Seed를 이용한 마이크로어레이 데이터 클러스터링과 유전자 온톨로지를 이용한 클러스터의 해석

강은미[0] 신미영[1] 정호열[2] 박선희[1] 조환규[3]
한국전자통신연구원[0,1], 국립보건원 유전체 연구소[2], 부산대학교[3]
{emkang, shinmy, shp}@etri.re.kr[0,1], hyjung@ngri.re.kr[2], hgcho@pusan.ac.kr[3]

Electronics and Telecommunications Research Institute

Eunmi Kang[0] Miyoung Shin[1] H.Y. Jung[2] S.H. Park[1] H.G. Cho[3]
Bioinformatics Research Team of ETRI[0,1]
National Genome Research Institute, National Institute of Health[2]
Dept. Computer Science and Engineering of Pusan National University[3]

요 약

마이크로어레이 칩 실험을 통하여 대량으로 생산되는 유전자 발현 데이터는 여러 가지 클러스터링 방법을 적용하여 분석할 수 있으며, 생성된 클러스터들 또한 여러 가지 방법으로 해석 할 수 있다. 본 논문에서는 기존의 클러스터링 방법들을 응용한 seed 클러스터링 방법을 제안하고, 생물학적 온톨로지인 Gene Ontology를 기반으로 클러스터를 해석한다. 본 논문에서는 효과적인 유전자 발현 데이터 클러스터링 방법과 생물학적 지식을 바탕으로 클러스터를 해석, 평가하는 방법을 보여준다.

1. 서 론

최근 유전자에 관한 정보를 얻기 위하여 수천, 수만 개의 유전자를 한 번에 실험할 수 있는 마이크로어레이 칩을 많이 사용한다. 이 실험을 통하여 다양한 조건에서 서로 다른 발현 양상을 보이는 유전자 발현 데이터가 대량으로 생산되고 있으며, 데이터의 분석과 해석에 매우 관심이 높다.

유전자 발현데이터를 분석하는 가장 기본적인 방법으로는 유사한 발현패턴을 갖는 유전자끼리 묶는 클러스터링이 있다. 클러스터링 방법에는 발현 데이터가 유사한 유전자들을 이웃하는 트리 형태로 구성하는 계층적 클러스터링 방법과 K개의 유사한 발현 패턴 그룹인 클러스터로 나누는 군집형 클러스터링 방법이 있다. 계층적 클러스터링 방법은 클러스터링 결과를 트리 모양인 댄드로그램으로 시각화하여 전체적인 발현패턴을 파악하기는 좋으나, 데이터를 특정 K개의 클러스터로 나누기 어렵다. 군집형 클러스터링 방법인 K-means나 SOM은 전체 데이터를 분석자가 원하는 K개의 클러스터로 나누지만, 클러스터링 결과가 초기치의 영향을 많이 받는다는 단점이 있다.

그리고 여러 클러스터링 방법을 적용하여 생성한 클러스터들을 해석하는 것도 매우 중요하다. 일차적인 해석 단계로 클러스터에 속하는 유전자들의 공통적인 특징을 파악할 수 있는데, 최근 생물학적 온톨로지인 Gene Ontology[1]나 MIPS를 이용한 방법[2]이 있다.

본 논문에서는 기존 클러스터링의 결과를 응용하여 클러스터링하는 seed 클러스터링 방법을 제안하였다. 그리고 생물학적 온톨로지인 Gene Ontology를 이용하여 seed 클러스터링과 K-*means*클러스터링 방법으로 생성한 클러스터들을 해석하고 비교해보았다.

2. Seed 클러스터링

Seed 클러스터링은 발현 데이터가 매우 유사한 유전자들은 여러 가지 클러스터링 방법에서 같이 묶여서 나타나는 특징에 기초하여 제안한 알고리즘이다. 즉 클러스터링 알고리즘이나 파라메터에 민감하지 않으며, 안정적으로 같은 클러스터에 나타나는 유전자들의 집합을 조사하여 이를 클러스터링에 이용하자는 것이다. 안정적으로 같은 클러스터에 나타나는 유전자 집합들은 그 자체로도 충분히 의미가 있으며, 이 유전자들의 발현 데이터를 적절히 이용하여 군집형 클러스터링의 초기치로 이용하면 초기치에 민감한 군집형 클러스터링의 단점을 보완할 수 있다.

Seed 클러스터링은 다음의 세 단계로 이루어진다. 첫번째, 잘 알려진 클러스터링 방법을 사용하여 클러스터링 한다. 이때 다양한 파라메터를 적용함으로써 여러 가지 클러스터링 결과를 얻을 수 있다. 두 번째, 여러가지 클러스터링 결과 중에서 같은 클러스터에 같이 나타나는 유전자 집합들을 추출하여 seed를 생성한다. 세번째, 두 번째 단계에서 추출한 seed를 클러스터링 방법의 초기치로 설정하고 클러스터링 한다. 각 단계별 구현과 알고리즘은 다음과 같다.

① 기존의 클러스터링 방법을 이용

먼저 알려져 있는 클러스터링 방법을 사용하여 클러스터링 한다. 계층적 클러스터링 방법으로는 Hierarchical 클러스터링을 사용하였다. Hierarchical 클러스터링은 특정 K개의 하위 클러스터를 생성하지 않는다. 이를 보완하기 위하여 그림 1에서처럼 Hierarchical 클러스터링의 결과를 시각화하는 댄드로그램에 세가지 편리한 인터페이스 기능을 추가하였다.. 댄드로그램에서 노드들이 연결된 유사도 값을 보고 적절한 값에서 트리를 절단할 수 있는 절단 기능[3]과 이 절단에 의해 생성되는 하위 클러스터 발현 패턴의 시각화 기능, 그리고 절단하여 생성되는 하위 트리들 중에서 원하는 부분 또는 전체를 하위 클러스터 데이터 집합으로 등록할 수 있는 클러스터 등록 기능을 제공한다. 이를 이용하여 분석자는 댄드로그램에서 절단선을 조정, 생성되는 하위 클러스터들의 패턴을 보고 원하는 하위 클러스터들을 생성할 수 있다.

군집형 클러스터링 방법인 K-means이나 SOM의 경우에는 초기 파라메타 입력값을 지정함으로써 분석자가 원하는 개수만큼의 하위 클러스터를 생성할 수 있으므로 기존의 알고리즘을 그대로 사용한다.

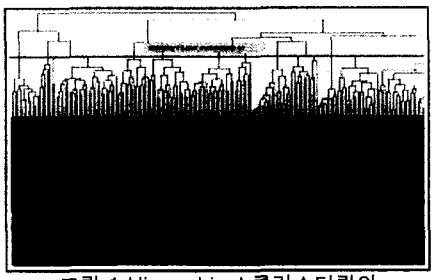

그림 1 Hierarchical 클러스터링의
댄드로그램에서 하위클러스터 생성

② 하위 클러스터에서 공통으로 나타나는 유전자들을 조사

①에서 생성한 하위 클러스터에서 같이 나타나는 유전자들을 파악한다. 예를 들어 그림2 에서 A클러스터링 결과와 B클러스터링 결과를 비교해보면 유전자 g1, g2, g8은 같은 클러스터 내에서 같이 나타난다. g6, g7 도 마찬가지이다. 이렇게 같은 클러스터 내에서 같이 나타나는 유전자들의 발현 데이터의 평균치를 계산하여 새로운 가상의 유전자 seed의 발현 데이터 값을 생성한다. 여러 개의 seed를 생성하는 경우에는 같은 클러스터에서 같이 나타나는 유전자들의 집합의 크기 순서대로 선택된다.

③ seed를 이용한 군집 클러스터링

②에서 생성한 seed들을 K-means나 SOM같은 군집 클러스터링의 초기치로 사용하여 클러스터링 한다. 이때 클러스터의 개수는 seed의 개수와 동일하며 클러스터링 과정 중에서 서로 다른 seed가 속한 클러스터가 합쳐지면 그 바로 전 단계에서 클러스터링을 멈춘다.

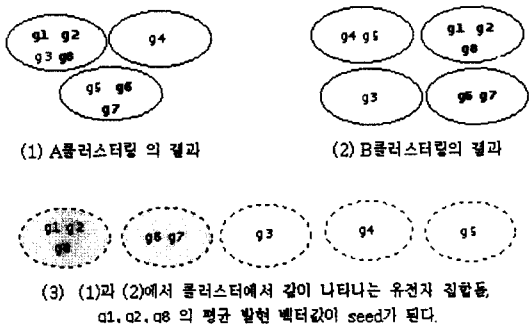

그림 2 Seed 추출 과정

3. 생물학적 온톨로지를 이용한 클러스터의 해석과 평가

여러 가지 클러스터링 방법을 사용하여 클러스터를 생성한 이후에 그 클러스터가 어떤 의미를 갖는지 또 클러스터가 잘 묶였는지 대해 해석할 필요가 있다. 클러스터에 속한 유전자들이 공통적으로 갖는 특징을 파악하여, 클러스터를 해석할 수 있으며 클러스터 내의 또는 클러스터 간의 유사도(homogeneity)나 분할도(separation)을 수학적으로 계산하여 클러스터가 잘 묶였는지를 평가할 수 있다.

그러나 클러스터에 속한 유전자들이 공통적으로 기지는 특징이라 하더라도, 그 특징이 대부분의 유전자에서 나타나는 것이라면 클러스터의 대표 특징이라 하기 어렵다. 또한 클러스터가 잘 묶였는지 수학적인 계산법으로 평가하는 경우, 클러스터의 묶인 정도를 수치화 함으로써 비교 가능하다는 장점이 있으나, 각 클러스터에 속한 유전자 데이터의 생물학적 의미를 반영하지 못한다. 만일 클러스터의 수학적인 평가척도는 나쁘지만, 같은 기능을 하는 유전자들로 묶여있다면 분석의 관점에 따라 잘 묶여진 클러스터로 해석할 수 있기 때문이다.

이에 대한 방안으로 체계화 되어있는 생물학적 온톨로지를 이용하여서 클러스터의 특징을 해석하는 방법이 있다[4]. 본 논문에서는 대표적인 생물학적 온톨로지인 Gene Ontology(이하 GO)를 이용하여 클러스터의 대표 특징을 파악하고, 대표 특징의 p-value를 계산하여봄으로써 특징의 유의미성을 통계학적인 관점에서 평가하였다.

클러스터의 해석에 사용하는 GO는 서로 다른 바이오 데이터베이스에 있는 gene product에 대한 일관성 있는 주석정보의 필요에 의해 시작된 프로젝트로서 종(organism)에 독립적이고, biological processes, cellular component와 molecular function의 세가지 카테고리로 구조화된 생물학적 온톨로지이다. 여기에 사용된 GO 용어 간에는 부모-자식의 관계가 설정되어 있으며, DAG라는 전체 구조로 되어 있다.

GO를 이용하여 클러스터의 대표특징을 파악하는 방법은 먼저 클러스터에 속하는 유전자들을 gene product로 대응시킨 후, gene product와 연결되는 GO 용어의 분포를 조사한다. 따라서 클러스터에 속하는 유전자들이 어떤 GO 용어에 많이 속하는지, 어떤 대표 GO 용어로 요약되는지 파악

할 수 있다. 또한 대표 GO 용어가 우연히 뽑힐 확률인 p-value를 다음과 같이 계산함으로써 통계적인 유의미성을 검증해 볼 수 있다.

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

G : 주어진 종 내에서 전체 유전자의 개수
C : 주어진 GO 용어를 주석정보로 가지는 유전자 개수
n : 클러스터 내 유전자의 개수
k : 클러스터 내에서 주어진 GO 용어를 주석정보로 가지는 유전자의 개수

위 식의 의미는 n개의 유전자로 이루어진 클러스터 내에서 주어진 GO 용어를 주석으로 가지는 유전자의 개수가 k개 이상인 경우의 확률을 구하는 것이다. 이 확률이 작을수록 우연히 k개의 유전자가 해당 GO 용어를 가지지 어렵다는 뜻이다. 즉, 클러스터를 대표하는 GO 용어의 p-value값이 작을수록 통계적으로 유의미하다.
클러스터를 대표하는 GO 용어가 GO에서 상위에 있을수록 전체 유전자에서 GO 용어에 속하는 유전자들이 많아지므로 p-value가 낮아지게 된다. 따라서, 클러스터를 대표하는 GO 용어의 p-value값이 작을수록 클러스터가 잘 묶인 것으로 해석할 수 있다

4. 실험 결과

실험데이터로는 2000년 Mol.Bio에 발표된" Genomic expression programs in the response of yeast cell to environment changes" 에서 공개한 데이터를 사용하였다[5]. 조건 173개에 총 유전자 6152개 유전자 발현 데이터 중에서 missing value가 전혀 없는 유전자 755개를 선별하여 사용하였다. 또한 논문[5]에 따르면 분석결과 비슷한 특징으로 묶이는 17개의 클러스터가 있다. 이를 바탕으로 군집형 클러스터링 방법인 Kmeans와 본 논문에서 제안한 방법인 seed Kmeans 클러스터링 방법을 비교하였다. 클러스터링에서 유전자 사이의 거리는 모두 유클리디언 거리를 사용하였다. Kmeans의 경우에는 클러스터의 개수는 17개 반복횟수는 100회를 주었다. Seed Kmeans의 경우에는 먼저 complete linkage를 사용하여 Hierarchical 클러스터링 한 후 34개의 하위 클러스터를 생성하였고, Kmean로부터 34개의 클러스터 생성하였다. 그리고 두 결과로부터 17개의 seed를 생성하고 이를 다시 클러스터링 하였다.
Kmeans와 Seed Kmeans의 클러스터링 결과는 GO를 이용하여 클러스터별로 대표 GO 용어에 대한 p-value를 계산해주는 프로그램[4]을 이용, 이를 비교하여보았다
그림3에서 보이듯이 seed Kmeans 클러스터링에서 생성한 클러스터의 p-value값이 GO의 세가지 카테고리에서 모두 Kmeans 보다 작다. 이는 GO측면에서 클러스터를 해석하였을 때 seed kmeans 클러스터링이 더 잘 된 클러스터링이며, 더 구체적인 GO 용어로 묶인 클러스터 들을 생성한다고 해석할 수 있다.

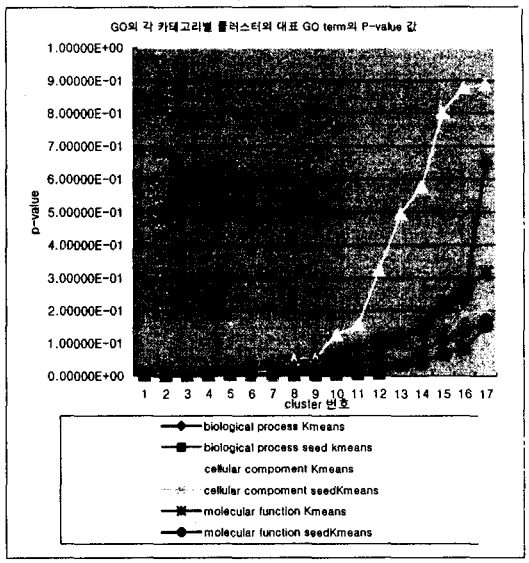

그림 3 GO를 이용한 클러스터링 방법의 비교

대량의 유전자 발현 데이터를 분석할 때 먼저 데이터를 클러스터링 함으로써 유전자 발현 데이터에서 나타나는 특징적인 클러스터들을 파악할 수 있다. 여러 클러스터링 방법 중에서 본 논문에서 제시한 seed 클러스터링 방법은 클러스터 알고리즘이나 파라메타에 민감하지 않은 유전자 집합들을 클러스터링에 이용하여 클러스터링함으로써 구체적인 특징들로 묶여지는 클러스터들을 생성할 수 있다. 또한 생물학적 온톨로지인 GO를 사용하여 클러스터의 대표 특징을 파악함으로써 기존의 생물학적 기반이 배제된 수치적 클러스터 평가가 아닌, 생물학적 지식에 기반을 둔 클러스터를 해석이 가능하였다.

참고문헌

[1] M. Ashburner et al, Gene Ontology: tool for the unification of biology, Nature Genet. 25: 25-29 ,2000
[2] A. Clare and and R. D.King, How well do we understand the clusters found in microarray data?, In Silico Biology 2 0046, 2002
[3] J.W. Seo and B. Shneiderman, Interactively Exploring Hierarchical Clustering Result, IEEE Computer vol.35 number 7 80-86, 2002
[4] S.G. Lee et al., Extraction of biological contexts and ontological DAG structures from gene groups using GO term distribution, GIW 2003, 2003
[5] A.P Gasch et al., Genomic expression programs in the response of yeast cells to environmental changes, Mol.Biol.Cell 11 4241-4257, 2000