응용통계연구 제 19권 1호, 2006년, pp. 13–32

시간 경로 마이크로어레이 자료의 군집 분석에 관한 고찰*

손인석[1] 이재원[2] 김서영[3]

요 약

생물학자들은 시간에 따라 발현 수준이 변화하는 유전자의 군집화를 시도하고 있다. 지금까지는 마이크로어레이 자료의 군집분석에 관한 연구의 경우 군집 방법 자체를 비교하는 연구가 주를 이루었다. 그러나 군집화 이전에 의미있는 변화를 보이는 유전자 선택에 따라 군집화 결과가 달라지기 때문에, 군집 분석에 있어서 유전자 선택 단계도 중요하게 고려되어야 한다. 따라서, 본 논문에서는 시간 경로 마이크로어레이 자료를 군집분석하는데 있어서 유전자 선택, 군집 방법 선택, 군집평가 방법 선택 등 3가지 요인을 고려한 폭 넓은 비교 연구를 하였다.

주요용어: 시간 경로 마이크로어레이 자료, 유전자 선택, 군집 분석, 군집 타당성

1. 서론

DNA 마이크로어레이는 수많은 유전자의 발현 수준을 동시에 관찰할 수 있는 바이오 기술로서, 생물학 또는 의학계에 걸쳐 종양 분류와 같은 광범위한 문제를 연구하는 데 많은 도움을 줄 수 있는 새롭고 유망한 기술로 각광받고 있다. 마이크로어레이 자료 분석의 핵심 목표 중의 하나는 시공간적으로 변화하는 유전자들의 발현패턴의 차이를 확인하고, 유전자들 사이의 상호작용과 기능의 연계성을 파악하는데 있다(Chu 등, 1998; Spellman, 1998). 특히, 유전자들의 시간에 따른 발현수준의 변화를 고려함으로써 발현패턴에 기초한 유전자들의 그룹을 찾고자 하는 것이 연구의 목표 중 하나라 할 수 있다. 이에 유사한 유전자 그룹을 찾기 위한 방법으로 통계적 군집분석이 널리 사용되고 있다. 그러나 시간경로자료의 경우 일반자료와는 달리 매 시점마다에서의 발현수준의 변동이 크기 때문에 동일한 군집방법이라도 일반 마이크로어레이 자료에서와 시간경로자료에서의 군집 수행능력이 다를 것이다. 따라서 본 논문은 시간에 따른 발현수준들의 변동을 고려함으로써 다양한 군집 알고리즘들의 특성을 비교하고자 하는데 연구의 목표를 두었다.

마이크로어레이 자료의 군집을 위한 많은 방법(Goldstein 등, 2002)들이 제안되었다. Eisen 등(1998)의 계층적 군집(hierarchical clustering)방법의 적용을 시작으로,Tamayo 등

* 본 연구는 Korea Science and Engineering Foundation Grant (R14-2003-002-01002-0)에 의해 지원받았고, 김서영은 한국과학재단 목적기초연구(R08-2003-0000-10572-0)지원으로 수행되었음.

1) (130-701) 서울시 성북구 안암동 5가 1번지, 고려대학교 통계학과, 박사과정
E-mail : sis46@korea.ac.kr
2) (130-701) 서울시 성북구 안암동 5가 1번지, 고려대학교 통계학과, 교수
E-mail : jael@korea.ac.kr
3) (교신저자)(500-757) 광주광역시 북구 용봉동 300, 전남대학교 기초과학연구소, 전임연구원
E-mail : gong@chonnam.ac.kr

(1999)은 Self-Organizing Map(SOM)을 적용하여 마이크로어레이 자료를 분석하는 방법을 제시하였다. Hastie 등(2000)은 계층적 군집분석과 주성분 분석을 혼합한 Gene Shaving방법과 함께 찾아진 군집간의 교호작용을 모형화하기 위하여 Tree Harvesting방법을 제시하였다. Laura와 Owen(2002)은 유전자들이 하나 이상의 군집에 속하거나 어떤 군집에도 속하지 않을 수도 있다는 2단계 군집알고리즘인 모형기반 군집(Model-based Clustering)방법을 제안하였다. McLachlan 등(2002)은 혼합 모형에 기반한 군집방법을 제안하였고, Ghosh와 Chinnaiyan(2002)은 군집 결과의 신뢰도를 평가하기 위해 혼합모형에 기반한 방법을 사용하였다. 또한 Barash와 Friedman(2002)의 베이지안 군집(Bayesian Clustering)방법이나 Waddell와 Kishino(2000)의 비정칙 분해(Singular Value Decomposition)를 이용한 방법 등도 사용되었다. 최근에는 Luan과 Li(2003)가 B 스플라인 함수에 기초한 혼합효과모형을 이용한 군집방법을 제안하였고, Kasturi 등(2003)은 시간에 따른 유전자들의 발현패턴을 분석하기 위해 정보이론(theoretic information)에 근거한 군집방법을 제안하였다. Hong과 Li(2004)는 B 스플라인 함수에 기반을 둔 계층적 모형에 의한 군집방법을 사용하였다.

한편, 군집 알고리즘의 신뢰성 및 타당성 평가를 위한 다양한 군집분석 방법이 연구되었다. Kerr 등(2001)은 재표본(resampling) 기법에 의한 분석방법을 제안하였고, 이들은 재표본하는 과정에서 잔차와 선형모형을 사용하였다. Chen 등(2002)은 군집 결과에 대해 동질성(Homogeneity)과 이질성(Separation)과 같은 물리적인 특성으로 여러 가지 군집 알고리즘의 성능을 비교하였다. Yeung 등(2001)은 Figure Of Merit (FOM)의 개념을 소개함으로써 잭나이프(Jackknife) 방법에 의해 군집결과를 평가하는 방법을 사용하였고, Datta와 Datta(2003)는 3가지의 타당성 기준을 사용하여 군집 알고리즘에 의해 생성된 그룹들의 일치성을 평가하였다.

이처럼 군집분석 방법은 마이크로어레이 자료와 같은 대용량의 자료로부터 유용한 정보를 추출하기 위한 필수적인 도구로 사용되고 있다. 그러나 다양하게 적용되고 있는 군집방법들 각각은 서로 다른 특성을 가질 뿐만 아니라 군집분석에 사용된 유전자에 따라 혹은 군집결과를 평가하는 평가방법에 따라서 각 방법들의 수행능력 결과는 다르게 나타난다. 지금까지의 군집 방법에 대한 많은 연구들은 군집방법 자체에 대한 비교가 주를 이루었다. 그러나 군집화 이전의 유전자 선택 방법(Dudoit 등, 2002; Smyth 등, 2003)에 의해 군집 결과가 달라지기 때문에 유전자 선택 단계도 군집 방법과 동시에 고려될 필요가 있다. 유전자 선택 방법으로 Fold Change와 T-통계량, B-통계량(Lonnstedt와 Speed, 2002), 그리고 SAM(Tusher 등, 2001)을 사용하였다. Cyanobacterium sp. PCC 6803(Hihara 등, 2001) 자료를 적용하여 4 가지 유전자 선택 방법과 7가지 군집방법에 따라 군집분석을 수행하고, 6가지 타당성 평가방법에 따라 각 군집 알고리즘을 평가함으로써 폭 넓은 비교 연구 결과를 제시하였다.

2. 군집 분석에서의 요인

2.1. 유전자 선택

마이크로어레이 실험에서 각 유전자에 대한 자료는 두 개의 발현강도로 측정된다. 즉 (R, G)는 mRNA샘플에 부여된 적색(Cy3)과 녹색(Cy5) 형광물질에 의해 측정된 발현수준을 나타낸다. 이때 두 발현강도 비의 로그 값을 $M = logR/G$라 하면, $\bar{M}$ 는 유전자에 대한 반복된 어레이에서 값의 평균, s는 M에 대한 표준편차를 나타낸다. 유전자 선택 방법으로 다음과 같이 4가지 방법을 고려한다.

(1) T-통계량

T-통계량은 강한 모수적 가정을 필요로 하는 통계량으로서, 일반적으로 반복의 수가 많을 때 널리 사용되는 통계량이다. 그러나 현실적으로 마이크로어레이 실험은 반복이 적으며, 정규분포라는 모수적 가정을 만족한다는 확신도 따르지 않는다. 특히, 마이크로어레이 실험의 경우 유전자의 발현수준이 낮은 경우에는 그룹 간 평균 차이에 비해 표준오차가 너무 작기 때문에 통계량이 커지는 결과를 초래함으로써 위양성률(false positive rate)을 증대시킬 수 있다. T-통계량은

$$t = \frac{\bar{M}}{s/\sqrt{n}}$$

와 같다. 여기서 $\bar{M}$ 는 평균이고, s는 표준편차이다.

(2) SAM

SAM(Significant Analysis of Microarrays, Tusher et al., 2001)은 T-통계량을 변형한 통계량으로서 강한 모수적 가정에 의존하지 않고 경험적 분포를 사용하는 비모수적 방법이다. T-통계량의 발현 수준이 낮은 유전자에서 평균에 비해 분산이 너무 작기 때문에 생기는 문제점을 보완한 방법으로, T-통계량의 분모인 분산에 일정한 상수(s_0)를 더해줌으로써 분산안정을 유도한다. SAM 통계량은 다음과 같다.

$$d = \frac{\bar{M}}{(s + s_0)/\sqrt{n}}.$$

여기서 s는 표준편차이고, s_0는 분산 안정화 상수이다. 이때 상수 s_0는 SAM 통계량의 변동계수(coefficient of coefficient)를 최소화하는 값으로 선택한다. SAM은 이론적 가정보다는 경험적 방법을 사용하기 때문에 표본의 수가 적은 경우에 보다 효율적인 것으로 알려져 있다. 또한 T-통계량의 단점을 상당히 개선하였으며 자료가 정규분포를 따를 때는 T-통계량과 거의 유사한 수행 결과를 나타낸다.

(3) B-통계량

SAM과 마찬가지로 유전자의 낮은 발현수준에서 분산 안정화를 위해 개발된 방법이다.

Lonnstedt and Speed(2002)는 각 유전자에 대해서 경험적 베이지안 방법에 의해, 사후 로그 오즈를 추정하여 이를 분산 안정화 상수로 사용하였다. B 통계량은 다음과 같다.

$$B = log\frac{p}{1-p}\frac{1}{\sqrt{(1+nc)}}\left[\frac{a+s^2+\bar{M}^2}{a+s^2+\frac{M^2}{1+nc}}\right]^{v+n/2}.$$

여기서 p는 유전자의 사전 비율이고, c는 평균에 대한 상위모수(hyper parameter)이다. a와 v는 분산에 대한 상위모수로서, 조건부 정규성 가정하에서 모수 a와 v가 추정된다. B-통계량 값이 0보다 크면, 유전자들이 다르게 발현될 가능성이 50%를 넘는다는 것을 의미한다. B-통계량의 단점은 유의한 유전자에 대한 사전비율 p를 사전에 결정해야 한다는 것과 원 자료가 정규분포를 따른다면 이 통계량에 의한 유전자 선택에는 다소 문제가 발생한다는 것이다. 한편, 유의한 유전자를 선택할 때 연구자가 B-통계량에 대한 임계값(cut-off value)을 임의로 정할 수 있어 유의한 유전자에 대한 위양성률 조절할 수 있고 유전자 수가 상당히 많을 때 위음성률(false negative rate)을 줄일 수 있다는 장점을 가진 것으로 알려져 있다.

(4) Fold Change

$\bar{x}_{i1}, \bar{x}_{i2}$가 각각 첫 번째와 두 번째 조건 하에서 i번째 유전자의 평균 발현수준이라고 할 때, 각 유전자에서 $|\bar{x}_{i1}/\bar{x}_{i2}| \geq m$이거나 $|\bar{x}_{i1}/\bar{x}_{i2}| \leq 1/m$이면 m 배수 (m-fold)하에서 유의하게 다르게 발현되는 유전자라고 판단한다.

2.2. 군집화 방법

(1) 계층적 군집

계층적 군집 방법(Eisen 등, 1998)은 사전에 고정된 군집의 개수 대신 계층적으로 군집을 생성한다. 초기 수준에서는 관찰치 각각에 대해서 각자의 군집을 형성하고, 다음 단계에서 가장 가까운 두 군집을 결합하여 더 큰 하나의 군집을 형성함으로써 최종적으로 모든 관찰치를 포함하는 하나의 군집을 형성한다. 계층적 군집 방법은 가장 일반적으로 사용되는 병합적(agglomerative) 군집방법으로 간단한 나무구조 방법이다.

(2) Diana

Diana(Datta 와 Datta, 2003)는 분리적(Divisive)인 계층적 군집방법이다. Diana는 초기에 모든 관찰치들이 하나의 군집을 이룬 후, 이를 유사성 있는 개체들끼리 군집을 나누는 방법이다. Diana는 많은 관찰치를 포함하거나 소수 개의 군집으로 분류하고자 하는 경우에 유용한 군집방법으로 알려져 있다.

(3) K-means

K-means 방법(Hartigan 등, 1979)은 사전에 군집 수를 고정하고, 초기 군집을 중심으로

하여, 그룹 내 총제곱합(total within-class sum of squares)을 최소화하기 위해 관찰치들을 다양한 군집으로 할당한다. 최소값을 찾기 위해 복잡하고 반복적인 수치 알고리즘이 사용되고 있다.

(4) Fanny

Fanny 방법(Datta와 Datta, 2003)은 fuzzy 논리를 이용하고 각 관찰치에 대해서 각 그룹에 할당될 확률 벡터가 생성된다. Fanny는 원하는 군집의 개수를 정한 후에 비유사성 거리의 가중 평균이 되는 목적함수를 최소화하는 모든 유전자 확률 벡터를 계산한다. 가장 높은 확률을 가지는 그룹에게 유전자를 할당함으로써 최종 군집이 결정된다.

(5) Fuzzy c-means

Fuzzy c-means 방법(Guthke 등, 2000)은 패턴 인식에서 자주 사용된다. 이 알고리즘은 K-means와 유사한 hard c-means방법을 퍼지화한 군집방법이다. 이 알고리즘은 각 군집중심과 fuzzy c-분할 행렬을 동시에 구할 수 있는 자기 조직화(self-organization) 및 무 관리자 학습의 대표적인 방법이다.

(6) PAM

PAM(Partitioning Around Medoid)은 분할법에 해당되는 군집화 방법으로서 Kaufman과 Rousseeuw (1990)가 제안한 방법이다. 이 알고리즘은 관찰치들의 대표값인 메도이드(medoid)를 이용한다. 이 k개의 메도이드를 찾은 후에 각각의 메도이드와 가장 가까운 점들을 할당시켜 군집을 형성한다. k-메도이드를 찾는 목적은 관찰치의 점들과 가장 가까운 메도이드와의 거리의 합을 최소화하는 데에 있다. 이 방법은 K-means 방법보다 로버스트하고 효율적으로 계산하는 경향이 있다고 알려져 있다.

(7) 모형 기반 군집

Laura와 Owen(2002)이 제안한 모형 기반 군집(model based clustering)방법은 모든 자료를 혼합분포에서 온 것이라고 가정한다. i번째 관찰치의 그룹 수준을 γ_i이라고 하자. $f_j(.;\Theta_j)$를 그룹 j에 속하는 관측치의 밀도함수라고 하고, Θ_j는 미지의 모수라고 하고, 발현 프로파일들 (profiles) $x_1, \cdots, x_n$의 우도는

$$L(\Theta, \gamma) = \Pi_{i=1}^{n} f_{\gamma_i}(x_i, \Theta_{\gamma_i})$$

와 같고, 알려지지 않은 그룹 수준 γ_i는 $L(\Theta, \gamma)$를 최대화시키는 최대우도 방법에 의해 얻어진다.

2.3. 타당성 평가 방법

(1) Homogeneity와 Separation

Chen 등(2002)은 동질성과 이질성과 같은 군집 결과의 물리적인 특성으로 군집 알고리즘의 성능을 평가하였다. 동질성(Homogeneity)은 각 유전자의 발현 프로파일과 각 유전자가 속하는 군집의 중심사이의 평균거리를 계산하며

$$H_{avg} = \frac{1}{N}\sum_{i} D(g_i, C(g_i))$$

와 같다. 여기서 g_i는 i번째 유전자, $C(g_i)$는 g_i가 속하는 군집의 중심, N은 유전자 총 개수, D는 거리함수를 나타낸다. 이질성(Separation)은 군집 중심들 간의 가중치 평균 거리를 계산하며

$$S_{avg} = \frac{1}{\sum_{i\neq j} N_{C_i} N_{C_j}} \sum_{i\neq j} N_{C_i} N_{C_j} D(C_i, C_j)$$

와 같다. 여기서 C_i와 C_j는 i번째와 j번째 중심이고, N_{C_i}와 N_{C_j}는 i번째와 j번째 군집에 소속된 유전자의 개수이다. 본 자료의 경우, 한 시점씩을 제외한 자료에 대해 전체를 각각 군집분석을 수행하고 각 군집결과에 대한 Homogeneity와 Seperation을 구한 후 전체를 평균함으로써 Homogeneity와 Seperation을 구하였다. 이때 Homogeneity 는 척도 값이 작을수록, Separation은 척도 값이 클수록 좋은 군집 알고리즘으로 평가된다.

(2) Aggregate figure of merit (FOM)

Yeung 등(2001)은 군집결과의 평가를 위한 다음과 같은 방법을 제안하였다. 유전자 발현자료가 l시점에 걸쳐서 관측되었다고 하고, 시점을 $t_1, t_2, \cdots, t_l$이라고 하자. 각 시점 $i = 1, 2, \cdots, l$과 유전자 $g = 1, 2, \cdots, M$ 에 대해서 K개의 군집 $C_1, C_2, \cdots, C_k$이라고 할때, $R(g, i)$는 t_i번째 시점에서 유전자 g의 발현 프로파일을 나타내고, $C_j(i)$은 C_j군집에 있는 t_i번째 시점에 포함된 유전자들의 평균발현 프로파일을 나타낸다.

$$FOM(i, K) = \sqrt{\frac{1}{M}\sum_{j=1}^{k}\sum_{g\in C_j}(R(g,i) - C_j(i))^2},$$

Aggregate figure of merit(FOM), $FOM(K) = \sum_{i=1}^{l} FOM(i, K)$는 K군집에 대한 총 예측력의 추정이다. 이 척도는 값이 작을수록 좋은 알고리즘으로 평가한다.

(3) Average proportion of non-overlap measure (V1)

유전자 발현자료가 l시점에 걸쳐서 관측되었다고 하고, 시점을 $t_1, t_2, \cdots, t_l$이라고 하자. 각 시점 $i = 1, 2, \cdots, l$ 과 유전자 $g = 1, 2, \cdots, M$에 대해서, $C^{g,i}$은 t_i번째 시점에서 관찰값이 제외된 자료에 대해서 군집분석을 수행했을 때 유전자 g가 포함된 군집이라 하고,

$C^{g,0}$는 모든 시점에 걸쳐 군집분석을 수행했을 때의 유전자 g가 포함된 전체군집이라 하고, K는 군집의 수라 하자. 위의 기호는 식 2.1, 2.2, 2.3에 적용되며, Average proportion of non-overlap 척도는 다음과 같다.

$$V_1(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\left(1 - \frac{n(C^{g,i} \cap C^{g,0})}{n(C^{g,0})}\right) \tag{2.1}$$

이 척도는 한 번에 한 시점을 제외시켰을 때의 자료에 대해 군집 분석을 수행한 결과와 모든 시점에 걸친 전체 데이터에 대해 군집분석을 수행했을 때의 결과가 일치하지 않은 유전자들의 평균비율을 계산한다.

(4) Average distance between means measure (V2)

$$V_2(K) = \frac{1}{Ml}\sum_{g=1}^{M} d(\bar{x}_{C^{g,i}}, \bar{x}_{C^{g,0}}) \tag{2.2}$$

여기서 $\bar{x}_{C^{g,0}}$은 $C^{g,0}$에 있는 유전자들의 평균 발현 프로파일을 나타내고, $\bar{x}_{C^{g,i}}$는 $C^{g,i}$에 있는 유전자들의 평균 발현 프로파일을 나타낸다. 이 척도는 전체 데이터와 시점마다 발현 수준을 제거한 데이터에 대해서 각각 군집 분석을 수행한 후, 각각의 결과로부터 동일한 군집 내에 포함되어 있는 유전자들의 평균간의 거리(평균거리)를 계산한다.

(5) Average distance measure (V3)

$$V_3(K) = \frac{1}{Ml}\sum_{g=1}^{M}\sum_{i=1}^{l}\frac{1}{n(C^{g,0})n(C^{g,i})} \times \sum_{g\in C^{g,0}, g'\in C^{g,i}} d(x_g, x'_g) \tag{2.3}$$

여기서 $d(x_g, x'_g)$는 유전자 g와 g'발현 프로파일간의 거리이다. 이 척도는 한 번에 한 시점을 제외시켰을 때의 군집 결과와 전체 시점에 걸친 자료에 대한 군집 결과에서 각 유전자들의 발현 프로파일간의 평균거리를 계산한다. V_1, V_2, V_3와 같은 타당성 평가 척도들은 자료에서 한 시점의 관측치들을 제외했을 때의 군집에 대한 안정성 또는 일관성을 비교하는 방법으로, 척도값들이 작을수록 좋은 군집 알고리즘으로 평가된다.

3. 자료 분석 결과

본 연구에서는 Hihara 등(2001)의 cDNA 자료에 대해 유전자 선택방법을 고려하여 다양한 군집 평가방법에 의해 군집 알고리즘의 성능을 평가하였다. 유전자 선택 방법으로 T-통계량, Fold change, SAM, B-통계량 등 4가지 방법을 사용하였고, 선택된 유전자 자료

에 대해서 계층적 군집, Diana, K-means, Fanny, Fuzzy c-means, PAM, 모형 기반 군집방법 등 7가지 군집 알고리즘들을 적용하여 분석하였다. 각 군집 결과에 대해서 Homogeneity, Separation, FOM 등 6가지 군집 평가방법을 적용하여 다양한 측면에서 군집 알고리즘의 성능을 평가하였다.

3.1. 자료 및 방법

Hihara 등(2001)의 Cyanobacterium sp. PCC 6803 자료는 광합성을 하는 Cyanobacteria를 낮은 빛의 환경에서 유지시키다가 높은 빛의 환경으로 변화시킨 후 그 경과 시간에 따른 유전자의 발현 양상을 관찰한 실험 결과이다. 광합성 유기체들은 빛의 강도 변화에 잘 순응해야한다는 특징을 가지고 있다. Cyanobacterium sp. PCC 6803은 낮은 빛에서 높은 빛으로의 변화에 영향을 받는 유기체로서, 전체 유전자의 서열이 알려져 있고, 모든 ORF를 이용할 수 있기 때문에 빛의 변화에 따른 광합성 실험에 적합하다. 낮은 빛과 높은 빛에 노출된 모든 RNA에 대해서 각각 Cy3(빨간색), Cy5(녹색)의 형광물질로 염색하여 합성한 후 DNA 마이크로어레이를 통해 각 유전자의 발현수준을 측정하였다. 낮은 빛 조건에 있는 Cyanobacteria를 대조군(Cy3)으로 하여 높은 빛(Cy5)에 15분, 1시간, 6시간, 15시간 노출시킨 Cyanobacteria에서 3079개의 ORF의 발현정도를 측정하였다. 각 시간에서 측정된 반복수는 15분에서 6번, 1시간에서 6번, 6시간에서 4번, 15시간에서 4번이다.

.자료 전처리는 Moon 등(2002)과 동일하게 하였다. 첫째, Background 보정을 하였고, 둘째, Cy3 또는 Cy5의 강도가 2000보다 작은 유전자는 제거함으로써 912개의 유전자를 선택하였으며, 셋째, 유전자 912개에 대해 Global Mean Normalization을 실시하였다. 각 시점에서 낮은 빛(대조군)과 높은 빛에서의 발현강도의 차이가 있는 지를 검정함으로서 각 검정방법에 의해 유의한 유전자를 선발하였다. 즉, Fold Change 의 경우 한 시점에서 2개 이상 반복에서 2 Fold 이상 차이 나는 유전자를 유의한 발현차이를 갖는 유전자로 선택하였다. T-통계량과 B-통계량은 유의수준을 각각 0.0001을 사용하고, SAM은 $\Delta = 1.4$ 를 임계치로 사용하여 최소한 한 시점에서 유의하다고 판단되는 모든 유전자를 최종 분석을 위한 유전자로 선택하였다.

3.2. 분석결과

3.2.1. 유전자 선택 결과

표 3.1은 분석에 사용된 912개의 유전자 중에서 각 통계량에 의해 선택된 유의한 유전자 개수를 나타낸다. B-통계량과 T-통계량은 유의수준 0.0001하에서 각각 119개와 90개, SAM은 임계값 기준 1.4에서 112개를 유의한 발현차이를 갖는 유전자로 선택하였고, Fold change 방법은 2 fold 기준에 의해 119개를 선택하였다. 통계량마다 적용되는 임계값 기준이 다르기 때문에 유전자의 개수를 대략 100여개 정도로 선택하고자 하였다.

또한 표 3.1에는 제시되지 않았지만, SAM과 B-통계량에 의해서만 유의하다고 선택된 유전자가 각각 8개 (7%)와 10개 (8%)인 반면, Fold change 와 T-통계량 의해서만 유의하다고 선택된 유전자는 각각 21개 (8%)와 19개 (21%)에 해당된다. 유전자 선택을 위한 임계값

표 3.1: 유전자 선택방법에 따라 선택된 유전자 개수

선택방법	Fold change	B	T	SAM
유전자 개수	119	119	90	112

기준이 서로 다르긴 하지만, 예를들어 T-통계량의 경우 4가지 통계량 중에서 가장 적은 개수의 유전자를 선택했음에도 불구하고, SAM이나 B-통계량에 비해 상대적으로 많은 비중을 차지한다. 이로부터 Fold change와 T-통계량은 다른 방법에 비해 상대적으로 위양성 또는 위음성 판단 가능성이 높다는 것을 예측할 수 있다. Fold change는 자료의 변동을 전혀 고려하지 않을 뿐만 아니라, 유전자 선택기준이 매우 주관적이기 때문에 유의한 발현차이를 갖는 유전자 선택에 있어서 가장 원시적인 방법에 해당된다. T-통계량은 마이크로어레이 자료 분석에서 유의한 발현차이를 갖는 유전자를 선택하는데 자주 사용되고 있는 방법이지만, 강한 분포적 가정과 많은 반복수를 필요로 하기 때문에 오히려 실험 회수가 적은 마이크로어레이 자료 분석에 적합하지 않은 경우가 빈번히 발생하게 된다.

3.2.2. 군집 평가방법에 따른 알고리즘 성능

표 3.2는 유전자 선택 방법에 의해 선택된 유전자 셋에 대해 다양한 군집방법을 적용하여 분석하고, 여러 가지 평가방법에 의해 각 알고리즘의 성능을 평가한 것이다. 표 3.2의 결과를 토대로 다양한 유전자 선택방법과 군집 평가방법들을 고려하여 군집 알고리즘들의 특성과 성능을 비교하였다. 표 3.2의 통계량 내에서 군집 평가방법별로 가장 우수하다고 평가된 알고리즘들을 진하게 표시하였다.

(1) Homogeneity

Homogeneity은 유전자와 군집 중심 간의 평균거리가 작을수록 우수한 성능을 갖는 군집 알고리즘으로 평가한다. 동질성에 의한 척도 값이나 척도 값의 순서에 의해 비교한 결과 PAM, Diana, K-means, FCM은 우수한 성능을 갖는 알고리즘으로 평가된 반면에 모형기반, Fanny, 계층적 방법은 성능이 별로 좋지 않은 것으로 평가되었다. Homogeneity 측도 값이나 순위 입장에서 PAM, 모형기반 군집방법, K-means, Diana는 유전자 선택방법에 대해서 상당히 안정적인 특성을 지닌 반면, Fanny는 유전자 선택방법에 따라 민감한 반응을 보이는 것으로 나타났다. 이처럼 Homogeneity 평가방법은 군집 내에 속한 유전자들의 유사성 정도를 측정하기 때문에, PAM과 FCM과 같이 유사한 유전자를 갖는 소수개의 군집으로 쪼개는 분할적 알고리즘들이 약간 우세하게 평가된 것으로 볼 수 있다.

(2) Seperation

Separation은 서로 다른 군집 중심들 간의 가중치 평균거리가 클수록 우수한 성능을 갖는 알고리즘으로 평가한다. 이 방법에 의해 Diana와 계층적 군집방법은 측도값이나 측도

값 순위에 면에서 우수한 성능을 갖는 알고리즘으로 평가된 반면, 모형기반, Fanny, FCM은 가장 성능이 낮은 방법으로 평가되었다. 또한 Diana, 계층적, PAM 과 같은 방법은 유전자 선택 방법에 따라 안정적인 특성을 보였지만, 모형기반, Fanny 와 같은 방법은 유전자 선택에 따라 Seperation 측도값이나 순위 면에서 상당히 민감한 군집결과를 보여주었다. Seperation은 서로 이질적인 유전자들 간의 거리를 측정하는 방법으로 유전자와 군집 간 비유사성 거리에 의해 군집을 쪼개는 Diana와 계층적 방법과 같은 알고리즘들을 우세하게 평가하는 경향이 있다. 또한 상대적으로 Homogeneity에서 우세한 성능을 보였던 PAM이나 K-means, FCM과 같은 방법들은 이 척도에 의해서는 군집 성능 면에서 그다지 좋지 않은 방법으로 평가되었다.

이처럼 Seperation과 Homogeneity는 군집 간 이질성과 군집 내에서 동질성이라는 상반된 군집 특성을 이용하여 알고리즘의 성능을 평가하기 때문에 군집을 쪼개는 방법에 따라 서로 다르게 평가된다고 볼 수 있다.

(3) FOM

FOM은 잭나이프(jackknife; Efron, 1982) 접근법에 의한 평가 방법으로 척도 값이 낮을수록 우수한 군집 알고리즘으로 평가된다. FOM에 의해서 PAM과 FCM은 우수한 군집 성능을 갖는 알고리즘으로 평가되었고, Diana와 Fanny는 척도 값이나 척도의 순위 면에서 보통으로 평가되었다. 그러나 모형기반, K-means, 계층적 방법은 별로 좋지 않은 평가 결과를 보여 주었다. 2.3절에서 설명한 바와 같이 FOM은 K개의 군집에 대한 총 예측력의 추정치로서, 전체 m개의 시점 중에서 m-1개의 시점에 해당되는 유전자들의 프로파일에 대해 각 군집 알고리즘을 적용하고, 나머지 한 시점의 발현프로파일은 군집 알고리즘의 예측력을 평가하는데 사용되는 원리이다. 따라서 시간 경로 마이크로어레이 자료와 같이 특정 시점에서 특별한 발현 양상을 나타낼 수 있는 경우에 군집 알고리즘은 그 특정 시점에 영향을 받을 수 있다. 또한 PAM이나 Fanny는 유전자 선택방법에 로버스트한 군집 결과를 보였지만, 모형기반이나 K-means, 계층적 방법 등은 유전자 선택에 대해서 상당히 민감한 군집 결과를 보여주었다. 특히 K-means 방법은 초기 군집 중심을 랜덤하게 선택한 후, 각 유전자를 군집 중심과 유전자간 거리에 의해 가까운 군집에 할당하고, 모든 유전자가 할당된 후에 k개의 새로운 군집 중심이 계산되기 때문에 초기 군집의 중심 설정뿐만 아니라 차후 군집에 할당되는 유전자 셋에 따라서 군집 결과가 영향을 받을 수 있다. 이러한 이유에서 FOM과 같이 모형설정과 예측에 사용되는 시점이 다른 경우에 K-means는 유전자 선택방법에 따라 민감한 결과를 제시한다고 보여진다.

(4) V1

V1은 leave-one out 교차 타당성 방법으로 척도 값이 작을수록 우수한 알고리즘으로 평가한다. 이 척도에 의해 Diana, 계층적 방법과 모형기반 방법이 우수한 성능을 갖는 알고리즘으로 평가되었고, FCM, Fanny, K-means는 좋지 않은 방법으로 평가되었다. V1은 전체 시점에 대해서 분석한 결과와 한 시점을 제외하고 분석한 군집 결과의 비일치적 평균비율

을 측정하기 때문에 이 척도에 의해 우수한 평가를 받은 알고리즘들은 시점에 대해서 매우 안정적인 결과를 제시한다고 볼 수 있다. 한편 FCM과 Fanny는 퍼지 알고리즘으로서 각 유전자에 대해서 모든 군집에 할당 가능한 확률을 계산하고, 최종적으로 가장 큰 확률을 갖는 군집으로 할당되는 원리를 사용한다. 따라서 각 유전자가 모든 군집에 할당될 비율이 유사할 경우, 아주 적은 확률차이로 군집 소속이 달라지기 때문에 특히 비일치적 비율을 사용하는 이 척도에 대해서 매우 불안정한 군집 결과를 나타낼 수 있다. 또한 이 척도에 의해, 계층적, PAM, K-means 방법은 유전자 선택 방법에 대해서 로버스트한 반면, FCM, 모형기반, Fanny 방법은 상당히 민감한 결과를 보였다. 특히, 분포적 가정을 필요로 하는 모형기반 방법은 성능 면에서는 우수하게 평가되었지만, 유전자 선택방법에 대해서는 매우 민감하게 작용하였다.

(5) V2

V1과 마찬가지로 교차타당성 방법을 적용하고, 척도값이 낮을수록 우수한 알고리즘으로 평가한다. 이 척도에 의해 PAM과 Diana는 우수한 군집 알고리즘으로 평가되었고, 모형기반과 K-means는 성능 면에서 별로 좋지 않은 것으로 평가되었다. 또한 PAM, Diana, K-means는 유전자 선택에 안정적인 반면에, FCM과 모형기반 방법은 유전자 선택에 약간 민감한 반응을 보였고, Fanny와 계층적 방법은 상당히 민감한 결과를 보였다. 이 척도는 전체 시점을 사용한 군집결과에 대해서 각 군집에 속한 유전자들의 평균 프로파일과 한 시점을 제외한 군집 결과로부터 각 군집에 속한 유전자들의 평균 프로파일 간의 유사성 정도를 평균거리 개념을 적용하여 군집결과를 평가하기 때문에 거리개념에 의한 유사성을 이용하는 PAM, Diana, 계층적 알고리즘들이 우수한 방법으로 평가되는 경향이 있다.

(6) V3

V3는 V1, V2와 같이 교차타당성 평가방법으로 척도값이 작을수록 우수한 알고리즘으로 평가한다. 이 척도는 Diana와 K-means 방법을 가장 우수한 성능을 갖는 알고리즘으로 평가하였고, FCM과 모형기반 방법이 뒤를 따랐으며 PAM과 Fanny는 가장 좋지 않은 방법으로 평가하였다. 또한 Diana, PAM, Fanny는 유전자 선택에 대해 다소 로버스트한 반면, 계층적과 모형기반 방법은 유전자 선택에 대해 약간 민감한 반응을 보였다. 이 척도는 전체 시점에 대해서 분석한 결과와 한 시점을 제외시키고 분석한 결과와의 프로파일간 평균거리를 계산한다. 즉, 전체 시점과 한 시점을 제외시킨 군집 분석 결과로부터 각 군집에 속한 개별 유전자들 간의 거리를 계산하기 때문에 분석에 사용된 유전자들의 발현 프로파일의 영향력을 어느 정도 잘 반영할 수 있을 것으로 판단된다.

이처럼 V1, V2, V3은 모두 leave-one out에 의한 평가방법으로서 PAM과 Diana는 이 세 가지 척도들에 의해서 성능도 우수하고 유전자 선택방법에 대해서도 안정적인 방법으로 평가되었다. 그러나 Fanny와 모형기반에 의한 알고리즘들은 성능도 좋지 않고, 유전자 선택방법에 대해서도 상당히 민감한 방법으로 평가되었다. 이들 세 척도들은 위치(location)나 척도(scale) 변화에 불변하기 때문에 위치 또는 척도에 관련된 발현 패턴을 잘 구분할 수 없

다는데서 그 이유를 찾을 수 있을 것이다.

지금까지의 분석 결과로부터 PAM과 Diana는 평가방법에 상관없이 우수한 성능을 갖는 알고리즘으로서 유전자 선택방법에도 안정적인 방법으로 평가되었다. 또한 FCM과 계층적 방법은 성능 면에서는 보통이였지만 유전자 선택에 대해서는 다소 민감한 반응을 보였고, 모형기반, Fanny 방법은 대체로 성능이 열등하고 유전자 선택방법에 대해서도 민감한 반응을 보이는 것으로 평가되었다.

4. 결론 및 토의

군집분석은 자료 탐색과 유전자 그룹화를 위한 첫 번째 분석 단계로 마이크로어레 자료 분석에서 자주 사용되는 방법 중 하나이다. 본 연구에서는 Cyanobacterium sp. PCC 6803 (Hihara 등, 2001) 자료를 적용하여 4가지 유전자 선택 방법과 7가지 군집 알고리즘에 따라 군집분석을 수행하고, 6가지 타당성 평가방법에 의해 각 군집 알고리즘을 평가함으로써 폭넓은 비교 연구 결과를 제시하였다. 결론적으로 다양한 군집 알고리즘의 성능은 유전자 선택방법과 군집 평가방법에 따라 매우 영향을 받는 것으로 나타났다. 본 연구는 마이크로어레이 자료의 군집 분석을 위한 알고리즘의 선택에 있어서 몇 가지 가이드라인을 제시하고자 하였다.

4.1. 군집 알고리즘의 성능-총체적 결과 측면에서

각 군집 알고리즘들은 군집 평가방법과 유전자 선택방법에 따라 서로 다른 결과를 보였지만, 본 연구의 분석결과들을 토대로 몇 가지 공통성을 찾을 수 있었다. 군집 알고리즘의 성능에 대한 종합적인 결과를 군집 평가방법과 유전자 선택방법에 따라 표4.1과 같이 요약하였다. 표4.1의 앞에서부터 첫 번째 열은 군집평가방법을 나타내고, 두 번째 열은 유전자 선택방법의 특성을 나타낸다. 표 3.2의 결과로부터 유전자선택방법별로 각 군집 평가방법 내에서 군집방법들을 비교하여 순위나 평가수치의 변화정도에 따라 유전자 선택방법의 특성을 '로버스트'와 '민감'으로 분류하였다. 대체로 순위가 선두(대략 1, 2등)에서 후미(대략 6, 7등)로의 변동이 클 경우 '민감', 그렇지 않을 경우는 '로버스트'로 평가하였다. 또한 7개 군집 방법 중 각 방법들의 우선순위가 각 유전자선택방법별로 선두그룹을 많이 차지할 경우 '우수'한 군집방법으로, 후미그룹을 많이 차지할 경우 '나쁜' 군집방법으로 분류하였다.

PAM과 Diana는 평가방법에 상관없이 대체로 우수한 알고리즘으로서 유전자 선택방법에 대해서도 로버스트한 특성을 나타내었다. PAM은 K-means의 확장된 방법으로 K-means에 비해 노이즈나 이상치에 로버스트한 방법으로, 샘플 크기가 작은 자료에 대해서는 효과적이지만 샘플 크기가 큰 경우에는 덜 효과적인 것으로 알려져 있다. Diana는 분리적 계층적 방법으로 군집의 크기가 크거나 적은 개수의 군집을 찾고자 하는 경우에 적합한 방법이다. 따라서 마이크로어레이 자료 분석에 적합한 방법으로 알려져 있다 (Kim 등, 2005).

FCM과 계층적 군집 알고리즘은 성능면에서는 7가지 방법 중 보통이었지만, 유전자 선

표 3.2: 4가지 통계량에 따라 선택된 유전자 자료에 대한 군집 평가 결과.

	Fold change						
	K-means	PAM	Fanny	Diana	계층적	Fuzzy c-means	모형기반
homogeneity	0.81	**0.80**	0.91	0.82	0.88	0.85	0.82
separation	3.58	3.58	3.55	3.69	**3.70**	3.48	3.57
FOM	21.42	**17.42**	19.23	19.55	25.34	18.81	18.16
V1	0.24	0.24	0.34	**0.16**	0.21	0.32	0.18
V2	0.53	0.54	0.63	**0.46**	0.68	0.65	0.51
V3	0.40	**0.30**	0.42	0.35	0.53	0.42	0.34
	B-통계량						
	K-means	PAM	Fanny	Diana	계층적	Fuzzy c-means	모형기반
homogeneity	**0.76**	**0.76**	0.82	0.80	0.81	0.77	0.86
separation	3.48	3.48	3.50	3.59	**3.68**	3.47	3.17
FOM	**21.94**	25.42	23.52	24.40	30.85	24.08	26.74
V1	0.20	0.15	0.18	0.12	**0.01**	0.27	**0.01**
V2	0.55	0.42	**0.35**	0.42	0.38	0.55	0.50
V3	0.23	0.44	0.51	0.33	0.36	**0.18**	0.40
	T-통계량						
	K-means	PAM	Fanny	Diana	계층적	Fuzzy c-means	모형기반
homogeneity	0.77	**0.74**	0.78	0.76	0.83	**0.73**	0.89
separation	3.49	3.50	3.43	3.57	**3.69**	3.51	3.35
FOM	16.42	**13.29**	14.49	16.18	20.33	16.01	16.10
V1	0.24	0.35	0.96	**0.08**	0.11	0.44	0.24
V2	0.55	**0.39**	0.59	0.46	0.48	0.60	0.75
V3	**0.17**	0.50	0.45	0.43	0.54	0.18	0.28
	SAM-통계량						
	K-means	PAM	Fanny	Diana	계층적	Fuzzy c-means	모형기반
homogeneity	**0.75**	**0.75**	0.76	0.76	0.81	**0.75**	0.88
separation	3.59	3.55	3.52	3.62	3.77	3.54	**3.93**
FOM	35.01	**20.06**	23.54	20.66	22.99	22.76	21.79
V1	0.55	0.17	0.44	0.34	0.06	**0.04**	0.33
V2	0.49	**0.40**	0.49	0.46	0.48	0.66	0.73
V3	0.22	0.50	0.47	0.23	0.41	0.34	**0.18**

V1 : Average proportion of non-overlap measure, V2 : Average distanace between means measure, V3 : Average distance measure

택에 대해서는 다소 민감하였다. FCM은 퍼지정도(fuzziness)를 반영하는 모수에 따라 분석 결과가 상당히 민감한 반응을 보이는 것으로 알려져 있다 (Dembele와 Kastner, 2003; Kim 등, 2005). 특히, Dembele와 Kastner(2003)은 FCM 알고리즘을 마이크로어레이 자료 분석에 적용할 경우 공통적으로 사용되는 퍼지모수, 2는 유전자 자료 분석에 적절하지 않다는 것을 주장하고, 최적의 퍼지모수를 찾는 방법을 제안한 바 있다. FCM은 각 유전자가 모든 군집에 할당될 확률을 계산하고, 각 유전자내에서 군집에 할당될 확률의 합이 0이 된다는 제약조건이 따르기 때문에, 찾고자 하는 군집의 수가 많은 경우에는 상당히 불리하게 작용할 수 있다. 병합적 계층적 방법은 계산이 간단하기 때문에 마이크로어레이 자료 분석에 널리 이용되고 있지만, 군집의 크기가 작거나 많은 군집의 패턴을 찾고자 하는 경우에 유리한 방법으로 마이크로어레이 자료의 경우에는 분리적 계층적 방법에 비해 덜 효과적이다.

Fanny와 모형기반 방법들은 군집 성능도 좋지 않을 뿐더러 유전자 선택에 대해서도 상당히 민감한 결과를 보여주었다. K-means는 성능은 그다지 좋지는 않지만 대체로 안정적인 결과를 보여주었다. 한편, K-means는 계층적 알고리즘처럼 마이크로어레이 자료의 군집분석에 널리 적용되고 있지만, 계층적 군집 알고리즘과 함께 본 연구에서 뿐만 아니라 Datta와 Datta(2003)등 많은 문헌 등에서도 그 우수성이나 적합성은 증명되지 않았다는 것이다. 모형기반 알고리즘은 주로 다변량 정규분포를 가정할 수 있는 경우에 적합한 방법이지만, 모형에 결측치나 이상치가 포함되어 있는 경우에는 군집 결과에 치명적인 영향을 미치는 것으로 알려져 있다 (Yeung 등, 2001).

4.2. 군집 알고지름 성능-유전자 선택방법 측면에서

한편, 실제로 마이크로어레이 자료를 분석하고자 하는 연구자는 여러 가지의 통계량을 동시에 적용하여 유전자를 선택하지는 않는다. 연구자의 분석 목적에 따라 다르겠지만, 오히려 선행연구나 연구자가 즐겨 사용하는 방법에 의해 먼저 유전자를 선택한 후, 적절한 군집 알고리즘을 적용하여 분석하는 것이 일반적인 절차라고 볼 수 있다. 따라서 본 절에서는 4.1절에서 제시한 방향과는 달리 연구자가 유전자를 특정 방법에 의해 선택하였을 경우, 이에 적합한 군집 알고리즘의 선택에 대해 몇 가지 방향을 제시하였다. 군집 평가방법을 고려하여 유전자 선택방법별로 군집 알고리즘들의 성능을 요약하면 표 4.2와 같다. 표 4.2의 결과를 모두 설명하기에는 너무 양이 많고, 결과들에 대한 설명 방법이 매우 유사하기 때문에 구체적인 설명이 필요한 부분에 대해서만 부분적으로 설명하였다.

표 4.2에서 연구자가 Fold change에 의해 유전자를 선택하고 Diana에 의해 분석하고자 할 경우, V1, 또는 V3와 같은 방법들에 의해 평가한다면 Diana에 대한 우수한 평가 결과를 얻을 수 있다. 또 PAM 알고리즘을 적용하고자 한다면 Homogeneity, FOM, V3에 의해 평가되는 것이 바람직하다. 그러나 Fold change 방법은 유전자들 간 변동을 전혀 고려하지 않기 때문에 실제 유전자 선택에 있어서 권장할 만한 방법은 아니다.

연구자가 T-통계량에 의해 유전자를 선택하고, PAM에 의해 군집 분석하고자 한다면, FOM이나 V2 평가방법을 적용하는 것이 바람직하다. 한편 표 4.2를 보면 T-통계량 내에서 Diana는 단 한번만 가장 우수한 방법으로 평가되었지만 표 3.2의 이 통계량 내에서 척도값이나 척도의 순서 면에서 보면 대체로 우수한 알고리즘으로 평가되었다. T-통계량은 마이

표 4.1: 군집 평가방법과 유전자 선택에 따른 군집 알고리즘 성능 비교

		군집방법 능력	
	유전자선택	우수	나쁨
homogeneity			
	로버스트	PAM, Diana, K-means	모형기반
	민 감	FCM	계층적, Fanny
seperation			
	로버스트	Diana	PAM, K-means
	민 감	계층적	모형기반, Fanny, FCM
FOM			
	로버스트	FCM	Fanny
	민 감	PAM	모형기반, 계층적, K-means, Diana
V1			
	로버스트	계층적	FCM, Fanny
	민 감	모형기반, Diana	K-means, PAM
V2			
	로버스트	PAM, Diana	FCM, 모형기반, K-means
	민 감	계층적	Fanny
V3			
	로버스트	Diana	PAM, Fanny
	민 감	K-means, 모형기반	계층적, FCM

크로어레이 자료의 분류 분석에 관련된 연구들에서 사전에 유의한 유전자를 선택하기 위해 자주 사용되는 방법 중 하나이다. 그러나 이 통계량은 발현수준이 낮은 유전자에 대해서 집단 간 평균차이에 비해 표준오차가 너무 작기 때문에, 위양성 또는 위음성적 오류가 크게 발생할 수 있다. T-통계량을 사용하는 연구자는 이런 문제점을 염두 해 두어야 할 것이다.

만약, 연구자가 B-통계량에 의해 유전자를 선택하고, Seperation이나 V1에 의해 군집 결과를 평가하고자 한다면, 계층적 알고리즘을 적용하여 분석하는 것이 타당할 것이다. K-means로 분석하고자 한다면 Homogeneity나 FOM에 의해 우수한 군집 알고리즘으로 평가될 수 있다. B-통계량은 T-통계량의 발현 수준이 낮은 유전자에 대한 분산 안정성 문제를 보완한 방법이지만, 여전히 분포적 문제를 내포하고 있다. SAM에 의해 유전자를 선택할 경우, 6가지 평가방법 중 3가지 방법에 의해 가장 우수한 방법으로 평가된 PAM을 적용하는 것이 적절하다 할 것이다. 게다가 표 4.2에 의하면 Diana는 B와 SAM 내에서 각 평가방법에 의해 한 번도 가장 우수한 방법으로 평가되지는 않았지만 표 3.2에 따르면 이들 통계량 내에서 평가방법에 상관없이 대체로 우수한 성능을 보이는 것으로 나타났다. SAM은 B

와 마찬가지로 T-통계량의 분산 안정성 문제를 보완한 방법으로 재표본에 의한 비모수적 방법에 해당된다. 따라서 SAM은 분포적 가정에 자주 위배되는 마이크로어레이 자료 분석에 유용한 방법으로 권장할 만하다.

Kim 등(2005)에 따르면 유전자 선택 통계량들은 서로 다른 특성을 갖기 때문에 유의한 발현차이를 갖는 유전자를 선택하는 것이 연구의 목적이라면 통계량 선정에 매우 신중할 필요가 있을 것이다. 자세한 통계량의 특성에 대한 비교는 Kim 등(2005)를 참고할 수 있다. 또한 선택된 유전자에 따라 분석결과가 달라질 수 있다면 연구자의 분석 취지를 잘 반영할 수 있는 유전자 선택방법을 선택하는 것은 더욱 중요한 문제라고 볼 수 있다.

표 4.2: 유전자 선택에 따른 군집 알고리즘 평가

	평가방법 별 가장 우수한 군집 방법					
유전자 선택	Homogeneity	Seperation	FOM	V1	V2	V3
Fold change	PAM	계층적	PAM	Diana	Diana	PAM
B	K-means	계층적	K-means	계층적	Fanny	FCM
	PAM			모형기반		
T	FCM	계층적	PAM	Diana	PAM	K-means
SAM	K-means	모형기반	PAM	FCM	PAM	모형기반
	PAM					
	FCM					

4.3. 논의 및 고찰

현재 마이크로어레이 자료의 군집분석을 위한 상용 프로그램은 너무나 다양한 방법들이 존재한다. 오히려 연구자들이 군집 알고리즘의 선택에 있어서 혼란을 일으킬 정도이다. 또한 시간경로 마이크로어레이 자료는 시간의 변화에 따른 유전자들의 발현패턴을 찾거나 유전자 발현양상의 최적 시점을 찾고자 하는 경우에 사용되고 있는 실험방법이다. 시간경로 자료에 대해서 군집분석을 수행하고자 하는 경우, 본 논문에서 언급한 방법들을 포함하여 다양한 통계적 알고리즘들이 사용되고 있다. 그러나, 일반적인 마이크로어레이 자료에 비해 시간에 따른 변동이 존재하기 때문에 군집방법이 일반 자료의 경우와는 다르게 작용할 수 있다. 따라서 본 연구는 시점에 따른 변동을 고려한 다양한 군집방법들의 수행능력을 비교함으로써 그 안정성을 파악하고자 하였다. 그 결과 PAM과 Diana는 유전자 선택방법과 평가방법에 상관없이 비교적 안정적인 군집방법을 제시하였다. 각 유전자 선택방법별로 군집 알고리즘의 선택을 위한 방향을 표 4.2에 제시하였다. 그러나 각 군집방법들은 저마다의 특성을 지니고 있기 때문에 어떤 방법을 최적이라고 단언하는 것은 어려운 문제라 할 수 있다. 본 연구를 토대로 군집 분석에 대한 좀더 명확한 가이드라인을 제시하기 위해서는 다양한 자료의 특성을 고려한 추후 연구가 더 필요할 것으로 생각된다. 추후 연구로서 다음과 같은 문제들을 고려해 볼 필요가 있을 것이다. 첫째, 마이크로어레이 자료는 유

전자가 동일한 조직으로부터 관찰된 것인가 아니면 서로 다른 조직으로부터 관찰된 자료가 혼합된 것인가; 둘째, 자료에 얼마나 많은 결측치나 이상치가 포함되어 있는가 그리고 이에 대한 처리는 어떻게 할 것인가; 셋째, 군집 분석의 가장 중요하고 어려운 쟁점이 되는 군집의 개수를 몇 개로 할 것인가; 이런 사항들에 의해 군집 알고리즘들은 매우 다르게 작용할 수 있다는 것이다. 이처럼 다양한 생물학적 현상을 반영할 수 있는 통계적 문제를 직시함으로써 마이크로어레이 자료로부터 유용한 정보를 도출할 수 있을 것이다.

본 논문은 기능유전체 연구자들에게 각자의 실험상황이나 자료성격에 적합한 군집방법을 제시함으로써 부적절한 방법의 적용으로 인한 연구결과의 오류를 최소화시키는데 기여하고자 하였다. 뿐만 아니라 유전체 자료의 분석을 연구하는 수학, 전산학 및 통계학 전문가에게는 기존의 군집 방법들이 자료형태나 연구 설계에 따라 어떠한 장단점이 있는가를 정리, 요약하는 계기가 될 수 있을 뿐만 아니라, 새로운 군집방법을 개발하는 데에 필요한 유용한 정보와 모티브를 제공할 수 있을 것이다.

참고문헌

Barash, Y. and Friedman, N. (2002). Context-Specific Bayesian Clustering for Gene Expression Data, *Journal of Computational Biology*, **9**, 169-191.

Chen, G. et al. (2002). Evaluation and comparison of clustering algorithms in analyzing ES cell gene expression data, *Statistica Sinica*, **12**, 241-262.

Chu, S., DeRisi, J. et al., (1998). The transcriptional program of sporulation in budding yeast, *Science*, **282**, 699-705.

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

Dudoit, S., Yang, Y. H., Speed, T. and Callow, M. J. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, *Statistica Sinica*, **12**, 111-139.

Efron, B. (1982). *The jackknife, the bootsrap, and other resampling plans*, Society for industrial and applied mathematics.

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.

Goldstein, D. R., Conlon, E. and Ghosh, D. (2002). Statistical issues in the clustering of gene expression data, *Statistica Sinica*, **12**, 219-240.

Ghosh, D. and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments, *Bioinformatics*, **18**, 275-286.

Guthke, R., Schmidt-Heck, W., Hahn, D. and Pfaff, M. (2000). Gene expression data mining for functional genomics, *Proceedings of European Symposium on Intelligent Techniques* (EIST 2000), Aachen, Germany, 170-177.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*. Vol 28. 100-108.

Hastie, T., Tibshirani, R. et al. (2000). Gene shaving as a method for identifying distinct sets of genes with similar expression patterns, *Genome Biology*, **1**, research003.

Hihara, Y., Kamei, A., Kanehisa, M., Kaplan, A. and Ikeuchi, M. (2001). DNA microarray analysis of cyanobacterial gene expression during acclimation to high light, *The Plant Cell*, **13**, 793-806.

Hong, F. and Li, H. (2004). B-spline Based Empirical Bayes Methods for Identifying Genes with Different Time-course Expression Profiles. submitted.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*, New York, John Wiley.

Kasturi, J., Acharya, R. and Ramanathan, R. (2003). An information theoretic approach for analyzing temporal patterns of gene expression, *Bioinformatics*, **19**, 449-458.

Kerr, M. K. and Churchill, G. A. (2001). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments, *Proc. Natl Acad. Sci.*, **98**, 8961-8965.

Kim, S. Y., Choi, T. M. and Bae J. S. (2005). Fuzzy types clustering for microarray data, *International Journal of Computational Intelligence*, bf 2, 12-15.

Kim, S. Y., Lee, J. W. and Bae J. S. (2006). Effect of data normalization on fuzzy clustering of DNA microarray data., *BMC Bioinformatics*, To appear.

Kim, S. Y., Lee, J. W. and Shon, I. S. (2006). Comparison of various statistical methods for identifying differential gene expression in replicated microarray data, *Statistical Methods in Medical Research*, **15**, 1-18.

Laura, L. and Owen, A. (2002). Plaid models for gene expression data, *Statistica Sinica*, **12**, 61-86.

Lonnstedt, I. and Speed, T. P. (2002). Replicated microarray data, *Statistica Sinica*, **12**, 31-46.

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.

McLachlan, G. J., Bean, R. W. and Peel, D. (2002). A mixture model based approach to the clustering of microarray expression data, *Bioinformatics*, **18**, 1-10.

Moon et al. (2002). Mice Lacking Paternally Expressed Pref-1/Dlk1 Display Growth Retardation and Accelerated Adiposity, *Molecular and Cellular Biology*, **22**, 5585-5592.

Smyth, G. K., Yang, Y. H. and Speed, T. (2003). *Statistical issues in cDNA microarray data analysis*, in Functional Genomics: Methods and Protocols, eds.

Spellman, P. T., Sherlock, G., Zhang, M. Q. et al., (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, *Mol. Biol. Cell.*, **12**, 3273-3297.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S. and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps:methods and application to hematopoietic differentiation, *Proceedings of the National Academy of Sciences*, **96**, 2907-2912.

Tusher, V., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, **98**, 5116-5124.

Waddell, P. and Kishino, H. (2000). Cluster inference methods and graphical models evaluated on NC160 microarray gene expression data, *Genome Informatics*, **11**, 129-140.

Yeung, K., Haynor, D. R. and Ruzzo,W. L. (2001). Validating clustering for gene expression data, *Bioinformatics*, **17**, 309-318.

Yeung, K. Y., Fraley, C. Murua, A, Raftery, E. and Ruzzo, W. L. (2001). Model based

clustering and data transformations for gene expression data, *Bioinformatics*, **17**, 977-987.

[2005년 7월 접수, 2005년 11월 채택]

A Review of Cluster Analysis for Time Course Microarray Data*

In Suk Sohn[1] Jae Won Lee[2] Seo Young Kim[3]

ABSTRACT

Biologists are attempting to group genes based on the temporal pattern of gene expression levels. So far, a number of methods have been proposed for clustering microarray data. However, the results of clustering depends on the genes selection, therefore the gene selection with significant expression difference is also very important to cluster for microarray data. Thus, this paper present the results of broad comparative studies to time course microarray data by considering methods of gene selection, clustering and cluster validation.

Keywords: Time course microarray data, Gene selection, Cluster analysis, Cluster validation.

* This work was supported by Korea Science and Engineering Foundation Grant (R14-2003-002-01002-0) and Seo Young Kim was supported by grant No. R08-2003-0000-10572-0 from the Basic Research Program of the Korea Science and Engineering Foundation

1) Ph.D. Student, Department of Statistics, Korea University, Seoul 136-701, Korea
E-mail : sis46@korea.ac.kr

2) Professor, Department of Statistics, Korea University, Seoul 136-701, Korea
E-mail : jael@korea.ac.kr

3) (Corresponding author) Researcher, Research Institute for Basic Science, Chonnam National University Gwangju 500-757, Korea.
E-mail : gong@chonnam.ac.kr