유전자 발현 데이터에 적용한 거시적인 바이클러스터링 기법

(Macroscopic Biclustering of Gene Expression Data)

안 재 균* · 윤 영 미† · 박 상 현‡

Jaegyoon Ahn · Youngmi Yoon · Sanghyun Park

요 약 마이크로어레이 데이터는 유전자의 집합이 어떠한 조건 혹은 샘플의 집합 하에서 얼마나 발현되는지를 수치화한 2차원 행렬 데이터이다. 바이클러스터는 마이크로어레이의 샘플의 부분 집합과 이 샘플 부분 집합 하에서 일정한 증감 패턴을 보이는 유전자의 부분 집합을 말한다. 이렇게 같은 패턴을 보이는 유전자의 부분 집합은 일정한 정도의 유의 수준으로 비슷한 기능을 한다고 말할 수 있다. 따라서 바이클러스터링 알고리즘은 같은 기능에 연관된 유전자의 집합과, 이 기능이 발현되고 있는 조건의 집합을 밝혀내는데 있어서 매우 유용하다. 본 논문에서는 다항식 시간 복잡도를 유지하면서, 높은 기능적 상관관계를 가지는 바이클러스터를 밝혀 낼 수 있는 알고리즘을 제안한다. 이 알고리즘은 1) 마이크로어레이 데이터에 심한 노이즈가 있을 경우 패턴으로 인식하지 못하는 기존 알고리즘과 달리, 노이즈 레벨이 심하더라도 거시적으로 비슷한 모양을 보이는 패턴을 찾아내는 방식을 이용하여 숨어있는 패턴들을 찾아낼 수 있고, 2) 바이클러스터 상호간에 오버랩을 허용하며, 또한 다양성이 보장되는 복수의 바이클러스터를 찾아내며, 3) 찾아진 유전자 부분 집합의 기능적 상관관계가 매우 높은 특성을 지니고, 4) 유전자 및 샘플의 순서와 상관없이 결정적인(deterministic) 결과를 도출한다. 또한 본 논문에서는 알고리즘이 찾아낸 바이클러스터의 기능적 상관관계의 정도와, 비교 알고리즘이 찾아낸 바이클러스터의 기능적 상관관계의 정도를 유전자 온톨로지(Gene Ontology)를 통해서 측정함으로써 비교하고 있다.

키워드: 데이터 마이닝, 바이클러스터링, 유전자 표현형 데이터 분석, 마이크로어레이 데이터 분석, 노이즈

Abstract A microarray dataset is 2-dimensional dataset with a set of genes and a set of conditions. A bicluster is a subset of genes that show similar behavior within a subset of conditions. Genes that show similar behavior can be considered to have same cellular functions. Thus, biclustering algorithm is a useful tool to uncover groups of genes involved in the same cellular process and groups of conditions which take place in this process. We are proposing a polynomial time algorithm to identify functionally highly correlated biclusters. Our algorithm identifies 1) the gene set that has hidden patterns even if the level of noise is high, 2) the multiple, possibly overlapped, and diverse gene sets, 3) gene sets whose functional association is strongly high, and 4) deterministic biclustering results. We validated the level of functional association of our method, and compared with current methods using GO.

Key words: Knowledge discovery, Data mining, Biclustering, Gene expression data analysis, Microarray analysis, Noise

1. 서론

마이크로어레이 데이터는 유전자의 집합이 어떠한 조건 혹은 샘플의 집합 하에서 얼마나 발현되는지를 수치화한 2 차원 행렬 데이터이며, 그 예는 표 1 과 같다. 마이크로어레이 분석의 목적 중 하나는 마이크로어레이에 참여하는 유전자의 기능을 밝히는 것이다. 이를 위해서 기존의 많은 분석 방법은 마이크로어레이의 모든 조건 하에서 유전자의 발현값을 조사함으로써, 기능적 상관관계를 가지는 유전자를 클러스터링하는 방법을 취했다[1][9].

그러나 모든 조건에서 특정 기능과 관련된 유전자의 발현 정도를 관찰할 수 있는 것은 아니다. 즉, 우리는 어떤 유전자의 부분 집합이 특정한 실험적 조건 집합 하에서 상관관계를 가진다고 기대할 수 있지만, 다른 조건에서는 이 유전자의 부분 집합의 상관관계는 사라질 수 있다 [2]. 서로 간에 상관관계가 있는 유전자의 부분 집합을 밝혀내는 것은 유전자 집합의 기능을 밝히고, 나아가 유전자 제어 네트워크를 밝히는 중요한 역할을 할 수 있다. Cheng [3]은 마이크로어레이 행렬 데이터에서 서로 간에 밀접한 상관관계를 가지는 유전자의 집합과 샘플의 집합으로서 구성되는 부분 행렬을 찾는 데이터 마이닝 기법을 바이클러스터링이라고 명명했고, 이러한 부분 행렬을 바이클러스터라고 명명했다.

바이클러스터링 프로세스는 NP-Hard임이 증명되었고 [3], 지금까지 제시된 많은 바이클러스터링 알고리즘은

*본 연구는 한국과학재단의 생물정보학연구개발사업(2008-2004103)의 지원을 받아 수행되었습니다.

* 준회원 : 연세대학교 컴퓨터과학과 석사과정

† 종신회원 : 가천의과학대학교 IT학과 교수

‡ 종신회원 : 연세대학교 컴퓨터과학과 교수 (교신저자)

논문접수 : 2009년 1월 29일, 심사완료 : 2009년 4월 1일