생물학적 기능 데이터베이스 기반 유전자 군집화 단계에서의 최적 군집개수 추정방법 탐구

백교훈[1] 김재영[2] 신미영[3]
[1]경북대학교 전자공학과, [2]경북대학교 정보통신학과, [3]경북대학교 전자전기컴퓨터학부
{khbaek, widebrowboy}@ee.knu.ac.kr, shinmy@knu.ac.kr

Estimation of Optimal Number of Gene Expression Data Clusters based on Biological Knowledge

Kyo-Hoon Baek[1],Jae-YoungKim[2],Mi-YoungShin[3]
[1]Dept. of Electronic Engineering, Kyungpook National University
[2]Dept. of Information Communication, Kyungpook National University
[3]School of Electrical Engineering & Computer Science, Kyungpook National University

마이크로어레이 기술의 개발은 한 실험에 하나의 유전자를 대상으로 하던 기존 실험방식으로부터 유전체 단위의 대규모 유전자를 대상으로 실험할 수 있게 해줌으로써 기능유전체학 발전에 크게 기여하였다. 이와 함께 마이크로어레이 실험으로부터 생성된 유전자 발현데이터를 이용해 유전자의 기능을 규명하기 위한 시도가 이루어져 왔고 현재에도 활발한 연구들이 진행 중에 있다. 그 중에서도 특히, 계층형 군집화 방법을 이용한 분석[1]이나 자기조직화 방법[2], 그리고 k-평균 방법[3] 등을 이용한 군집분석방법은 유전자의 상호기능 및 의미분석 등에 있어서 성공적인 결과를 제시해왔다. 하지만 이러한 방법들은 유전자 발현데이터만을 군집분석과정에 이용하므로 분석데이터 주체인 해당 생물의 세포활동과 같은 "실제적인 생물 특성"을 반영할 수 없으며, 효모 세포주기(Yeast cell-cycle) 발현데이터와 같은 시계열 데이터를 다루는 문제에 있어서, 유전자의 각 시간에 따른 특징간의 다중적이고 연속된 상호관계를 생물학적으로 해석하는 것에 한계가 있다[4].

최근 이를 보완하기 위한 방법으로 생물학적 기능정보를 유전자 분석과정에 적용하는 연구가 활발히 진행되고 있으며, 이와 더불어 이미 알려진 유전자 기능정보를 발현데이터에 대한 군집분석과정에서 고려하여 최적의 유전자 군집개수를 추정하는 연구가 이뤄지고 있다. 기존의 관련연구로는 MIPS CYGD Functional Catalogue(FunCat)[5]를 생물학적 기능 데이터베이스로 사용하여 효모 발현데이터의 군집개수 추정 및 군집별 대표 유전자기능 추정방법을 제시한 Okada et al.[6]의 연구가 있으며, 또 다른 연구로는 FunCat과 Gene Ontology(GO)[7]를 이용하여 군집에 포함된 유전자들의 생물학적 기능간의 상관도를 통해 적절한 군집의 개수를 추정하는 Datta et al.[8]의 연구를 예로 들 수 있다. 그 밖에도 유전자 발현값과 그 유전자가 지닌 생물학적 기능정보간의 유사성을 연구한 Sevilla et al.[9]과, 분산팽창계수(VIF: Variance Inflation Factor)를 이용한 군집화 및 FunCat을 통한 결과분석을 다룬 Horimoto et al.[10] 등의 관련연구가 있다. 하지만 이러한 연구에서 보여준 군집개수 추정치들은 실제적인 생물학적 실험을 통해 규명해 놓은 클래스(state of nature)의 개수와 차이가 나며, 결과군집들의 각 클래스별 정확도면에서도 많은 개선이 요구되는 상황이다.

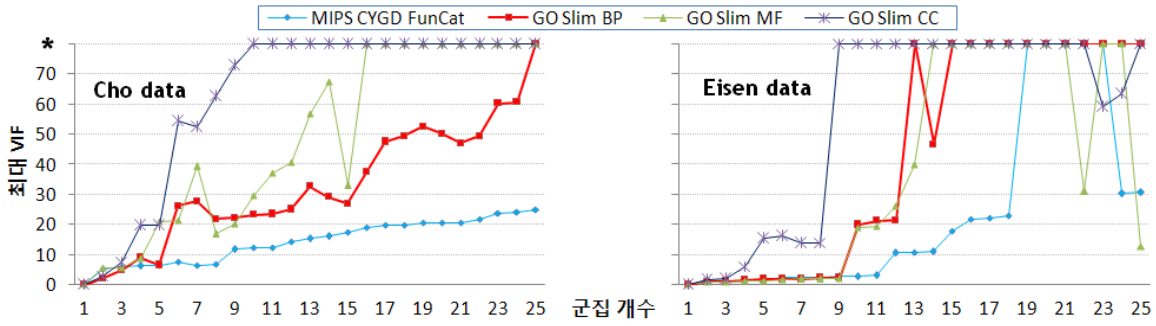
본 논문에서는 효모 유전자의 기능정보를 특화시켜 제공해주는 Yeast GO Slim[11]을 통계학적인 방법론과 접목시킴으로써 생물학적인 최적의 군집개수 k^*를 추정하고, 이를 통해 실시한 군집화 결과에 대한 성능을 분석하였다.

GO Slim은 Saccharomyces Genome Database(SGD)에서 제공하는 생물학적 기능 데이터베이스로서, 모든 생물종을 통틀어 분류해 놓은 방대한 양의 GO 기능항목 중에서 특정 생물 종에 해당하는 기능항목만을 모아놓은 특성화된 GO이다. 특히 본 논문에서 사용한 Yeast GO Slim은 세 종류의 생물학적 기능분류체계에 따라 Biological process(BP) 34개, Molecular function(MF) 23개, Cellular component(CC) 25개의 효모 유전자 기능에 해당하는 GO 기능항목들로 분류되어있고 이러한 기능항목들은 이후 최적의 생물학적 군집개수 k^*를 추정할 때 사용되는 유전자기능벡터의 구성에 이용한다.

실험은 두 종류의 효모유전자 발현데이터인 Cho데이터[12]와 Eisen데이터[1]를 대상으로 진행되었다. Eisen데이터의 총 10개 군집 중 "The ribosome and translation"에 해당하는 군집의 유전자 이름이 수록되어있지 않아 9개 군집만을 분석대상으로 삼았다[10]. 분석데이터에 내재된 생물학적인 최적 군집개수를 추정하기위한 과정 중 군집화 방법으로는 상관계수를 거리로 한 k-평균 방법을 사용했다. k-평균 방법은 군집분석 실행 시마다 임의적인 군집결과를 나타내는 특성이 있으며, 군집분석 실행 반복횟수를 100회 이상으로 설정하여 적용시킴으로써 이러한 군집결과의 변동가능성을 줄였다. 군집분석과정에서는 나누고자하는 전체 군집개수 k를 60개에서 2개까지 변경시켜가며 적용하여 각각의 군집결과를 얻고, 이러한 군집들에 대응하는 유전자기능벡터를 구성했다. 그 후 각 k에서 해당 유전자기능벡터의 분산팽창계수를 계산하여 군집간 공선성이 존재하지 않을 때의 k를 최적의 생물학적 군집개수 k^*로 선택했다. 각 군집별 분산팽창계수의 최대값이 10.0 미만일 때 공선성이 존재하지 않는다고 해석했다[13][14].

유전자기능벡터는 Yeast GO Slim의 BP, MF, CC를 바탕으로 각각 구성하였고, 비교분석을 위해 MIPS CYGD FunCat에도 같은 방법을 적용하였다. 생물학적 기능 데이터베이스에 따른 다섯 종류의 유전자기능벡터를 실제 군집화과정에 적용했을 때 Yeast GO Slim BP가 두 종류의 실험 데이터 모두에서 실제 규명된 클래스의 개수와 동일하게 추정되었다. 추정된 k^*를 k-평균 방법에 적용하여 군집분석을 실시한 결과, 기존의 FunCat으로 추정한 k^*가

적용된 군집결과[6]나 단순히 유전자 발현값을 대상으로 분산팽창계수를 통해 군집을 분석한 결과[10]보다 더욱 개선된 분류성능을 볼 수 있었다.



GO Slim과 FunCat을 통해 추정한 생물학적 군집개수의 비교

GO Slim과 FunCat의 k*를 적용한 Cho데이터 군집결과 비교

결과 군집	Cho데이터 세포주기단계	GO-Slim k*=5 (분류 정확도)	FunCat k*=8 (분류 정확도)
1	Early G1	49/67 *73.1%*	48/67 *71.6%*
2	Late G1	112/135 *83.0%*	112/135 *83.0%*
3	S	35/75 *46.7%*	25/75 *33.3%*
4	G2	35/52 *67.3%*	19/52 *36.5%*
5	M	52/55 *94.5%*	49/55 *89.1%*

GO Slim과 FunCat의 k*를 적용한 Eisen데이터 군집결과 비교

결과 군집	Eisen데이터 규명된 군집	GO-Slim k*=9 (분류 정확도)	FunCat k*=11 (분류 정확도)
1	B	11/11 *100%*	11/11 *100%*
2	C	27/27 *100%*	27/27 *100%*
3	D	14/14 *100%*	12/14 *85.7%*
4	E	17/17 *100%*	17/17 *100%*
5	F	22/22 *100%*	22/22 *100%*
6	G	15/15 *100%*	15/15 *100%*
7	H	8/8 *100%*	8/8 *100%*
8	J	5/5 *100%*	5/5 *100%*
9	K	16/16 *100%*	14/16 *87.5%*

위 실험결과를 통해, 효모 발현데이터의 군집분석에서 Yeast GO Slim의 BP를 통해 군집개수를 추정하는 것이 기존의 MIPS CYGD FunCat 기반 추정방법보다 생물학적인 실험을 통해 규명된 클래스의 개수를 추정하는데 더욱 향상된 결과를 보이는 것을 알 수 있었다. 또한 위 결과에서는 나타내지 않았지만, k-평균 방법이 군집개수 k를 잘 정해주기만 한다면 계층형 군집화 방법보다 더 좋은 군집 분류결과를 보인다는 것을 실험을 통해 알 수 있었다.

향후 연구에서는 효모 이외에 다른 종들을 대상으로 기존의 통계적인 분석에서 얻을 수 있었던 결과보다 생물학적 의미를 더욱 잘 반영할 수 있도록 본 논문의 연구내용을 확장 적용할 예정이며, 이를 통해 유전자 기능분석 및 생물학적 해석에 있어서 유전자에 대한 이해를 증진시킬 수 있는 연구를 진행할 예정이다.

◎ 참고문헌

[1] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 95(25):14863—8, 1998.
[2] Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpeting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA, 96(6):2907—12, 1999.
[3] Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM. Systematic determination of genetic network architecture. Nat Genet, 22(3):281—5, 1999.
[4] Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, Tyers M, Boone C, Friend SH. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science, 287(5454):873—80, 2000.
[5] Mewes HW, Hani J, Pfeiffer F, Frishman D. MIPS: a database for protein sequences and complete genomes. Nucleic Acids Res, 26(1):33—7, 1998.
[6] Okada Y, Sahara T, Mitsubayashi H, Ohgiya S, Nagashima T. Knowledge-assisted recognition of cluster boundaries in gene expression data. Elsevier Artif Intell Med, 35(1-2):171—83, 2005.
[7] The Gene Ontology Consortium. (http://www.geneontology.org/GO.contents.doc.shtml)
[8] Datta S, Datta S. Methods for ecaluating clustering algorithms for gene expression data using a reference set of functional classes. BMC Bioinformatics, 7:397, 2006.
[9] Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A. Correlation between Gene Expression and GO Semantic Similarity. IEEE/ACM Trans Comput Biol Bioinform, 2(4):330—8, 2005.
[10] Horimoto K, Toh H. Statistical estimation of cluster boundaries in gene expression profile data. Bioinformatics, 17(12):1143—51, 2001.
[11] Hirschman JE et al. Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. Nucleic Adids Res, 34(Database issue):D442—5, 2006.
[12] Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. Bioinformatics, 17(10):977—87, 2001.
[13] Fox J. Regression diagnostics: An Introduction. CA: Quantitative Applications in the Social Science 1991.
[14] Myers RH. Classical and modern regression with applications. Boston: Duxbury Classic, 1986.