응용통계연구 제 18권 2호, 2005년, pp. 371-379

시간경로 유전자 발현자료의 군집분석에서 이질적인 시계열의 탐지를 위한 패턴일치지수*

손영숙 [1] 백장선 [2]

요 약

본 논문에서는 피어슨 상관계수를 이용한 시간경로 유전자 발현자료의 군집분석에서 군집의 대표적인 패턴에서 벗어나는 이질적인 패턴을 보이는 시계열을 탐지하기 위한 패턴일치지수를 제안하고, 이를 마이크로어레이 실험으로 부터 얻어진 혈청 시간경로 유전자 발현자료에 적용하여 유용성을 검토해 본다.

주요용어: 마이크로어레이, 시간경로 유전자 발현자료, 피어슨 상관계수, 패턴일치지수, 계층적 군집분석

1. 서 론

마이크로어레이(microarray) 기술은 대용량 유전체 분석 시스템으로서 수천 혹은 수 만개 유전자의 발현수준을 동시에 관측실험 할 수 있도록 한 생명공학의 가장 강력하고 획기적인 수단으로서 받아들여지고 있다. 실험대상인 생명체의 유전자들을 고밀도로 심어 놓은 마이크로어레이(광학현미경 슬라이드)에 연구대상인 특정 암, 특정 병원균, 특정조직, 혹은 특정세포 등으로부터 얻은 cDNA 타겟(target)을 혼합(hybridization)시키면, 이들이 서로 결합하여 화학반응을 일으켜 녹색과 빨간색의 색상변화로 나타난 발현(expression) 현상을 화상(image) 자료분석하여 수치화된 마이크로어레이 자료를 얻게 된다. 실제 마이크로어레이 자료분석에서는 보통 빨간색과 녹색발현 값의 로그비(log ratio)로 변환된 자료가 사용된다.

마이크로어레이 유전자 실험이 시간의 흐름에 따라 연속적으로 수행되면서 얻어지는 시계열자료를 마이크로어레이 시간경로(time course) 유전자 발현자료(gene expression data)라고 한다. 많은 생물학적 시스템은 동적(dynamic) 시스템이기 때문에 시간경로자료의 분석은 주어진 생물학적 과정이 전개되면서 유전자의 발현수준이 시간에 따라 어떻게 변화하는지를 파악할 수 있게 한다. 또한, 비슷하게 발현되는 시간경로를 가지는 유전자들의 군집화(clustering)는 어떤 유전자들이 같은 생물학적 과정에 종속되는가에 관한 정보를 줄 수 있을 뿐만 아니라 유전자들의 알려지지 않은 기능들을 예측가능하게 하여 같은 메카니즘에 의해 규제되는 유전자들을 식별케 하는 데 도움을 준다.

* 본 연구는 산업자원부 지방기술혁신사업(RTI04-03-03) 지원으로 수행되었음.

1) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수
E-mail: ysson@chonnam.ac.kr

2) (500-757) 광주광역시 북구 용봉동 300, 전남대학교 통계학과, 교수
E-mail: jbaek@chonnam.ac.kr

Hoon 등 (2002), Peddada 등 (2003), Schliep (2003), 그리고 Luan 과 Li (2003)에서는 마이크로어레이 시간경로 유전자 발현자료의 군집화를 위하여 선형 스플라인함수모형, 순서제약추론법, 은닉마코브모형(hidden Markov model), 그리고 B-스플라인함수를 설명변수로 가지는 혼합모형과 같은 고급의 확률모형의 추론에 기초하고 있다. 이에 반해 시간경로자료의 군집화를 위하여 일반적으로 아주 간단히 그리고 자동적으로 사용되어지는 유사성측도로서 피어슨 상관계수 (Pearson correlation coefficient)가 있다.

시간경로자료는 시간의 흐름에 따른 시계열의 진행방향과 변화량이 비슷하거나, 시계열의 상승(up)-하강(down)-정체(duration)의 흐름이 같은, 혹은 최소 및 최대발현값을 주는 시점이 같은 유전자들을 동일한 군집으로 본다. (Peddada 등 (2003), Schliep (2003), Luan 과 Li (2003)의 자료분석 결과 참조). 이런 관점에서 볼 때, 시계열의 패턴과는 무관하게 오직 거리가 가까운 유전자들을 같은 군집으로 소속시키는 유클리디안 거리(Euclidean distance), 마하라노비스 거리(Mahalanobis distance), 민코우스키 거리(Minkowski distance) 보다는 상관계수 거리(correlation distance)가 마이크로어레이 시간경로자료의 군집화를 위한 측도로서 더 적절하다. 그러나 Peddada 등 (2003, pp835, Fig.1)의 간단한 예에서 보듯이 매우 적은 시점을 가지는 시간경로자료의 연관성측도로서 상관계수는 완전하지 않다. 즉, 서로 다른 패턴을 가지는 두 유전자들의 상관계수가 같은 패턴을 가지는 상관계수보다 더 높게 나타날 수 있다. 즉, 두 시계열이 소수개의 시점들에서는 기울기의 방향이 다를지라도 대부분의 시점에서 기울기가 비슷하다면 피어슨 상관계수는 높지만 한 두 시점에서 시계열의 방향이 바뀌어져 시계열의 패턴은 다를 수 있다.

본 논문의 제 2장에서는 시계열의 상승-하강-정체 패턴, 그리고 최소 및 최대발현값을 주는 시점의 일치 정도를 수량화한 패턴일치지수(pattern consistency index)를 제안한다. 제 3장에서는 인간의 혈청(serum)에 대한 시간경로 유전자 발현자료의 군집분석에서 제안된 패턴일치지수를 이용하여 군집내 이질적인 패턴을 보이는 시계열을 탐지할 수 있음을 보였다.

2. 패턴일치지수

p개의 유전자들 중 유전자 $i(i = 1, 2, \ldots, p)$의 n개의 시점 $t_1, t_2, \ldots, t_n$ 에서의 시간경로자료를 $x_{i,t_1}, x_{i,t_2}, \ldots, x_{i,t_n}$이라 놓자. 그러면 유전자 i 와 유전자 j의 피어슨 상관계수(Pearson correlation coefficient)는 다음과 같이 정의된다.

$$R_{i,j} = \frac{\sum_{k=1}^{n}(x_{i,t_k} - \bar{x}_i)(x_{j,t_k} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{i,t_k} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{j,t_k} - \bar{x}_j)^2}}, \quad i, j = 1, 2, \ldots, p, \; i \neq j,$$

여기서 $\bar{x}_i = \frac{1}{n}\sum_{k=1}^{n} x_{i,t_k}$, $\bar{x}_j = \frac{1}{n}\sum_{k=1}^{n} x_{j,t_k}$, 그리고 $-1 \leq R_{i,j} \leq 1$.

유전자 i의 시간경로점 (t_k, x_{i,t_k})과 인접하는 점 $(t_{k+1}, x_{i,t_{k+1}})$을 통과하는 직선의 기울기를

$$slope(i, t_k, t_{k+1}) = \frac{x_{i,t_{k+1}} - x_{i,t_k}}{t_{k+1} - t_k},$$

그리고 직선의 상승-하강-정체의 정보를 가지는 함수를

$$L_{i,t_k,t_{k+1}} = \begin{cases} 1, & slope(i,t_k,t_{k+1}) > 0, \\ -1, & slope(i,t_k,t_{k+1}) < 0, \\ 0, & slope(i,t_k,t_{k+1}) = 0 \end{cases}$$

과 같이 나타내자. 이제, 유전자 i와 유전자 j의 상승-하강-정체 패턴의 일치도를 다음과 같이 정의 한다.

$$A_{i,j} = \frac{1}{n-1}\sum_{k=1}^{n-1} I(L_{i,t_k,t_{k+1}} = L_{j,t_k,t_{k+1}}),$$

여기서 $I(D)$는 사상 D가 참이면 1의 값을, 거짓이면 0의 값을 가지는 지시함수(indicator function)이고, $0 \le A_{ij} \le 1$ 이다.

다음으로 유전자 i의 최소발현값 시점 및 최대발현값 시점을 각각 T_i^{min} 그리고 T_i^{max}라 놓고, 유전자 i와 유전자 j의 최소 및 최대발현값 시점이 일치하는 지의 정보를 가지는 함수를 다음과 같이 정의한다.

$$M_{i,j} = \begin{cases} 1, & T_i^{min} = T_j^{min} \text{ 그리고 } T_i^{max} = T_j^{max} \text{ 인 경우}, \\ 0.5, & T_i^{min} = T_j^{min} \text{ 혹은 } T_i^{max} = T_j^{max} \text{ 중 하나만 성립하는 경우}, \\ 0, & T_i^{min} \neq T_j^{min} \text{ 그리고 } T_i^{max} \neq T_j^{max} \text{ 인 경우}. \end{cases}$$

피어슨 상관계수 R_{ij}를 사용하여 시간경로자료를 군집화하였을 때 각 군집 내에 비슷한 패턴을 가지는 유전자들이 모였는지를 평가하기 위하여 어느 군집내에 소속된 유전자 i와 j의 패턴일치지수를 다음과 같이 정의하자.

$$P_{i,j} = w_1 \cdot A_{i,j} + w_2 \cdot M_{i,j},$$

여기서 $0 \le P_{ij} \le 1$ 이고, w_1 과 w_2는 합이 1이 되는 음이 아닌 실수로서, 상승-하강-정체 패턴 혹은 최소발현값 시점 및 최대발현값 시점의 일치 등의 요인이 P_{ij} 에서 차지하는 비중을 나타낸다.

3. 예 제

NCBI(National Center for Biotechnology Information: http://www.ncbi.nlm.nih.gov/)의 마이크로어레이 데이타베이스에서 인간의 혈청에 대한 시간경로 유전자 발현자료(Accession No. GDS145)는 혈청에 대한 인간 섬유아세포의 반응에 관여하는 전사프로그램의 분석을 위한 실험자료로서 Cycloheximide가 존재하는 상태에서 섬유아세포에 혈청이 첨가된 후 9,983개 유전자의 발현값들이 0, 0.5, 2, 4시간(hr)이 지난후 관측되어 로그비 형태의 자료로 보존되어 있다.

마이크로어레이 자료분석에서 다루는 유전자의 수는 보통 수천에서 수만 개에 이르지만, 모든 유전자를 군집분석에 사용하지는 않으며 그 중 유용한 정보를 주는 유전자를

선별하여 군집분석한다. 유전자선별(gene filtering)은 생물학적으로 그리고 통계학적으로 의미있는 정보를 주는 유전자를 추출하는 과정이다. 예를 들면, Hoon 등 (2002)은 남조세균(Cyanobacteria)의 3,079 개 유전자 발현자료 중 선별한 유전자 90 개만을 군집분석에 사용하였고, Peddada 등 (2003)은 유방암 세포의 1,900 개 유전자 발현자료 중 유의한 유전자 50 개만을 선택하여 군집분석에 사용하였다. 본 논문에서는 유전자선별을 위하여 MATLAB(2003)의 Bioinformatics Toolbox의 유전자 선별함수들을 사용하였다. 먼저, 결측치가 없는 총 9,983 개 유전자들의 분산 들 중 가장 작은 10백분위수(percentile)에 해당하는 분산보다 더 작은 분산을 가지는 유전자를 제외한 8,985 개 유전자를 선택하였다. 그 중에서, 각 유전자의 4 개의 절대발현값(absolute expressive level)들 중 최대값이 전체 유전자들의 $8,985 \times 4 = 35,940$ 개 절대발현값들의 중앙값(median)보다도 더 작은 절대발현값을 가지는 유전자를 제외한 8,411개 유전자를 선택하였다. 마지막으로 8,411개 유전자들의 엔트로피(entropy)의 중앙값보다도 더 작은 엔트로피를 가지는 유전자를 제외한 3,769개 유전자가 최종 선택되었다. 분산과 엔트로피가 작은 유전자를 제거한다는 것은 그 유전자의 발현값들이 관측시점의 변화에 따른 변동이 매우 작아 시간경로상 생물학적 특성이 뚜렷하지 않은 유전자들을 제외시키기 위함이다. 물론 이상에서 사용된 유전자선별을 위한 방법 혹은 기준값은 얼마든지 달리하여 사용할 수 있음을 밝혀둔다. 최종 선별된 3,769 개 유전자들을 9개 군집으로 군집분석한 결과는 군집내에 무수히 많은 이질적인 유전자패턴을 보였다. 이 경우는 군집의 수를 9개 보다는 훨씬 크게 하여 군집내 동질성을 어느정도 유지한 후 본 논문에서 소개한 패턴일치지수를 그대로 활용할 수 있다. 본 연구에서 제안된 측도의 유용성에 대한 설명을 한정된 지면에서 보다 효과적으로 하기 위하여 3,769 개 유전자목록 중에서 다음의 순서, 즉, $[(3769 \times i)/50], i = 1, 2, \ldots, 50$, 여기서 $[x]$는 x보다 크지 않은 정수, 에 해당하는 유전자들을 추출하는 계통추출법에 의해 50 개의 유전자만이 최종적으로 본 실험의 군집분석에 사용되었다.

표 3.1은 분석에 사용된 50 개 유전자들의 식별자(id)이다. 피어슨 상관계수를 유사성 측도로 사용하여 완전연결법(complete linkage)에 의한 계층적 군집분석(hierarchical clustering) 을 적용하여 얻게 된 9 개 군집의 결과가 그림 3.1 및 그림 3.2에 보여진다. 표 3.2에는 각 군집별 피어슨 상관계수 및 패턴일치지수($w_1 = w_2 = 0.5$)의 평균(표준편차)이 계산되어 있다. 표 3.2와 그림 3.1에서 군집 1과 군집 3은 각 패턴일치지수의 평균(표준편차)이 1(0)으로서 각 군집내 모든 유전자들의 상승-하강-정체 패턴이 일치하고, 또한 최소 및 최대발현값 시점이 일치한다. 표 3.3에는 군집 2, 5, 8, 9의 군집내 각 유전자쌍의 상관계수 및 패턴일치지수가 계산되어 있다. 이때 유전자 쌍 $(\cdot, \cdot)$에 제시된 번호는 표 3.1에서 50 개 유전자 식별자에 편의상 붙인 일련번호로서 해당 식별자를 갖는 유전자를 나타낸다고 하자.

군집 2에서 유전자쌍 (9,16)의 상관계수는 0.9944로서 매우 높으나 패턴일치지수는 0.5833으로서 낮다. 유전자 16이 포함된 유전자쌍 모두 0.5833의 패턴일치지수를 갖는다. 유전자 16은 그림 3.1의 군집 2의 그림에서 점선으로 표시된 유전자이다. 군집내 다른 유전자들이 첫번째 시간구간에서 상승패턴을 보이고 최소값 발현시점이 0hr 인 반면 유전자 16은 첫번째 시간구간에서 하강패턴을 보이고 최소값 발현시점이 0.5hr이다. 군집 5의 경우 그림 3.1에서 점선으로 표시된 유전자 12가 이질적임은 표 3.3의 상관계수 및 패턴일치지수

표 3.1: 군집분석에 사용된 50 개 유전자들의 식별자(id)

no	Gene id	no	Gene id	no	Gene id	no	Gene id	no	Gene id
1	T95670	11	AA00420	21	W67698	31	N21084	41	H21397
2	R63646	12	AA04317	22	AA01312	32	N71440	42	H77748
3	H18657	13	AA05325	23	AA04781	33	W70084	43	N57554
4	H60294	14	T65308	24	AA01138	34	AA02578	44	N80766
5	N30669	15	R35829	25	AA03931	35	AA00467	45	W69987
6	N64157	16	H25223	26	AA04572	36	AA04020	46	W96169
7	N91556	17	H30200	27	5038	37	AA05620	47	AA05925
8	W69211	18	H91693	28	R02280	38	T54621	48	AA03198
9	AA01579	19	N64039	29	R54786	39	R11230	49	AA04581
10	AA05602	20	N79537	30	R99947	40	R81826	50	9992

표 3.2: 군집내 유전자쌍의 상관계수 및 패턴일치지수의 평균(표준편차)

군집 번호	군집 크기	상관계수(R_{ij}) 평균(표준편차)	패턴일치지수(P_{ij}) 평균(표준편차)
1	5	0.9671(0.0203)	1.0000(0.0000)
2	6	0.9215(0.0792)	0.8611(0.2033)
3	3	0.9802(0.0131)	1.0000(0.0000)
4	5	0.8488(0.0841)	0.6000(0.2215)
5	4	0.7030(0.3244)	0.7083(0.3195)
6	8	0.8793(0.0886)	0.5387(0.2855)
7	12	0.8592(0.1129)	0.6035(0.3450)
8	4	0.9335(0.0644)	0.8750(0.1369)
9	3	0.7917(0.1389)	0.8889(0.0962)

로도 확인할 수 있다. 그림 3.1의 군집 8에서 점선으로 표시된 유전자 26은 최대 발현시점이 0hr인 반면 나머지 유전자들은 2hr이다. 그림 3.1의 군집 9에서 점선으로 표시된 유전자 49는 마지막 시간구간에서 상승패턴을 보임으로서 0.8333의 패턴일치지수를 갖는다.

위에서 소개된 절차에 의해 군집 4, 6, 7로 부터 각각 이질적인 유전자들을 추출한 결과가 그림 3.2에 보여진다. 여기서 그림 (a)에 보여지는 유전자들은 그림 (b)의 이질적인 유전자들을 각 군집으로 부터 골라내고 남은 유전자들을 나타낸다. 그림 (a)에 포함된 유전자쌍들의 패턴일치지수의 평균(표준편차)은 1(0)이다. 유전자 8, 28, 33, 35, 48을 포함하는 군집 4에서 그림 (a)에 포함된 유전자는 28과 48이다. 유전자 7, 10, 11, 20, 23, 24, 25, 34를 포함하는 군집 6에서 그림 (a)에 포함된 유전자들은 7, 20, 34이다. 유전자 2, 3, 5, 13, 15, 21, 31, 38, 39, 40, 41, 50을 포함하는 군집 7에서 그림 (a)에 포함된 유전자들은 5, 13, 15, 31, 40, 41, 50이다.

표 3.3: 군집내 유전자쌍의 상관계수 및 패턴일치지수

Cluster 2			Cluster 5		
유전자 쌍	R_{ij}	P_{ij}	유전자 쌍	R_{ij}	P_{ij}
(9,16)	0.9944	0.5833	(6,12)	0.4339	0.4167
(9,17)	0.9056	1.0000	(6,36)	0.9984	1.0000
(9,27)	0.9866	1.0000	(6,37)	0.9987	1.0000
(9,32)	0.9971	1.0000	(12,36)	0.3992	0.4167
(9,45)	0.7782	1.0000	(12,37)	0.3883	0.4167
(16,17)	0.8991	0.5833	(36,37)	0.9995	1.0000
(16,27)	0.9711	0.5833	Cluster 8		
(16,45)	0.7797	0.5833	(26,29)	0.9276	0.7500
(17,27)	0.9510	1.0000	(26,30)	0.8359	0.7500
(17,32)	0.9353	1.0000	(26,43)	0.8807	0.7500
(17,45)	0.9692	1.0000	(29,30)	0.9719	1.0000
(27,32)	0.9935	1.0000	(29,43)	0.9888	1.0000
(27,45)	0.8457	1.0000	(30,43)	0.9959	1.0000
(32,45)	0.8239	1.0000	Cluster 9		
			(22,46)	0.9301	1.0000
			(22,49)	0.6523	0.8333
			(46,49)	0.7929	0.8333

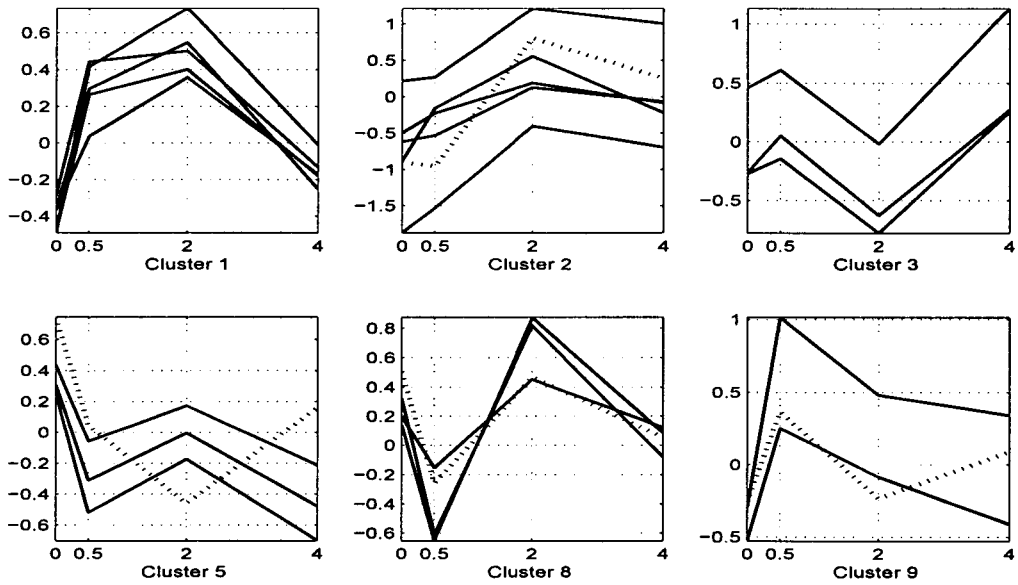

그림 3.1: 혈청 시간경로자료의 계층적 군집결과: 군집 1,2,3,5,8,9

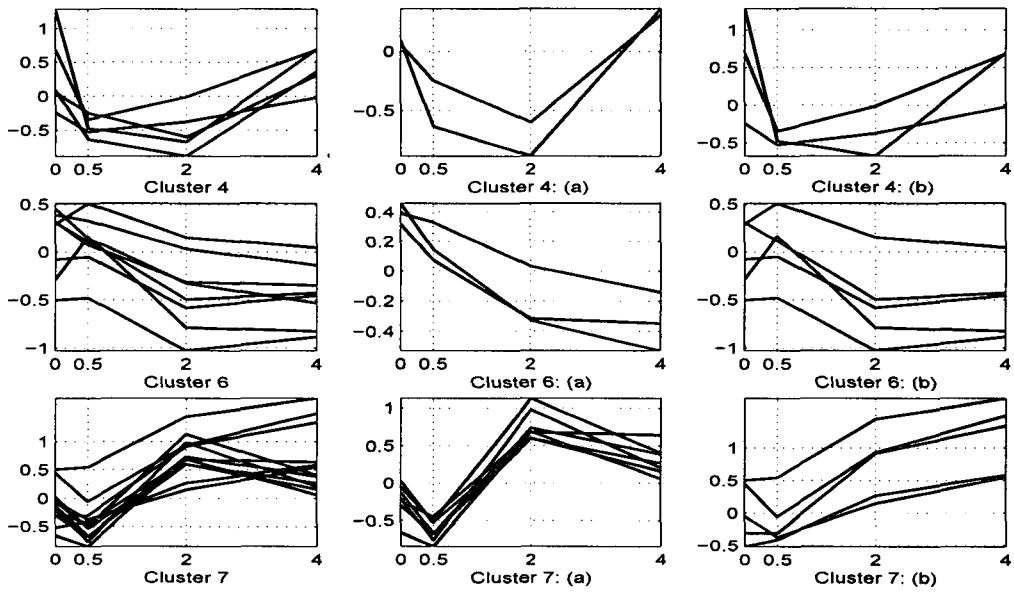

그림 3.2: 혈청 시간경로자료의 계층적 군집결과:
군집 4,6,7에서 이질적인 유전자(b)를 추출한 결과(a)

4. 결 론

계층적 군집분석 혹은 K-means 군집분석은 군집의 수가 정해지면 모든 시계열은 군집들 중 하나의 군집에 반드시 속하게 된다. 군집에 속하는 하나의 시계열은 사용된 유사도를 기준으로 볼 때, 다른 군집보다 소속 군집에 대해 상대적으로 보다 높은 유사도 값을 보일지라도 소속 군집내에서는 매우 이질적일 수 있다. 이질적인 시계열은 골라내어 군집의 생물학적인 기능면에서도 이질적인지를 검토해 볼 수 있다.

시간경로 유전자 발현자료의 군집분석에서는 비슷한 패턴의 생물학적 동적시스템을 가지는 유전자들을 군집화 하는데 목적이 있으므로, 연구목적에 따라 시계열의 움직임 패턴이 같은 유전자들, 시간에 따른 시계열의 변화의 방향과 변화량이 비슷한 유전자들, 혹은 최소 및 최대발현값을 주는 시점이 같은 유전자들을 동일한 군집으로 볼 수 있다. 이러한 관점에서 볼 때 피어슨 상관계수는 완전하지 않다. 본 논문에서 제안된 패턴일치지수는 피어슨 상관계수가 간과할 수 있는 두 시계열의 상승-하강-정체의 일치 정도와 최소 및 최대 발현 값을 주는 시점의 일치 정도를 측정케 한다. 최종적으로 분할된 군집에 속하는 모든 유전자 쌍에 대한 패턴일치지수 값을 검토하여 군집의 대표적인 패턴으로 부터 벗어나는 이질적인 유전자를 골라낼 수 있다. 그러나 최종적인 판단은 생물학적 지식 기반 하에서 행해야 함은 자명한 사실이다.

참고문헌

Hoon, M. J. L., Imoto, S., and Miyano, S. (2002). Statistical analysis of a small set of time-ordered gene expression data using linear splines, *Bioinformatics*, **18**, 1477-1485.

Luan, Y. and Li, H. (2003). Clustering of time-course gene expression data using a mixed-effects model with B-splines, *Bioinformatics*, **19**, 474-482.

The MATH WORKS Inc. (2003). *Bioinformatics Toolbox: For Use with MATLAB, User's Guide, Version 1*, The Math Works Inc., Natick.

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 experments using order-restricted inference, *Bioinformatics*, **19**, 834-841.

Schliep, A., Schonhuth, A., and Steinhoff, C. (2003). Using hidden Markov models to analyze gene expression time course data, *Bioinformatics*, **19**, i255-i263.

[2004년 12월 접수, 2005년 1월 채택]

A Pattern Consistency Index for Detecting Heterogeneous Time Series in Clustering Time Course Gene Expression Data *

Young Sook Son[1] Jangsun Baek[2]

ABSTRACT

In this paper, we propose a pattern consistency index for detecting heterogeneous time series that deviate from the representative pattern of each cluster in clustering time course gene expression data using the Pearson correlation coefficient. We examine its usefulness by applying this index to serum time course gene expression data from microarrays.

Keywords: Microarray, Time course gene expression data, Pearson correlation coefficient, Pattern consistency index, Hierarchical clustering.

* This work was supported by grant No. RTI04-03-03 from the Regional Technology Innovation Program of the Ministry of Commerce, Industry and Energy(MOCIE).

1) Professor, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Gwangju, 500-757, KOREA.
E-mail: ysson@chonnam.ac.kr

2) Professor, Department of Statistics, Chonnam National University, 300, Yongbong-dong, Gwangju, 500-757, KOREA.
E-mail: jbaek@chonnam.ac.kr