생물학 관련 문헌으로부터 상호작용 정보 자동 추출

정의헌[0] 김민경 박현석
세종대 컴퓨터공학부, 이화여자대학 컴퓨터공학부
{juhpeter[0]}@empal.com minkykim@ewha.ac.kr neo@ewha.ac.kr

Automatic Extraction of protein-protein interaction information from biological literature

Euiheon Jeong[0] Minkyung Kim Hyunseok Park
School of Computer Engineering, Sejong University, School of Computer Engineering,
Ewha Womans University

요 약

본 논문에서는 생물학 관련 문서에서 단백질 간의 상호작용을 추출하는 방법에 대한 전반적인 기술 동향을 소개하고, 현재 구현된 상호작용 정보 자동추출 시스템의 연구 결과에 대해 기술한다. 일반적으로 이미 알려진 단백질들의 관계를 추출함에 있어서는 단백질의 이름에 대한 특성 구분과 표현의 의미적 해석등에 NLP 기법을 사용하여, 사용자 정의에 따른 룰을 생성하는 방법과 데이터 마이닝 기법을 적용하여, 단백질간의 관계를 자동적으로 추출하는 방법, 또한 위의 이 두가지 방법을 병행하는 방법이 현재 연구되고 있다.

이 논문에서는 자연언어처리 기법과 머신러닝 기법(SVM)을 이용하여, 단백질간의 상호작용에 관한 일반 생물 정보 문헌에서 추출하고, 그 성능을 테스트 해 보겠다.

1. 서 론

바이오인포매틱스 또는 프로테오믹스 및 기타 관련 생물학 연구에서 중요한 문제로 인식되는 것 중 하나는 지속적으로 증가하고 있는 생물학적인 정보의 분석이다. 특히 일반 생물학 문헌에서는 컴퓨터가 분석할 수 없는 free text 형태로 정보가 축적되고 있기때문에, 그것들에 대한 분석은 많은 시간과 노력이 들 수 밖에 없다. 데이터를 사용자가 원하는 정보에 맞게 특정 형태로 변환하고, 변환된 데이터를 이용해서 원하는 정보를 추출하고, 이용하려는 연구가 주를 이루어 수행되고 있다. 뿐만 아니라 정보를 추출하기전, 전처리 단계(transformation of free text) 에서의 다양한 시도와 정보추출 단계에서도 보다 정확한 정보를 얻기 위해 여러가지 방법들이 이용되고 있다. 특히 이 논문에서는 단백질간의 상호작용에 대한 정보를 일반 생물학 문헌에서 자동 추출하고 분석하게 함으로써, 생물학 연구에 유용할 것으로 기대된다.

단백질간의 상호작용에 대한 정보를 얻기 위해서는 첫번째로 단백질 이름의 구분이 중요한데, 유전자와 단백질의 이름을 찾아내기 위해서 통계적인 방법과 knowledge-based strategies 를 이용하고 있는 연구도 있었으며 [1], 미리 만들어진 단백질 이름에 대한 사전을 사용함으로써[3] 단백질 이름을 구분하고 있다.

두번째로는 위에서 뽑아낸 유전자/단백질 이름과 interact, activate, inhibit, bind 등의 동사정보를 바탕으로 두 개의 엔티티들의 관계에 대한 정보를 정하는 데에 있어서는 관계절등과 같은 구문이 포함된 복잡한 문장이나, 부정문이 섞인 경우, 그리고 대명사문제 등과 같은 해결하기 어려운 문제들이 있을 수 있다. 간단한 rule을 적용하는 방식이나[2], 부정문 해석에 대해서 간단한 정규표현식을 사용하는 연구도 있었다[3].

또 다른 방향으로는 단백질들 간의 관계에 대한 정보 표현을 정형화함으로써, 단백질들 간의 관계를 정형화된 원소를 채워나가는 방법을 사용하거나 [4] , 단계별로 미리 만들어진 템플릿의 원소들을 채워 나감으로써 원하는 정보를 찾아내는 경우도 있다[5].

그러나 아직 각 연구마다 해결하지 못하는 문제점들이 있고, 그러한 문제해결을 위해 서로 다른 방법들을 제시하고 있는 사정이다.

다음 장에서는 기존의 연구들에 대해 바탕이 되는 기술들을 살펴보고, 3장에서는 그러한 기술들을 응용한 연구 동향들을 살펴볼 것이며, 4장에서는 본 논문의 시스템 구현 및 시험결과를 보고, 5장에서 결론을 기술하도록 한다.

2. 텍스트 정보 분석에 대한 기반 기술

일반 생물정보 문헌에서 정보를 얻기 위해 사용되는 기술들로는 파서, 태거 등과 같이 자연언어 처리 기법을 사용하는 경우가 있고, 문서 안에 포함된 단어들의 빈도등과 통계적인 수치로 문서분류를 하는 등 텍스트 마이닝 기법도 사용되며, 자연언어 처리 기법을 적용하여, 기계적으로 학습 시키는 방법들이 이용되고 있다.

자연언어처리 기술은 주로 free text 형식의 문헌을 파싱하고, 태깅하여 각 문장성분 들에 대해 구문 분석과 의미분석에 대한 주석을 붙임으로써, 정보추출의 기본이 되는데 사용되며, 더 나아가서는 문맥상의 규칙들을 생성하여, 문장의 의미 분석을 보다 정확하게 도와준다. 예를

들어, Medline과 같은 문헌정보로부터 유전자 단백질의 이름[7]이나, activate, inhibit, bind 등의 동사정보를 바탕으로 그 주어나 목적어를 자동적으로 찾아 냄으로써 유전자나 유전자 산물들의 상호 작용 정보를 추출하여 짧은 시간에 수십만 건의 유전자 또는 단백질 상호작용 데이터베이스를 구축하는 것이 가능해 졌다. 물론 최종확인은 기존의 wet lab (in vitro/in vivo)' 에서 완성시켜야 하겠지만, 그 최종 후보자 target 을 찾아내기까지의 실험은 dry lab (In Silico)' 에서 이루어질 수 있다는 것이다.

바이오인포매틱스에서 다루는 문헌정보는 이 PubMed의 문헌과, PubMed의 Abstract를 Indexing 해 놓은 Medline의 문헌이 중심이 된다.

이렇게 자동화된 Text Processing [8] 정보들은 여러 분석용 시스템과 데이터 베이스를 구축하거나, 기존 시스템(데이터 베이스, 검색엔진) 에서의 정보 추출 , 정보 검색시 high-quality 서비스를 지원하며, 현재 실험내용을 그 간의 연구결과 중에서 찾아 비교하거나 검증하는 등 전체적으로 performance의 향상에 도움을 준다.

텍스트 마이닝 기술은 관련이 있는 정보를 자연언어 텍스트로부터 추출하는 것과 추출된 엔티티 간의 관계들 검색과도 연관된다. 텍스트 분류(Text classification)는 텍스트 마이닝 분야 중에서 기초적인 기술 중 하나이며, 그것은 콘텐츠가 하나의 셋으로 나누어지는 것을 의미한다. 텍스트 분류 중에 SVM(Support Vector Machine) 은 현재 가장 좋은 성능을 보이는 솔루션 중의 하나이다.

머신러닝 기술은 전처리 된 문장 구성성분들의 규칙들을 이용하여, 각 규칙들에 대해 scoring 하는 방법으로 기계 학습을 시킴으로써, 보다 정확한 정보추출을 자동으로 할 수 있게 해주거나, 텍스트 마이닝 기법에 더하여 자연언어처리 기술을 사용하지 않고 문서 분류를 하는데 사용하기도 한다.

2. Protein-Protein Interaction 정보 추출연구 동향

첫번째 방법으로는 간단한 NLP(자연언어처리) 기술을 사용하여 복잡한 문제를 간단히 해결하려 하였다.[3]

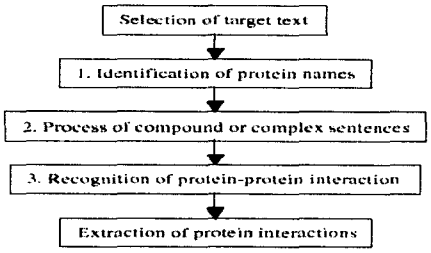

그림 1

위의 시스템 구성도에 첫번째의 단백질 이름을 식별하는 단계에서는 사전을 사용하고 있으며, 두번째는 간단한 정규표현식을 사용하여 문장들을 나누어, 좀더 복잡한 문장의 구조를 단순하게 나누고 있다.

여기서도 템플릿을 만들어 구성성분을 채우고, 거기에

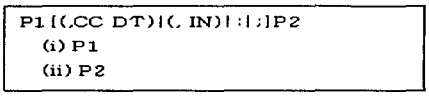

그림 2

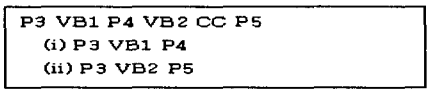

그림 3

또한 부정문의 경우도 아래와 같이 정규 표현식을 사용하여 해결하고 있다.

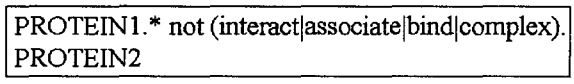

그림 4

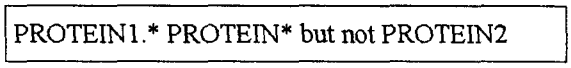

그림 5

하지만, 여기에서의 문제점은 단백질 이름이 동의어나 복잡한 형태의 이름을 구분하려는 시도를 사전 에만 창조하여 해결하고자 한다. 그리고 관계절과 같은 좀더 복잡한 형태의 문장은 의미해석을 못하고 있다.

또 다른 방법으로는 일반적인 목적의 파서를 생물학 분야에 적용함으로써 full 파서 자체의 사용을 테스트 해보는 Term recognizer [5] 와 shallow parser 를 사용하여, 메모리를 많이 사용하고, 애매한 해석의 경우 때문에 패턴매칭에 적절하지 못한 Full Parser 의 문제를 해결하고, 길고 복잡한 단백질의 이름은 파서에 사전을 사용하여 이름 분석에 절감되는 시스템의 효율 문제를 보완한 것이다.[4]

Shallow parser 사용은 메모리의 사용을 약 30%의 절감을 가져다 주었으며, 파싱시간의 경우는 순수 full parser 를 사용하였을 때 보다 약 7배정도의 속도향상이 있었다.

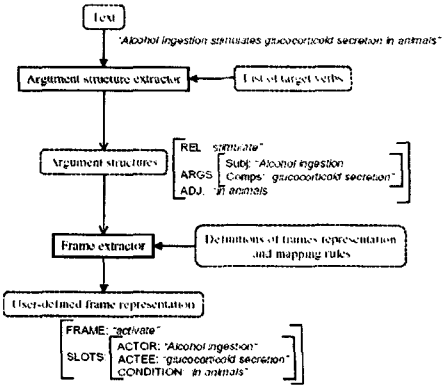

그림 6

mapping rule 를 적용하여, Frame 이라는 새로운 템플릿을 채우고 있다. 복잡한 문장이나 능동/수동의 문제는 Full parser를 이용하여 해결하고 있다. 그러나, 전치사구의 적용범위나, 파싱할 때의 기준 (구두점, 등위접속사) 같은 것들의 애매성 때문에 생기는 파싱의 실패가 문제점으로 확인되었다. 그러나 잘못된 파싱의 결과들을 후처리를 사용해서 그 문장들의 부분적인 정보들을 이용할 수 있다고 주장하였다.

3.시스템 구현 및 시험결과

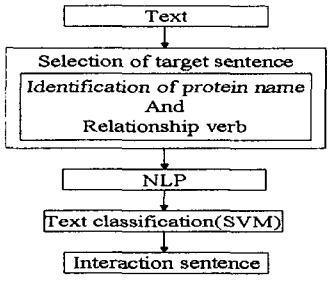

그림 7

테스트에 사용될 생물학 문헌들은 swissprot과 bind의 각 protein information에 reference 된 NCBI의 Pubmed id로부터 그 abstract(27877) 를 사용하였고, protein name 들의 구분은 swissprot [9]과 bind[10] 시스템에 있는 251810 개의 이름들을 사용하였다. 파서는 partial parser 인 FDG 3.7 을 사용하였으며, SVM 엔진은 LIBSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm) 인 자바 패키지로 구성된 라이브러리를 사용하였다.

SVM 트레이닝 데이터는 activation(983/500), bind(1887/500), associate(594/500), inhibit(629/500), interact(629/500), regulate(930/500) 개의 위의 동사들이 포함되고, 동사를 전후로 단백질 이름들이 포함된 문장, 즉 relationship을 가지고 있다고 생각되는 문장 / 그렇지 않은 문장들을 구분하여 각 동사별로 들을 선택하여 SVM training을 시켰다. 그리고 테스트 결과는 평균 실제로 interaction 관계가 포함된 문장 70%의 acuuracy를 보였다. Acuuracy 가 낮은 이유는, 부정문이 포함되거나, 관계절이 삽입된 경우등이 있어서 문장을 정확히 찾아내지 못하는 경우가 있다.

4.결론

서열 데이터베이스는 알려진 모든 서열이 데이터베이스화 되어 있는데 반해서, 상호작용은 논문으로 흩어져서 존재하고 있다.

그래서 분산되어 있는 상호작용 관련 text들을 학자들이 이용하거나, 데이터베이스화 하는데 많은 시간과 노력이 든다. 이러한 이유로 문서들을 자동으로 분류하여, 논문 정보의 이용을 쉽게 만드는 것에서 본 연구의 의의를 찾을 수 있겠다.

Protein-protein interaction 관계를 일반 생물 문헌 정보에서 추출하기 위해서는 , 역시 단백질 이름을 구분하는데 있어서, 일반 생물 데이터 베이스를 사용하여, 현재 사용되고 있는 단백질 이름의 리스트를 사용하였다. 역시 이 문제는 앞으로 더 연구되어야 할 부분중에 하나이며, machine learning (SVM)기법을 사용하는 것이, text classification 하는 것이 유용한 방법임을 알았다.

앞으로 이번 시스템에 대한 성능 향상을 위하여 text classification 뿐만 아니라, 자연언어처리 기법을 머신러닝 기법더하여, abstract 뿐만 아니라 full article 로부터 정보를 추출할 계획이다

5. 참고문헌

[1] Lorraine Tanabe and W. John Wilbur (2002) Tagging gene and protein names in biomedical text Bioinformatics 2002 18: 1124-1132.

[2] Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A.: Automatic extraction of biological information from scientific text: protein-protein interactions. In: Proceedings of the 5th Int. Conference on Intelligent Systems for Molecular Biology. (ISMB99) AAAI Press(1999), pp. 60-67

[3] Ono, T., Hishigaki, H., Tanigami, A., Takagi, T.: Automated extraction of information on protein-protein interactions from the biological literature. In: Bioinformatics. 17(2) (2001)

[4] Yakushiji, A., Tateisi, Y., Miyao, Y., Tsujii, J.: Event Extraction from Biomedical Papers Using a Full Parser. In: Proceedings of the Pacific Symposium on Biocomputing(2001)

[5] Tu Minh Phuong, Doheon Lee* , Kwang Hyung Lee :Learning Rules to Extraction Protein Interaction from Biomedical Text pakdd 2003 148-158

[6]. C. Nobata, N. Collier, and J. Tsujii. Automatic Term Identi.cation and Classi.cation in Biology Texts. In *Proc. NLPRS*, 1999.

[7] Fukuda, K., Tsunoda, T., Tamura, A. and Takagi, T.(1998) *Toward information extraction : idetifying protein* names from biological papers. Pacific Symp. Biocomputing 3: 705-716

[8] Salton, G.(1989). Automatic text Processing. Addison-Wesly (*Addison-Wesly series in Computer Science*), Reading, Massachusetts, USA The interactive Fly

[9] Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res.,28,45-48

[10] Barder,G.D., Donaldson,I., Wolting,C., Ouellette,B,F., Pawson,T. and Hogue,C.W. (2001) BIND-The Biomolecular Interaction Network Database.Nucleic Acids Res., 29, 242-245