유전자의 기능분류를 위한 클러스터링 알고리즘 연구

한석현, 이강만*
강릉원주대학교 컴퓨터공학과
e-mail:shhan@cs.gwnu.ac.kr

Research for clustering algorithm for the functional classification of genes

Seok-Hyeon Han, Gangman Yi*
Dept of Computer Science & Engineering, Gangneung-wonju National University

요 약

차세대 유전정보 분석기 시퀀서의 개발은 양질의 시퀀싱 데이터를 증가시켰다. 수많은 유전정보는 유전자 분석의 새로운 연구 방향을 제시하였다. 본 논문은 유전자 분석 중에서 기존의 유전정보를 활용하여 유전자의 기능예측을 하고자 한다. 클러스터링 알고리즘의 정확도를 높이기 위해서 본 논문에서는 데이터 유사성 조절이 가능한 클러스터링 알고리즘을 적용하였다. 그 결과 데이터 유사성 조절을 할 경우에 그렇지 않을 경우보다 유전자 기능 예측의 정확도가 높아졌다. 따라서 제안된 데이터 유사성 조절 기법은 유전자 기능을 예측하는 방법에 정확도를 높일 수 있을 것으로 기대된다.

1. 서론

생물 분야에서 DNA의 유전자 시퀀스를 생성하는 시퀀서의 등장은 알려지지 않은 유전자 기능 분석으로 발전하였다. 차세대 유전자 분석 시퀀서인 NGS의 등장으로 유전자 시퀀스 관련 데이터가 대량으로 증가하였고, 양적으로 증가한 데이터는 많은 연구 분야에 접목되었다[1].

생명체의 DNA는 시퀀서 머신을 통해 초기 작은 사이즈의 리드데이터로 표현된다. 표현된 시퀀스 데이터는 구조적 모양과 생물체의 조상의 계통적 특성의 차이점으로 일련의 패턴을 보인다. 일정한 패턴은 생물체가 가진 유전자의 기능을 유추하는 가장 핵심적인 단서이다.

조상으로부터 이어져오는 계통의 기능과 유전자의 구조적 모양의 차이점은 다양한 방법으로 구분할 수 있다. 생물학적 실험으로 유전자의 기능적 분석이 완료되면, 유전자들은 family로 분류가 가능하다[2]. 분류의 가장 대표적인 방법론은 기능적으로 연관된 유전자 클러스터링이다. 기존 컴퓨터과학 분야에서 연구된 클러스터링 기법들은 생물의 유전자 기능분석에 응용되어 사용된다[3]. 생물 분야와 접목된 알고리즘 방법론은 복잡한 계산의 문제점을 해결하는 것에서부터 시간과 경제적 비용을 줄이는 새로운 연구방법으로 발전되었다.

본 논문의 방법론은 유전자 데이터베이스인 pfam을 이용한다[4]. pfam을 통해서 지도학습에 사용할 단백질들의 family정보를 확인한다. 서로 다른 단백질들은 시퀀스 매칭을 적용하여 상관도를 측정한다[5]. 상관도가 측정된 단백질의 정보를 이용하여 클러스터링 분석을 진행한다. 데이터 사이의 상관도가 존재하는 단백질 데이터들을 모아서 클러스터를 구성한다. 구성이 완료된 클러스터의 데이터들 중에서 상관도가 높은 유의미한 인자들만 남겨놓고 상관도가 낮은 무의미한 인자들을 제거하여 클러스터를 조절한다. 조절이 완료된 클러스터의 데이터는 지도 학습된 데이터를 이용하여 알려지지 않은 단백질의 최종적인 family를 결정한다.

본 논문은 다음과 같이 구성된다. 2장에는 클러스터링 방법의 문제점과 해결 방법 등에 대하여 기술한다. 3장에서는 방법론을 적용한 실험과 결과에 대해서 기술하고, 4장에서는 본 방법론의 기대효과와 결론을 기술한다.

2. 방법론

유전자의 기능을 판단하는 여러 방법 중 하나는 단백질의 시퀀스 정보를 확인하여 기능을 예측하는 것이다. 시퀀스 정보는 구조와 계통의 유사성을 통해서 family로 분류된다[6]. 시퀀스 유사성이 높을수록 해당 단백질들이 공통된 조상이나 동일한 기능을 가질 가능성이 매우 높다. 이러한 특성을 이용하여 클러스터링 알고리즘을 적용한다. 알려지지 않은 단백질의 family를 결정하기 위해서는 유사성이 높은 알려진 단백질들의 family를 이용한다. 가장 통계적으로 많이 나오는 family를 해당 단백질의 family로 결정한다.

2.1 데이터 유사성

유전자 데이터인 단백질 시퀀스는 문자열로 구성되어있

* 교신저자 : 이강만 (gangman@cs.gwnu.ac.kr)

다. 문자열 데이터는 시퀀스 매칭을 통해서 시퀀스의 패턴을 확인할 수 있다. 서로 다른 단백질 시퀀스의 상관도는 시퀀스의 패턴이 얼마나 유사한 정도의 값으로 결정된다.

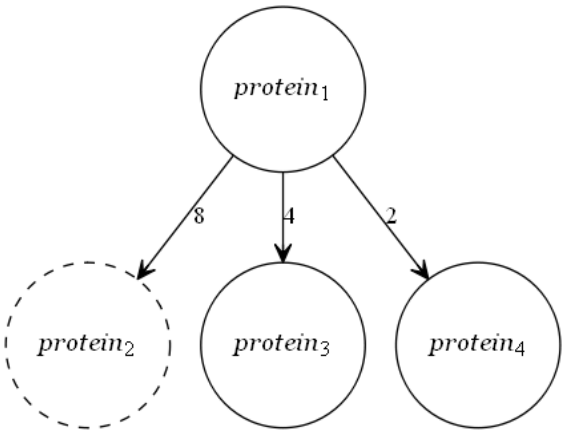
시퀀스 패턴을 확인하는 방법은 두 개의 시퀀스 패턴 중에서 최장으로 공통된 부분을 찾는 것이다. 최장 공통 시퀀스를 찾는 방법에는 동적계획법이 이용된다[7]. 동적계획법은 데이터의 처리에 필요한 메모리 공간이 시퀀스 길이에 비례한다. 이것은 데이터의 양이 증가할 때 계산에 필요한 시간이 증가하는 문제점을 발생시킨다. 이러한 문제점을 해결하기 위해서 동적계획법에서 사용하는 전체 시퀀스를 대상으로 정렬하는 방식에서, 지역적인 시퀀스를 대상으로 휴리스틱한 값을 통해 정렬하는 BLAST를 이용한다[5].

모든 단백질들은 BLAST[5]를 통해서 서로의 상관도를 확인한다. 상관도는 시퀀스 유사성의 정도인 *e*-value를 이용하여 측정한다. *e*-value는 서로 다른 단백질의 시퀀스의 유사성이 높을수록 0에 가까운 값을 가진다.

2.2 데이터 유사성 조절

서로 다른 단백질들로 구성된 클러스터는 BLAST[5]를 통해서 시퀀스 유사성이 존재하는 단백질들로 구성된다. 각각의 알려지지 않은 단백질을 기준으로 구성된 클러스터는 많은 수의 단백질들을 포함한다. 이 단백질들 중에서 유사성이 높은 유의미한 단백질을 선별하고, 유사성이 낮은 무의미한 단백질을 클러스터에서 모두 제거하여 클러스터를 재구성한다.

(그림2)는 유사성이 낮은 데이터 제거하는 것을 나타낸 그림이다. $protein_1$을 기준으로 연결된 많은 데이터 중에서, 유사성이 높은 유의미한 데이터인 $protein_2$, $protein_3$, $protein_4$를 선별한다. 이 중에서 가장 유사성이 낮은 데이터인 $protein_2$을 제거하는 그림을 나타낸 것이다.



(그림 2) 데이터 제거

3. 실험

실험의 정확도를 측정하는 비용을 줄이기 위해서 기존의 이미 기능이 결정된 pfam데이터베이스를 활용한다. 전체 pfam데이터베이스 중에서 10분의 1을 테스트 셋으로 구성한다.

클러스터링 알고리즘 모델의 유효성을 점증하기 위해서, 테스트 셋 이외의 데이터는 학습 데이터로 활용한다. 그리고 테스트 셋은 클러스터링 알고리즘을 통해서 결과를 도출하고, 도출된 결과와 테스트 셋을 비교하여 제안된 모델의 결과와 예측된 결과와의 정확도를 판단한다.

3.1 데이터 구성

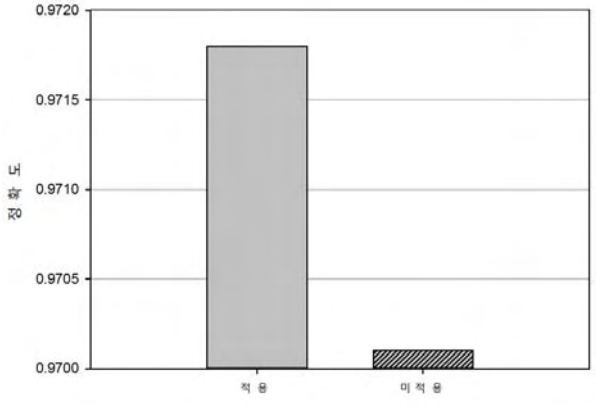
본 연구의 데이터는 유럽 생물정보학 연구소에서 제공하는 pfam데이터베이스[4]을 이용한다. 전체 단백질은 약 200만개이며, family는 약 3000개이다. 하나의 family는 최소한 100개 이상의 protein으로 구성되어 있으며, 테스트 셋의 단백질은 약 20만개다.

3.2 실험결과

본 실험에서 알려지지 않은 단백질을 기준으로 클러스터를 선별한다. 시퀀스 유사성이 높아 선별되는 데이터의 최대 *cut_off*는 10으로 제안한다. 그리고 선별된 데이터 중에서 가장 최악의 데이터 하나를 제거한 데이터들로 클러스터를 구성하여 실험한다.

(그림 3)은 실험 결과를 나타낸다. 유사성이 높은 유의미한 데이터만으로 클러스터를 구성하였을 경우에 정확도는 97.18%이고, 반대로 유사성이 낮은 데이터도 함께 클러스터를 구성할 경우에는 97.01%이다.

실험 결과에 따르면 데이터 유사성 조절을 적용할 경우 약 0.1%의 정확도가 증가하였고, 결과적으로 제안된 클러스터 조절 알고리즘을 적용할 경우가 적용하지 않을 경우보다 더 좋은 결과를 나타내었다.



(그림 3) 실험 결과

4. 결론

본 논문의 목적은 알려지지 않은 새로운 단백질의 family를 결정하는 것이다. family결정은 단백질의 유전자 기능을 추측하고, 종의 계통을 확인 할 수 있도록 한다.

family의 정보를 확인하기 위해서 pfam데이터베이스[4]의 이미 기능적으로 알려진 단백질 시퀀스 정보를 활용한다.

실험을 통해서 클러스터링 알고리즘에서 제안된 데이터 유사성 조절을 적용하였다. 유사성이 높은 데이터만을 이용하여 클러스터링 할 경우 결과 값이, 유사성이 낮은 데이터도 함께 클러스터링한 경우보다 정확하게 나왔다. 이러한 결과는 클러스터링을 통해 데이터를 분석할 경우 불필요한 노이즈들을 제거하는 하나의 방법이 될 수 있다.

따라서 본 연구에서 제시한 데이터 유사성 조절의 적용은 또 다른 클러스터링 알고리즘이나, 여러 목적의 연구에도 활용할 수 있을 것이라 예상한다.

사사

이 논문은 2015년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NRF-2013R1A1A2063006)

참고문헌

[1] Michael, L. Metzker, “Sequencing technologies-thenext generation”, Nature Reviews Genetics, 11, 31-46, 2010

[2] Cath H. Wu, et al, “Protein family classification and functional annotaion”, Computational Biology and Chemistry, 27, 37-47, 2003

[3] Chen, Yonghui et al. “SEQOPTICS: A Protein Sequence Clustering System.” BMC Bioinformatics 7.Suppl 4, 2006, S10. PMC. Web. 10 Mar. 2015

[4] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Grif-fiths-Jones S, Howe KL, Marshall M, Sonnhammer ELL “The Pfam Protein Families Database” Nucl Acids Res, 30, 276-280, 2002

[5] Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ “Basic Local Alignment Search Tool” Journal of Molecular Biology, 215, 1990

[6] Barker W.C, Pfeiffer F, George D.G, “Superfamily classification in PIR-International Protein Sequence Database”, Methods in Enzymology, 266, 59-71. 1996

[7] Smith T, Waterman M “Identification of common molecular subsequences” J Mol Biol, 147, 195-197, 1981