Clustering gene expression data using Non -Negative matrix factorization

Non-negative matrix factorization을 이용한 마이크로어레이 데이터의 클러스터링

Min-Young Lee, Ji-Hoon Cho, In-Beum Lee*

Department of Chemical Engineering, POSTECH, Pohang, Korea

*To whom correspondence should be addressed. E -mail: iblee@postech.ac.kr

Abstract

마이크로어레이 (microarray) 기술이 개발된 후로 연관된 유전자 클러스터 (cluster)를 찾는 문제는 깊이 연구되어왔다. 이 문제는 핵심적인 과제 중 하나는 생물학적으로 타당한 클러스터의 수를 결정하는 데 있다. 본 논문은 최적의 클러스터 수를 결정하는 기준을 제시하고, non-negative factorization (NMF)를 이용해 클러스터 centroid의 패턴을 찾는 방법을 제안한다. NMF에 의해 발견된 각각의 패턴은 생물학적 프로세스의 특정 부분으로 해석될 수 있다. NMF는 factor matrix의 entity를 non-negative로 제약 (constraint)하고, 이 제약은 오직 additive combination만 허용하기 때문에 이러한 부분적인 패턴을 찾아낼 수 있다. NMF의 유용성은 이미지 분석과 텍스트 분석에서 이미 입증되어 있다. 본 논문에서 제안한 방법에 의해 위의 패턴과 유사한 발현 패턴을 갖는 유전자를 모을 수 있었다. 제안된 방법은 human fibroblast 데이터와 yeast cell cycle 데이터에 적용해 성능을 입증하였다.

Introduction

최근 DNA 마이크로어레이 기술의 발전으로 세포의 다양한 생물학적 메커니즘에 관한 연구를 도울 수 있는 데이터가 축적되었다. 이러한 데이터에서 가치 있는 정보를 추출하기 위한 다양한 데이터 마이닝 기법이 개발되어 왔다. 그 중, 클러스터링은 일반적으로 서로 관련된 유전자 클러스터를 찾는데 널리 사용되어 왔다.

Hierarchical clustering (HC) [1]는 비슷한 패턴을 갖는 유전자나 샘플의 클러스터를 찾는 가장 일반적인 방법이다. 이 방법은 사용하기 쉽고, 클러스터가 형성되는 과정이나 클러스터의 발현 패턴을 해석하는 것을 도와주는 그래프 (dendrogram)를 제공한다. 하지만 데이터의 구조를 트리 모양으로 제한하고 두 샘플간의 거리를 측정하는 기준에 민감하다는 단점이 있다. 또 클러스터의

개수는 dendrogram의 특정 부분을 잘라서 결정하는데, 이때 주관적인 판단이 개입될 가능성도 있다. Self-organizing map (SOM) [2]은 multi-dimensional 데이터에서 중요한 feature를 찾아내어 클러스터링을 돕는다. SOM은 데이터의 nonlinear 성질을 반영하는 장점이 있지만, learning rate나 window function같은 parameter에 민감하고, 계산시간이 상당히 길다.

Matrix decomposition 방법은 데이터에서 서로 다른 패턴을 찾아내기 때문에 클러스터링에 사용될 수 있다. Singular value decomposition (SVD) 또는 principle component analysis (PCA) [3]는 데이터에서 orthogonal basis vector를 찾아낸다. 하지만 이 vector는 생물학적으로 해석하기가 어렵다. Bayesian decomposition (BD) [4]는 non-orthogonal basis vector를 찾아낸다. 이 vector는 생물학적인 의미를 지니기는 하지만 이를 찾기 위해서는 시스템에 대한 지식이 필요하고 dimension을 reduction할 때 rank를 정하기가 어려운 단점이 있다.

Non-negative matrix factorization [5, 6]은 PCA나 BD처럼 matrix decomposition을 위해 개발된 방법이다. 하지만 PCA나 BD와는 달리 NMF는 데이터의 부분적인 특징을 갖는 basis vector (part-based representation)를 추출한다. 이 vector들로 데이터를 전체적으로 설명할 수 있게 된다. NMF는 얼굴 이미지 분석에서는 basis vector가 코나 눈처럼 얼굴의 부분적인 이미지로 추출되었다 [5]. 마찬가지로 마이크로어레이 데이터에서는 이런 basis vector는 생물학적인 프로세스의 일부가 된다고 생각할 수 있다. 이미 Brunet et al. [7]은 NMF를 샘플을 클러스터링하는 문제를 연구했다. 이전 연구에서는 클러스터의 개수를 결정하기 위해 consensus matrix와 cophenetic correlation coefficient를 이용했다.

이러한 matrix decomposition 방법의 공통된 문제는 최적의 클러스터의 개수를 찾는데 있다. SVD는 일반적으로 eigenvalue와 eivenvalue의 개수를 플롯해서 그래프의 기울기가 0에 가까울 때의 개수를 클러스터의 개수로 결정한다 [3]. 하지만 어느 지점에서 기울기가 0에 가까운지 결정하는 것은 상당히 주관적인 문제이다. BD를 사용하는 방법에서는 여러 데이터 베이스에서 필요한 정보를 얻고 Markov chain, wavelet decomposition으로 분석을 해서 클러스터의 개수를 결정한다 [4]. 따라서 전체적인 과정은 상당히 복잡하다. NMF를 이용한 방법에서는 Brunet et al. [7]이 consensus clustering 방법으로 클러스터의 개수를 결정했지만, 유전자를 클러스터링할 때는 matrix의 크기가 커지기 때문에 계산이 어렵게 된다.

본 논문에서는 NMF를 이용해 유전자를 클러스터링하는 방법을 제안한다. NMF의 결과로 추출된 패턴은 클러스터의 centroid에 해당한다. 또 클러스터의 개수를 결정하는 기준을 제시한다. 클러스터의 개수는 factorization rank에 해당한다. 제안된 방법은 human fibroblast와 yeast cell cycle 데이터에 적용해 성능을 평가한다.

Methods

NMF algorithm

NMF의 목적은 다음과 같은 factorization matrix를 찾는 것 이다.

$$\mathbf{V} \approx \mathbf{WH} \qquad (1)$$

여기서 **V**는 $n \times m$ non-negative matrix이고 **W**와 **H**는 각각 $n \times r$, $r \times m$ matrix이다. **W**와 **H**를 계산하기 위해 다음의 divergence를

const function으로 사용한다.

$$D(\mathbf{V}\|\mathbf{WH}) = \sum_{i,j} \mathbf{V}_{ij} \log \frac{\mathbf{V}_{ij}}{(\mathbf{WH})_{ij}} - \mathbf{V}_{ij} + (\mathbf{WH})_{ij} \quad (2)$$

위 식은 $\mathbf{V} = \mathbf{WH}$ 일 때 0의 값을 갖고 그 외에는 항상 0보다 크다. **W**와 **H**의 초기값은 랜덤으로 지정되고 다음의 multiplicative rule에 의해 업데이트된다.

$$\mathbf{H}_{kj} \leftarrow \mathbf{H}_{kj} \frac{\sum_i \mathbf{W}_{ik} \mathbf{V}_{ij} / (\mathbf{WH})_{ij}}{\sum_{p=1}^{n} \mathbf{W}_{pk}}$$

$$\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik} \frac{\sum_j \mathbf{H}_{kj} \mathbf{V}_{ij} / (\mathbf{WH})_{ij}}{\sum_{q=1}^{m} \mathbf{H}_{kq}} \quad (3)$$

여기서 i는 1에서 n (유전자의 개수), j는 1에서 m (샘플의 개수), k는 2에서 r (NMF의 factorization rank)까지 이다. 자세한 알고리즘과 증명은 [6]의 논문에 있다. 최적의 **W**와 **H**는 divergence의 값이 더 이상 변하지 않을 때 이다. 이 rule은 오직 additive combination만 허용하기 때문에 **W**와 **H**는 sparse matrix가 된다. 위의 식으로부터 **W**의 column은 얼굴 이미지 분석에서 코나 입으로 표현되는 basis vector가 되고 **H**의 row는 원래 데이터 **V**의 reduced dimensional representation이 된다. 즉, 전체적인 유전자 발현 패턴은 몇 개의 패턴의 조합으로 간주 될 수 있다. 결과적으로 **H**의 각 row는 클러스터의 centroid가 되는 것이다. 또 **W**의 i번째 row에 대해 k번째 값이 가장 크면 i번째 유전자는 k번째 클러스터에 속한 다고 할 수 있다.

Criterion for determining the number of clusters

Brunet et al. [7]의 연구에서 factorization rank r은 클러스터의 개수로 간주되었다. 그들은 r 값을 증가시키면서 NMF를 여러 번 반복하여 consensus matrix를 만들고 이로부터 cophenetic correlation coefficient를 계산하였다. 클러스터의 개수는 coefficient 값이 감소할 때의 rank로 결정되었다. 하지만 유전자를 클러스터링할 때는 consensus matrix의 크기가 커지기 때문에 계산하기가 어려워서 이 방법은 유전자 클러스터를 찾는 데는 적합하지 않다.

본 논문에서 제시하는 방법은 다음과 같다. 여기서 factorization rank r은 최적의 클러스터 개수로 간주된다. Factorization rank k를 2에서 r까지 증가시키면서 클러스터의 centroid와 각 클러스터에 속하는 유전자 사이의 거리를 측정해서 모두 합산하여 이를 with-in cluster scatter WS라고 정의한다. 클러스터링이 적절한 결과를 보인다면 이 값은 작을 것 이다. 하지만 이 WS는 클러스터의 개수가 증가할수록 감소한다. 따라서 이 효과를 제거하기 위해 랜덤 데이터를 만들고 위와 같은 방법으로 WS_{ref}를 계산한다. 랜덤 데이터는 일정한 패턴을 갖지 않기 때문에 WS와 WS_{ref}의 비율은 클러스터링이 적절하게 될 때 최소값을 갖게 된다. 제안된 기준은 다음과 같다.

$$Ratio = \frac{WS}{WS_{ref}} = \frac{\sum_{k=1}^{r} \sum_{i \in k\text{th cluster}} \left\| \mathbf{v}_i^k - \mathbf{c}^k \right\|^2}{\sum_{k=1}^{r} \sum_{i \in k\text{th cluster}} \left\| \mathbf{v}_{ref,i}^k - \mathbf{c}_{ref}^k \right\|^2} \quad (4)$$

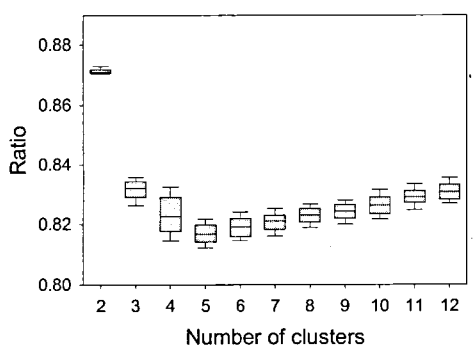
여기서 $\mathbf{v}_i^k$는 k번째 클러스터에 속하는 i번째 유전자이고 $\mathbf{c}^k$는 k번째 클러스터의 centroid이다. 하첨자 ref는 랜덤 데이터의 경우를 나타낸다. 거리를 계산할 때 $\mathbf{v}_i^k$와 $\mathbf{c}^k$는 평균은 0이고 분산은 1로 조정되었다.

Biological data

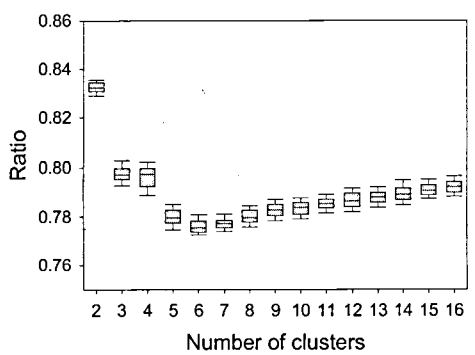
본 논문에서 제안된 방법은 human fibroblast [8] 데이터와 yeast cell cycle [9]데이터에 적용되었다. Iyer et al. [8]에 의해 만들어진 Human fibroblast 데이터는 517개의 유전자에 대해 13개 샘플의 발현 패턴으로 이루어져 있다. Yeast cell cycle 데이터는 Spellman et al. [9]에 의해 만들어졌고 763개의 유전자와 17개의 샘플로 이루어져 있다. Yeast 데이터는 $\log_2$ base로 되어 있어서 non-negative matrix로 전환하기 위해 liner scale로 조정했다.

Results

제안된 방법을 human fibroblast 데이터와 yeast cell cycle 데이터에 적용할 때 NMF가 초기값에 의존하는 것을 고려하여 100번 실행을 하고 ratio의 평균값을 얻어 클러스터의 개수를 결정했다.



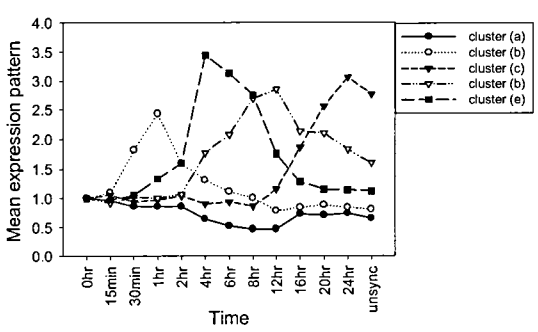
(a)



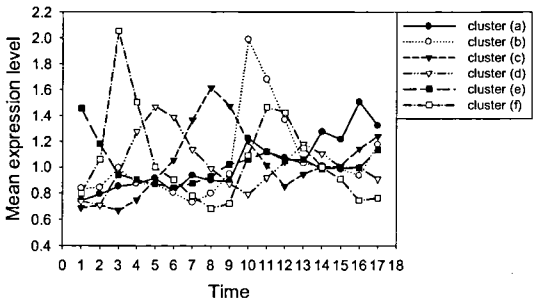
(b)

Fig. 1. ratio value(boxplot of 100 times run) (a) fibroblast 데이터: 최적 클러스터 개수는 5로 결정되었다. (b) yeast cell cycle 데이터: 최적 클러스터 개수는 6으로 결정되었다.

Fig. 1(a)에서 fibroblast에 적용한 결과를 보면 적절한 클러스터의 개수가 5로 결정되었다. 이는 이전의 graph-theoretic approach [10]나 CLICK 알고리즘 [11]을 이용한 연구와 같은 결과이다. 결정된 클러스터의 평균 유전자 발현 패턴은 Fig. 2(a)에 있다. 이는 Iyer et al. [7]의 논문에서 제시된 결과와도 일치한다. 클러스터 (a)에는 p57Kip2, wee1-like protein kinase, p27Kip1, cyclin A 과 alpha importin를 포함하여 263개의 유전자가 있다. 클러스터 (b)는 58개의 유전자, 클러스터 (c)는 furin을 포함하여 73개의 유전자를 갖는다. 클러스터 (d)는 metallothionein 1A 와 tumor associated antigen L6를 포함하여 77개의 유전자를 갖고 클러스터 (e)는 NET1을 포함해 47개의 유전자를 갖는다. 각각의 클러스터는 Iyer et al.의 논문에서 tissue remodeling 클러스터, angiogenesis 클러스터, unidentified role 클러스터, cytoskeletal reorganization 클러스터에 해당한다. 전체적인 클러스터링 결과는 Fig. 4(a)에 heat map의 형태로 나와있다.



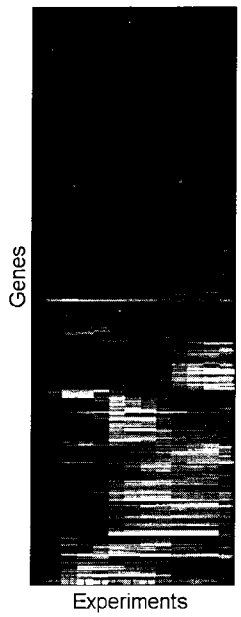
(a)



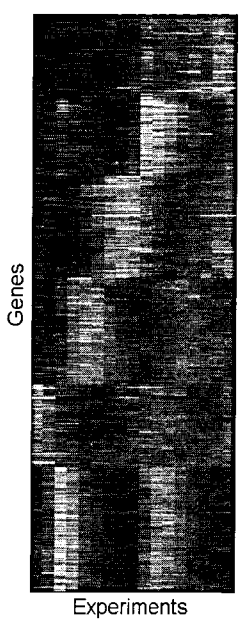
(b)

Fig. 2. 각 클러스터의 평균 유전자 발현 패턴. (a) fibroblast 데이터 (b) yeast cell cycle 데이터

Yeast cell cycle 데이터에서는 Fig. 1(b)에서 보듯이 최적의 클러스터 개수는 6개로 제안된다. 이는 Bayesian decomposition을 이용한 Moloshok et al. [4]의 연구결과와 일치한다. 또 각 클러스터의 평균 유전자 발현 패턴도 매우 유사하고 각 클러스터에 포함된 유전자도 대부분 일치한다. 결과는 Fig. 2(b)에서 볼 수 있다. 각 클러스터를 살펴 보면 클러스터 (a)는 M phase에서 G1 phase로 이동될 때 큰 peak를 보여서 Moloshok et al.의 논문에서 cluster M/G1에 해당한다. 이 클러스터는 Spellman et al. [9]에 의해서 발견된 CDC6를 비롯하여 전체 109개의 유전자를 포함하고 있다. 클러스터 (b)는 G1 phase에서 발현이 많이 되기 때문에 클러스터 G1a와 유사하며 CLN2, RNR1, CDC9, RAD27와 SMC3을 포함해 Moloshok et al.의 연구에서와 거의 동일한 165개의 유전자가 포함되어 있다. 클러스터 (c)는 클러스터 S/G2와 비슷하며 CLB4와 WHI3를 비롯하여 140개의 유전자가 있다. 클러스터 (d)는 클러스터 M에 해당하며 DBF2, CLB2, CDC5, CDC20와 SWI5를 포함하고 있다. 전체적인 클러스터의 패턴은 Fig. 4(b)에서 heat map의 형태로 알 수 있다.



(a)

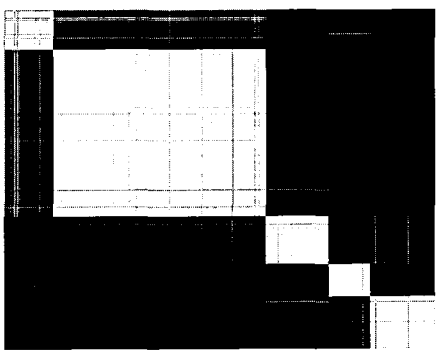


(b)

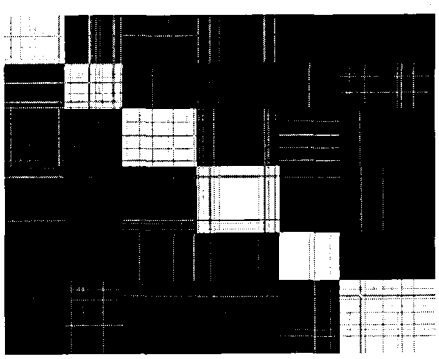
Fig. 3. 클러스터링 결과의 heat map. (a) fibroblast 데이터 (b) yeast cell cycle 데이터.

NMF는 위에서 이미 언급된 것처럼 초기값

에 의존하기 때문에 제안된 결과가 초기값에 따라 불안정하게 달라지는 지를 알아보기 위해 Brunet et al. [7]이 제안한 것과 같은 방법으로 consensus matrix를 만들었다. 클러스터링을 20번 반복하여 만든 consensus matrix는 Fig. 4에서 알 수 있다.



(a)



(b)

Fig. 4. 20번 실행했을 때의 Consensus matrix. (a) fibroblast 데이터 (b) yeast cell cycle 데이터

그림에서 알 수 있듯이 제안된 방법은 상당히 안정적인 결과를 제시해주는 것을 알 수 있다.

즉, 본 논문에서 제안한 방법은 최적의 클러스터의 개수를 찾고, 각 클러스터의 centroid와 연관된 유전자를 안정적으로 찾아준다. 또 각 클러스터는 이전 연구와 비교하여 대부분 같은 유전자를 포함하고 있다.

Discussion

NMF는 이미지 분석과 텍스트 분석의 영역에서 효과적으로 사용된 방법이다. 본 논문에서는 NMF를 이용하여 클러스터의 centroid를 추출하고 각 클러스터와 연관된 유전자를 효과적으로 찾는 알고리즘을 제시하였다. 또 최적의 클러스터 개수를 결정하는 방법도 제시하였다. 이전에 제안된 방법들은 많은 parameter와 주관적인 판단을 필요로 하기 때문에 상당히 복잡하고 불안정한 결과를 낼 가능성이 있다. 하지만 본 논문에서 제안된 방법은 factorization rank r 하나의 parameter만 결정하면 되고 consensus matrix처럼 큰 matrix를 계산할 필요가 없기 때문에 훨씬 간단하고 빠르게 안정적인 결과를 보여줄 수 있다.

본 논문에서 제안된 방법은 human fibroblast 데이터와 yeast cell cycle 데이터에 적용해 성능을 평가하였다. 제안된 기준으로부터 클러스터의 개수가 타당한 값으로 결정되었고 각 클러스터의 centroid로 추출되었다. 동시에 각 클러스터와 연관된 유전자도 찾을 수 있었다. 따라서 본 논문은 복잡한 마이크로어레이 데이터를 분석하여 유용한 정보를 얻기 위한 다른 연구나 실험에 기초가 되는 정보를 제공하여 생물학적 메커니즘을 규명하는데 도움을 줄 수 있을 것으로 기대된다.

References

[1] Eisen, M., Spellman, P., Brown, P. and Botstein, D., Cluster Analysis and Display Genome-wide Expression Patterns, *Proc. Natl.*

Acad. Sci. 95, 1998, 14863–14868.

[2] Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Dmitrovsky, E., Lander, E.S. and Golub, T.R., Interpreting Patterns of Gene Expression with Self-organizing maps: Methods and Application to hepatopoietic differentiation, *Proc. Natl. Acad. Sci.* 96, 1999, 2907–2912.

[3] Alter, O., Brown, P.O. and Botstein, D., Singular Value Decomposition for Genome-wide Expression Data Processing and Modeling, *Proc. Natl. Acad. Sci.* 97, 2000, 10101–10106.

[4] Moloshok, T.D., Klevecz, R.R., Grant, J.D., Manion, F.J., Speier, W.F. 4th, and Ochs, M.F., Application of Bayesian Decomposition for Analyzing Microarray Data, *Bioinformatics* 18, 2002, 566–575.

[5] Lee, D.D. and Seung, S.H., Learning the Parts of Objects by Non-negative Matrix Factorization, *Nature* 401, 1999, 788-791.

[6] Lee, D.D. and Seung, H.S., Algorithms for Non-negative Matrix Factorization, *Adv. Neural Info. Proc. Syst.* 13, 2001, 556–562.

[7] Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P., Metagenes and Molecular Pattern Discovery using Matrix Factorization, *Proc. Natl. Acad. Sci.* 97, 2004, 4164–4169.

[8] Iyer, V.R., Eisen, M.B., Ross, D.T., Schuler, G., Moore, T., Lee, J.C., Trent, J.M., Staudt, L.M., Hudson, J., Jr., Boguski, M.S., Lashkari, D., Shalon, D., Botstein. D. and Brown, P.O., The Transcriptional Program in the Response of Human Fibroblasts to Serum, *Science* 283, 1999, 83-87.

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

[10] Xu, Y., Olman, V. and Xu, D., Clustering Gene Expression Data using a Graph-theoretic approach: an Application of Minimum Spanning Tree, *Bioinformatics* 18, 2002, 536-545.

[11] Sharan, R. and Shamir, R., CLICK: A Clustering Algorithm with Applications to Gene Expression Analysis, *Proceedings of AAAI-ISMB*, 2000, 307-316.

[12] Duda, R.O., Hart, P.E. and Stork, D.G. Pattern Classification, Second Edition, 2001, John Wiley and Sons, New York.