Time Control Microarray 자료의 군집 분석에 관한 고찰

손인석[1], 이재원[2]

<요약>

생물학자들은 시간 패턴에 따라 발현 수준이 변화하는 유전자의 군집화를 시도하고 있다. 지금까지는 군집 방법의 비교 연구가 주로 진행되어 왔으나, 군집화 이전의 유전선택 방법에 따라 군집화 결과가 달라지기 때문에 유전자 선택 단계도 같이 고려되어야 한다. 따라서 본 연구에서는 Time Control Microarray 자료를 가지고 군집 분석을 하는데 있어서 유전자 선택, 군집 분석 방법의 선택, Validation 방법의 선택 등 3가지 요인별로 보다 폭 넓은 비교 연구를 하였다.

주요용어 : Microarray, Gene selection, Clustering analysis, Validation

1. 서론

cDNA 마이크로어레이는 수 많은 유전자의 발현 수준을 동시에 관찰 연구하는 새롭고 유망한 바이오기술로서 생물학 또는 의학계의 넓은 영역에 걸쳐 적용이 증가하고 있는 추세이다. 마이크로어레이 데이터 분석의 핵심적인 목표중의 하나는 유전자들이 서로 연계되어 특정한 기능을 하는 유전자를 구별하거나 시공간적으로 유전자가 발현되는 패턴의 차이를 구분하는 데에 있다.

Eisen et al. (1998)의 계층적 군집분석(hierarchical clustering)의 적용을 시작으로, Tamayo et al. (1999)은 Self-Organizing Map (SOM)을 사용하여 마이크로어레이자료를 분석하는 방법을 제시하였다. Hastie et al. (2000)은 계층적 군집분석과 주성분 분석을 혼합한 gene shaving 방법을 제안하였으며, 또한 찾아진 몇 개의 군집간의 교호작용을 모형화하기 위하여 tree harvesting을 제시하였다. Tree harvesting은 각 군집을 구성하는 유전자들의 발현양상들을 평균하여 각 군집의 발현을 하나의 벡터로써 표현한다. 그리고 알려진 결과변수(例:생존시간)를 반응변수로 하고 각 군집의 발현 정도를 설명변수로 하여 다중회귀모형을 구성하되 각 군집간의 교호작용도 반영할 수 있는 모형이라 주목을 받고 있다. Model based clustering 방법으로 Laura and Owen(2000)은 Plaid model을 사용하였는데 이는 2-way clustering alogorithm으로서 유전자들은 하나 이상의 군집에 속할 수도 있으며 어떤 군집에도 속하지 않을 수도 있다. 최근에는 Pan et al.(2002)이 normal mixture model-based clustering을 제안한 바 있으며, Ghosh and Chinnaiyan(2002)는 clustering 결과의 신뢰도를 평가하기 위한 mixture model based 방법을 사용하기도 하였다. 이 밖에도 Bayesian Clustering의 적용 (Barash and Friedman, 2001)이나 Singular Value Decomposition을 이용한 방법등(Alter et al.,2000; Kishino and Waddell, 2000)도 사용되고 있다.

Kerr et al.(2001)는 군집 알고리즘의 신뢰성 평가를 위해 재표본에서 잔차와 선형 모델

1) 고려대학교 통계학과 박사과정, (130-701) 서울특별시 성북구 안암동 5가 1번지
2) 고려대학교 통계학과 교수, (136-701) 서울특별시 성북구 안암동 5가 1번지

(ANOVA)을 사용하였고, Chen et al.(2002)는 동질성(Homogeneity)과 이질성(Separation) 같은 군집 결과의 물리적인 특성으로 많은 군집 알고리즘의 성능을 비교하였다. Yeung et al.(2001)은 Figure of Merit (FOM)의 개념을 소개하였고, Datta et al(2003)는 군집 알고리즘에 의해 생성된 그룹들의 일치성을 체크하는 목적으로 3개의 서로 다른 Validation 기준을 소개하였다.

지금까지는 군집 방법의 비교 연구가 주로 진행되어 왔으나, 군집화 이전의 유전선택 방법에 따라 군집화 결과가 달라지기 때문에 유전자 선택 단계도 같이 고려되어야 한다. 따라서 본 연구에서는 Cyanobacterium sp. PCC 6803 (Hihara et al., 2001) 자료를 가지고 군집 분석을 하는데 있어서 유전자 선택, 군집분석 방법의 선택, Validation 방법의 선택 등 3가지 요인별로 보다 폭 넓은 비교 연구를 하였다.

2. 군집 분석에서의 요인

2.1 유전자 선택

. t-test

t 통계량은 다음과 같다.

$$t = \frac{\overline{M}}{s/\sqrt{n}} \quad , \tag{1}$$

여기서 $\overline{M}$은 반복된 어레이에서의 유전자들의 발현강도비의 로그값인 M값의 평균이고, s는 M값의 표준편차이다.

. SAM

SAM(Significant Analysis of Microarrays, Tusher et al., 2001) 통계량은 다음과 같다.

$$d = \frac{\overline{M}}{(s + s_0)/\sqrt{n}}, \tag{2}$$

여기서 s는 M값의 표준편차이고, s_0는 유전자 발현이 작은 분산에 의존하지 않게 하기 위해서 더해주는 양의 상수이다.

. B-statistic

Lonnstedt and Speed (2002)은 모수적이고 경험적인 베이즈적인 접근을 시도하였으며, 각 유전자에 대해서 사후 로그 오즈를 추정하는 B-statistic를 사용한다. 조건부 정규성 가정 하에서 파라미터 a와 v가 추정되어 지며, B 통계량은 다음과 같다.

$$B = \log \frac{p}{1-p} \frac{1}{\sqrt{(1+nc)}} \left[\frac{a + s^2 + \overline{M^2}}{a + s^2 + \frac{\overline{M^2}}{1+nc}} \right]^{v+n/2}, \tag{3}$$

여기서 p 는 유전자의 사전 비율 p이고, a 와 v 는 분산에 대한 상위모수 이며, c는 평균에 대한 상위모수이다.

. Fold Change

$\overline{x_{i1}}$ 와 $\overline{x_{i2}}$은 각 두 조건 하에서 n-번째 유전자의 평균 발현이다. 즉, 각 유전자의 값이 $|\overline{x_{i1}} / \overline{x_{i2}}| \geq t$ 이거나 $|\overline{x_{i1}} / \overline{x_{i2}}| \geq 1/t$ 이면 유의한 유전자라고 말한다.

2.2 군집화 방법

. Hierarchical clustering

계층적 군집 방법(Eisen et al, 1998)은 사전에 고정된 군집의 개수 대신 계층적 군집을 생성한다. 초기 수준에 근거하여 각 관찰치 각자의 군집을 형성하고, 각 다음 수준에서 가장 가까운 두 군집을 결합하여 더 큰 하나의 군집을 형성한다. 계층적 군집 방법은 가장 흔하고 간단한 나무구조 방법이다.

. K-means

K-means 군집 방법은 사전에 군집 수를 고정하고, 초기 군집을 중심으로하여, 그룹 내 총 제곱합(total within-class sum of squares)을 최소화하기 위해 관찰치를 다양한 군집으로 할당한다. 복잡하고 반복적인 수치 알고리즘은 이 최소값을 찾기위해 사용되었다.

. Diana

Diana(Datta et al.,2003)은 하나의 군집에서 여러 개의 군집으로 분열하는(Divisive) 군집방법이다. Diana은 초기에 모든 관찰치들이 하나의 군집을 이룬 후 다시 유사성 있는 것끼리 군집을 나누는 방법이다.

. Fanny

Fanny 군집 방법(Datta et al., 2003)은 fuzzy logic을 이용하고 각 관찰치에 대해서 확률 벡터가 생성된다. 최종 선택된 군집은 가장 높은 확률을 가진 관찰치를 그룹에 할당하므로써 결정된다. k를 원하는 군집의 총 개수라고 하고, Fanny는 일종의 비상사성 걸리의 가중 평균이 되는 목적함수를 최소화하는 모든 유전자 확률 벡터를 계산한다. 최종 선택된 군집은 가장 높은 확률을 가지는 그룹에게 유전자를 할당하므로써 생성되어진다.

. Model-based clustering

Model based clustering(Laura and Owen, 2000)은 모든 데이터를 Mixture 분포에서 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 하자. 발현 profiles $x_1, \ldots, x_n$ 우도는 다음과 같다.

$$L(\theta, \gamma) = \prod_{i=1}^{n} f_{\gamma_i}(x_i, \theta_{\gamma_i}) \tag{4}$$

알려지지 않은 그룹 수준 r는 γ 와 θ 를 게 최대화시키는 최대우도 방법에 의해 얻어진다.

. PAM (Partitioning around Medoids)

PAM 군집 방법은 분할법(partitioning method)에 해당되는 군집화 방법으로서 Kaufman and Rousseeuw(1990)가 제안한 방법이다. 이 알고리즘은 관측개체들 중의 대표값을 이용하는데, 그 대표값을 메도이드(medoid)라고 한다. 이 k개의 메도이드를 찾은 후에 각각의 메도이드와 가장 가까운 점들을 할당시켜 군집을 형성한다. k-메도이드를 찾는 목적은 관측개체의 점들과 가장

가까운 메도이드와의 거리의 합을 최소화하는 데에 있다. 이 방법은 k-평균 군집화 방법 보다 더 로버스트하고 효율적으로 계산하는 경향이 있다고 알려져 있다.

. Fuzzy c-means clustering

Fuzzy c-means 방법(Guthke et al., 2000) 은 패턴 인식에서 자주 사용된다. 이 알고리즘은 하드c-Means(HCM) 클러스터링의 퍼지 모델의 결과로 알고리즘의 결과로 클러스터 중심과 퍼지c-분할 행렬을 동시에 구할 수 있는 자기 조직화, 무 관리자 학습의 대표적인 예이다. 퍼지 ISODATA는 이러한 퍼지c-Means 클러스터링에 발견적 특성을 가미한 알고리즘이다.

2.3 Validation 방법

. Homogeneity and Separation

동질성(Homogeneity)은 각 유전자의 발현 Profile과 각 유전자가 속하는 군집의 중심사이의 평균거리를 계산한다. 즉,

$$H_{avg} = \frac{1}{N} \sum_{i} D(g_i, C(g_i)) , \tag{5}$$

여기서 g_i는 i 번째 유전자, $C(g_i)$는 g_i가 속하는 유전자의 군집의 중심, N는 유전자 총 수, D는 거리함수 이다. 이질성(Separation)는 군집 중심들 간의 가중치 평균 거리로서 계산한다.

$$S_{avg} = \frac{1}{\sum_{i \neq j} N_{ci} N_{cj}} \sum_{i \neq j} N_{ci} N_{cj} D(C_i, C_j) , \tag{6}$$

여기서 C_i 와 C_j 는 i 번째와 j 번째 중심이고, and N_{ci} 와 N_{cj} 는 i 번째와 j 번째 군집에 있는 유전자의 개수이다.

. The aggregate figure of merit

$$FOM(t, k) = \sqrt{\frac{1}{n} \sum_{i=1}^{k} \sum_{i \in C_i} (R(x, t) - \overline{x}C_i(t))^2}, \tag{7}$$

여기서 $R(x, e)$는 시점 t에서 유전자 x의 발현 Profile을 나타내고, $\overline{x}C_i(t)$은 $C_i(t)$ 에 있는 유전자들의 평균 발현 Profile을 나타낸다.

각 각의 m 조건에 속하는 유전자를 사용한다. The aggregate figure of merit, $FOM(k) = \sum_{e=1}^{m} FOM(e, k)$ 는 k 군집에 대한 총 예측력의 추정이다.

. The average proportion of non-overlap measure

$$V_1(K) = \frac{1}{Ml} \sum_{g=1}^{M} \sum_{i=1}^{l} \left(1 - \frac{n(C^{g,i} \cap C^{g,0})}{n(C^{g,0})}\right), \tag{8}$$

이 척도(measure)는 전체 데이터와 시간마다 발현 수준을 제거한 데이터에 근거한 군집 분석을 사용하여 같은 군집에 들어 있지 않은 유전자들의 (평균) 비율을 계산한다.

. The average distanace between means measure

$$V_1(K) = \frac{1}{Ml} \sum_{g=1}^{M} d(\overline{x}C^{g,i}, \overline{x}C^{g,0}), \quad (9)$$

여기서 $\overline{x}C^{g,0}$은 $C^{g,0}$에 있는 유전자들의 평균 발현 Profile을 나타내고, $\overline{x}C^{g,1}$은 $C^{g,1}$에 있는 유전자들의 평균 발현 Profile을 나타낸다. 이 척도는 전체 데이터와 시간마다 발현 수준을 제거한 데이터에 근거한 군집 분석을 사용하여 같은 군집에 들어 있는 모든 유전자들의 평균 발현 비율(log transformed) 간의 (평균)거리를 계산한다.

. The average distance measure

$$V_3(K) = \frac{1}{Ml} \sum_{g=1}^{M} \sum_{i=1}^{l} \frac{1}{n(C^{g,0})n(C^{g,i})} * \sum_{g \in C^{g,0}, g' \in C^{g,i}} d(x_g, x_{g'}), \quad (10)$$

여기서 $d(x_g, x_{g'})$는 유전자 g와 g' 발현 Profile 간의 거리이다.

이 척도는 전체 데이터와 시간마다 발현 수준을 제거한 데이터에 근거한 군집 분석을 사용하여 같은 군집에 들어 있는 모든 유전자들간의 평균 거리를 계산한다.

3. 실제 자료 분석

본 연구에서는 3가지 요인(유전자 선택, 군집 분석 방법의 선택, Validation 방법의 선택)에 따라 군집화 결과가 어떻게 달라지는 비교 연구를 하기 위해서 4가지 유전자 선택 방법(t-test, Fold chagne, SAM, B-statistic), 7가지 군집 방법(Hierachical clustering, K-means, Diana, Fanny, Model-bayed clustering, PAM, Fuzzy c-means clustering), 6가지 Validataion 방법 (Homogeneity, Separation, FOM, The average proportion of non-overlap measure, The average distance between means measure, The average distance measure)를 사용하였다.

먼저 데이터 전처리는 M.J.L et al.(2002)와 동일하게 하였다. 첫째, Background 보정을 하였고, 두 번째, Cy3이나 Cy5 강도가 2000보다 작은 유전자는 제거하였으며, 셋째, 유전자 912개를 Global Mean Normalization을 실시하였다. Fold Change 분석(Hihara et al., 2001)은 한 시점에서 2개 이상 반복에서 2 Fold Change에 벗어나는 유전자를 유의하다고 하였으며, T-test(DE hoon et al., 2002)을 유의수준 0.001로, B-statistc은 p를 0.001로 하였으며, SAM은 delta를 1.4로 하였다. 유전자 선택으로 선택된 Data Set을 가지고 Validation 방법으로 군집 방법을 비교 연구 하였다.

Cyanobacterium sp. PCC 6803 데이터(Hihara et al, 2001)는 광합성을 하는 Cyanobacteria을 낮은 빛(Low Light)에서 높은 빛(Hight Light)에 노출 시간에 따른 유전자의 발현 양상을 보았으며, 3079 ORF 발현 수준을 낮은 빛 조건에 있는 Cyanobacteria와 높은 빛에 노출된 Cyanobacteria을 15분, 1시간, 6시간, 15시간에서 측정하였다. 각 시간에서 측정된 수는 15분에서 6번, 1시간에서 6번, 6시간에서 4번, 15시간에서 4번이다.

참고 문헌

1. Banfield,J.D. and Raftery,A.E. (1993) Model-based Gaussian and non-Gaussian clustering. *Biometrics*, **49**, 803-822.
2. Chen,G. *et al.*, (2002) Evaluation and comparison of clustering algorithms in analyzing ES

cell gene expression data. *Statistica Sinica*, **12**, 241-262.
3. Chu,S., DeRisi,J. *et al.*, (1998) The transcriptional program of sporulation in budding yeast. *Science*, **282**, 699-705.
4. 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.
5. DR Goldstein, E Conlon, D Ghosh (2002), "Statistical issues in the clustering of gene expression data", *Statistica Sinica*, 12(1):219-240.
6. MJL de Hoon, S Imoto, S Miyano (2002), "Statistical analysis of a small set of time-ordered gene expression data using linear splines", *Bioinformatics*, **18**, 1477-1485.
7. 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. USA*, **95**, 14863-14868.
8. GK Smyth, YH Yang, T Speed (2003), "Statistical issues in cDNA microarray data analysis", *in Functional Genomics: Methods and Protocols*, eds. MJ Brownstein and AB
9. Hartigan,J.A. (1975) *Clustering Algorithms*. Wiley, New York.
10. Hartigan,J.A. and Wong,M.A. (1979) A *k*-means clustering algorithm. *Applied Statistics*,
11. 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.
12. Kerr,M.K. and Churchill,G.A. (2001) Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. *Proc. Natl Acad. Sci. USA*, **98**, 8961-8965.
13. Lonnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31-46.
14. McLachlan,G.J., Bean,R.W. and Peel,D. (2002) A mixture modelbased approach to the clustering of microarray expression data. *Bioinformatics*, **18**, 1-10.
15. MJL de Hoon, S Imoto, S Miyano (2002), "Statistical analysis of a small set of time-ordered gene expression data using linear splines", **Bioinformatics**, **18**:1477-1485.
16. Sandrine Dudoit, Yee Hwa Yang, Terry Speed, Matthew J Callow (2002) "Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments", Statistica Sinica, 12(1):111-139.
17. Spellman,P.T. *et al.*, (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell.*, **12**, 3273-3297.
18. Susmita Datta, Somnath Datta (2003), "Comparisons and validation of statistical clustering techniques for microarray gene expression data", *Bioinformatics*, **19**:459-466.
19. Venables,W.N. and Ripley,B.D. (1998) *Modern Applied Statistics with S-Plus, (3rd corrected printing)*, 2nd edn, Springer, New York.
20. 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.
21. Waddell,P. and Kishino,H. (2000) Cluster inference methods and graphical models evaluated on NC160 microarray gene expression data. *Genome Informatics*, **11**, 129-140.
22. Yeung,K., Haynor,D.R. and Ruzzo,W.L. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**, 309-318.