The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 335-348

Consensus Clustering for Time Course Gene Expression Microarray Data[1)]

Seo Young Kim[2)] and Jong Sung Bae[3)]

Abstract

The rapid development of microarray technologies enabled the monitoring of expression levels of thousands of genes simultaneously. Recently, the time course gene expression data are often measured to study dynamic biological systems and gene regulatory networks. For the data, biologists are attempting to group genes based on the temporal pattern of their expression levels. We apply the consensus clustering algorithm to a time course gene expression data in order to infer statistically meaningful information from the measurements. We evaluate each of consensus clustering and existing clustering methods with various validation measures. In this paper, we consider hierarchical clustering and Diana of existing methods, and consensus clustering with hierarchical clustering, Diana and mixed hierachical and Diana methods and evaluate their performances on a real microarray data set and two simulated data sets.

Keywords : Consensus clustering, Time course gene expression data, Validation measures.

1. 서론

최근, DNA에 있는 유전자 정보들을 분석해 내기 위한 고도의 생명공학 실험 기술인 cDNA 마이크로어레이 기술은 대량의 유전자 정보를 동시에 관찰할 수 있게 하였다(Brown 과 Botstein, 1999). 이러한 실험 자료들을 생물학자 또는 의학, 약학 관련 연구자들이 연구할 수 있는 의미 있는 정보로 조직화하고 분석하는 과정에 통계학이 중요한 위치를 차지하고 있다. 생물학적인 측면에서 수많은 유전자 전체의 발현정도를 연구하는 데에는 기존의 실험 방법들보다 cDNA 나 올리고뉴클레오타이드(oligonucleotide) 칩이 많은 장점을 가지고 있다. 이러한 DNA 마이크로어레이 기술은 돌연변이 검색, 질병의 진단 또는 유전자 발현의 청사진을 밝혀내는데 많은 기여를 할 것으로 예상되고 있고, 또한 과학 기술적인 측면뿐만 아니라 인류의 건강과 생명의 신비를 해석하는데 아주 중요한 역할을 할 것으로 기대를 모으고 있다.

1) This work was supported by Korea Research Foundation Grant (KRF-2002-075-C0005).
2) Researcher, Research Institute for Basic Science, Chonnam National University, Gwangju 500-757, Korea, E-mail: gong@chonnam.ac.kr.
3) Professor, Department. of Statistics, Chonnam National University, Gwangju 500-757.

측정된 유전자 발현 자료는 유전자들의 수가 많고 생물학적 진행과정이 복잡하기 때문에, 유전자들의 군집은 이러한 자료들을 분석하는데 가장 널리 사용되고 있는 통계적 방법 중의 하나이다. 유사한 발현 프로파일을 갖는 유전자들을 군집화하는 것은 알려지지 않는 유전자들의 기능을 예측할 수 있고, 동일한 메카니즘에 의한 유전자들의 부분적인 집합들을 확인할 수 있게 한다.

마이크로어레이 기술의 중요한 응용분야 중 하나는 연속적인 시점에 따른 유전자들의 발현 양상을 연구하는 것이다. 유사한 발현 프로파일을 공유하는 유전자들은 유사한 기능을 갖을 뿐만 아니라 동일한 생물학적 조절 작용을 받는 것으로 고려되고 있다. 따라서 이러한 마이크로어레이 자료는 유전자들 간의 상호작용 및 유전자 기능과 경로 확인에 대한 식견을 제공할 것이다. 시간 경로 (time course) 유전자 발현 자료의 특징 중 하나는 주어진 유전자에 대해 유전자의 발현수준이 시간에 의존적이라는 것이다. 게다가 유전자 발현 수준들이 시간에 걸쳐 진화하기 때문에 시간은 유전자 발현 수준에 영향을 미치는 중요한 요인이 될 수 있다는 것이다. 이와 관련된 선행연구 (DeRisi, Iyer 와 Brown, 1997; Chu 와 DeRisi 등, 1998; Cho 등, 1998)로서, 주로 시각적 기법을 이용한 분석 방법이 기능이 유사한 소수의 유전자들의 그룹으로 분류하는데 매우 효과적으로 사용되었다. 그러나 이 방법은 많은 노력을 필요로 하고, 매우 주관적인 경향이 있을 뿐만 아니라 더 복잡하고 대량의 자료 분석에는 적합하지 않다는 특징을 가지고 있다. 이어서 발현비에 의한 단순한 정렬이라든가 상관거리의 형태를 이용한 방법들 (Spellman 등, 1998; Eisen 등, 1998)이 유전자들을 확인하는데 사용되었다. 이러한 체계적 방법은 소수의 시점을 갖는 자료로부터 유전자 발현 측정치들을 분석하는데 있어서 통계적 고려가 결여되어 있었다. De Hoon 등 (2002)은 최대우도 방법 (maximum likelihood method)을 사용한 선형 스플라인 함수 (Linear spline function)에 기반을 둔 방법을 적용하였고, Luan 과 Li (2003)는 B-스플라인 함수 (B-spline function)를 갖는 혼합효과모형을 이용하여 시간 경로 자료를 분류하였다. 또한 Peddada 등(2003)은 순서 제약 (order-restricted) 추론 방법에 근거한 알고리즘을 제안하였고, Datta 와 Datta(2003)는 다양한 군집방법을 타당성 평가 위주로 비교하였다.

최근에는 샘플링 변동에 관련된 군집 결과들의 안정성을 평가하기 위해 원시 자료로부터 가상의 자료셋을 생성하는 재표본(resampling)과 교차 타당성(cross validation)기법들이 효과적으로 사용되고 있다(Bhattacharjee 등, 1998, Dudoit 와 Fridlyand, 2002, Jain 과 Moreau, 1988, Levine 와 Domany, 2001). 특히 Bhattacharjee 등(2001)은 계층적 군집(Hierachical clustering: HC)에 의한 결과를 평가하고 군집 안정성을 평가하기 위한 붓스트랩핑에 의한 사용을 소개하였고, Monti 등(2003)은 유전자 발현자료로부터 재표본에 기반을 둔 방법으로 군 발견과 군집 평가를 위한 일치군집(Consensus Clustering: *CC*) 방법을 제안하였다. 이 방법은 군집을 나누기 위한 새로운 유사성행렬을 사용한 것으로, 재표본 방법에 의해 매번 원시 자료로부터 가상의 자료셋을 구성한다. 구성된 자료에 대해 초기 군집 알고리즘을 이용하여 반복적으로 군집 분석을 수행하고, 각 가상의 자료셋으로 부터 각 개체가 같은 군집에 할당되는 비율이 어느 정도인가를 계산하여 이를 새로운 유사성 거리행렬로 사용하는 방법이다. 따라서 매번 동일한 군집 알고리즘에 의해 군집분석을 반복 수행하기 때문에 이 방법은 초기 군집알고리즘의 선택에 상당히 의존하게 되는 특징을 가지고 있다. 이러한 문제점을 보완하기 위해 Kim 과 Lee (2004, submitted)은 몇 가지의 군집 알고리즘을 혼합하는 확장된 *CC* 방법을 제안하고, 이 방법이 기존의 방법들에 비해 수행능력이 우수하다는 것을 제시하였다.

본 연구에서는 *CC* 알고리즘을 소수 개 시점을 갖는 시간 경로 유전자 발현 자료에 대해 적용

하고, *CC* 알고리즘이 시간에 대한 영향력을 잘 반영함으로써 유전자 발현 패턴을 효과적으로 분류할 수 능력을 평가하고자 한다. *CC* 방법이 시간 경로 자료에 얼마나 효과적으로 사용될 수 있는지 평가하기 위해 거리행렬을 구성하는 과정에서 한 시점을 제거해 나가는 교차 타당성(cross validation) 방법 위주로 기존의 군집 알고리즘들과의 비교 결과를 제시하였다. 본 논문의 구성은 다음과 같다. 2장에서는 *CC* 알고리즘을 설명하고, 3장에서는 초기 군집방법으로 사용될 병합적 계층방법인 HC 와 분리적 계층방법인 Diana 알고리즘을 소개하고 타당성 평가 측도들을 설명한다. 4장에서는 실제 분석에 사용될 유전자 발현 자료와 모의실험 방법을 설명한다. 5장에서 분석 결과를 위주로 다양한 군집 알고리즘의 수행 능력을 비교하고, 마지막으로 6장에서 간단하게 결론을 내린다.

2. 일치 군집 알고리즘

2.1. CC 알고리즘

본 절에서는 Monti 등(2003)의 알고리즘을 위주로 설명함으로써 독자들의 이해를 돕고자 한다. 먼저, 관심 있는 자료셋을 $D=\{e_1, e_2, \cdots, e_N\}$라 하자. 이때 군집의 목적은 상호배타적이고 서로 겹치지 않는 군집들의 부분으로 관측된 유전자들을 쪼개는 것이다. D의 K개 군집으로의 분할 P를 $P=\{P_1, P_2, \cdots, P_K\}$이라 하면, $\cup_{k=1}^{K} P_k = D$ 이고, 서로 다른 모든 i, j $P_i \cap P_j = \emptyset$이다. *CC* 알고리즘에 의한 군집 절차는 다음과 같다.

(1) 재표본 방법과 초기 군집 알고리즘(HC, Diana)이 선택되었다고 하자.

(2) 원시자료로부터 구성된 H개의 가상의 재표본 자료셋을 $D^{(1)}, D^{(2)}, \cdots, D^{(H)}$라 하자.

(3) 먼저, 임의의 자료셋 $D^{(h)}$에 대한 $N \times N$ 연결행렬(connectivity matrix)을 $M^{(h)}$라 하면,

$$M^{(h)}(i,j) = \begin{cases} 1, & \text{유전자 } i,j\text{가 같은 군집에 속할 때} \\ 0, & \text{그렇지 않으면} \end{cases},$$

이고, 이때 지시행렬(indicator matrix)은 다음과 같다.

$$I^{(h)}(i,j) = \begin{cases} 1, & \text{유전자 } i,j\text{가 같은 자료셋 } D^{(h)}\text{에 동시에 선택되었으면} \\ 0, & \text{그렇지 않으면} \end{cases}$$

재표본에 의한 자료셋을 사용하여 군집분석을 하기 때문에 반드시 지시행렬은 필요하다. 대부분 붓스트랩이나 부분적 샘플링에 의한 자료는 원시 자료로부터 모든 개체를 포함하지 않게 되는 것이 일반적이다.

(4) 단계 3에서 행렬들을 이용하면, 일치행렬(consensus matrix) C는 다음과 같이 구성된다.

$$C(i,j) = \frac{\sum_{h=1}^{H} M^{(h)}(i,j)}{\sum_{h=1}^{H} I^{(h)}(i,j)}, \quad h = 1, 2, \cdots, H,$$

이때, $C(i,j) = C(j,i)$이고, $0 \le C(i,j) \le 1$ 으로 행렬 C의 (i,j) 번째 엔트리가 0 또는 1

인 것은 임의의 두 유전자가 완전히 일치함을 의미한다.

(5) 모든 군집 수 K에 대해 단계3-단계4를 반복하고, 군집의 개수 결정 기준에 의해 최적의 군집의 개수를 정한다.

(6) 마지막으로, 최적의 군집 개수에 해당되는 일치행렬 $C^{(K)}$에 대해서 $1-C^{(K)}$을 계산하고 이 거리행렬에 기반한 최종 군집분석을 수행함으로써 각 유전자들을 각 군집으로 할당한다.

2.2 혼합 *CC* (Mixed Consensus Clustering) 알고리즘

2.1절에서 소개된 *CC* 알고리즘은 반복적으로 구성된 가상의 자료셋에 대해 초기에 선택된 군집방법으로부터 일치행렬을 구성하여 이를 거리행렬로 사용하기 때문에, 군집결과는 초기 군집방법에 매우 의존하게 될 것으로 예상된다. 따라서 각 군집방법의 고유한 특성을 완화시킬 수 있는 방법으로 두 가지 이상의 방법을 혼합하여 군집 할당의 근거가 되는 일치행렬을 구성하는 방법을 고려할 수 있겠다 (Kim 과 Lee, 2004 submitted). 방법의 핵심은 두 개의 군집방법에 의해 구성되는 각각의 일치행렬을 혼합하자는 것이다. 초기 군집알고리즘으로 병합적 계층적 방법인 HC와 분리적 계층적 방법인 Diana를 사용할 것이다. 이는 병합적 방법과 분리적 방법을 적절하게 혼합하자는 취지에서 사용된 것이다. 2.1절의 알고리즘과 유사한 기법을 사용하고, 단, 일치행렬 대신에 혼합된 일치행렬 (mixed consensus matrix)을 사용한다는 것만 다르다. 초기 알고리즘으로 HC와 Diana를 사용한 경우라면, 2.1절의 *CC* 알고리즘의 수행절차에서 3, 4, 6번은 다음에 주어진 방법에 따라 수행하고, 나머지 절차는 동일하다.

(3-1) 자료셋 $D^{(h)}$에 대한 $M_{hc}^{(h)}$ $M_{diana}^{(h)}$를 구하고, 지시행렬 $I_{hc}^{(h)}$, $I_{diana}^{(h)}$를 각각 구성한 다음, 각각의 일치행렬 $C_{hc}^{(h)}$, $C_{diana}^{(h)}$를 구성한다.

(4-1) 혼합된 일치행렬, *MC*를 구한다. 이때 *MC*의 모든 엔트리(entry)는 두 일치행렬의 평균 엔트리를 취한다. 즉,

$$MC^{(h)} = mean(C_{hc}^{(h)}, C_{diana}^{(h)}).$$

이때, $MC(i,j)=MC(j,i)$이고, $0 \le MC(i,j) \le 1$ 으로 행렬 *MC*의 (i,j)엔트리가 0, 1인 것은 임의의 두 유전자가 완전히 일치함을 의미한다.

(6-1) 마지막으로, 최적 군집 개수에 해당되는 일치행렬 $MC^{(K)}$에 대해서 $1-MC^{(K)}$를 계산하고 이 거리행렬에 기반한 최종 군집분석을 수행함으로써 각 유전자들을 각 군집으로 할당한다.

3. 군집알고리즘과 타당성 측도

3.1. 군집알고리즘

계층적 군집방법에는 단계적으로 발현양상이 비슷한 유전자들이나 같은 종류의 암 환자들로 이

루어진 군집을 형성해 나가는 병합적 방법(agglomerative method)과 하나의 군집에서 보다 큰 비유사성을 갖는 유전자들을 분리해 나감으로써 더 작은 군집을 형성해 가는 분리적 방법(divisive method)이 있다. 우리는 병합적 방법으로 UPGMA (Unweighted Pair Group Method with Arithmetic Mean), 즉 평균연결법(average linkage method)을 사용하고 두 유전자간의 거리는 유클리디안 거리(Euclidean distance) 또는 $d(x,y)=1-|corr(x,y)|$를 사용한다. 이때 $corr(x,y)$는 두 유전자 x,y의 발현프로파일 간의 상관계수를 의미한다. 그리고 대표적 분리방법으로 Diana를 사용한다. 분리적 군집방법은 하나의 군집을 소수의 큰 군집들로 나누고자 할 때 효과적인 방법으로 알려져 있다.

3.2. 타당성 평가 측도

시간 경로 마이크로어레이 자료에 대해 CC 알고리즘 및 혼합 알고리즘의 수행능력을 평가하기 위해 4 가지 평가 측도들을 사용한다. 많은 경우에 있어서 생물학자들은 군집의 개수에 대한 사전 정보를 어느 정도 알고 있다고 한다(Datta 와 Datta, 2003). 타당성 평가방법의 기본 아이디어는 군집결과의 일치성에 대한 보장이 이루어져야 한다는 것이다. 유전자 발현자료가 l 시점에 걸쳐서 조사되었다고 한다면, 시점을 $t_1, t_2, \cdots, t_l$이라 하자. 각 시점 $i=1,2,\cdots,l$ 에 대해서 시점 t_i의 관측값을 제외함으로써 군집 알고리즘을 반복적으로 수행한다(Leave-one out). 이때 유전자 $1 \le g \le G$에 대해서, $C^{g,i}$은 t_i번째 시점에서 관측값이 제외된 자료에 대해서 군집 분석을 수행했을 때 유전자 g가 포함된 군집이라 하고, $C^{g,0}$는 모든 시점에 걸쳐 군집분석을 수행했을 때 유전자 g가 포함된 전체군집이라 하자. 군집 알고리즘의 능력을 평가하기 위해 다음과 같은 4 가지 타당성 평가 측도들을 사용한다.

(1) 겹치지 않는 평균비율(the average proportion of non-overlap measure)

$$V_1(K)=\frac{1}{Gl}\sum_{g=1}^{G}\sum_{i=1}^{l}\left(1-\frac{n(C^{g,i}\cap C^{g,0})}{n(C^{g,0})}\right).$$

이 측도는 한 번에 한 시점을 제외시켰을 때의 자료를 이용하여 군집 분석을 수행한 군집결과와 모든 시점에 걸쳐 군집분석을 수행했을 때의 군집결과가 일치하지 않은 유전자들의 평균비율을 계산한다.

(2) 평균간 평균거리(the average distance between means measure)

$$V_2(K)=\frac{1}{Gl}\sum_{g=1}^{G}\sum_{i=1}^{l} d\left(\bar{x}_{C^{g,i}},\ \bar{x}_{C^{g,0}}\right).$$

여기서 $\bar{x}_{C^{g,i}}$는 군집 $C^{g,i}$내에 포함되어 있는 유전자들 간의 평균발현 프로파일을 나타내고, $\bar{x}_{C^{g,0}}$는 군집 $C^{g,0}$에 포함되어 있는 유전자들 간의 평균발현 프로파일을 나타낸다. 이 측도는 한 번에 한 시점에서의 발현수준을 제외시켰을 때, 같은 군집 내 유전자들의 평균과 전체자료에 대한 군집결과에서 같은 군집 내 유전자들의 평균들 간의 평균거리를 계산한다.

(3) 평균거리(the average distance measure)

$$V_3(K)=\frac{1}{Gl}\sum_{g=1}^{G}\sum_{i=1}^{l}\frac{1}{n(C^{g,0})n(C^{g,i})}\times\sum_{g\in C^{g,0},\,g'\in C^{g,i}} d(x_g, x_{g'}).$$

여기서 $d(x_g, x_{g'})$은 유전자 g, g'들의 발현 프로파일들 간의 거리를 나타낸다. 이 측도는 한 번에 한 시점을 제외시켰을 때의 군집결과와 전 시점에 걸친 자료에 대한 군집 결과에서 각 유전자들의 발현 프로파일간의 평균거리를 계산한다.

(4) 수정된 Rand 지수의 평균 (the average of the adjusted rand index)

n개 관측값에 대해 2개 분류가 가능하다고 하자. 이때 두 분류 중 R 개 그룹을 갖는 분류는 $U=\{u_1, u_2, \cdots, u_R\}$, C개 그룹을 갖는 분류를 $V=\{v_1, v_2, \cdots, v_C\}$라 하면 두 분류의 일치도는 <표 1>과 같은 분할표로 나타낼 수 있다.

<표1> n 개체들에 대한 분류 분할표

	v_1	v_2	$\cdots$	v_C	
u_1	n_{11}	n_{12}	$\cdots$	n_{1C}	$n_{1.}$
u_2	n_{21}	n_{22}	$\cdots$	n_{2C}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
u_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	$n_{R.}$
	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.C}$	n

여기서 n_{ij}는 두 그룹 u_i와 v_j ($i=1, \cdots, R$, $j=1, \cdots, C$)에 동시에 속하는 개체들의 수를 나타낸다. 이때 $n_{i.}=\sum_{j=1}^{C} n_{ij}$ 와 $n_{.j}=\sum_{i=1}^{R} n_{ij}$ 는 각각 <표 1>에서 행의 합과 열의 합을 나타낸다. 두 분류의 일치도를 나타내는 수정된 Rand 지수(Huber and Arabie, 1985)는 다음과 같다.

$$Rand=\frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2}-[1/\binom{n}{2}]\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{(1/2)[\sum_{i=1}^{R}\binom{n_{i.}}{2}+\sum_{j=1}^{C}\binom{n_{.j}}{2}]-[1/\binom{n}{2}]\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}.$$

모든 시점에서의 자료를 적용했을 때 분류결과와 시점을 하나씩 제외시켜가면서 분류했을 때의 수정된 Rand 지수를 구하고, 모든 시점에 대해서 평균 수정된 Rand 지수 측도값을 구한다, 즉

$$V_4(K)=\sum_{i=1}^{l} Rand_i/l.$$

위의 V_1, V_2, V_3는 Datta 와 Datta (2003)에서 사용되었으며 측도값이 낮을수록 우수한 알고리즘으로 평가하고, V_4는 측도값이 높을수록 우수한 군집 알고리즘으로 평가한다.

4. 유전자 발현 자료 및 모의실험

Sporulation 자료

Sporulation 자료(Chu 등, 1998)는 총 6114개의 유전자를 포함하는 DNA 마이크로어레이를 이용하여 실험하였다. mRNA 수준은 포자형성(sporulation)이 진행되는 동안 7 시점 간격으로 측정되었다. 각 유전자의 mRNA 발현수준 대 포자형성의 중간에 도달하기 바로 직전 성장 세포에서의 mRNA수준의 비(ratio)가 측정되었고, 각 비율 자료(R)는 $\log_2$ 변환되었다. Chu 등(1998)은 포자형성이 진행되는 동안 $\log_2 R$의 평균제곱근이 1.13 수준보다 작은 유전자는 유의하게 발현되지 않는 유전자로 간주하여 분석에서 제외시켰다. 최종적으로 포자형성이 진행되는 동안 양적으로 발현되는 즉, $\Sigma_t \log R > 0$ 을 만족하는 유전자 513개가 분석에 사용되었다.

모의실험 자료

Quackenbush(2001) 와 Datta와 Datta(2003)에서 사용된 모의실험 방법을 적용하였다. 모의실험은 10개 시점에 걸쳐 9개의 다른 패턴들을 갖도록 설계되었다. 각 시점에서 각 패턴이 갖는 평균 발현 수준값들을 다르게 설정하였다. 모의실험 자료는 이들 각 시점별 패턴이 취하는 평균 발현 수준값에 독립적인 랜덤 변수를 더해 줌으로서 생성되었다. 전체적으로 10개 시점에 걸쳐 각각 9개의 군집을 생성하고 한 시점에서 각 군집에는 50개씩의 유전자가 포함되어, 총 10*9*50=4500개의 유전자가 생성되었다. 즉, 유전자의 발현 프로파일을 생성하기 위해 각 군집은 450개 유전자가 포함되어 있다. 두 경우에 대한 모의실험을 다음과 같이 시행하였다. 모의실험 1은 평균이 0이고 표준편차가 0.2를 갖는 정규분포로부터 랜덤 변수를 생성하여 10개 시점에 걸쳐 각각 9개 군집으로부터 4500개의 유전자가 생성되었다. 모의실험 2는 4500개중 절반은 평균이 0, 표준편차가 0.4인 정규분포로부터 생성되었고, 나머지는 위치와 척도 모수가 각각 -0.2와 0.2를 갖는 지수분포로부터 랜덤 변수가 생성되었다.

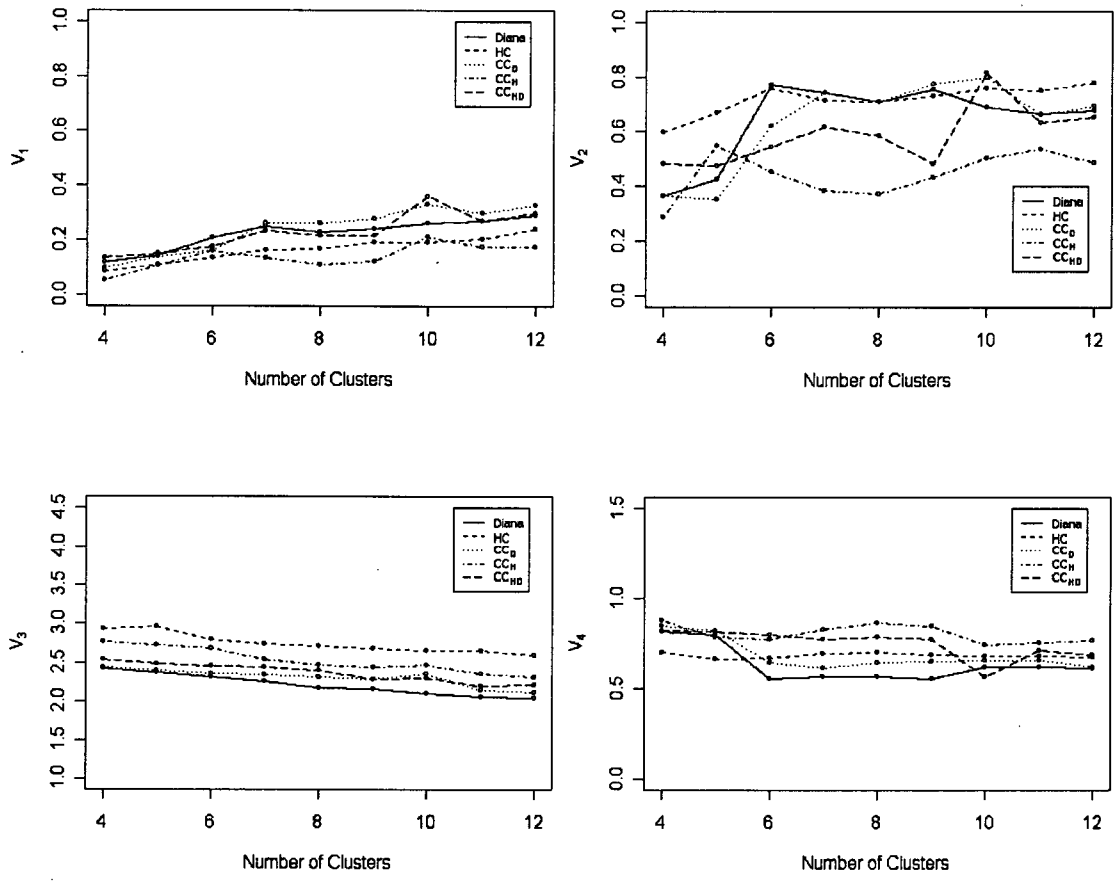
5. 분석결과

실제 유전자 발현 자료와 모의실험 자료에 대해 5가지의 군집 알고리즘을 적용하고, 4가지의 타당성 평가 측도값을 계산하였다. 5가지 군집 알고리즘은 HC, Diana, 그리고 HC를 초기 알고리즘으로 한 CC(CC_H), Diana를 초기 알고리즘으로 한 CC(CC_D), HC와 Diana를 혼합한 CC (CC_{HD})이다. 원시자료 셋으로부터 80% 비복원 랜덤 추출(without random sampling)하여 재표본 자료셋을 구성하였다.

5.1 Sporulation 자료의 분석결과

4개에서 12개의 군집 개수에 걸쳐서 타당성 평가를 위한 측도값을 계산하였다. 그 결과를 <그림1>에 도시하였다. V_1의 평가에 의하면 HC, CC_H가 상대적으로 낮은 측도값을 보여 우수한 수행능력을 나타내고, V_2에서는 CC_H, CC_{HD}가 낮은 측도값을 나타내고, V_3에서는 $Diana, CC_D$ 가 가장 낮은 측도값을 나타냄으로서 우수한 알고리즘으로 평가될 수 있다. V_4에서는 CC_H,

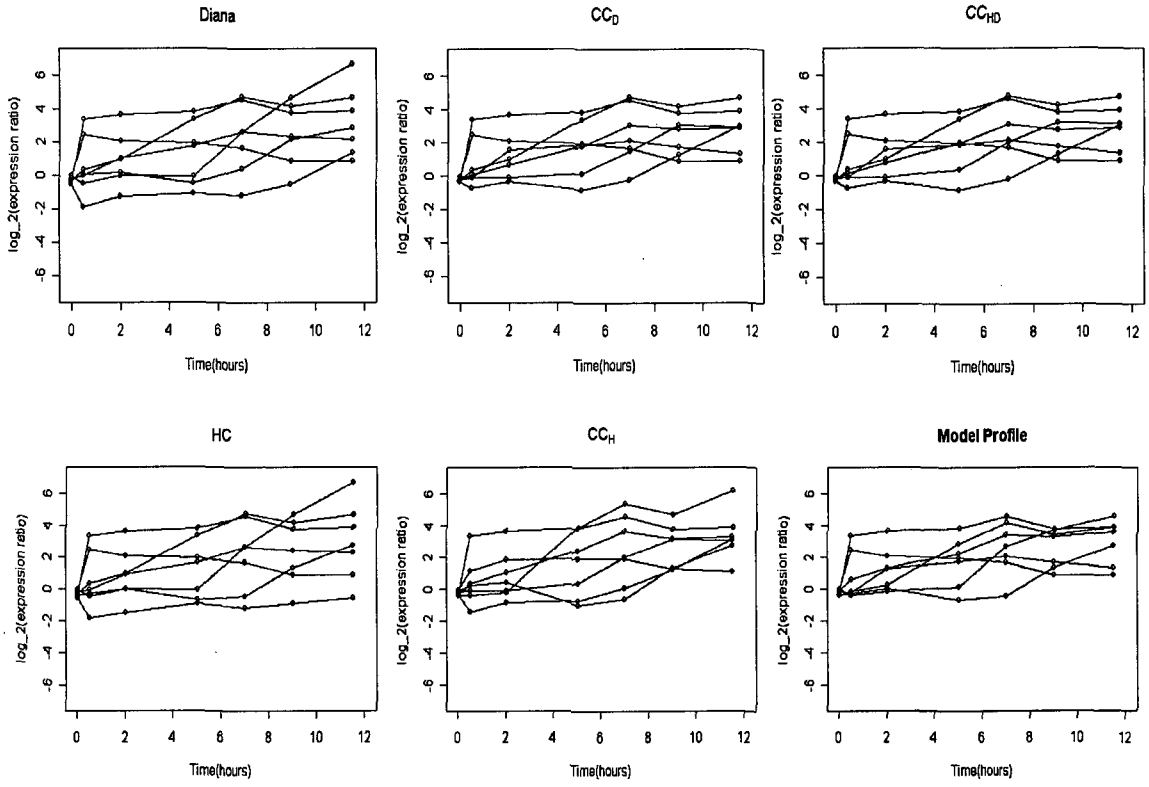
CC_{HD}가 우수한 수행능력을 보이는 것으로 나타났다. 각 평가방법에 따라 알고리즘의 우수성 정도는 다르게 평가되었지만, 전체적으로 CC 방법이 일반적인 방법(HC 또는 Diana)보다 더 우수한 능력을 보였다. CC 중에서는 CC_H, CC_{HD}가 우수한 능력을 보인다는 것을 알 수 있다. 전체적으로 4가지 타당성 평가 측도 하에서 CC_H와 CC_{HD}가 안정적이고 군집의 개수에 대해서도 로버스트한 수행능력을 나타내는 것으로 판단되었다.



<그림 1> Sporulation 자료에 적용된 다양한 군집 알고리즘의 타당성 측도

한편, Chu 등(1998)은 포자형성과정에서 발현되는 유전자 중 각 시점별로 발현패턴을 가장 잘 반영할 것으로 고려되는 소수 개의 유전자를 엄선하였다. 7개의 시점에 대해서 엄선된 유전자들이 나타내는 발현 프로파일을 모델 프로파일(model profile)이라 한다. 이 모델 프로파일은 각 시점에 대해서 7가지 발현 패턴을 가지고 있다고 한다. 5가지 군집 알고리즘이 실제로 모델 프로파일의 패턴을 얼마나 잘 나타내는가를 비교하였다. 각 알고리즘에 대해서, 각 시점에서 7개의 군집의 수를 갖도록 군집분석을 수행하였다. 7개 시점에 걸쳐 각 군집 알고리즘에 의해 생성된 7개 군집 각각에 해당되는 유전자들의 로그 발현비의 평균값을 그림으로 도시하였다. <그림 2>는 5개 군집

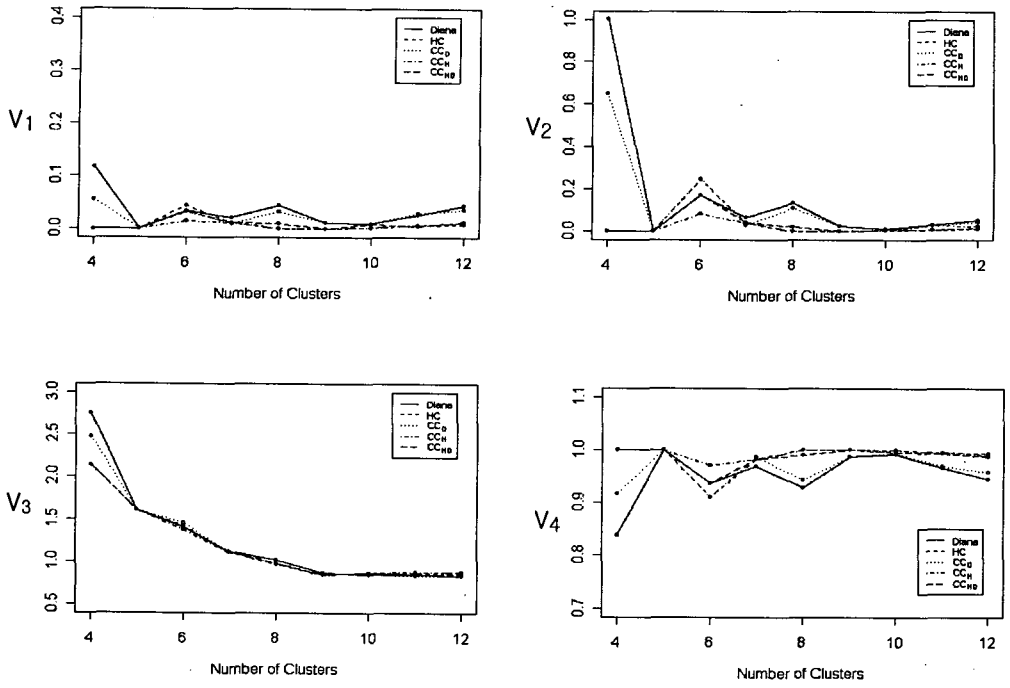
알고리즘에 의해 생성된 평균 발현 프로파일과 모델 프로파일을 나타낸 것이다. 시각적 판단에 의해 CC_{HD}의 발현 프로파일이 실제 모델 프로파일과 가장 근접하다는 것을 알 수 있다. 다음으로 CC_D가 모델프로파일과 유사한 패턴을 나타낸다. 한편, 기존의 방법인 *Diana*나 *HC*보다는 CC_D, CC_H가 모델프로파일에 더 근사한 패턴을 갖는다는 것을 알 수 있다.



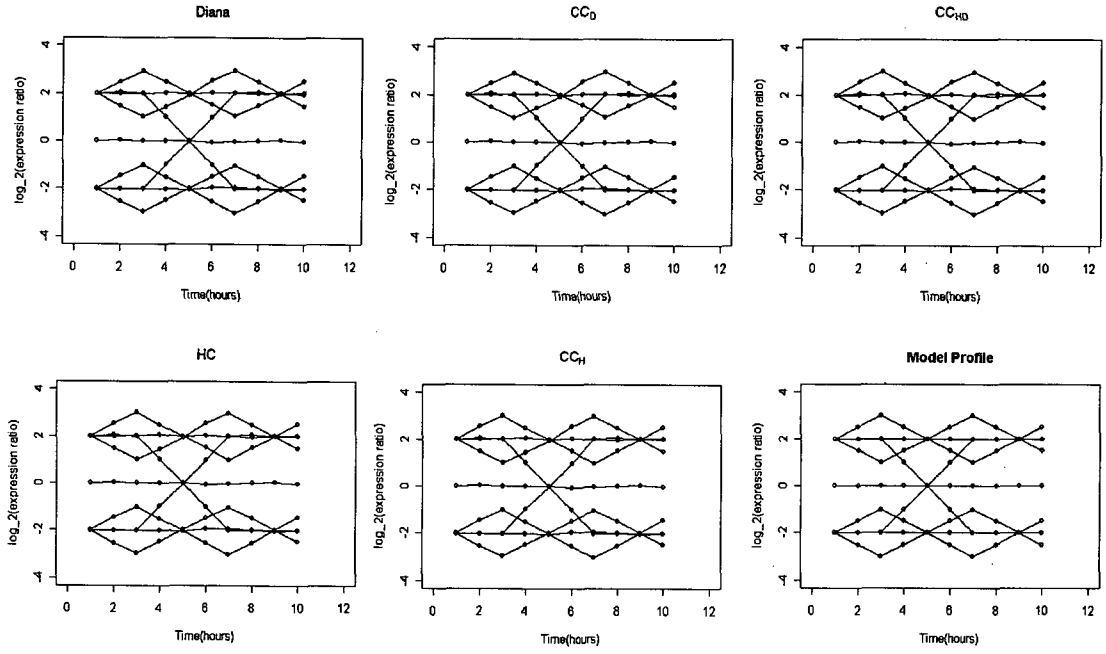
<그림 2> Sporulation자료에 대한 다양한 군집알고리즘의 평균프로파일과 모델프로파일

5.2 모의실험 자료의 분석결과

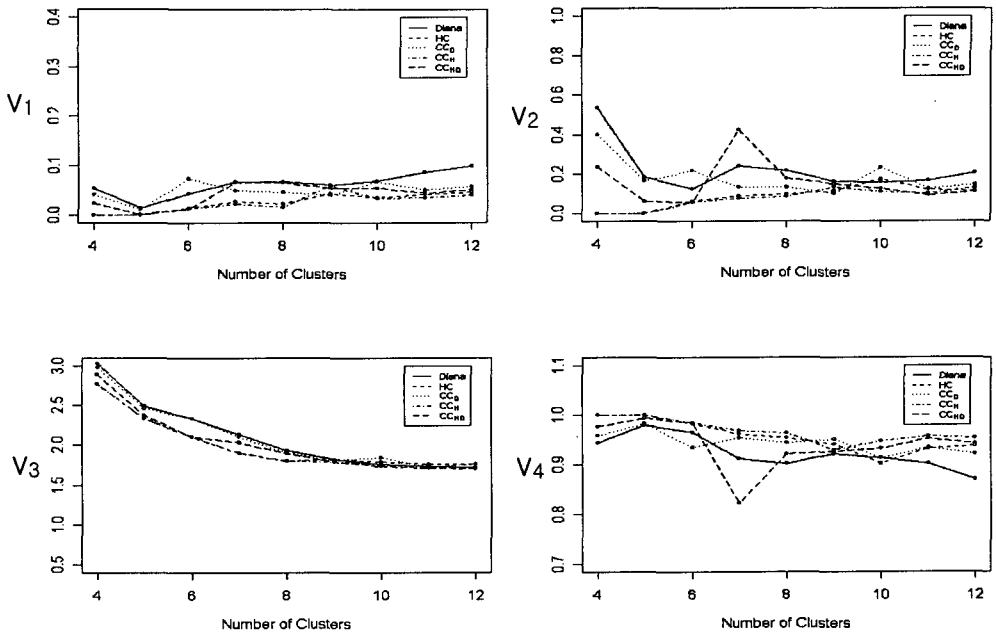
두 가지 모의실험 자료에 대해 분석하였다. <그림 3>과 <그림 4>는 모의실험 1에 대한 타당성 측도와 발현 프로파일을 나타낸다. <그림 3>으로부터 4가지 타당성 측도에 의해 CC_{HD}, CC_H가 가장 안정적인 평가 결과를 나타낸다. 특히, 이 모의실험자료 1의 경우 실제 군집의 개수는 9개인데, 군집의 수가 9개일 때는 CC_{HD}가 V_1, V_2, V_3에서 가장 낮은 측도값을 갖고, V_4에서는 가장 높은 측도값을 갖는 것으로 나타났다. 즉, 4가지 타당성 측도에 의한 평가 결과 CC_{HD}가 가장 안정적이고 실제 군집의 수를 정확하게 추정한다는 것을 알 수 있다. <그림 4>의 모델 프로파일과 5가지 군집 알고리즘에 의한 프로파일을 비교한 결과 정규분포에 의한 랜덤변수를 추가한 이 자료의 경우 5가지 알고리즘 모두 모델 프로파일과 거의 유사한 패턴을 갖는다는 것을 알 수 있다.



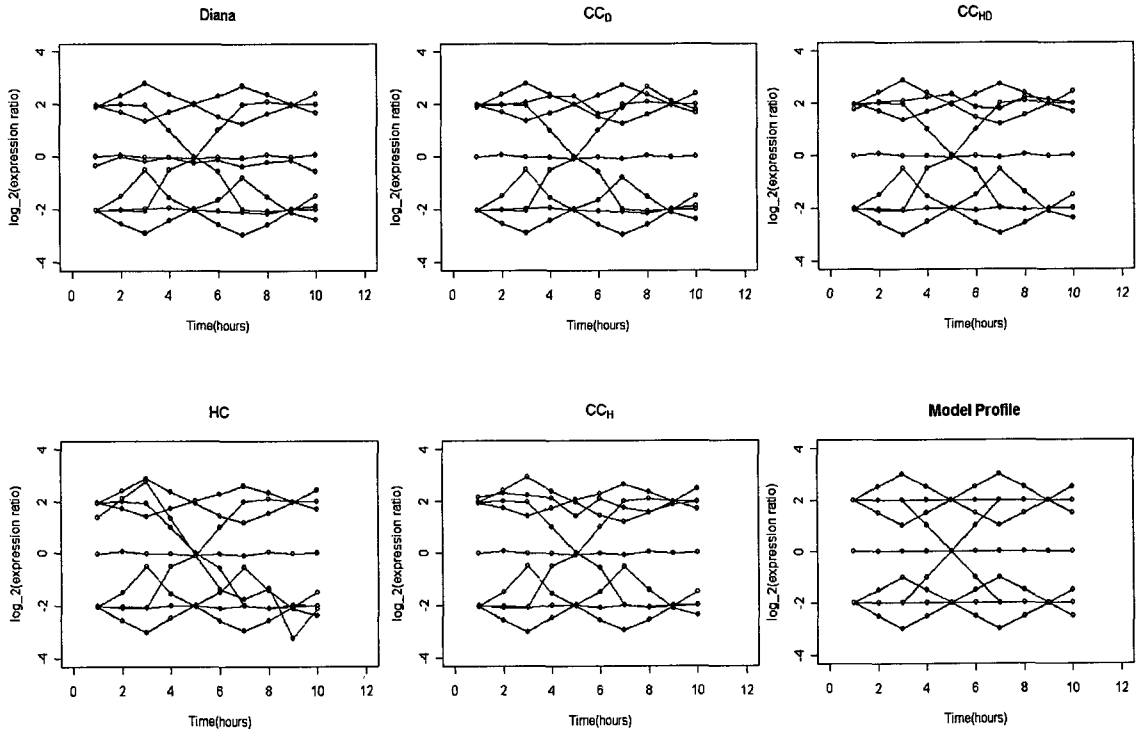
<그림 3> 모의실험 1에 적용된 다양한 군집알고리즘의 타당성 측도



<그림 4> 모의실험 1에 적용된 다양한 군집알고리즘의 평균프로파일과 모델프로파일



<그림 5> 모의실험 2에 적용된 다양한 군집알고리즘의 타당성 측도



<그림 6> 모의실험 2에 적용된 다양한 군집알고리즘의 평균프로파일과 모델프로파일

<그림 5>와 <그림 6>은 모의실험 2에 대한 타당성 평가 측도와 발현프로파일을 나타낸다. 모의실험 1과 유사하게 HC, CC_H가 4가지 측도에 대해서 안정적인 평가결과를 보여주었다. 또한 <그림 6>으로부터 CC가 기존의 방법보다 모델 프로파일과 더 유사한 패턴을 보이고, 특히 CC_{HD}가 가장 모델 프로파일과 유사하다는 것을 알 수 있다. 종합적으로 CC가 기존의 방법인 HC와 Diana에 비해 전체 군집의 개수에 걸쳐 안정적인 결과를 보여주었다. CC 내에서는 CC_{HD}가 우수한 평가결과를 보여주었다.

6. 결론 및 논의

본 논문에서는 시간 경로 마이크로어레이 유전자 발현 자료와 두 개의 모의실험 자료에 일치 군집 알고리즘을 적용하고, 다양한 타당성 측도에 의해 기존의 알고리즘과 비교하였다. 기존의 알고리즘으로는 일반적으로 사용되는 계층적 군집방법인 HC와 Diana를 사용하였다. 한편 HC와 Diana를 초기 알고리즘으로 하는 일치 군집과 이 두 가지 방법을 혼합한 일치 군집방법을 동시에 사용하여 이들의 수행능력을 비교하였다. 일치 군집(Monti 등, 2003) 알고리즘은 일치행렬의 누적분포함수라든가 시각적 그래프 등을 이용해서 자동적으로 최적 군집의 개수를 찾을 수 있는 잇점을 가지고 있다. 그러나 본 연구는 최적 군집의 개수를 찾는 문제보다는 시간 경로 발현 자료에서 시간의 영향력을 잘 반영함으로써 유전자들의 발현 패턴을 얼마나 잘 찾을 수 있는가에 초점을 두었다. 실제 시간 경로 유전자 발현 자료와 두 개의 모의실험 자료에서 기존의 방법보다는 일치 군집 방법이 전체적의 시점에 걸쳐 안정적인 수행결과를 보였다. 또한 일치 군집 방법 내에서는 HC와 Diana를 혼합한 CC_{HD}가 가장 우수한 수행능력을 나타냈다. 혼합된 CC 방법은 원래 CC 알고리즘이 초기 군집 알고리즘에 상당히 의존적이라는 점을 보완한 것으로, 혼합된 일치 행렬을 유사성 행렬(similarity matrix)로 사용함으로써 시간 경로 자료에 대해서 CC 방법이 지니고 있는 초기 알고리즘에 대한 의존성을 해결할 수 있다는 것을 알 수 있다.

시간 경로 유전자 발현 자료는 시간의 변화에 따른 생물학적 동적 진행과정에서 시간의 효과를 반영한 유전자들의 생물학적 발현 패턴을 찾는다거나 가장 활발한 진행이 나타나는 시점을 찾는 연구에서 자주 사용되고 있다. 그러나 일반적인 시계열 자료와는 달리 시점의 개수가 소수 개이고 시점간의 간격이 좁기 때문에 시계열 자료에서 사용되는 일반적인 기법들을 적용하는 것은 바람직하지 않다. 비모수적 기법들을 이용한 연구결과들이 발표되고는 있지만, 여전히 시간의 영향을 잘 반영함으로써 유전자들의 발현패턴을 효과적으로 찾을 수 있는 다양한 통계적 기법들이 개발되어야 할 것이다. 이와 더불어 대부분의 군집 분석에서 군집의 개수는 입력변수로 요구되고 있다. 따라서 최적 군집의 개수를 결정하는 문제 또한 안정적인 군집 알고리즘에 비추어 중요한 문제라 하겠다.

참고문헌

[1] Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, B., Gillette, M., Loda, M., Weber, G., Mark, E.J., Lander, E.S., Wong, W., Johnson, B.E., Golub, T.R., Sugarbaker, D.J., and Meyerson, M. (2001).

Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinomas sub-class. *Proceeding of the National Academy of Sciences*, 98, 13790-13795.

[2] Brown, P.O., Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *The chipping forecast,* 21, 33-37.

[3] Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J. and Davis, R.W. (1998). A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2, 65-73.

[4] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P.O. and Herskowitz, I. (1998). The transcriptional program of sporulation in budding yeast. *Science*, 282, 699-705.

[5] Datta. S. and Datta, S. (2003). Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics,* 19, 459-466.

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

[7] DeRisi, J.L., Iyer, V.R. and Brown, P.O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278, 680-686.

[8] Dudoit, S., and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Bilogy*, 3, research0036.

[9] Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceeding of the National Academy of Sciences*, 95, 14863-14868.

[10] Huber, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2, 193-218.

[11] Jain, A.K. and Moreau, J. (1988). Bootstrap techniques in cluster analysis. *Pattern Recognition*, 20, 547-568.

[12] Kim, S. Y. and Lee, J. W. (2004). Ensemble clustering method based on the resampling similarity measure for gene expression data. Submitted.

[13] Levine, E. and Domany, E. (2001). Resampling method for unsupervised estimation of cluster validity. *Neural Computation*, 13, 2573-2593.

[14] 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.

[15] Monti, S., Tamayo, P., Mesirov, J. and Golub, T. (2003). Consensus Clustering: A resampling based method for class discovery and visualization of gene expression microarray data. *Kluwer Academic Publishers.*

[16] 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.

[17] Quackenbush, J. (2001). Computional analysis of microarray data. *Nature Reviews,*

Genetics, 2, 418-427.

[18] Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D. and Futcher, B. (1998). Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* 9, 3273-3297.

[2004년 12월 접수, 2005년 4월 채택]