응용통계연구 (2011)
24(6), 1077–1094
DOI: http://dx.doi.org/10.5351/KJAS.2011.24.6.1077

효모 마이크로어레이 유전자 발현 데이터에 대한 유전자 선별 및 군집분석

이경아[1] · 김태훈[2] · 김재희[3]

[1]덕성여자대학교 정보통계학과, [2]덕성여자대학교 PrePharmMed학과
[3]덕성여자대학교 정보통계학과

(2011년 9월 접수, 2011년 11월 채택)

요 약

마이크로어레이 유전자 발현 데이터인 yeast cdc15에 대해 시계열 데이터의 특성을 반영한 푸리에 계수를 이용한 검정통계량과 FDR 다중비교법을 이용하여 차별화된 유전자를 선별한 후 선별된 유전자들에 대해 모형기반 군집방법, K-평균법, PAM, SOM, 계층적 Ward 군집방법과 Fuzzy 군집방법을 실시하였다. 군집방법에 따른 특성을 알아보고 군집화 결과와 내부유효성 측도로 연결성 측도, Dunn 지수와 실루엣 값을 살펴본다. 또한 GO분석을 통한 생물학적 의미도 파악해본다.

주요용어: 계층적 군집방법, 모형기반 군집방법, 실루엣, 연결성 측도, 자기 조직화 지도(SOM), 푸리에 계수, Dunn 지수, Fuzzy 군집방법, K-평균법, PAM.

1. 서론

군집분석은 가장 잘 알려진 자율학습(unsupervised learning)의 예이다. 자율학습이란 목표패턴이 주어지지 않고 입력패턴에 근거하여 학습을 진행하는 방법으로 구조화되지 않은 다변량 데이터를 분석하는데 가장 많이 쓰이는 방법이다.

마이크로어레이 유전자 발현 데이터는 비슷한 성질을 갖는 유전자들을 군집화 함으로써 특정한 기능이나 공통성을 찾고자한다. 유전자 데이터의 경우 그 수가 많기 때문에 분석을 할 때에도 시간과 비용 등의 어려움이 많고 분석 후에 결과가 좋지 않은 경우도 발생하게 된다. 이러한 문제를 해결하기 위해 변수변환을 이용하여 데이터의 차원수를 줄이거나 선별(screening)과정을 통해 유전자를 선별하여 선별된 유전자만을 가지고 분석을 하는 방법이 있다.

여러 연구자들이 유전자 데이터에 대한 선별방법과 군집분석방법을 제안하였다. 유전자선별방법에 대한 연구로 Hero 등 (2004)은 다양한 기준을 가진 DNA 마이크로어레이 실험을 통해 얻은 유전자들 중 다르게 발현되는 유전자를 확인하는 통계적 방법에 대해 연구하였다. Eckel 등 (2004)은 마이크로어레이 유전자 발현 데이터에 대해 데이터의 차원수를 줄이는 효율적인 방법에 대해 연구하였다. Datta와 Datta (2005)는 p-값을 이용하여 각 유전자를 비교하여 다르게 발현되는 유전자를 찾는 통계적 검정을 실시하였다. Ma (2006)는 8개의 통계량에 기초한 관리된 유전자 선별에 대해 일치성과 복제 가능성에 대해 조사하였다.

[3]교신저자: (132-714) 서울시 도봉구 쌍문동 419, 덕성여자대학교 정보통계학과, 교수.
E-mail: jaehee@duksung.ac.kr

유전자 데이터에 대한 군집분석 개발은 다양한 방법으로 활발히 연구되고 있다. Törönen 등 (1999)은 자기 조직화 지도(SOM)방법을 이용하여 군집분석을 하였다. Getz 등 (2000)은 초모수적(super paramagnetic) 군집분석을 하였고 Tusher 등 (2001)은 마이크로어레이의 유의한 분석 방법에 대해 연구하였다. Gasch와 Eisen (2002)는 퍼지(fuzzy) K-평균법을 이용하여 군집분석을 하였다. Zhang 등 (2003)은 이산 푸리에 변환(discrete Fourier transform)을 이용한 군집분석 결과를 보여주었다. Dudoit 등 (2003)은 마이크로어레이 실험에서의 다중비교 검정에 대한 기존 방법들에 대한 비교 연구하였다. Serban과 Wasserman (2005)은 비모수적 추정과 변수 변환을 통한 군집화에 대한 방법을 연구하였다. Kim 등 (2006)은 푸리에 프로화일을 이용하여 2단계 군집방법을 적용하였으며 Kim과 Kim (2008)은 미분 푸리에 계수를 사용한 군집분석을 제안하였다. 최근에 Bickel (2011)은 분포를 모르는 게놈-스케일(Genome-scale) 선별에 대한 조절된 관측 유의수준 추정에 대한 연구를 하였다.

2. 유전자 선별

수천수만 개의 유전자들에 대한 분석은 시간과 비용이 소요되므로 의미있는 유전자들을 선별하여 효율적인 통계적 분석이 필요하다.

본 연구에서는 여러 개의 시점(time point)에서 얻은 유전자 발현 데이터에 대해 차원축소 방법일 뿐만 아니라 기저함수에 대한 정보를 포함하는 푸리에 계수를 이용하여 유전자를 선별하고 선별된 유전자에 대한 군집분석을 하고자한다.

2.1. 푸리에 계수

유전자 발현데이터에 대해 다음의 모형을 고려하고자 한다. i번째 곡선에 있는 u번째 데이터를 Y_{iu}라고 하면

$$Y_{iu} = f_i(t_{iu}) + \epsilon_{iu}, \quad i = 1, 2, \ldots, n, \ u = 1, 2, \ldots, m \tag{2.1}$$

와 같은 모형을 갖는다. 여기서 $E(\epsilon_{iu}) = 0$이고 $\mathrm{Var}(\epsilon_{iu}) = \sigma^2$이다. 유전자 발현 데이터인 Y_{iu}는 t_{iu}시간에서 i번째 데이터의 로그값이다. 곡선 f_i가 smooth 함수집합에 속한다고 가정하면 f_i는

$$f_i(t) = \sum_{j=1}^{\infty} \phi_{ij} b_j(t) \tag{2.2}$$

와 같이 정의된다. 여기서 b_j는 orthonormal basis system이고 푸리에 계수(Fourier coefficient) ϕ_{ij}는 다음과 같이 정의된다.

$$\phi_{ij} = \int f_i(t) b_j(t) dt. \tag{2.3}$$

f_i를 J개의 푸리에 계수를 이용하여 다음과 같이 추정할 수 있다.

$$f_i(t) \approx \sum_{j=1}^{J} \phi_{ij} b_j(t), \quad 1 \le J \le m. \tag{2.4}$$

표본 푸리에 추정량은 다음과 같다.

$$\hat{\phi}_{ij} = \frac{1}{m} \sum_{r=1}^{m} Y_{ij} b_j(t_r), \tag{2.5}$$

여기서 $t_r = r/m$, $t_r \in [0,1]$이다.

푸리에열(Fourier series)의 함수는 코사인(cosine) 함수를 기초로 다음과 같이 나타낼 수 있다.

$$f_i(t) = \phi_{i0} + \sum_{j=1}^{\infty} \phi_{ij}\sqrt{2}\cos\pi jt. \tag{2.6}$$

J개의 표본 푸리에 계수로 f_i를 추정하면

$$\hat{f}_i(t) = \phi_{i0} + \sum_{j=1}^{J} \hat{\phi}_{ij}\sqrt{2}\cos\pi jt \tag{2.7}$$

이고 표본 푸리에 계수(sample Fourier coefficient) $\hat{\phi}_{ij}$는

$$\hat{\phi}_{ij} = \frac{1}{m}\sum_{r=1}^{m} Y_{ir}\sqrt{2}\cos\pi jt_r \tag{2.8}$$

이다.

2.2. 선별과정(Screening)

마이크로어레이 실험의 목표는 의미가 있는 유전자를 찾아내는 것이다. 군집분석에서 다르게 발현된 유전자를 찾아내지 못하는 경우가 있는데 이는 유전자들이 자신이 속한 군집 안에서 의미있게 변하지 않기 때문이다. 이러한 문제를 해결하기 위해 각각의 유전자를 검정하여 의미있는 유전자만을 선별해 분석에 이용하고자 한다. i번째 유전자가 다르게 발현되지(differentially expressed) 않았다는 귀무가설 하에 검정을 실시한다. 이 경우에는 적어도 하나의 유전자가 잘못 군집화 되었을 확률인 FWER(familywise error rate)와 활동성(active) 유전자라고 밝혀진 것 중 비활동성(inactive) 유전자인 부분의 평균인 FDR(false discovery rate)를 제어해야한다. Benjamini와 Hochberg (1995)에 의한 FDR의 절차는 다음과 같다.

$P_{(g)} \le g\alpha/n$이면 모든 귀무가설($H_{(g)0} : f_i(\cdot) = c$, $g = 1, 2, \ldots, k$)을 기각한다. 여기서 $P_{(g)}$는 각 가설에 대한 p-값 중에 순서대로 늘어놓은 값이다.

Kim과 Hart (1998)의 T_S통계량

$$T_S = \max_{1\le k\le m-1} \frac{1}{k}\sum_{j=1}^{k} \frac{m\hat{\phi}_j^2}{\hat{S}(0)} \tag{2.9}$$

을 이용하여 Benjamini와 Hochberg (1995) FDR 과정을 적용하고자한다 (Kim 등, 2011). 여기서 $\hat{S}(0) = \hat{\gamma}(0) + 2\sum_{k=1}^{m-1}\hat{\gamma}(k)$로 0에서 스펙트럼으로 추정한다.

자기공분산 함수는 Yule-Walker 추정방법을 통해

$$\hat{\gamma}(k) = \frac{1}{m}\sum_{u=1}^{m-k} \left(\hat{\epsilon}_u - \bar{\hat{\epsilon}}\right)\left(\hat{\epsilon}_{u+k} - \bar{\hat{\epsilon}}\right) \tag{2.10}$$

이고

$$\hat{\gamma}(0) = \frac{1}{m}\sum_{u=1}^{m} \left(\hat{\epsilon_u} - \bar{\hat{\epsilon}}\right)^2 \tag{2.11}$$

이다. 검정통계량 T_S의 값은 근사적 분포를 통해 다음과 같이 p-값을 계산할 수 있다. 여기서 각각의 $m\hat{\phi}_j^2/\hat{S(0)}$는 근사적으로 카이제곱 분포를 따른다.

$$P(T_s \leq C) \approx \exp\left(-\sum_{j=1}^{\infty} \frac{p(\chi_j^2 > jC)}{j}\right), \tag{2.12}$$

여기서 χ_j^2은 자유도가 j인 카이제곱 분포를 따르는 확률변수이다.

3. 군집분석 방법

3.1. 모형기반 군집방법

모형기반 군집방법은 Fraley와 Raftery (2002)에서 데이터 $\mathbf{y} = (y_1, \ldots, y_n)$를 다음과 같은 밀도함수를 갖는 혼합모형이라고 가정한다.

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k\left(y_i \theta_k\right), \tag{3.1}$$

여기서 θ_k는 모수벡터이고 τ_k는 관측벡터가 k번째 군집 $(1, \ldots, G)$에 속할 확률이다. EM 알고리즘을 이용하여 가능도 함수인 $f(y)$를 가장 크게 하도록 모수를 추정한다. EM 알고리즘은 E-단계와 M-단계로 나눌 수 있다. E-단계에서는 i개체가 k번째 군집에 속할 조건부 확률을 다음과 같이 계산한다.

$$z_{ik} = \frac{\tau_k \phi\left(y_i | \mu_k, \Sigma_k\right)}{\sum_{j=1}^{G} \tau_k \phi\left(y_i | \mu_j, \Sigma_j\right)}. \tag{3.2}$$

M-단계에서는 조건부 확률 z_{ik}를 이용하여 모수를 추정한다. 각 개체가 최대 확률로 해당 군집에 속할 때 EM 알고리즘 결과로 수렴하게 된다. 모형을 선택할 때 BIC 값이 최소가 되는 군집의 수를 최종모형으로 선택할 수 있다. 모형기반 군집방법은 Fraley와 Raftery (2006)에서 설명하고 있는 R 프로그램의 MCLUST 패키지를 사용하여 분석할 수 있다.

3.2. K-평균법

K-평균법은 매우 효율적이며 대규모의 데이터를 군집화 할 때 자주 쓰인다. K-평균법의 알고리즘을 살펴보면 다음과 같다. n개의 데이터벡터를 $\mathbf{y}_i$, $i = 1, 2, \ldots, n$이라 하고 군집의 개수를 K라 한다. K개의 군집과 각 군집에 개체들의 초기 무작위 할당(random assignment)을 이용하여 무게중심 값($\bar{\mathbf{y}}_k$, $k = 1, 2, \ldots, K$)를 구한다. 군집 무게중심 값을 이용하여 각 개체의 제곱-유클리드(squared-Euclidean) 거리

$$\text{ESS} = \sum_{k=1}^{K} \sum_{c(i)=k} \left(\mathbf{y}_i - \bar{\mathbf{y}}_k\right)' \left(\mathbf{y}_i - \bar{\mathbf{y}}_k\right) \tag{3.3}$$

를 구한다. 여기서 $\bar{\mathbf{y}}_k$는 k번째 군집의 무게중심이고 $c(i)$는 $\mathbf{y}_i$를 포함하는 군집이다. 각 개체를 가장 가까운 군집에 재할당하고 군집 무게중심을 업데이트한다. 이러한 과정을 반복한다.

K-평균법은 비계층적 군집방법으로 초기 군집 선정 후 전체 개체를 개의 군집으로 나누는 방법이기 때문에 군집의 정확한 개수를 알아야하고 초기 군집선정에 따라 최종 군집결과가 영향을 받는다. K-평균법은 계층적인 군집방법과 함께 쓰일 수 있는데 계층적 군집방법으로 군집의 개수를 정하고 K-평균법을 이용하여 개체를 재배치 할 수 있다. K-평균법은 수치화된 자료에만 사용할 수 있다. K-평균법에 대한 설명은 김재희 (2011)와 Izenman (2008)을 참고하였다.

3.3. PAM 방법

PAM 방법은 Kaufman과 Rousseeuw (1990)이 제안한 방법으로 K-medoid 군집화 알고리즘을 수정한 방법으로 K-평균법과 비슷하지만 K-평균법에서 군집 무게중심 대신 이상점과 결측값에 덜 민감한 medoid를 군집의 대표값으로 사용한다. 또한 제곱유클리드 거리 대신 비유사성에 근거한 거리(dissimilarity-based distance)를 이용한다.

관측값과 가까운 medoid 간의 거리 합을 최소화하도록 medoid M^*를 구한다. 즉 K개 군집의 medoids $M = (m_1, m_2, \ldots, m_k)$일 때

$$M^* = \text{argmin}_M \sum_i \min_k d(\mathbf{y}_i, m_k) \tag{3.4}$$

K개 군집이 되도록 medoids를 구한 후 각 개체를 가장 가까운 medoid가 있는 군집으로 분류한다. K-평균법에 비해 비유사도(dissimilarity)가 큰 편이나 이상값에 덜 민감한 방법이므로 군집결정시 K-평균법에 비해 이상점의 영향력이 작아진다.

3.4. 자기 조직화 지도(SOM)

SOM(Self-organizing-map)는 Kohonen (1998)이 전개한 방법으로 저차원의 구조로 정렬된 신경망(neural network)과 뉴런(neuron)의 집단이 자율적으로 반복하는 과정이다. 훈련과정은 네트워크에 속한 뉴런의 무게벡터(weight vector) w가 작은 확률변수 값으로 할당되어 초기화된다. 각각의 훈련과정은 세단계로 구성된다. 입력공간(input space)에서 랜덤하게 선택한 입력벡터 제시하고 네트워크를 평가하고 무게벡터를 업데이트 한다. 패턴을 발표한 후 입력패턴과 무게벡터의 유클리드 거리를 계산한다. 가장 짧은 거리를 갖는 뉴런을 k로 표시한다. 뉴런 i가 특정 공간이웃 $N_t(l)$ 내에 속하면 그 무게는 다음과 같은 규칙에 따라 업데이트 한다.

$$w_i(l+1) = w_i(l) + \alpha(l)[x(l) - w_i(l)], \quad \text{if } i \in N_t(l), \tag{3.5}$$

$$w_i(l+1) = w_i(l), \quad \text{if } i \notin N_t(l). \tag{3.6}$$

공간이웃 N_t와 무게 adoptation α의 크기는 반복할수록 단조감소한다.

3.5. 계층적 군집방법

계층적 군집방법에는 병합(agglomorative)과 분리(divisive) 두가지 종류가 있으며 서로 반대방향으로 군집화가 진행된다. 병합 군집화는 각 개체를 자신의 군집에 속한다고 놓고 군집화를 시작하여 하나의 군집이 남을 때까지 병합하는 방법으로 흔히 아래-위(bottom-up) 방법이라고 불린다. 분리 군집화는 위-아래(top-down) 방법이라 불린다.

이 연구에서는 관측벡터간 거리는 유클리드 거리로 정의하고, 군집간 거리는 편차제곱합을 이용하여 정의한 Ward 군집 방법을 고려하고자한다. Ward 방법은 군집내 제곱합 증분과 군집간 제곱합을 고려한 방법으로 군집간 정보의 손실을 최소화하도록 군집화를 한다. 여기서 군집간의 정보는 편차제곱합 ESS(error sum of squares)로 나타내며 군집간 편차제곱합의 증분을 최소화 하도록 군집을 형성된다.

계층적 군집구조의 표현 예로는 덴드로그램(dendrogram)이 있다. 덴드로그램에서 보면 군집분리 기준값인 절개선(cutting line)에 따라 군집의 개수가 달라진다. 유사성을 가진 유전자들을 같은 군집으로 군집화 하고, 거리가 가까운 두 집단을 연결하는 방법이다. 계층적 군집방법의 단점은 한 번 군집화 된 경우 다시 번복될 수 없다는 것이다 (김재희, 2011).

3.6. 퍼지 군집방법

퍼지(Fuzzy) 군집방법은 개체들이 K개의 군집에 속할 각각의 확률을 이용하여 할당하는 방법이다. 개체 i가 k번째 군집에 속할 힘(strength)을 u_{ik}라고 하면, u_{ik}는 모든 i와 $k = 1, 2, \ldots, K$에서 0보다 크거나 같은 확률값을 갖는다. 또한 모든 i에 대해 $\sum_{k=1}^{K} u_{ik} = 1$을 만족한다. 퍼지 군집방법이 K-평균법과 PAM 방법과 다른 점은 각 개체들이 오직 하나의 군집에만 할당된다는 것이다. 근접행렬(proximity matrix) $D = (d_{ij})$와 군집의 수를 정하고, 다음의 개체함수

$$\sum_{k=1}^{K} \frac{\sum_i \sum_j u_{ik}^2 u_{jk}^2 d_{ij}}{2 \sum_l u_{lk}^2} \tag{3.7}$$

를 작게 하는 알지 못하는 u_{ik}를 찾는다. 개체함수는 반복하는 알고리즘을 이용하여 제한조건 하에서 값을 작게 만들도록 해를 구한다.

4. 군집유효성 측도

4.1. 연결성 측도

Handl 등 (2005)이 제안한 방법으로 n개의 개체가 G개의 집단으로 군집화 된 경우의 연결성 측도(connectivity)의 정의는 다음과 같다.

$$\text{Conn}(\mathcal{C}) = \sum_{i=1}^{n} \sum_{j=1}^{p} \mathbf{y}_{i, nn_{i(j)}}, \tag{4.1}$$

여기서 $\mathcal{C} = C_1, C_2, \ldots, C_G$이며 p는 개체로부터 측정되는 변수의 수이며 $nn_{i(j)}$는 개체 i로부터 j번째 가까이 위치한 개체를 의미한다. 연결성 측도는 어떤 개체가 그 개체와 가까운 거리에 있는 개체들과 얼마나 같은 군집에 배치되어 있는지 알 수 있게 해주며 작은 값을 가질수록 군집화가 잘 됐다고 판단한다.

4.2. Dunn 지수

Dunn (1974)이 제안한 방법으로 Dunn 지수는 같은 군집에 속해 있는 두 개체간의 가장 큰 거리에 대한 서로 다른 군집에 속해 있는 두 개체간의 가장 작은 거리의 비를 나타낸다. 같은 군집에 속해 있는 두 개체간의 거리가 작을수록, 다른 군집에 속해 있는 두 개체간의 거리가 클수록 Dunn 지수는 커지므로 이 값이 클수록 군집화가 잘 되었다고 판단할 수 있다. Dunn 지수는

$$D(\mathcal{C}) = \frac{\underset{(i \in C_k, j \in C_l) \in \mathcal{C},\ C_k \neq C_l}{\text{MIN}} \left(\underset{}{\text{MIN}}\ \text{dist}(i, j) \right)}{\underset{C_m \in \mathcal{C}}{\text{MAX}}\ \text{diam}(C_m)} \tag{4.2}$$

으로 정의되며 여기서 C_m은 $\mathcal{C}$의 분할에서 거리가 가장 큰 두 개체가 있는 군집이며 $\text{diam}(C_m)$은 군집 C_m에서 가장 큰 두 개체의 거리를 나타낸다.

4.3. 실루엣 측도

Rousseeuw (1987)가 제안한 방법으로 실루엣(silhouette)은 내부유효성 측도로 널리 사용되고 있다. 유전자 j에 대해, 유전자 j와 유전자 j가 속하지 않은 다른 군집들의 평균 차이를 a_j라고 한다. 유전자

표 **5.1.** 3개 푸리에 계수로 유전자 선별 후 3개 푸리에 계수와 원래 유전자발현데이터로 각각 군집분석한 결과

군집방법		군집별 실루엣			실루엣		Connectivity	Dunn
		1	2	3	mean	med		
모형기반	군집내 유전자 수	10	19	20				
	푸리에 계수	0.24	0.32	0.32	0.21	0.19	25.27	0.15
	유전자발현값	0.15	0.14	−0.03	0.07	0.10		
	군집내 유전자 수	17	19	13				
	Original (tp = 24)	0.0006	0.21	0.33	0.17	0.23	25.92	0.32
K-평균법	군집내 유전자 수	19	20	10				
	푸리에 계수	0.32	0.32	0.24	0.30	0.29	25.27	0.15
	유전자발현값	0.15	−0.03	0.14	0.07	0.10		
	군집내 유전자 수	18	15	16				
	Original (tp = 24)	0.27	0.03	0.27	0.20	0.23	18.23	0.39
PAM	군집내 유전자 수	18	11	20				
	푸리에 계수	0.15	0.35	0.26	0.24	0.27	33.65	0.07
	유전자발현값	0.002	0.17	−0.002	0.04	0.10		
	군집내 유전자 수	23	17	9				
	Original (tp = 24)	0.18	0.34	0.06	0.21	0.19	26.41	0.31
SOM	군집내 유전자 수	10	27	12				
	푸리에 계수	0.24	0.11	0.36	0.19	0.18	27.27	0.15
	유전자발현값	0.10	−0.06	0.25	0.05	0.08		
	군집내 유전자 수	15	19	15				
	Original (tp = 24)	0.30	0.20	0.08	0.19	0.23	24.96	0.35
Ward	군집내 유전자 수	14	17	18				
	푸리에 계수	0.18	0.27	0.32	0.26	0.26	15.71	0.16
	유전자발현값	0.14	0.13	0.01	0.09	0.13		
	군집내 유전자 수	17	29	13				
	Original (tp = 24)	0.01	0.23	0.31	0.17	0.21	16.42	0.40
Fuzzy	군집내 유전자 수	14	18	17				
	푸리에 계수	0.19	0.30	0.31	0.27	0.30	28.75	0.11
	유전자발현값	0.05	0.13	−0.02	0.06	0.10		
	군집내 유전자 수	17	19	13				
	Original (tp = 24)	0.0006	0.21	0.33	0.17	0.23	25.92	0.32

j가 속하지 않은 군집 l이 있다고 가정할 때, 군집 l의 개체들과 유전자 j의 평균 차이를 b_{jl}이라고 한다. $b_j = \min_l b_{jl}$이라 하면 유전자의 실루엣(silhouette) 값은

$$S(i) = \frac{b_i - a_i}{\max(b_i, a_i)} \tag{4.3}$$

으로 정의되며 −1에서 1 사이의 값으로 나타난다. 실루엣 값은 개체들이 군집에 얼마나 잘 군집화 되었는지를 측정해주는 값으로 큰 값이 나올수록 군집화가 잘 되었다고 볼 수 있다 (김재희와 고윤실, 2009).

5. Yeast 데이터 분석

5.1. 효모 데이터 설명

분석에 이용한 데이터는 효모 마이크로어레이 발현 데이터로 cdc15 방법을 이용하여 값을 측정한 데이

표 **5.2.** 4개 푸리에 계수로 유전자 선별 후 4개 푸리에 계수와 원래 유전자발현데이터로 각각 군집분석한 결과

군집방법		군집별 실루엣			실루엣		Connectivity	Dunn
		1	2	3	mean	med		
모형기반	군집내 유전자 수	3	4	22				
	푸리에 계수	0.14	0.42	0.23	0.25	0.26	15.38	0.38
	유전자발현값	−0.21	0.05	0.12	0.07	0.11		
	군집내 유전자 수	1	10	18				
	Original (tp = 24)	0	0.17	0.15	0.15	0.18	18.14	0.32
K-평균법	군집내 유전자 수	5	19	5				
	푸리에 계수	0.28	0.21	0.2	0.22	0.22	23.93	0.22
	유전자발현값	−0.13	0.05	0.08	0.02	0.05		
	군집내 유전자 수	1	9	19				
	Original (tp = 24)	0	0.16	0.18	0.17	0.17	5.86	0.77
PAM	군집내 유전자 수	9	7	13				
	푸리에 계수	0.12	0.22	0.14	0.15	0.16	30.96	0.17
	유전자발현값	−0.07	0.03	0.04	0	0.02		
	군집내 유전자 수	10	12	7				
	Original (tp = 24)	0.03	0.09	0.06	0.06	0.06	35.73	0.25
SOM	군집내 유전자 수	5	20	4				
	푸리에 계수	0.22	0.21	0.39	0.24	0.25	19.16	0.22
	유전자발현값	−0.13	0.08	0.04	0.04	0.04		
	군집내 유전자 수	5	17	7				
	Original (tp = 24)	0.003	0.15	−0.03	0.08	0.08	12.83	0.33
Ward	군집내 유전자 수	6	19	4				
	푸리에 계수	0.32	0.09	0.41	0.18	0.2	6.41	0.33
	유전자발현값	−0.1	0.05	0.04	0.02	0.03		
	군집내 유전자 수	1	4	24				
	Original (tp = 24)	0	0.42	0.14	0.17	0.17	5.86	0.77
Fuzzy	군집내 유전자 수	6	19	4				
	푸리에 계수	−0.1	0.05	0.04	0.23	0.26	23.93	0.22
	유전자발현값	0.31	0.13	0.26	0.02	0.04		
	군집내 유전자 수	10	4	15				
	Original (tp = 24)	0.01	0.4	0.16	0.14	0.16	25.62	0.31

터이다. 효모는 인간을 포함한 고등생물체와 같은 진핵세포로 구성되는 생물 모델이기 때문에 그 유전자 발현 데이터 분석은 생물정보연구에 의미가 있다. cdc(cell division cycle)는 세포분열을 조절하는 유전자를 지칭하며, 세포 주기(cell cycle)는 세포가 분열을 시작하여 다음 세포 분열이 일어날 때 까지를 말하여 엄마세포의 분열과 그 후에 일어나는 딸세포의 분열 사이의 기간에 의해 정의될 수 있다.

5.2. 데이터 분석

Screening 분석 시 검정통계량에 사용된 푸리에 계수의 개수 $J = 3, 4, 5$에 대해 FDR을 통해 유전자 선별작업을 수행한 결과 전체 유전자수 4381개 중에서 $J = 3$일 경우에는 49개, $J = 4$일 경우에는 29개, $J = 5$일 경우에는 418개의 유전자가 선별되었다. 여기서 선별을 보수적으로 하기위해 스펙트럼 계산 시 유전자별로 표본분산과 표본공분산을 이용하였다. 각 경우에 대해 푸리에 계수의 개수를 변화시켜가며 푸리에 계수들로만 군집분석을 하고 또한 선별된 유전자의 24개 시점에서의 유전자 발현값들로도 군

표 **5.3.** 5개 푸리에 계수로 유전자 선별 후 5개 푸리에 계수와 원래 유전자발현데이터로 각각 군집분석한 결과

군집방법		군집별 실루엣			실루엣		Connectivity	Dunn
		1	2	3	mean	med		
모형기반	군집내 유전자 수	227	74	67				
	푸리에계수	0.35	−0.17	−0.26	0.16	0.3	197.79	0.02
	유전자발현값	0.19	−0.11	−0.22	0.07	0.15		
	군집내 유전자 수	204	175	39				
	Original (tp = 24)	0.08	0.21	0.01	0.13	0.16	226.97	0.08
K-평균법	군집내 유전자 수	42	157	219				
	푸리에계수	0.2	0.2	0.29	0.26	0.25	97.95	0.03
	유전자발현값	0.22	0.04	0.06	0.15	0.16		
	군집내 유전자 수	40	186	192				
	Original (tp = 24)	0.21	0.22	0.18	0.2	0.21	100.17	0.1
PAM	군집내 유전자 수	123	105	190				
	푸리에계수	−0.005	0.05	0.35	0.17	0.22	163.46	0.02
	유전자발현값	0.002	−0.003	0.19	0.09	0.11		
	군집내 유전자 수	186	196	36				
	Original (tp = 24)	0.21	0.17	0.21	0.19	0.2	119.32	0.1
SOM	군집내 유전자 수	132	126	160				
	푸리에계수	−0.07	0.35	0.02	0.09	0.12	108.42	0.03
	유전자발현값	−0.02	0.14	0.02	0.05	0.06		
	군집내 유전자 수	155	79	184				
	Original (tp = 24)	−0.03	0.18	0.04	0.04	0.04	102.14	0.09
Ward	군집내 유전자 수	250	128	40				
	푸리에계수	0.28	0.18	0.17	0.24	0.26	11.12	0.13
	유전자발현값	0.14	0.12	0.13	0.13	0.14		
	군집내 유전자 수	150	212	56				
	Original (tp = 24)	0.32	0.08	0.01	0.16	0.17	14.23	0.2
Fuzzy	군집내 유전자 수	145	162	111				
	푸리에계수	−0.08	0.03	0.34	0.07	0.1	131.23	0.02
	유전자발현값	−0.04	0.02	0.17	0.04	0.04		
	군집내 유전자 수	153	186	79				
	Original (tp = 24)	0.25	0.16	−0.006	0.16	0.18	157.48	0.1

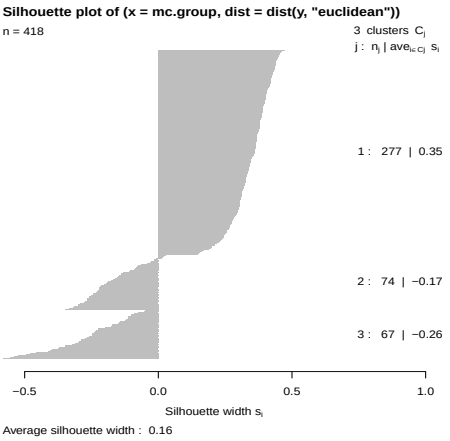
집분석을 실시하였다. 각 경우마다 모형기반 군집분석에서 BIC를 고려하면 최적군집 수는 3개로 결정된다. 표 5.1은 $J = 3$으로 유전자를 선별한 후 3개의 푸리에 계수(군집차원 = 3)로 군집분석 한 결과와 원래 유전자 발현데이터(군집차원 = 24)로 군집분석 한 결과이다. 표 5.2는 $J = 4$로 유전자를 선별한 후 4개의 푸리에 계수(군집차원 = 4)로 군집분석 한 결과와 원래 유전자 발현데이터(군집차원 = 24)로 군집분석 한 결과이다. 표 5.3은 $J = 5$로 유전자를 선별한 후 5개의 푸리에 계수(군집차원 = 5)로 군집분석 한 결과와 원래 유전자 발현데이터(군집차원 = 24)로 군집분석 한 결과이다. 선별과정에서 사용한 푸리에 계수의 개수, 검정통계량에서 사용한 푸리에 계수의 개수, 군집방법 등에 따라 군집결과의 차이를 보여준다. 표 5.1, 표 5.2, 표 5.3에서 군집분석별로 보면 첫 번째 행은 군집별 유전자 수이고 두 번째 행은 푸리에 계수로 형성된 군집에서 푸리에 계수들로 계산한 실루엣값이다. 세 번째 행은 푸리에 계수로 형성된 군집에서 원래 유전자 발현값으로 계산한 실루엣값으로 두 실루엣값은 차이를 보여준다. 네 번째 행은 24개 시점에서 측정된 유전자 발현값 모두를 이용하여 군집을 형성하고 이렇게 형성된 군

표 **5.4.** GO분석($J = 5$인 경우)

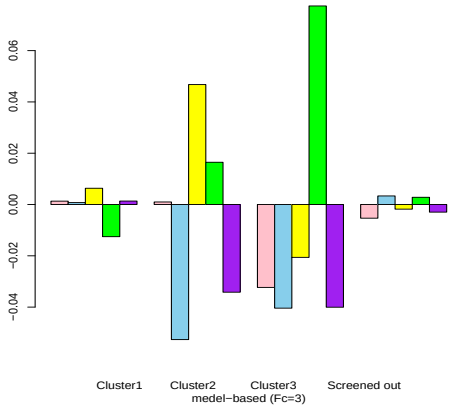
Cluster	Term	유전자 개수					
		모형기반	K-평균법	PAM	SOM	Ward	Fuzzy
Cluster A	Establishment of protein localization	33	0	24	0	0	0
	Protein transport	32	0	23	0	0	0
	Vesicle-mediated transport	33	0	22	0	25	0
	Intracellular transport	44	0	28	0	35	0
	DNA metabolic process	0	19	0	0	0	0
	Cellular response to stress	0	18	0	0	34	0
	DNA repair	0	17	0	0	19	0
	Response to DNA damage stimulus	0	17	0	0	22	0
	Regulation of transcription	0	0	0	21	35	0
	Cell cycle	0	16	0	19	0	18
	Phosphorus metabolic process	24	28	19	12	26	0
	Chromosome organization	0	11	0	0	23	12
	Chromatin modification	0	0	0	21	15	8
	Coenzyme metabolic process	0	0	0	7	0	8
Cluster B	Cell cycle	20	43	25	25	18	26
	DNA metabolic process	14	0	23	23	13	25
	Cellular response to stress	11	0	22	24	8	24
	Regulation of transcription	16	0	22	22	10	26
	Response to DNA damage stimulus	11	0	19	21	8	21
	Cell cycle phase	15	0	14	0	12	21
	Vesicle-mediated transport	0	31	0	14	0	15
	Phosphorus metabolic process	0	28	11	11	0	12
	Response to abiotic stimulus	0	23	0	8	0	0
Cluster C	Cell cycle	19	15	23	30	23	30
	Cellular response to stress	15	0	0	0	0	0
	DNA metabolic process	14	0	0	0	0	0
	Chromosome organization	11	8	0	0	0	0
	Translational elongation	0	9	11	0	0	0
	Organell fission	9	10	10	12	11	12
	Sexual reproduction	0	7	10	13	11	13
	M phase	9	10	14	16	13	17

집들에 따라 계산된 실루엣값을 다섯 번째 행에서 보여준다. 24개 시점 유전자 발현값 모두를 사용한 경우보다 푸리에 계수를 이용하여 군집을 형성했을 경우의 실루엣값이 더 높은 편임을 알 수 있다. 그러므로 푸리에계수를 이용했을 경우 차원축소효과와 더불어 더 높은 실루엣값을 얻음을 알 수 있다.

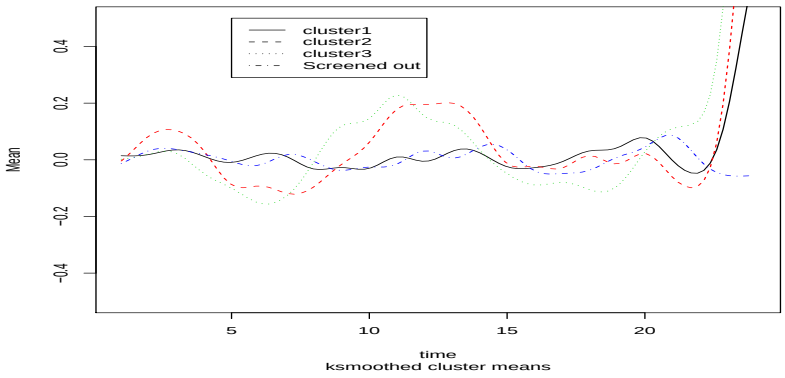
결과를 살펴보면 대체적으로 푸리에 계수를 이용하여 군집화한 결과가 원데이터를 이용하여 군집화 한 결과보다 내부유효성이 좋은 것을 알 수 있다. 먼저, 푸리에 계수로 군집화한 경우를 고려해보자. 표 5.1의 실루엣값을 살펴보면 모형기반 군집방법과 K-평균법, 연결성 측도와 Dunn 지수를 살펴보면 계층적 군집방법의 내부 유효성이 크다. 표 5.2의 실루엣값과 Dunn 지수를 살펴보면 모형기반 군집방법, 연결성 측도를 살펴보면 계층적 군집방법의 내부 유효성이 크다. 표 5.3의 실루엣값을 살펴보면 K-평균법, 연결성 측도와 Dunn 지수를 살펴보면 계층적 군집방법의 내부 유효성이 큰 것을 알 수 있다. 또한, 원래 유전자 발현값으로 군집화한 경우를 살펴보고자한다. 푸리에 계수를 이용하여 군집화하는 경우보



(a) 모형기반(5개 푸리에계수) 실루엣



(b) 모형기반 군집별 푸리에 계수 평균값

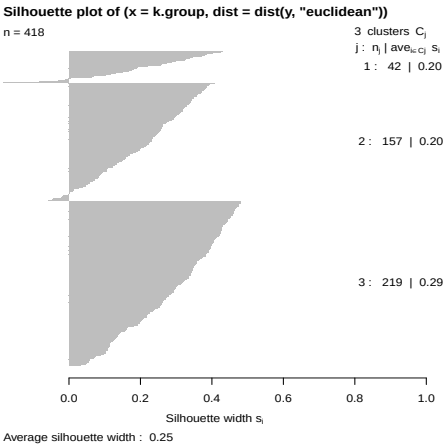


(c) 모형기반 군집별 평균

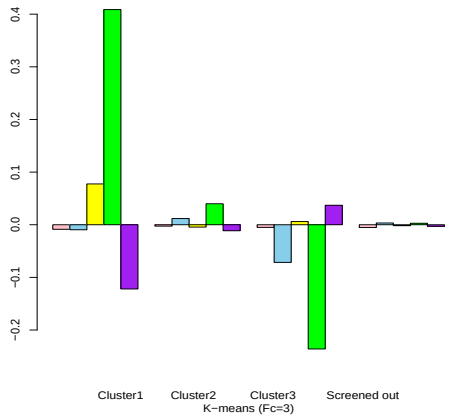
그림 5.1. 모형기반 군집분석 결과

다 실루엣값은 작은 편이지만 Dunn지수는 큰 편이다. 표 5.1의 실루엣값을 살펴보면 PAM 방법, 연결성 측도와 Dunn 지수를 살펴보면 계층적 군집방법의 내부 유효성이 크다. 표 5.2의 실루엣값과 Dunn 지수, 연결성 측도를 살펴보면 모두 계층적 군집방법의 내부 유효성이 크다. 표 5.3의 실루엣값을 살펴보면 PAM 방법, 연결성 측도를 살펴보면 계층적 군집방법, Dunn 지수를 살펴보면 모형기반 군집방법의 내부 유효성이 큰 것을 알 수 있다. 모형기반의 실루엣값과 생물학적 특성을 보여주는 패턴을 고려하여 $J = 5$인 경우를 최종모형으로 선택한 후 그룹별 특성을 보여주고자한다.

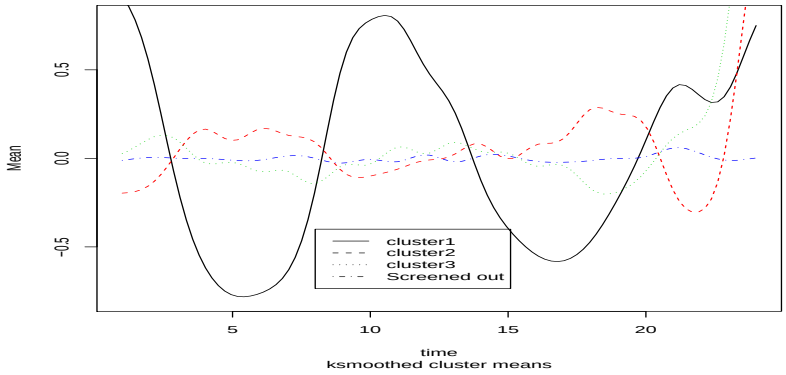
그림 5.1~그림 5.6에서는 푸리에 계수를 이용한 경우 각 군집분석 방법별로 (a)는 각 군집내 개체별 실루엣값과 더불어 군집별 평균 실루엣값, (b)는 군집별 푸리에계수 평균값, (c)에서는 군집별 24개 시점에서 유전자발현값들의 평균그래프를 보여준다. 이와 같은 그래프를 통해 각 군집의 특성을 파악할 수 있다. 그림 5.1은 모형기반 군집분석 결과, 그림 5.2는 K-평균법 군집분석 결과, 그림 5.3은 PAM 군집분석 결과, 그림 5.4는 SOM 군집분석 결과, 그림 5.5는 Ward 군집분석 결과, 그림 5.6은 Fuzzy 군집분석 결과이다. 군집별 평균그래프를 보면 모형기반은 비슷한 주기 모양의 이동된 군집을 보여주며



(a) K-평균법(5개 푸리에계수) 실루엣



(b) K-평균법 군집별 푸리에 계수 평균값



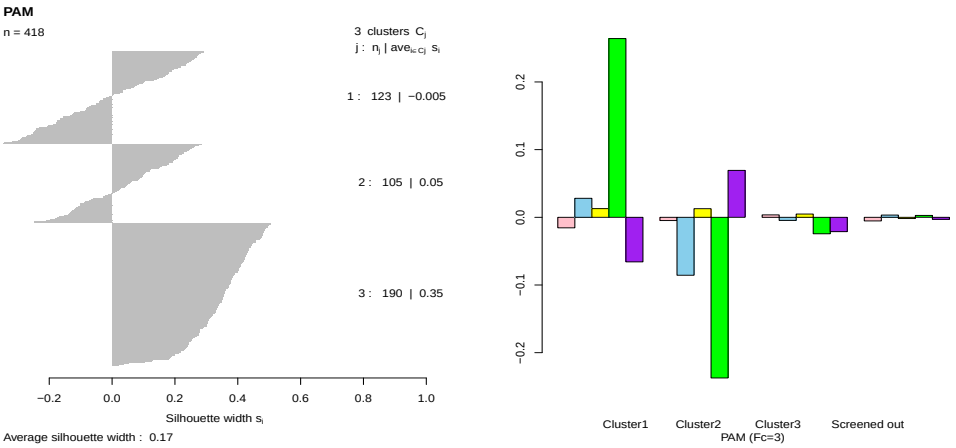
(c) K-평균법 군집별 평균

그림 5.2. K-평균법 군집분석 결과

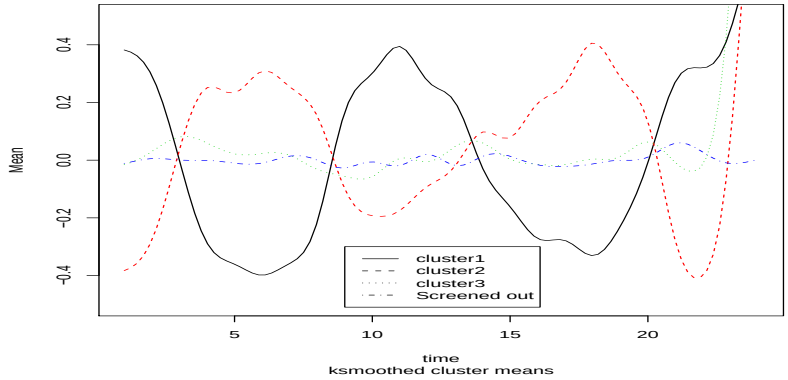
Ward, Fuzzy 방법 결과가 비슷하고 K-평균법, PAM, SOM 군집평균 그래프에서는 뚜렷한 주기성이 나타나는 그래프를 보여준다.

5.3. 생물학적 의미

검정통계량에 사용된 푸리에 계수의 개수 $J = 5$일때 나타나는 군집들의 유전자 발현 평균값은 모형기반 분석을 제외한 나머지 다섯가지 군집분석법 모두에서 시간에 따라 증가하고 감소하는 뚜렷한 패턴의 변화를 보여 주었다(그림 5.1(c), 그림 5.2(c), 그림 5.3(c), 그림 5.4(c), 그림 5.5(c), 그림 5.6(c)). 시간에 따른 유전자 발현 평균값의 패턴이 같은 군집에 속하는 유전자들은 각각이 암호화하는 단백질들을 동일한 시기에 생성함을 의미하며 이들 생성된 단백질들 중 일부는 공통적인 생물학적 기능에 관여한다고 할 수 있다. 각 cluster에 속하는 유전자들이 어떠한 생물학적 기능에 관여하는가를 예측하기 위하여 DAVID Bioinformatics Resources 6.7(National Institute of Allergy and Infectious Diseases(NIAID),



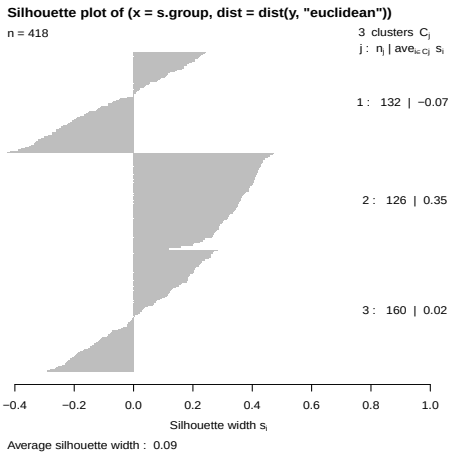
(a) PAM(5개 푸리에계수) 실루엣 (b) PAM 군집별 푸리에 계수 평균값



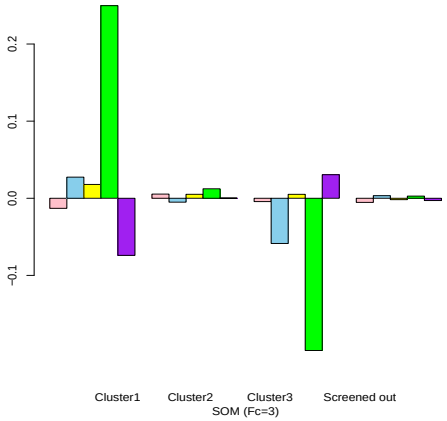
(c) PAM 군집별 평균

그림 5.3. PAM 군집분석 결과

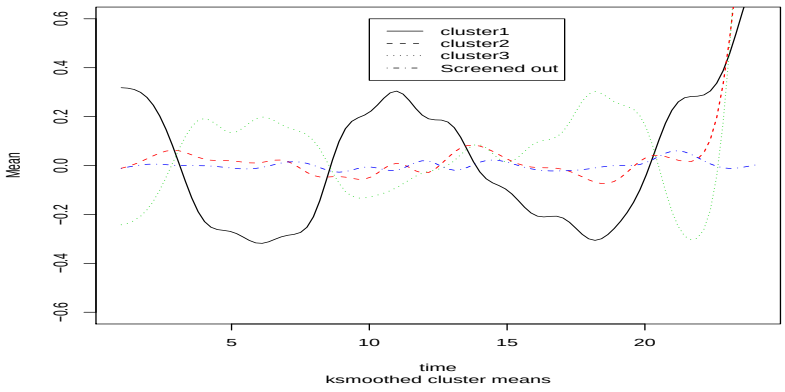
NIH) (Huang 등, 2009)을 이용한 GO(Gene Ontology) 분석을 하였다 (표 5.4). 각 통계기법 별로 군집당 유전자 개수가 가장 많은 4개의 GO term을 분석해 본 결과 Cluster A에서는 모형기반과 PAM이 유사성을 보이며 또한 SOM, Ward, Fuzzy 기법의 결과가 서로 유사함을 보여주었다. K-평균법에 의한 군집은 모형기반, PAM, SOM, Fuzzy 기법의 결과에서는 보이지 않는 DNA repair와 같은 stress 연관 유전자들이 발현되는 것을 알 수 있었다. Cluster B에 포함되는 유전자들의 기능을 분석한 결과는 K-평균법을 제외한 나머지 다섯가지의 군집분석법에서 유사한 결과들을 보여 주었다. 세포분열과 DNA 복제와 연관된 유전자들이 발현되는 것이 모형기반, PAM, SOM, Ward, Fuzzy 분석법에서 관찰된 반면 세포내 단백질 이동과 인(phosphorus)의 대사와 관련된 일련의 유전자 군이 K-평균법에 의한 군집에 포함되어 있음을 알 수 있었다. Cluster C에 속하는 유전자들은 모형기반을 제외한 나머지 다섯가지 군집분석 방법에서 유사하게 나타났는데 세포분열과 관련된 스트레스 연관 유전자들이 모형기반 분석법에 의한 군집에 포함되어 있는 반면 세포분열의 각 단계를 조절하는 M phase에 관여하는 유전자들



(a) SOM(5개 푸리에계수) 실루엣



(b) SOM 군집별 푸리에 계수 평균값

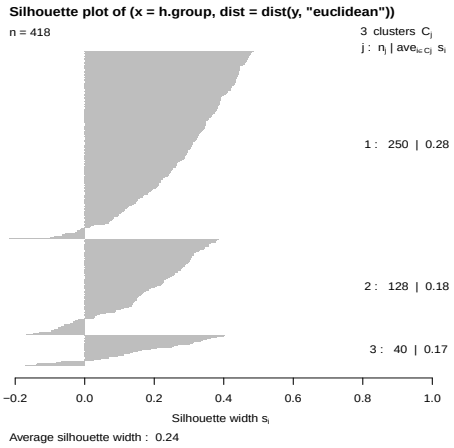


(c) SOM 군집별 평균

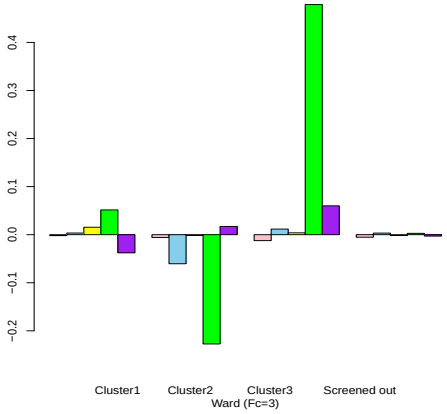
그림 5.4. SOM 군집분석 결과

과 단백질 합성에 관여하는 유전자들이 K-평균법, PAM, SOM, Ward, Fuzzy 통계 기법에 의한 군집들에 발견됨을 알 수 있었다.

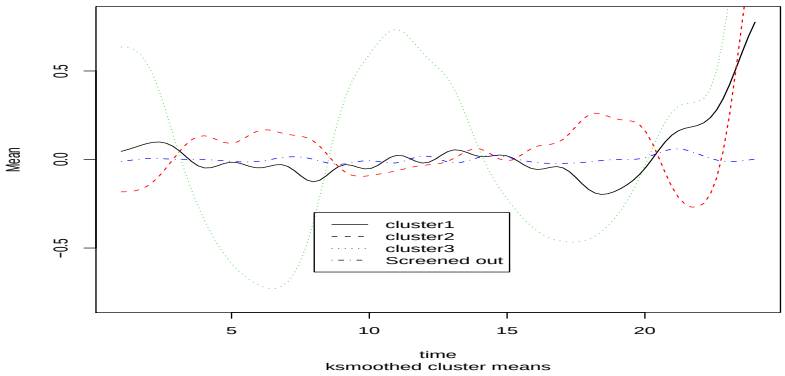
통계적인 분석으로 클러스트링된 유전자들의 발현 패턴은 기능이 아직까지 밝혀지지 않은 미지의 유전자들의 기능을 유추하고 또한 각 유전자들이 어떻게 협동적으로 상호조절하여 생체 기능을 활성화시킬 수 있는가에 대한 정보를 제공할 수 있다. 본 연구에서 제시된 통계적인 유전자 발현의 군집의 결과는 특정한 생물학적 process가 시간대별로 진행됨을 암시하며 또한 어떠한 유전자들이 그 과정에 참여하는지 보여준다. 또한 통계기법의 종류에 따라 특정 군집에 동일하게 나타나는 유전자들과 더불어 특이적인 발현 양상을 보이는 군집도 존재하여 여러 통계 기법의 활용에 따라 다양한 생물학적 결과를 예측할 수 있음을 보여준다. 그러나 통계적인 데이터가 제시하는 생물학적 process가 실제적으로 효모세포에서 일어나는가에 대한 실험적인 증명이 또한 필요하다 하겠다.



(a) Ward(5개 푸리에계수) 실루엣



(b) Ward 군집별 푸리에 계수 평균값



(c) Ward 군집별 평균

그림 5.5. Ward 군집분석 결과

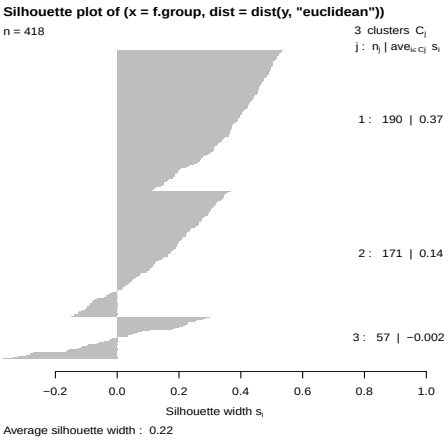
6. 결론

본 연구에서는 효모 cdc15 유전자 데이터를 이용하여 유전자 선별 작업을 한 후에 군집분석을 해보았다.

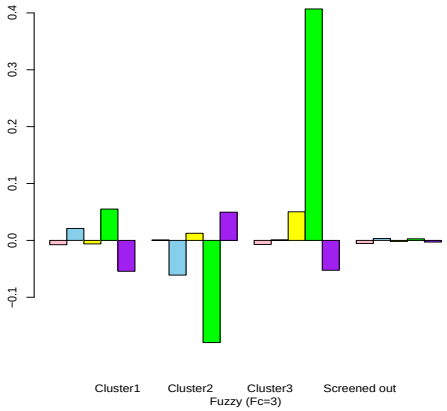
군집화 방법 중에서 분포를 가정하는 모형 기반 군집방법과 비계층적방법인 K-평균법, PAM, 자기 조직화 지도(SOM), 퍼지(fuzzy) 방법, 그리고 계층적 방법인 Ward 방법을 이용하였다.

선별방법으로는 시계열 특성을 고려한 검정통계량을 이용하여 FDR 다중검정법으로 유전자 선별을 한 후 군집화한 결과를 얻었다. 선별작업을 통해 군집시 고려한 유전자 수를 줄여 군집 계산시 효율성이 높아졌다. 유전자 선별작업 후 원데이터를 이용한 군집화와 푸리에 계수를 이용한 군집화를 실시하여 비교하였다.

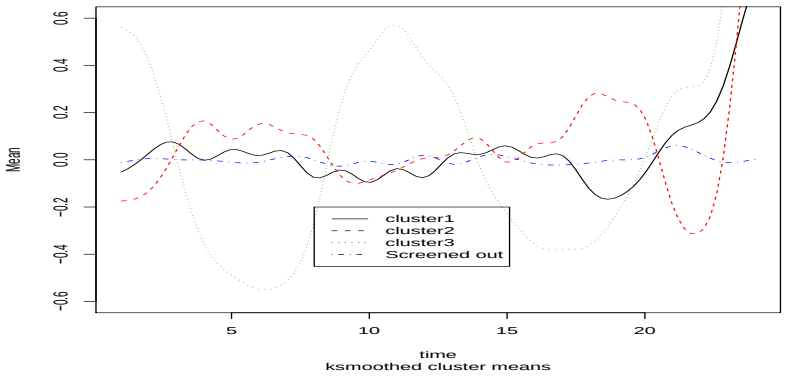
군집화는 군집방법과 군집의 개수에 따라 개체들의 군집화에 큰 차이가 있기 때문에 데이터의 특성과 상황을 고려하여 적절한 군집방법과 군집의 개수를 정하는 것이 중요하며 다각적인 연구와 분석이 필요하



(a) Fuzzy(5개 푸리에계수) 실루엣



(b) Fuzzy 군집별 푸리에 계수 평균값



(c) Fuzzy 군집별 평균

그림 5.6. Fuzzy 군집분석 결과

다. 또한 군집방법에 따라 서로 다른 결과를 보여주고 있기 때문에 생물학적 의미를 고려하여 결정할 수 있다.

참고문헌

김재희 (2011). <R 다변량 통계 분석>, 교우사, 서울

김재희, 고윤실 (2009). 군집분석 비교 및 한우 관능평가데이터 군집화, <응용통계 연구>, **22**, 745–758.

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: Series B*, **57**, 289–300.

Bickel, D. R. (2011). Estimating the null distribution to adjust observed confidence levels for genome-scale screening, *Bioinformatics*, **67**, 363–370.

Datta, S. and Datta, S. (2005). Empirical Bayes screening of many p-values with application to microarray studies, *Bioinformatics*, **21**, 1987–1994.

Dudoit, S., Shaffer, J. P. and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments, *Statistical Science*, **18**, 71–103.

Dunn (1974). Well-separated clusters and optimal fuzzy partitions, *Journal of Cybernetics*, **4**, 95–104.

Eckel, J. E., Gennings, C., Chinchilli, V. M., Burgoon, L. D. and Zacharewski, T. R. (2004). Empirical Bayes gene screening tool for time-course or dose-response microarray data, *Journal of Biopharmaceutical Statistics*, **14**, 647–670.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation, *Journal of the American Statistical Association*, **97**, 611–631.

Fraley, C. and Raftery, A. E. (2006). *MCLUST Version 3 for R: Normal mixture modeling and model-based clustering*, Technical Report No. 504.

Gentleman, R., Caray, V. J., Huber, W., Irizarry, R. A. and Dudoit, S. (2005). *Bioinformatics and computational biology solutions using R and bioconductor*, Spinger, New York.

Getz, G., Levine, E., Domany, E. and Zhang, M. Q. (2000). Super-paramagneic clustering of yeast expression profiles. *Physica*, **A279**, 457–464.

Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201–3212.

Huang, D. W., Sherman, B. T. and Lempicki, R. A.(2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources, *Nature Protocols*, **4**, 44–57.

Hero, A. O., Fleury, G., Mears, A. J. and Swaroop, A. (2004). Multicriteria gene screening for analysis of differential expression with DNA microarrays, *Journal on Applied Signal processing*, **2004**, 43–52.

Izenman, A. J. (2008). Modern Multivariate Statistical Techniques, *Spinger*, New York.

Kim, B. R., Littell, R. C. and Wu, R. (2006). Clustering periodic patterns of gene expression based on fourier appoximations, *Current Genomics*, **7**, 197–203.

Kim, J. and Hart, J. D. (1998). Test for change when the data are dependent, *Journal of Time Series*, **19**, 399–424.

Kim, J. and Kim, H. (2008). Clustering of change using Fourier coefficient, *Bioinformatics*, **24**, 184–191.

Kim, J., Ogden, R. T. and Kim, H. (2011). A method of identify differential expression profile with time-course gene data and Fourier transformation, *BMC Bioinformatics*, in revision.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, New York.

Kohonen, T. (1998). The self-organizing map, *Neurocomputing*, **21**, 1–6.

Ma, S. (2006). Empirical study of supervised gene screening, *BMC Bioinformatics*, **7**, 537.

Rousseeuw, P. T. (1987). Silhouettes: Graphical aid to the interpretation and validation of cluster analysis, *Journal of Computation Applied Math*, **20**, 53–65.

Serban, N. and Wasserman, L. (2005). CATS: Clustering after transformation and smoothing, *Journal of the American Statistical Association*, **471**, 990–999.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

Törönen, R., Kolehmainen, M., Wong, G. and Castrén, E. (1999). Analysis of gene expression data using self-organizing maps, *Federation of European Biochemical Societies*, **451**, 142–146.

Zhang, L., Zhang, A. and Ramanathan, M. (2003). Fourier harmonic approach for visualizing temporal patterns of gene expression data, *IEEE Computer Society Bioinformatics Conference*, **2**, 137–147.

Gene Screening and Clustering of Yeast Microarray Gene Expression Data

Kyunga Lee[1] · Taehoun Kim[2] · Jaehee Kim[3]

[1]Department of Statistics, Duksung Women's University
[2]Department of PrePharmMed Duksung Women's University
[3]Department of Statistics, Duksung Women's University

(Received September 2011; accepted November 2011)

Abstract

We accomplish clustering analyses for yeast cell cycle microarray expression data. To reflect the characteristics of a time-course data, we screen the genes using the test statistics with Fourier coefficients applying a FDR procedure. We compare the results done by model-based clustering, K-means, PAM, SOM, hierarchical Ward method and Fuzzy method with the yeast data. As the validity measure for clustering results, connectivity, Dunn index and silhouette values are computed and compared. A biological interpretation with GO analysis is also included.

Keywords: Connectivity, Dunn index, Fourier coefficient, FDR, Fuzzy, Microarray expression data, Model-based clustering, K-means, PAM, Silhouette, SOM, Ward method, yeast.

[3]Corresponding author: Professor, Department of Statistics, Duksung Women's University, 419 Ssangmun-Dong, Dobong-Gu, Seoul 132-714, Korea. E-mail: jaehee@duksung.ac.kr