유사한 두 모형 프로파일에 있는 유전자들이 원래 서로 같은 발현 프로파일을 나타내는 유전자들일 수 있다. 따라서 서로 충분히 유사하다고 생각되는 모형 프로파일들은 동일한 군집으로 그룹화 할 필요가 있다. 그러한 그룹화는 다음의 과정을 거쳐 진행된다.

우선 여기서 설명하는 그룹화의 대상은 유전자가 아니라 선택된 모형 프로파일임을 유의하자. $p_i \in P'(i = 1, \ldots, |P'|)$를 선택된 유의한 모형 프로파일집합 P'에 속한 모형 프로파일이라 하고 C_i는 p_i를 포함하는 군집이라고 하자. 최초의 군집설정은 $C_i = \{p_i\}$이다. 다음 단계에는 이미 C_i에 포함되어 있는 p_i와 가까운 거리에 있는 프로파일 $p_j(j \neq i)$를 C_i에 포함시킨다. 다시 말해서 실험자가 설정한 δ가 있을 때 모든 $p \in C_i$에 대해서 $d(p_j, p) \leq \delta$이면 p_j를 C_i에 포함시켜 최종적으로 그러한 p_j가 남아있지 않을 때까지 이러한 과정을 반복한다. 이러한 과정을 거쳐 모든 p_j에 대하여 군집을 설정한 후에 가장 많은 유전자를 포함하고 있는 군집을 선택하여 그 안에 포함되어 있는 p_i를 제외하고 남은 모형 프로파일을 대상으로 앞서 설명한 그룹화 작업을 다시 진행한다. 이러한 과정은 최종적으로 모든 p_i에 대해서 군집설정이 완료될 때 까지 진행된다.

3. 자료분석

Nuclear factor kappa B (NF-kappaB)는 전사 조절인자 (transcription regulation factor)로서 싸이토카인, 자외선, 세균과 바이러스 감염 등 세포 내외부의 다양한 자극에 의해 활성화 되는 특성을 지닌다. NF-kappaB가 비정상적으로 조절될 경우 다양한 감염성 질환, 면역기능 저해, 세포사멸등의 결과를 낳게 된다.

최근의 연구결과 NF-kappaB의 활성저해는 연골세포의 사멸을 유도할 수 있으며, 이러한 연골세포의 사멸현상은 연골세포 특이적 발현인자인 Nkx3.2의 과발현에 의해 효과적으로 억제될 수 있음이 보고된 바 있다(Park 등, 2007). 본 연구에서는 IKK (Inhibitor of KappaB Kinase)의 활성저해제인 Wedelolactone을 마우스의 연골세포에 처리하여 NF-kappaB 신호전달경로를 차단시키고 이로부터 유도되는 연골세포의 사멸 과정에서 유전자 발현의 변화 양상을 마이크로어레이 실험을 통하여 규명하고자 하였다. 대조군 연골세포 (control group)와 Nkx3.2 과발현 연골세포를 처리군으로 하여 약 38,000개의 유전자를 가지는 올리고뉴클레오타이드 마이크로어레이를 이용하여 그림 3.1의 실험설계를 경시적으로 분석하였다.

그림 3.1의 실험 디자인에서 'c'와 't'는 동일한 유전적 배경을 지니고 있는 세포주이며 'c'는 아무런 처리가 되지 않는 대조군(control group), 't'는 세포 사멸을 유도하는 물질을

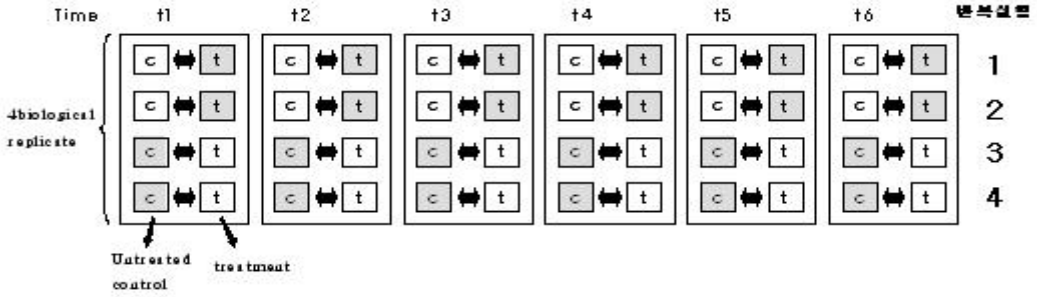

그림 3.1: 소규모 경시적 마이크로어레이 실험 디자인

처리(treatment)한 샘플을 나타낸다. 그림 3.1의 흰색 부분은 Cy3염료로 표지한 부분이고 어두운 부분은 Cy5염료로 표지한 부분으로서 염료교체(Dye-Swap)을 실시하였다. 이 실험의 목적은 세포 사멸과정에 저항적인 성향을 보이는 유전자를 알아내고 시간의 경과에 따른 양상을 분석하는 것이다.

전체 실험구성은 6개의 연속적인 관측시점(0hr, 0.5hr, 1hr, 2hr, 4hr, 6hr)에서 각 관측시점마다 4개의 생물학적 반복실험을 실행하였다. 따라서 총 24개의 마이크로어레이가 제작되었으나 이중 첫 번째 시점의 어레이중의 하나는 결측치의 비율이 70%를 초과하므로 분석에서 제외하였다. 자료의 표준화 방법은 Yang 등 (2002)의 국소회귀곡선(loess curve)을 이용한 프린트-팁(print-tip)별 표준화 방법을 사용하였고 각각의 유전자별로 결측치의 비율을 계산하여 20%가 넘는 유전자는 분석에서 제외하였다. 결측치 대체는 K최근접이웃(K-Nearest Neighbor) 방법을 사용하였고 중복유전자는 평균을 취하여 처리하였다. 최종적으로 12,576개의 유전자와 23개의 어레이로 이루어진 자료를 이용하여 세 가지 방법을 비교분석 하였다.

그림 3.1의 경시적 마이크로어레이 실험을 통하여 검정하고자 하는 귀무가설을 식(10)의 모형으로 구성하면 $H_0 : \beta_{ij} = \beta_{2j} = 0$가 되고, 식 (2.4)의 모형으로 구성하면 $H_0 : \delta_0 = \rho_0 = \cdots = \lambda_0 = 0$이 된다. 또한 STEM 방법에 의한 가설 검정은 2.3절에서 언급한 바와 같이 순열 검정을 실시하게 된다. 소규모 경시적 마이크로어레이 자료에 QR, maSigPro, STEM 방법을 각각 적용하여 그림 3.2과 같은 결과를 얻었다.

그림 3.2는 각각의 방법으로 검색한 특이발현유전자의 개수와 세 방법 간 검색 유전자의 포함관계를 나타낸다. 세 가지 방법에서 FDR은 모두 0.05로 설정하였다. 그림 3.2에서 특이한 점은 QR 방법과 maSigPro 방법 모두 유전자 프로파일의 양상을 인식하기 위해서 선형모형을 사용하고 있고 QR 방법이 오히려 더 간단한 이차 회귀모형을 사용하고 있음에도 불구하고 QR 방법이 maSigpro 방법에서 검색하지 못한 유전자를 많이 찾아냈다는 점이다. maSigPro 방법은 QR 방법이 검색한 유전자의 약 42%를 찾아내었으나 이차회귀 방법은 maSigPro 방법이 검색한 유전자의 60%를 찾아내었다. 직관적으로는 maSigPro 방법이 시점의 개수를 t라고 했을 때 $t-1$차 다항 회귀모형을 사용하므로 더욱 다양한 양상을 검색하는 것이 가능할 것이라고 생각할 수 있으나 유전자 개별 검정과정의 p값 계산에 사용

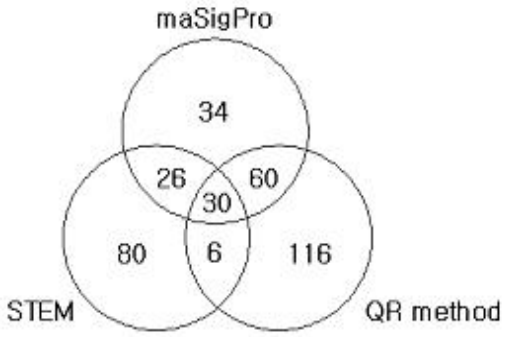

그림 3.2: 세 가지 방법에서 검색한 특이발현 유전자의 개수와 포함관계 (FDR = 0.05)

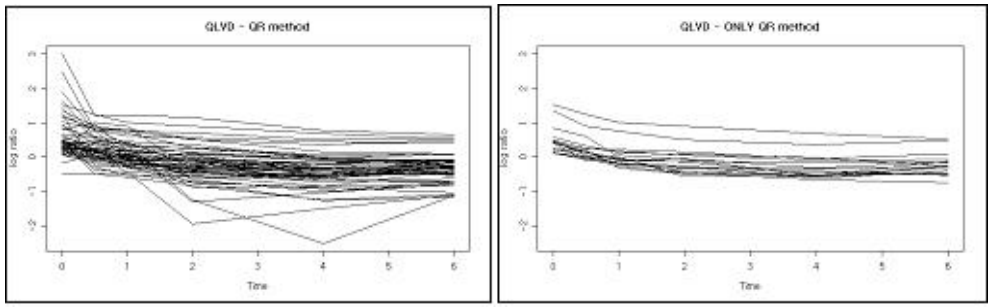

그림 3.3: QR 방법에서 검색한 유전자(왼쪽)와 그중에서 maSigPro 방법으로 검색한 유전자를 제외한 유전자의 프로파일 양상(오른쪽) 그림

되는 F통계량을 도출하는 과정에서 maSigPro 방법은 모형에 많은 변수가 투입됨으로 인하여 잔차의 자유도가 QR 방법보다 떨어진다. 즉, 유전자 프로파일의 양상이 비교적 단순한 경우 모형의 복잡성으로 인한 자유도의 손실로 인하여 p값이 QR 방법에 비하여 크게 계산되어 해당 유전자는 maSigPro 방법에서는 특이발현유전자가 아니라고 판명될 가능성이 있다. 따라서 유전자 프로파일이 시간에 따른 복잡한 양상을 보이는 것이 아니라 이차형식으로 충분히 표현 가능한 양상을 보이는 경우에는 QR 방법이 maSigPro 방법에서 유의하지 않다고 판단한 유전자들도 유의하다고 결정할 개연성이 있다는 것이다. 그림 3.3은 그러한 예를 보여준다.

그림 3.3의 왼쪽에 위치한 그림은 QR 방법으로 검색한 유전자 중에서 "이차선형 오목하향"양상으로 규정된 유전자 프로파일 그림이고 오른쪽에 위치한 그림은 그러한 유전자 중에서 maSigPro 방법으로 검색된 유전자를 제외한 후에 다시 구성된 유전자 프로파일 그림이다. 왼쪽 그림에서 비교적 복잡한 양상을 보이는 프로파일들은 maSigPro 방법에서도 검색되어 오른쪽 그림에서는 사라지고 남은 유전자 프로파일은 "이차선형 오목하향"양상

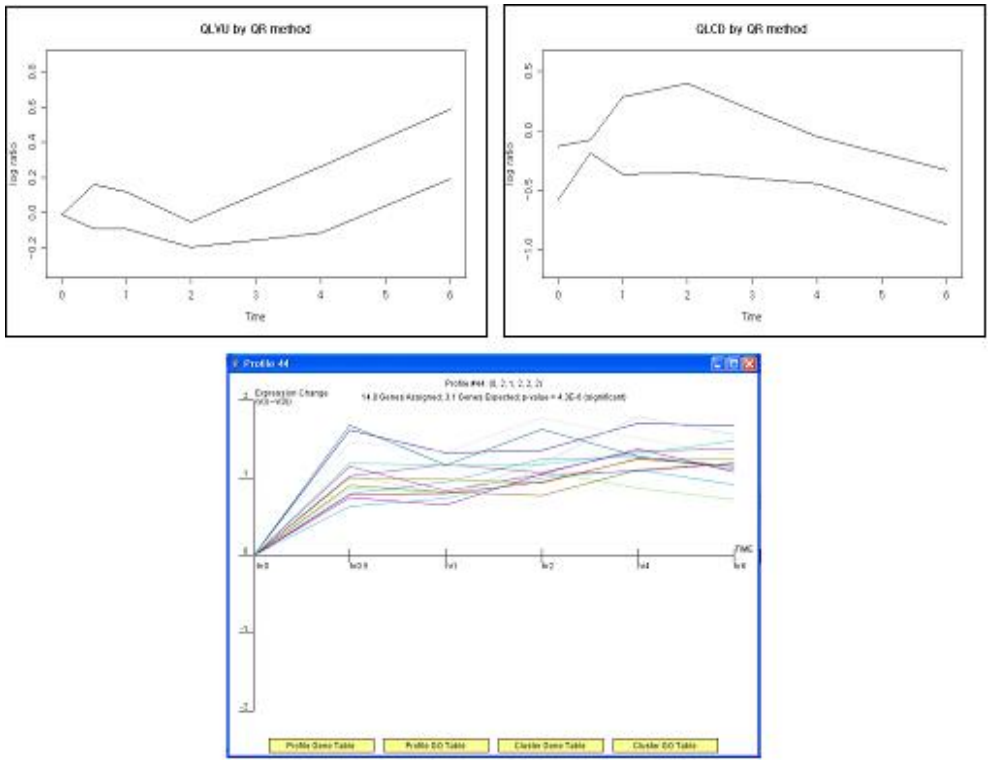

그림 3.4: QR 방법과 STEM 방법에서 검색된 유전자 프로파일 양상의 예 (STEM 방법에서는 소수의 양상이 나타나지 않는다.)

중에서도 완만한 굴곡을 지닌 이차식 형태의 양상을 보여준다. 이 유전자들은 maSigPro 방법에서는 다항회귀모형(3차 이상)으로 적합하였기 때문에 QR 방법에 비하여 과적합 문제로 인해 검색되지 못한 것이다. 그러나 오른쪽 그림에서의 유전자 프로파일이 실제 특이발현 유전자인지 알 수는 없기 때문에 QR 방법이 maSigPro 방법에 비해서 검정력이 좋다고 결론지을 수는 없다.

STEM 방법이 다른 두 방법과 비교하여 보이는 특징은 소수의 유전자 발현양상이 검색되지 않는다는 것이다. 그림 3.4는 STEM 방법의 이러한 특징을 보여준다. 그림 3.4에서 위에 위치한 그림은 QR 방법에서 검색된 유전자 양상 중에서 "이차선형 오목상향" 양상과 "이차선형 볼록하향" 양상으로 규정된 유전자 프로파일 그림으로서 각각 2개씩의 유전자를 포함하고 있다. 반면에 아래에 있는 그림은 STEM 방법에서 유의하다고 판명된 모형 프로파일 중에서 가장 적은 수의 유전자를 가지고 있는 것이지만 14개의 비교적 많은 수의 유전자를 포함하고 있다. 게다가 이 모형 프로파일은 이후에 그룹화 과정을 거쳐 다른 모형 프로파일과 묶여서 총 34개의 유전자를 포함하는 군집에 속하게 된다. 이러한 결과의 원인은 STEM 방법에서 유의한 모형 프로파일을 검정할 때 p값을 계산함에 있어서 순열 검정을 사용하기 때문이다. 즉, 모형 프로파일에 배정된 $t(m_i)$가 반복 순열로 인해 도출된 s_i^j의

분포와 비교하여 크면 클수록 p값이 작아지기 때문에 $t(m_i)$가 적은 모형 프로파일은 유의하다고 판명되기 힘들다. 이것은 QR 방법과 maSigPro 방법이 개별 유전자에 대해서 검정하는 것과 달리 모형 프로파일을 설정하고 그것에 대한 검정을 하기 때문에 발생하는 현상이라고 할 수 있다. 이러한 점 때문에 STEM 방법은 확률변동에 의한 잘못된 검정결과, 즉 위양성율을 줄일 수 있다는 장점과 더불어 어떠한 유전자가 뚜렷하게 특이 발현하는 프로파일 양상을 보인다 하더라도 그와 유사한 프로파일 양상을 가지고 있는 다른 유전자의 수가 적을 경우 검색이 힘들다는 단점을 동시에 가지고 있다.

지금까지 실제 자료를 통하여 세 가지 방법을 분석한 결과를 살펴보았는데 최종적으로 QR 방법은 212개, maSigPro 방법은 150개 그리고 STEM 방법은 142개의 유전자를 검색하였다. 하지만 실제로 어떠한 유전자가 특이발현 유전자인지 알 수 없는 상태이므로 유의하다고 판명된 유전자 중에서 실제로 유의한 유전자의 비율, 즉 검정력과 유의하다고 판명된 유전자 중에서 실제로 유의하지 않은 유전자의 비율, 즉 위양성율을 알 수 없다. 따라서 모의실험 자료를 생성하고 그것을 통해서 세 방법의 검정력과 위양성율을 계산하여 세 가지 방법을 비교할 필요가 있다.

4. 모의실험 자료 분석

모의실험 자료는 다음과 같이 생성하였다. 앞서 설명한 세 가지 방법으로 실제 자료에서 검색한 유의한 유전자의 개수는 그림 3.2에서 나타나듯이 총 352개이다. 이 유의한 352개의 유전자 집합을 G라고 하자. G를 바탕으로 하여 모의실험 자료 중 유의한 유전자 집합을 다음과 같이 생성하였다. G를 대상으로 K평균 군집분석($K = 10$)을 한 뒤 각 군집에서 평균벡터($m_i, i = 1, \ldots, 10$)와 공분산행렬($S_i, i = 1, \ldots, 10$)을 계산하였다. 가령 식 (4.1)은 실제 모의실험 자료에서 사용한 평균벡터와 공분산 행렬의 예이다.

$$m_8 = (1.001, 0.238, 0.229, 0.187, 0.085, 0.0008)^T, \tag{4.1}$$

$$S_8 = \begin{pmatrix} 0.0904 & 0.0062 & 0.0002 & 0.0017 & -0.0069 & -0.0017 \\ 0.0062 & 0.0497 & 0.0288 & 0.0182 & 0.0002 & -0.0016 \\ 0.0002 & 0.0288 & 0.0369 & 0.0195 & -0.0011 & -0.0046 \\ 0.0017 & 0.0182 & 0.0195 & 0.0215 & 0.0090 & 0.0051 \\ -0.0069 & 0.0002 & -0.0011 & 0.0090 & 0.0313 & 0.0223 \\ -0.0017 & -0.0016 & -0.0046 & 0.0051 & 0.0223 & 0.0338 \end{pmatrix},$$

단, X^T는 X의 전치 행렬을 나타낸다.

K평균 군집 분석은 G에서 확실한 양상을 구분한 뒤 평균벡터와 공분산 행렬을 계산해야 실제로 유의한 프로파일 양상을 나타낼 수 있기 때문에 시행하였고 K는 실제 자료에서 G의 양상을 뚜렷하게 구분하는 군집의 개수로서 적당하다고 판단되는 10으로 정하였다. 이를 통해서 각각의 m_i와 S_i를 모수로 하는 다변량 정규분포로부터 유의한 유전자 집합을 생성하였다. 모의실험 자료의 유의하지 않은 유전자 집합은 실제 자료에서 유의한 유전자 집합 G를 제외한 나머지로부터 평균벡터와 공분산 행렬을 계산하고 그것을 모수로 하는 다

표 4.1: 유의한 유전자 비율과 전체 유전자의 개수를 달리한 자료로부터 도출된 세 가지 방법의 검정력과 위양성율

		검정력			검정력		
		1%자료*	5%자료*	10%자료*	1%자료*	5%자료*	10%자료*
3520 자료	QR 방법	0.314 (11/35)	0.432 (76/176)	0.452 (159/352)	0.5 (11/22)	0.191 (18/94)	0.243 (51/210)
	maSigPro 방법	0.229 (8/35)	0.449 (79/176)	0.460 (162/352)	0.273 (3/11)	0.112 (10/89)	0.090 (16/178)
	STEM 방법	0.371 (13/35)	0.381 (67/176)	0.406 (143/352)	0.188 (3/16)	0.043 (3/70)	0.007 (1/144)
7040 방법	QR 방법	0.214 (15/70)	0.375 (132/352)	0.442 (311/704)	0.559 (19/34)	0.309 (59/191)	0.192 (74/385)
	maSigPro 방법	0.271 (19/70)	0.412 (145/352)	0.474 (334/704)	0.269 (7/26)	0.116 (19/164)	0.080 (29/363)
	STEM 방법	0.4 (28/70)	0.406 (143/352)	0.446 (314/704)	0.067 (2/30)	0.021 (3/146)	0.019 (6/320)
35200 자료	QR 방법	0.253 (89/352)	0.407 (716/1760)	0.471 (1659/3520)	0.480 (82/171)	0.256 (246/962)	0.199 (413/2072)
	maSigPro 방법	0.261 (92/352)	0.413 (726/1760)	0.461 (1621/3520)	0.118 (27/229)	0.132 (110/836)	0.087 (154/1775)
	STEM 방법	0.406 (143/352)	0.446 (785/1760)	0.442 (1556/3520)	0.083 (13/156)	0.021 (17/802)	0.011 (17/1573)

* 1% 자료, 5% 자료, 10% 자료 : 각각 1%, 5%, 10%의 유의한 유전자를 포함한 모의실험자료를 의미한다.

변량 정규분포로부터 생성하였다. 유의한 유전자의 비율에 따라 세 가지 방법의 결과가 어떤 차이를 보이는지 알아보기 위하여 1%, 5%, 10%의 유의한 유전자를 가지는 모의실험 자료를 생성하였고 각각 1%, 5%, 10%자료라고 언급하기로 한다. 세 자료 각각에 대해서 유전자의 개수는 3,520개, 7,040개, 35,200개의 세 경우를 고려하였다. 유전자 개수에 따라 변하는 자료를 각각 3520, 7040, 35200 자료라고 언급하기로 한다.

표 4.1은 유의한 유전자의 비율과 전체 유전자 개수를 달리한 모의실험 자료에 적용된 세 가지 방법의 검정력과 위양성율을 나타낸다. 검정력은 1% 자료의 경우 유전자의 개수와 상관없이 QR 방법과 maSigPro 방법의 검정력이 STEM 방법에 비해서 떨어진다. 유전자 개수가 3,520개이고 유의한 유전자의 비율이 각각 5%, 10%인 자료에서 STEM 방법의 검정력은 다른 두 가지 방법에 비해 약간 떨어지나 그 외 전체적으로 검정력은 세 가지 방법 모두 비슷한 수준을 보여주고 있다.

표 4.1에서 눈에 띄는 것은 QR 방법의 위양성율이 다른 방법에 비하여 월등히 높다는 것이다. 앞서 QR 방법이 maSigPro 방법에 비해서 단순한 유전자 프로파일 양상에 대해서 쉽게 유의하다고 결정하는 경향이 있음을 언급했다. 하지만 그로 인해서 실제로 유의한 유

전자뿐만 아니라 유의하지 않은 유전자도 다른 방법에 비해서 더 많이 검색된다는 것을 알 수 있다. 그리고 또 하나의 중요한 결과는 STEM방법이 다른 방법에 비해서 위양성율이 월등히 낮다는 점이다. 표 4.1의 모든 모의실험 자료에서 STEM의 위양성율이 가장 낮은데 이것은 앞서 언급한 STEM방법의 장점, 즉 확률변동으로 인한 개별 유전자 검정에서의 잘못된 귀무가설 기각의 우려가 없다는 점을 보여주는 결과라고 할 수 있다.

표 4.1의 결과를 토대로 정리하자면 유의한 유전자의 비율이 낮을 경우(1%)에는 STEM 방법이 다른 두 가지 방법보다 높은 검정력을 보여주었고 전체 유전자의 개수가 3,520개이고 유의한 유전자의 비율이 5%, 10%인 경우에는 STEM 방법의 검정력이 약간 떨어지지만 그 외의 모의실험 자료에서는 세 가지 방법이 전체적으로 비슷한 수준의 검정력을 가진 것으로 드러났다. 하지만 위양성율의 관점에서는 STEM 방법이 일관적으로 가장 좋은 결과를 보여주었다. 따라서 다른 두 가지 방법에 비해서 STEM 방법의 성능이 비교우위에 있다고 할 수 있다.

5. 결론 및 토의

본 연구의 목적은 앞서 언급한 세 가지의 방법 중에서 소규모 경시적 마이크로어레이 실험에 가장 적합한 것이 무엇인지 비교하는 것이었다. 실제 자료를 통하여 세 가지 방법의 유전자 검색에 있어서의 특징을 알아보았고 유의한 유전자의 비율과 전체 자료의 개수를 달리 한 모의실험 자료를 통해서 검정력과 위양성율을 비교해 보았다. 모든 모의실험 자료에서 일관적으로 우월한 검정력을 보이는 방법은 없었지만 위양성율의 관점에서 STEM 방법이 가장 우위에 있음을 밝혀냈다.

그러나, 세 가지 방법에 의하여 선택된 유전자 수가 동일하다고 가정해도 의미가 서로 다른 유전자가 선별된다면 이러한 현상도 통계적 방법에 따른 생물학적 해석이 달라질 수 있게 되는 것을 의미하게 된다. 이를 확인하기 위해 본 연구에서 선별한 세 가지 유전자군에 대한 생물학적 해석을 Onto-Express (Drăghichi 등, 2003)에 의한 pathway analysis를 시도 하였다. 동 분석의 결과 세 가지 분석 방법에 의하여 선별된 유전자군은 상당수 공통적인 분자 경로를 나타내고 있음을 알 수 있었다. 따라서 세 가지 통계적 방법에 의하여 검색된 특이 발현 유전자군간 생물학적 의미는 커다란 차이가 없는 것으로 결론을 내린다.

Peddada 등 (2003)은 순서통계량을 바탕으로 붓스트랩 표본추출 방법을 통한 검정방법을 이용하여 유의한 유전자 프로파일을 검색하고 군집분석하는 방법을 제안하였는데 유전자들이 배정될 프로파일을 미리 정의한다는 점에서 STEM방법과 유사한 면이 있다. Peddada의 방법을 포함하여 본 연구에서 비교한 소규모 경시적 마이크로어레이 분석 방법들과 다른 여러 일반적인 마이크로어레이 분석 방법들을 소규모 경시적 마이크로어레이 자료에 적용시켜 비교하는 것 또한 추후 연구과제로 남겨 놓는다.

참고문헌

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society*, Ser. B, **57**, 289–300.

Conesa, A., Nueda, M. J., Ferrer, A. and Talón, M. (2006). maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments, *Bioinformatics,* **22**, 1096–1102.

Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C. and Krawetz, S. A. (2003). Global functional profiling of gene expression, *Genomics*, **81**, 98–104.

Ernst, J., Nau, G. J. and Bar-Joseph, Z. (2005). Clustering short time series gene expression data, *Bioinformatics,* **21**, i159–i168.

Liu, H., Tarima, S., Borders, A. S., Getchell, T. V., Getchell, M. L. and Stromberg, A. J. (2005). Quadratic regression analysis for gene discovery and pattern recognition for non-cyclic short time-course microarray experiments, *BMC Bioinformatics,* **6,** 106.

Park, M., Yong, Y., Choi, S. W., Kim, J. H., Lee, J. E. and Kim, D. W. (2007). Constitutive RelA activation mediated by Nkx3.2 controls chondrocyte viability, *Nature Cell Biology,* **9**, 287–298.

Peddada, S. D., Lobenhofer, E. K., Li, L., Afshari, C. A., Weinberg, C. R. and Umbach, D. M. (2003). Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference, *Bioinformatics,* **19**, 834–841.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J. and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation, *Nucleic Acids Research,* **30**, e15.

[2007년 7월 접수, 2007년 11월 채택]

Statistical Analysis of a Small Scale Time-Course Microarray Experiment*

Keun-Young Lee[1] Sang-Hwa Yang[2] Byung-Soo Kim[3]

Abstract

Small scale time-course microarray experiments are those which have a small number of time points. They comprise about 80 percent of all time-course microarray experiments conducted up to 2005. Several statistical methods for the small scale time-course microarray experiments have been proposed. In this paper we applied three methods, namely, QR method, maSigPro method and STEM, to a real time-course microarray experiment which had six time points. We compared the performance of these three methods based on a simulation study and concluded that STEM outperformed, in general, in terms of power when the FDR was set to be 5%.

Keywords: Small scale time-course microarray, quadratic regression method, maSigPro, STEM, false discovery rate.

* This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korea government(MOST) (No. R01-2004-000-10057-0).

1) Post-Master Researcher, Dept. of Applied Statistics, Yonsei University, 134 Shinchon-Dong, Seodaemoon-Gu, Seoul 120-749, Korea.
E-mail: dung90@yonsei.ac.kr

2) Research professor, Cancer Metastasis Research Center, College of Medicine, Yonsei University, 134 Shinchon-Dong, Seodaemoon-Gu, Seoul 120-749, Korea.
E-mail: ysh@yumc.yonsei.ac.kr

3) Corresponding author. Professor, Dept. of Applied Statistics, Yonsei University, 134 Shinchon -Dong, Seodaemoon-Gu, Seoul 120-749, Korea.
E-mail: bskim@yonsei.ac.kr