계층적 군집화를 통한 이스트(*Yeast*) 단백질의 고차 상호작용 추출

엄재홍[0] 장병탁
서 울 대 학 교 컴 퓨 터 공 학 부
{jheom, btzhang}@bi.snu.ac.kr

Extraction of higher *Yeast* protein-protein interaction with hierarchical clustering from textual data

Jae-Hong Eom Byoung-Tak Zhang
School of Computer Science and Engineering, Seoul National University

요 약

본 논문에서는 텍스트 형태로 구성된 특정 생물에 대한 문헌 데이터에서 해당 생물의 주요 단백질간의 이진(binary) 관계를 추출하여 이들을 특징별로 계층적으로 군집화 함으로써 특정 현상을 나타내는 단백질간의 주요 관계를 추출하는 방법을 제시한다. 텍스트 데이터에서 단백질간의 이진관계는 기본적인 데이터마이닝 기법을 사용하여 연관규칙(association rule)의 형태로 추출하게 된다. 본 논문에서는 실험을 위해 PUBMED에서 추출한 *Yeast*의 주요 단백질간의 관계를 포함하고 있는 논문 데이터인 MEDLINE Abstract와 몇몇 공개 데이터베이스를 사용하였다. 실험 결과 SH3와 같이 기존에 알려진 단백질간의 단일 관계를 추출하는 것 이외에 이러한 관계들을 이용하여 클러스터링을 행한 결과 공통 현상에 작용하는 주요 단백질간의 관계들이 서로 군집화 됨을 확인 할 수 있었다. 또한 단순 이진관계가 아닌 클러스터링을 이용한 보다 상위 단계에서 단순 규칙들 간의 관계를 살펴봄으로써 단백질간의 이진관계를 추출하기 위한 데이터로 사용한 문헌 데이터에 나타나 있지 않은 1차 이상의 관계를 고찰 해 볼 수 있었다. 논문에서는 규칙 추출의 전체 과정과 함께 사용된 추출 시스템의 각 부와 데이터에 대한 설명을 다룬다.

1. 서론

정보화의 가속으로 말미암아 다양한 정보들에 대한 온라인 접근이 그 어느 때보다도 쉬워지고 있으며, 신문이나 잡지 학술 연구논문 등의 각종 정보들도 디지털화 되어 온라인에 전자문서의 형태로도 존재하게 되었다. 이러한 정보 흐름의 새로운 패러다임은 특히 연구개발을 하는데 있어서 예전과는 달리 지리학적 고립을 극복하는데 도움을 주었으며 또한, 서로 다른 지역에서 수행된 다양한 연구의 결과물들이 교류되는데 있어서의 시간차를 상당히 줄어주었다. 그러나 이와 같은 정보의 갑작스런 증가로 인한 정보 과부하(information overload)는 연구자들로 하여금 하루하루 생산되는 수많은 연구의 결과를 소화하기 힘들게 만들고 있다. 이러한 문제를 해결하고자 정보를 처리하는 다양한 방안이 연구되어왔다.

근래 완료된 인간게놈 프로젝트 이후 새롭게 부각되고 있는 생물정보학(bioinformatics)은 생물학 연구에 있어서 직면하는 엄청난 데이터를 정보기술을 이용하여 해결하려 하고 있는 새로운 연구 분야이다. 본 논문에서는 이러한 생물정보학 분야 중에서 문헌데이터마이닝(literature mining)을 다룬다.

생명체 내의 여러 가지 생화학적(biochemical) 현상을 이해하는데 있어서 단백질들 간의 상호작용은 매우 중요한 역할을 한다. 단백질들 간의 상호작용은 생명체의 특정 현상을 해석하기 위한 생물학적 진행(biological process)을 이해하는데 중요한 실마리를 제공하기 때문이다. 때문에 이러한 생물학적 현상 이해의 기본 실마리를 제공하는 단백질들 간의 상호작용을 추출하는 다양한 연구가 있어왔다. 이들 대부분의 연구들은 현재까지 대부분의 유전자와 단백질들의 관계가 파악되었고 또한 산업적으로 활용성이 큰 *Yeast*를 대상으로 이루어져 왔다.

Ito와 Uetz는 two-hybrid시스템을 이용하여 발아효모(budding *Yeast*)에 대하여 주어진 모든 단백질들 간의 상화 작용을 추출하였다[1][2]. 또한 이렇게 추출된 다수의 단백질간의 상호작용 정보에서 연관규칙(association rule)을 찾기 위한 데이터마이닝(data mining) 기법들을 적용한 연구도 수행되어 왔다. Agrawal등은 다양한 게놈(genome) 데이터에서 추출된 단백질(protein)의 시퀀스(sequence)나 구조(structure), 그리고 기능(function) 등의 이질적인 데이터에 대하여 데이터마이닝 기법을 적용하여 서로간의 연관 규칙을 추출하였다[3][4][5]. 이러한 데이터마이닝 기법은 DNA 마이크로어레이(microarray) 데이터를 분석하는 데에도 활용되었다[6]. 이처럼 데이터마이닝 기법은 생물정보학의 여러 분야에서 성공적으로 응용되어왔다. 이러한 데이터마이닝의 성공적 응용의 예로 Fellenberg등의 연구를 들 수 있다. Fellenberg등은 데이터마이닝 기법을 이용하여 단백질의 상호작용 데이터를 분석하는 통합 시스템을 구성하였다. 이 시스템은 당시까지 밝혀지지 않은 단백질 각각의 특징을 분석하는 목적으로 사용되었다[7]. 그러나 Fellenberg등의 시스템은 단백질 각각의 속성을 밝혔을 뿐 하나의 단백질이 다른 단백질과 작용하는 보다 일반적인 규칙을 발견하지는 못하였다. 이러한 한계를 해결하기 위하여 Oyama등은 데이터마이닝 기법을 이용하여 단백질간의 상호작용 데이터에서 연관

규칙을 찾는 연구를 수행하였다[8]. 하지만 Oyma등의 연구는 단순히 연관 규칙만 찾았을 뿐 최종적으로 수집된 연관규칙들 간의 상관관계를 분석하지는 못하였다. 따라서 본 논문에서는 Yeast *Saccharomyces cerevisiae*에 관한 문헌데이터를 기반으로 기존에 밝혀진 풍부한 단백질간의 상호작용 데이터를 추가로 활용하여 문헌데이터에서 단백질간의관계를 추출하고 추출된 작용 관계가 기존의 데이터에 없는 것인 경우 이를 기존 데이터와 함께 고려하여 이들 사이의 새로운 연관 규칙을 찾는 시스템을 다루기로 한다. 추가적으로, 시스템은 새로이 발견된 규칙들을 모두 고려하여 계층적 군집화(hierarchical clustering)을 행함으로써 보다 큰 관점에서 전체 단백질들 간의 상호작용이 갖는 성향(trend)를 살펴 볼 수 있도록 하고자 한다.

2. 연관규칙의 추출 및 군집화

2.1 연관 규칙의 추출

연관규칙(association rule)은 통계학의 집합 이론을 바탕으로 개별 항목 간의 연관성을 계산하여 그 결과를 일정한 규칙으로 생성한다. 데이터마이닝에서는 트랜잭션들 간의 연관성을 추론하는데 일반적으로 사용되는 방법이다. 연관규칙이 생산해 내는 많은 양의 연관 규칙 대부분이 실제 활용가치가 적은 편인데, 활용가치를 측정함으로써 가능한 한도 내에서 규칙의 수를 줄여줄 수 있어 사용자에게 보다 정제된 정보를 제공할 수 있다. 이러한 연관규칙의 효용성을 측정하는 방법에는 크게 지지도(Support), 신뢰도(Confidence), 리프트(Lift)의 3가지의 척도가 주로 쓰이고 있다.

지지도(Support)은 두 사건 A, B가 동시에 발생할 확률을 뜻한다. 다음으로 신뢰도(Confidence)은 사건 A 가 발생하였을 때 그 사건이 사건 B 를 포함하는 조건부확률을 뜻한다. 마지막으로 리프트(Lift)는 모집단 내에서의 특정 클래스에 대하여 모집단으로부터 의도적으로 조정된(biased sample) 샘플 내에서의 특정 클래스의 비율을 의미한다. 이는 연관(Association)을 통해 도출된 연관규칙이 임의로 추측하였을 때보다 얼마나 더 예측력을 가지고 있는지에 대하여 말해주는데, 이 값이 '1' 보다 크면 예측력이 있다고 평가할 수 있게 된다.

$$Support = \frac{P(Event\ A \cap Event\ B)}{P(\Omega)}$$

$$Confidence = \frac{P(Event\ A \cap Event\ B)}{P(Event\ A)}$$

$$Lift = \frac{P(class\ /\ sample)}{P(calss\ /\ population)}$$

실험에서는 위의 3가지 각각의 값에 대한 최소값을 실험적으로 정하여 문제의 복잡도를 낮추었다.

그림 1은 본 논문에서 사용한 시스템의 구조도를 나타낸다. 그림의 Interaction Extractor에서는 PUBMED의 문헌 데이터에 대하여 기본적인 POS Tagging과 함께 Vasileios등[9]의 방법을 사용하여 고유 단백질의 이름 및 이들 간의 관계를 추출하였다. 또한 SWISS-PROT등과 같은 기존의 데이터베이스를 활용하여 일부이기는 하지만 유용한 단백질들 간의 연관관계 정보를 추가로 활용하였다. 이렇게 추출된 단백질들 간의 상호작용 관계 규칙은 공통 구조를 갖는 하나의 테이블로 정규화 하였다. 정규화를 하는 과정에서는 그림 1에서 문헌데이터와 함께 추가로 고려하고 있는 MIPS등의 데이터를 참고하여 [8]와 유사한 방법으로 특징(feature)을 정의하여 테이블을 확장 하였다. 다만 문제의 복잡도를 낮추기 위해서 [8]의 접근방법을 간소화하여 최종적으로 약 1500여개의 특징을 사용하였다.

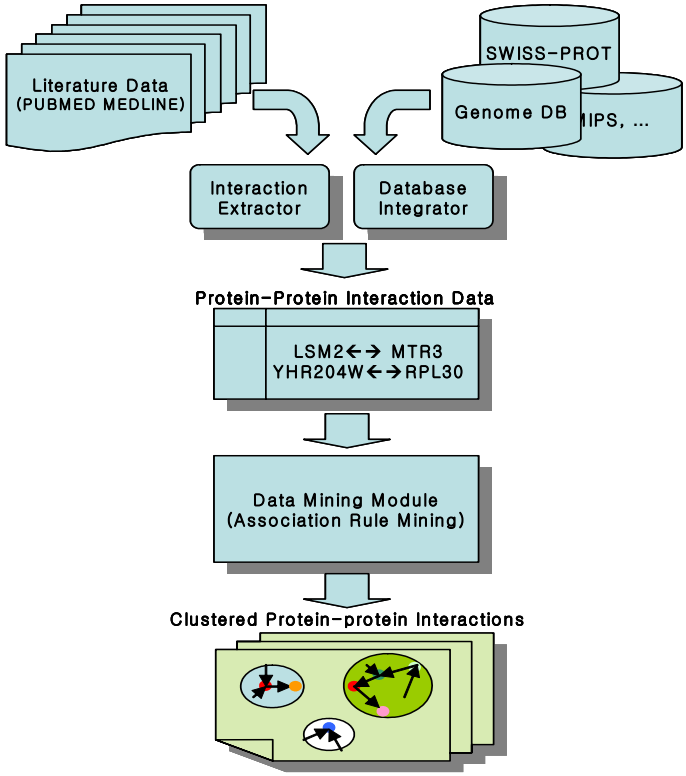

그림 1. 단백질간의 상호관계 추출 및 군집화 개요도

각각의 상호작용 규칙은 '단백질 A는 단백질 B와 상호작용한다.'와 같은 형태로 규칙화 하여 하나의 마이닝을 위한 트랜잭션으로 간주하였다. 이때 각 A, B에 해당하는 부분에 앞서 정의한 특징 필드를 두어 각 필드가 0(해당 특징 없음) 또는 1(해당 특징 존재)과 같이 표현하여 트랜잭션들을 구성하였다. 추출된 규칙의 점수 계산은 [8]와 유사한 방법으로 계산하였다. 단, [8]의 접근과는 달리 Lift의 값을 추가로 고려하여 각 규칙의 점수를 계산하였다. 연관 규칙은 기본적으로 데이터마이닝에서 사용하는 Apriori Algorithm을 따라 계산하였다.

2.2 계층적 군집화

추출된 단백질간의 상호작용 규칙을 앞에서 설명한 방법대로 정규화 하여 연관규칙을 찾은 후 각 연관 규칙의 공통 요소를 기준으로 계층적 군집화(Hierarchical clustering)를 수행하였다. 이것은 초기에 추출된 전체 규칙의 단편적인 지식뿐만 아니라 이들 간의 고차적 상관관계를 살펴보기 위함이다. 추출된 규칙 각각은 특정 단백질들 간의 상화작용 정보를 나타내주며, 연관규칙을 통해 발견된 새로운 규칙은 이전의 단순 단백질간의 상호작용과 같은 단계이지만 실험에 사용된 문헌이나 데이터베이스에서 직접적으로 다루어지지 않은 관계(Relation)를 나타내준다. 이들 모든 관계에 대하여 계층적 군집화를 수행하면 앞서 언급한 내용의 후자와 같은 문헌에 직접적으로 나타타지 않은 관계 문만 아니라 전체 관계들의 특정 성향(Trend)을 파악할 수 있기 때문에 [8]의 작업과는 차이를 갖는 부분이며 충분히 의미 있는 과정이라 생각한다. 계층적 군집화를 위한 각 단계의 특징 요소는 해당 시점에서 가장 많은 규칙들이 공통적으로 보유하고 있는 요소들을 기준으로 군집화 하였다. 계층적 군집화는 완전연결법을 이용하여 수행하였다. 완전연결을 위해 특정 단계에서 군집화 특징을 갖지 않는 연관규칙들은 기본적

으로 해당 규칙들 중 가장 적은 규칙들이 가지고 있는 공통 특징을 단 1개 갖는 것으로 간주하여 완전연결을 이루도록 하였다.

3. 실험 및 결과

3.1 실험 데이터 및 전처리

실험에서 사용한 *Yeast*에 대한 텍스트 데이터는 PubMed의 Medline[10]에서 수집한 논문 초록 데이터를 사용하였다. '*Yeast*'와 'Protein' 및 'Interaction'등의 키워드를 사용하여 검색한 결과에서 상위 8,673개의 문서를 사용하였다. 검색된 문헌 각각의 데이터에 대하여 Brill Tagger[11]를 이용하여 기본적인 POS Tagging을 수행하였다. Tagging된 문서에서 미리 정의한 단어사전을 참조하여 단백질간의 상호 작용을 추출하였다. 또한 추출작업과 병행하여 MIPS, YPD등의 데이터베이스에서 *Yeast*단백질의 이름정보 및 단백질간의 상호작용 정보를 추출하여 문헌에서 추출한 규칙과 함께 시스템의 입력으로 사용하였다. 이렇게 구성한 초기 상호작용 규칙이 6,784개였다.

연관규칙(Association rule)을 계산하여 추출된 2,012개의 규칙 중 Support와 Confidence 및 Lift를 고려하여 선택된 1,854개의 규칙이 초기의 상호작용 규칙 풀(Pool)에 추가되어 최종적으로 8,638개의 규칙이 구성되었다.

3.2 실험결과

아래의 표 1은 실험에 사용된 규칙들 및 실험을 통해 획득된 규칙들을 수치적으로 요약한 것이다.

항 목	규칙의 개수	비 고
초기 추출 규칙 (문헌)	2,035	Medline 문서
초기 추출 규칙 (DB)	4,749	DB Pool의 규칙들
시스템 입력 총 규칙 수	6,784	-
연관규칙의 총 개수	2,012	-
Cut off된 규칙의 개수	158	-
Clustering 입력 규칙	8,638	-
계층적 군집화 단계	12	Cluster의 level

표 1. 실험에 사용된 규칙들의 개수

규칙들을 군집화한 결과에 대한 고찰을 통해 다음과 같은 형태의 규칙들이 존재함을 확인 할 수 있었다.

(Relation) Protein A = Feature[Motif], Value[SH3] ⇔
Protein B = Feature[Amino Acid Pat.],
Value[RxxPxxP, PxxPxR]
S[42.2], C[45], L[1.65].

이것은 SH3라는 Motif 값을 가지는 단백질의 경우 Amino Acid Pattern이 'RxxPxxP' 또는 'PxxPxR'와 같은 어떤 단백질 B와 특정 관계를 갖는다는 것을 나타내고 있다. 실제로 다수의 생물학적 연구들을 통해서 위 규칙이 나타내는 바인딩(Binding) 현상이 보고 된바 있다. 위 규칙에서 'S[42.2], C[45], L[1.65]'는 각각 규칙의 Support와 Confidence 및 Lift 값을 나타낸다.

이처럼 실험을 통해 기존의 연구로 알려진 정보를 충분한 지지도(Support)와 신뢰도(Confidence) 및 리프트(Lift) 값을 가지고 재구성 할 수 있었다. 따라서 본 논문에서와 같은 접근방법을 *Yeast*가 아닌 다른 도메인에 대하여 적용 할 수 있을 것이며 그렇게 함으로써 생물학 분야의 특정 도메인에 대한 연구에 있어서 다량으로 제공되는 문헌 데이터를 보다 쉽게 요약하여 접근 할 수 있을 것이다.

4. 결론 및 향후 과제

본 논문에서는 *Yeast* 도메인에 대한 텍스트 데이터에서 *Yeast* 내의 단백질간의 상호작용 관계를 추출한 후 이를 계층적으로 클러스터링 하여 원본 데이터에서 제공하는 연관규칙 이외에 전체 규칙간의 상화작용 관계를 보다 큰 관점에서 살펴 볼 수 있는 방법을 제시하였다. 제시된 방법을 이용하여 *Yeast*에서 특정 현상에 작용하는 단백질들 간의 연관 관계가 전체적으로 그룹화 되는 것을 확인 할 수 있었다. 하지만, 추출된 연관규칙을 보다 쉽게 살펴 볼 수 있는 시각화(visualization)에 대한 추가 연구가 필요하다. 이것은 실제 텍스트 데이터를 모두 참조하지 않고 본 논문에서 제시한 시스템과 같은 방법으로 예를 들어 *Yeast*와 같은 특정 도메인에 관한 정보를 수집할 경우 시스템의 정확도문제와 연구자의 관심에 따라 직접 해당 연관규칙이 제시된 문헌을 참고할 필요가 있기 때문에 이를 위한 보다 편리한 접근 방법이 필요할 것이라 본다. 또한 특정 도메인에 지나치게 의존적이지 않은 메타 레벨의 규칙 추출 방법과 군집화에 있어서 비 계층적 방법인 k-평균법이나 SOM과 같은 매핑(mapping)방법을 이용한 군집화에 관한 추가 연구도 필요하다.

감사의 글

본 연구는 과학기술부 뇌신경정보학 사업 (BrainTech), 교육부 BK21-IT 프로그램 및 첨단기술 연구센터(AITrc), 국가지정연구실(NRL) 사업에 의하여 일부 지원되었음을 밝힙니다.

참고문헌

[1] Ito, T. *et al.*, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," *Proceedings of Natl Acad. Sci.* USA, pp. 4569-4574, 1998.
[2] Uetz, P. *et al.*, "A comprehensive analysis of protein-protein interaction in *Saccharomyces cerevisiae*," *Nature*, Vol. 403, pp. 623-627, 2000.
[3] Agrawal, R. *et al.*, "Mining association rules between sets of items in large databases," *Proceedings of ACM SIGMOD*, pp. 207-216, 1993.
[4] Satou, K. *et al.*, "Extraction of substructures of proteins essential to their biological functions by a data mining technique," *Proceedings of Int. Conf. Intell. Syst. Mol. Biol.*, Vol. 5, pp. 254-257, 1997.
[5] Satou, K. *et al.*, "Finding association rules on heterogeneous genome data," *Proceedings of Pacific Symposium on Biocomputing*, pp. 397-408, 1997.
[6] Zweiger, G. *et al.*, "Knowledge discovery on gene-expression microarray data: mining the information output of the genome," *Trends Biotech.*, Vol. 17 ,pp. 429-436, 1999.
[7] Fellenberg, M. *et al.*, "Interactive analysis of protein interaction data," *Int. Conf. Intell. Syst. Mol. Biol.*, Vol. 8, pp. 152-161, 2000.
[8] Oyama, T. *et al.*, "Extraction of knowledge on protein-protein interaction by association rule discovery," *Journal of Bioinformatics*, Vol. 18, No. 5, pp. 529-540, 2002.
[9] Vasileios H. *et al.*, "Disambiguating proteins, genes, and RNA in text: a machine learning approach," *Journal of Bioinformatics*, Vol.17, pp. 97-106, 2001.
[10] PubMed Medline, www.ncbi.nlm.nih.gov/PubMed/
[11] Brill Tagger, www.cs.jhu.edu/~brill/ ■