유전자 온톨로지의 자동 확장과 용어 분석 *

이진복[0] 박종철
한국과학기술원 전산학과 및 첨단정보기술연구센터
{jblee,park}@nlp.kaist.ac.kr

Automatic Gene Ontology Extension and Terminology Analysis

Jin-bok Lee[0] Jong C. Park
Computer Science Department and AITrc, KAIST

요 약

생물학 분야의 방대한 지식을 효율적으로 다루기 위하여 생물정보학이 주요한 연구 분야가 되었다. 이 중 특히 생물학 문헌에서 정보를 자동으로 추출하는 연구가 활발히 진행되고 있는데, 이러한 정보추출 결과를 이용하여 유전자 온톨로지와 같은 유용한 지식베이스를 자동으로 확장함으로써 폭발적으로 증가하는 생물학 분야의 연구 결과들을 지식베이스에 통합할 수 있다. 자동으로 확장된 온톨로지는 신뢰성을 보장하기 위한 검증 과정을 거쳐, 정보추출 시스템의 성능을 향상시키기 위한 지식베이스로 사용되게 된다. 본 연구에서는 단백질 간의 상호작용에서 나타나는 조건을 추출하는 시스템과 유전자 온톨로지를 이용하여 추출된 생물학 용어를 분석하는 시스템을 제안하고 유전자 온톨로지의 자동 확장 및 검증 시스템에 대하여 논의한다.

1. 서론

생물정보학(bioinformatics)은 전산학을 통하여 생물학 분야의 방대한 지식을 다루기 위한 학문이다. 생물학 문헌으로부터의 정보추출과 유용한 지식베이스(knowledge base) 구축에 대한 연구가 생물정보학 분야에서 활발히 진행되고 있다.

온톨로지는 어휘, 개념, 관계 등을 포함하는 특정 분야 지식의 의미 모델이다. 유전자 온톨로지(Gene Ontology, [1])는 진핵생물(eucaryote)의 유전자와 관련된 정보를 담고 있는 대표적인 온톨로지이다. 이 온톨로지는 생물학 전문가들에 의해서 수작업으로 구축되고 있는데, 구축된 정보에 대한 신뢰도가 높다는 장점이 있지만 구축되는 속도가 매우 느리다는 단점도 있다.

본 연구에서는 BioIE([2,3])에서 추출된 단백질 간의 상호작용 정보와 추가적인 정보들을 유전자 온톨로지를 이용하여 분석해내는 방법에 대하여 논의하고, 추출된 결과로 유전자 온톨로지를 자동으로 확장하고 검증하는 방법에 대하여 논의한다.

2. 관련연구

유전자 온톨로지([1])는 진핵생물의 유전자와 관련된 정보를 담고 있는데, 생물학적으로 다양한 목적에 대하여 다루는 생물학적 과정(biological process) 온톨로지, 생화학 수준에서 유전자 생산물에 대하여 다루는 분자 기능(molecular function) 온톨로지, 유전자 생산물의 위치에 대하여 다루는 세포 요소(cellular component) 온톨로지의 세 가지 세부 온톨로지로 구성되어 있다. 각 온톨로지는 'is-a'와 'part-of' 관계로 이루어진 방향성 비순환 그래프(directed acyclic graphs; DAGs) 형태를 갖게 된다. 현재 생물학적 과정, 분자 기능, 세포 요소 온톨로지는 각각 5733 개, 5216 개, 1101 개의 개념을 포함하고 있는데 생물학 분야 지식의 일부만이 정리되어 있는 상태이므로 효율적으로 확장시키는 방법이 요구된다.[1]

BioIE([2,3])는 생물학 문헌으로부터 단백질-단백질 상호작용 정보를 자동으로 추출해내는 시스템으로, 상호작용을 나타내는 어휘를 중심으로 양방향 점진적 파싱(bidirectional incremental parsing)을 통하여 단백질을 나타내는 용어를 인식해내고 결합범주문법(Combinatory Categorial Grammar; CCG)을 통해 문법성을 검증한다.[2] 폭발적으로 증가하는 생물학 분야의 문헌 정보를 정리하기 위해서는 대상으로 하는 정보를 자동으로 추출해주는

* 본 연구는 AITrc를 통해 과학재단의 지원을 받아 수행되었음
[1] 2002년 8월 16일에 배포된 유전자 온톨로지이다.
[2] BioIE는 MEDLINE의 요약문을 처리하고 있는데, 이 요약문은 http://www.ncbi.nlm.nih.gov/PubMed에서 제공된다.

시스템이 필수적이게 된다.

3. 단백질 상호작용 조건 정보추출

BioIE에서는 생물학 문헌으로부터 단백질에 해당하는 명사구들과 그것들 간의 상호작용 정보를 추출하고 있다. 그런데 생물학 문헌에는 그러한 정보 외에도 단백질 간의 상호작용이 일어나게 되는 조건(condition)에 대한 정보도 담겨 있다.

(1) HEC inhibits the proteolysis of metotic cyclin B *in vitro*. (PMID:9295362)

(2) *Autophosphorylated* DNA-PK dissociates from Ku:DNA. (PMID:8621537)

위의 (1)에는 'inhibit'이라는 상호작용에 대한 조건 정보인 'in vitro'가 포함되어 있는데, 이러한 종류의 조건은 문장 내의 부사구, 전치사구, 명사구, 종속절에서 주로 나타난다. (2)에는 'DNA-PK'를 수식해주는 조건인 'autophosphorylated'가 나타나고 있는데, 이러한 종류의 조건은 BioIE에서 추출된 명사구 내의 형용사, 전치사구, 관계절에서 주로 나타난다.

본 연구에서는 (1)에서와 같은 형태의 조건을 추출하는 시스템을 구현하였는데, 결합범주문법을 이용하여 대상 문장을 전체 파싱(full parsing)을 하는 것에 따르는 복잡도를 낮추기 위하여 BioIE에서 추출된 정보를 중심으로 수식관계에 있는 부분을 찾아주어 조건을 추출하고 있다. 이 방법을 효모(yeast)와 관련된 문헌에 대하여 테스트 한 결과, 현재 이 시스템의 정확률은 65%이고 재현률은 45%이다. 시스템의 정확률을 떨어뜨리는 주된 원인은 자연언어의 구조적 애매성에 의한 잘못된 조건의 추출이고, 재현률을 떨어뜨리는 주된 원인은 복잡하거나 특수한 문형이 나타나는 문장에 대한 파싱 실패인 것으로 분석되었다.

4. 유전자 온톨로지를 이용한 용어 분석

유전자 온톨로지는 생물학적 과정, 분자 기능, 세포 요소에 대한 내용을 담고 있는데, BioIE에서 추출되는 상호작용 정보와는 (그림1)과 같은 관계를 갖게 된다. BioIE에서 추출하는 명사구는 단백질과 조건 정보를 포함할 수 있는데, 각각은 유전자 온톨로지의 분자 기능과 세포 요소에 대응되게 된다.

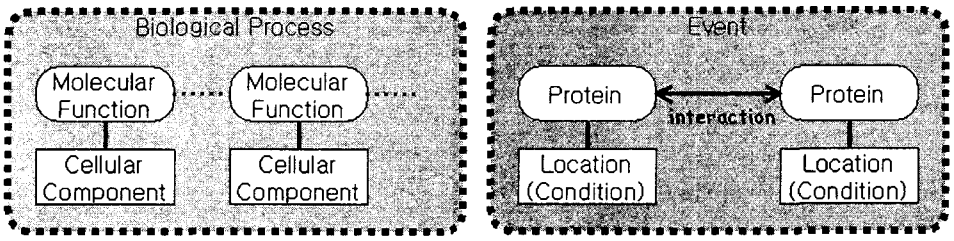
본 연구에서 구축한 용어 분석 시스템은 BioIE에서 추출된 명사구를 대상으로, 각 명사구 내에서 분자 기능에 해당하는 부분과 세포 요소에 해당하는 부분을 유전자 온톨로지의 용어를 이용하여 분석해낸다.

(3) *focal adhesion* kinase

(4) collagen IV of *basement membrane*

위의 (3)과 (4)는 BioIE에 의하여 추출된 명사구의 예이다. (3)에서 'focal adhesion'과 'kinase'는 각각 유전자 온톨로지의 분자 기능과 세포 요소에 포함되어 있는 용어이므로, (3)은 'focal adhesion'이라는 장소에 있는 'kinase'를 의미한다고 분석해낼 수 있다. (4)에서는 'of'라는 전치사가 나타나면서 분자 기능인 'collagen'이 세포 요소인 'basement membrane'보다 앞에 나타나게 되는 경우인데, 이러한 경우에 대해서도 명사구 내의 전치사 구조를 고려함으로써 분석을 진행한다.

이 용어 분석 시스템은 효모와 사이토카인(cytokine)에 관련된 문헌에 대하여 테스트를 하였는데, BioIE의 추출 용어 32984 개 중에서 유전자 온톨로지의 분자 기능과 세포 요소의 용어로 분석 가능한 것은 191 개로 0.58%에 불과하다. 이는 유전자 온톨로지가 다루고 있는 용어가 충분하지 못하기 때문으로 분석된다. 따라서 유전자 온톨로지를 효율적으로 확장시키는 방법이 필요하다.



(그림1) 유전자 온톨로지와 단백질간 상호작용의 비교

5. 유전자 온톨로지 확장 및 검증

유전자 온톨로지는 생물학적 과정, 분자 기능, 세포 요소 각각의 개념을 'is-a', 'part-of' 관계만으로 표현하고 있다. 본 절에서는 BioIE의 결과로 나오는 다양한 상호작용 관계를 이용하여 온톨로지를 확장하고 검증하는 시스템에 대하여 논의한다.

유전자 온톨로지를 확장하는 첫 번째 단계에서는 유전자 온톨로지에서 제공하는 정보를 관계형 데이터베이스(relational database)로 구축한다. 두 번째 단계에서는 정보추출 시스템에서 얻어진 단백질 상호작용 정보와 조건 정보를 데이터베이스에 통합하는데, 이 과정에서 단백질에 해당하는 용어에서 나타나는 동의어(synonym)/약어(acronym) 정보도 추출하여 통합한다. 세 번째 단계에서는 확장된 온톨로지 데이터베이스의 용어들의 형태소 분석, 구조 분석을 통하여 그 용어들 간의 연결성(inter-connectedness)을 높인다.

(5) peptidase *[functor-of]* peptide

(6) an IGF-I receptor monoclonal antibody *[same-to]* a monoclonal antibody to the IGF-I receptor

위의 (5)에서 'peptidase'는 'peptide'와 '-ase (enzyme)'로 구성된 용어라는 것을 형태소 분석으로 알아낼 수 있으므로 'functor-of'라는 관계를 찾아주게 된다. (6)에서 'to'라는 전치사에 대한 구조 분석을 통하여 두 용어가 'same-to'라는 관계를 갖는다고 찾아주게 된다.

앞의 세 단계를 거쳐서 자동으로 유전자 온톨로지를 확장할 수 있는데, 이 온톨로지가 포함하는 정보의 신뢰도를 높이기 위한 검증 과정이 필요하게 된다. 검증 시스템은 다음과 같은 정보를 검색해준다.

(7) In conclusion, our results suggest that *PKC epsilon* stimulates *Raf-1* indirectly by inducing the production of autocrine growth factors. (PMID:9416835)

(8) However, using coexpression experiments in *Sf-9 cells* and transiently transfected A293 cells we did not obtain any evidence for a direct activation of *Raf-1* by *PKC epsilon*. (PMID:9416835)

위의 (7)과 (8)의 문장에 대하여 정보추출 시스템은 'PKC epsilon'과 'Raf-1'에 대한 관계를 각각 'stimulate'와 'activate'의 두 가지로 추출하게 된다. 검증 시스템은 같은 용어들에 대하여 두 개 이상의 관계 정보가 존재할 경우에 잠재적으로 존재할 수 있는 오류를 제시해 줌으로써 확장된 유전자 온톨로지의 검증을 도와준다.

6. 결론 및 향후계획

본 연구에서는 유전자 온톨로지를 이용하여 생물학 분야의 문헌으로부터의 정보 추출 방법에 대하여 논의함과 동시에, 정보 추출 시스템을 이용하여 유전자 온톨로지를 자동으로 확장하는 방법에 대하여 논의하였다

유전자 온톨로지와 같이 신뢰도가 높은 자원을 바탕으로 지식베이스를 확장하는 시스템으로는 Pfam([4])이 있는데 HMM을 이용하여 자동으로 단백질군(protein family) 데이터베이스를 확장하나 적절한 검증 과정을 거치지 않고 있는 실정이지만, 본 연구에서는 자동으로 확장된 지식베이스를 검증하는 방법을 제안하였다.

자동으로 확장되고 검증이 되어 신뢰도가 높아진 지식베이스를 이용하여 추가적인 지식 발견(knowledge discovery)이 가능할 것으로 기대되는데, 실험 환경이나 단백질 간의 상호작용이 일어나는 장소에 따라 다르게 보고되고 있던 정보나 일부분만 보고되고 있는 정보들에 대한 발견 등을 위한 지식 발견 시스템을 현재 개발 중이다. 그리고 대용량의 정보를 포함하는 지식베이스를 효율적으로 시각화(visualization)하여 정보의 검색, 지식의 발견 등을 편리하게 해주는 시스템도 개발 중에 있다.

7. 참고 문헌

[1] The Gene Ontology Consortium. Creating the Gene Ontology Resource: Design and Implementation. Genome Research, 11(8):1425-1433, August 2001.

[2] J. C. Park, H. S. Kim, and J. J. Kim, Bidirectional Incremental Parsing for Automatic Pathway Identification with Combinatory Categorial Grammar. In Proceedings of PSB, 2001.

[3] J. C. Park. Using Combinatory Categorial Grammar to Extract Biomedical Information. IEEE Intelligent Systems, 16(6):62-67, 2001.

[4] A. Bateman et al. The Pfam Protein Families Database. Nucleic Acids Research, 30(1):276-280, 2002.

[5] J.-B. Lee and J. C. Park. Text Data Mining for Automatic Gene Ontology Extension. Second Meeting of the SIG on Text Data Mining, ISMB, 2002.