Logical representation of ontological terminologies in biomedical domain

생물의료분야의 온톨로지 용어의 논리적 표현 기법

Jung-jae Kim[1],, Jin-Bok Lee[1], Hye-Jin Min[1], Ji-yong Jung[1], and Jong C. Park[1*]

[1] Computer Science Division and AITrc, KAIST, Daejeon, Korea

*To whom correspondence should be addressed. E-mail: park@cs.kaist.ac.kr

Abstract

본 논문은 대량의 생물의료분야 문서에서 단백질 이름을 자동으로 인식하고 각 단백질의 특성을 문서에서 자동으로 파악하여 기존의 온톨로지와 연계시키는 방법을 제안한다. 온톨로지 용어가 문서에서 다양한 형태로 발견되기 때문에, 이들을 논리적 표현으로 자동 변환하고, 문서에서 단백질의 특성을 설명하는 문장들을 추출 및 분석하여 온톨로지 용어의 논리적 표현과 비교하였다. 문서에서 단백질 특성을 인식할 때, 약어 처리 및 조응 현상 해결 등의 자연언어처리 기법을 이용하는 방법을 제안하였다.

Introduction

대량의 생물의료분야의 문서들이 MEDLINE과 같은 전문 데이터베이스에 축적되면서 이들 문서에서 자동으로 정보를 추출하는 시스템들이 개발되고 있다 (Hirschman *et al.*, 2002). 이 시스템들은 유전자, 단백질, 세포 이름 등 개체명 뿐만 아니라, 단백질간의 결합정보, 단백질의 세포 내 위치정보, 단백질의 생물현상에서의 기능정보 등 주로 단백질과 관련된 관계정보도 추출하고 있다 (Park *et al.*, 2001, Park, 2001). 이를 위해서 문서 내에 나타난 단백질 이름을 자동으로 인식하는 연구가 많이 진행되었지만 (Collier *et al.*, 2000), 인식된 단백질 이름이 주로 RGS4와 같이 축약형태로 표현되어 있기 때문에 해당 단백질의 일반적인 특성을 파악하지 못하는 문제점이 있다. 본 논문에서는 인식된 단백질 이름을 기존의 생물의료분야 온톨로지와 연계시킬 때 발생하는 용어상 표현문제를 해결하는 방법을 제안한다.

생물의료분야 온톨로지에는 Gene Ontology, MeSH, UMLS 등이 있는데, Gene Ontology에 포함된 클래스의 경우, calmodulin binding activity (GO:0005516)와 같이 클래스에 포함된 단백질 등의 개체들이 공통적으로 가지는 기능을 서술하는 용어로 정의되어 있다. 반면 MeSH의 경우 주로 DNA Topoisomerase와 같이 전문용어로 표현되어 있고, 각 용어에 대한 정의는 다음과 같이 영어, 즉 자연언어로 기술되어 있다.

This work was supported by the Korea Science and Engineering Foundation through AITrc.

<DNA Topoisomerase>
MeSH definition: An enzyme catalyzing ATP-independent breakage of single-stranded DNA, followed by passage and rejoining of another single-stranded DNA. This enzyme class brings about the conversion of one topological isomer of DNA into another, e.g., the relaxation of superhelical turns in DNA, the interconversion of simple and knotted rings of single-stranded DNA, and the intertwisting of single-stranded rings of complementary sequences. (From Enzyme Nomenclature, 1992) EC 5.99.1.2.

이러한 온톨로지 용어들은 문서에서 그대로 활용되지 않고 주로 함축적인 의미만을 담고 있기 때문에 해당 단백질의 특성을 파악하여 온톨로지와 연계시키는데에는 어려움이 있다. 예를 들어 예제1[2]에서는 RGS4가 calmodulin binding activity의 특성을 가지는 것으로 기술되어 있지만 이 온톨로지 용어가 이와 같은 형태 그대로 문서에 나타나지는 않는다. 본 논문에서는 이러한 온톨로지 용어들을 논리적 표현으로 자동 변환하여 문서에서 단백질의 기능을 설명하는 부분과 자동 매칭하는 방법을 제안한다.

예제1) Protein: RGS4
GO:0005516 - calmodulin binding activity
PMID: 10747990
Text: Indeed, Ca2+/calmodulin binds a complex of *RGS4* and a transition state analog of Galpha i1-GDP-AIF4-.

Methods

본 논문에서는 문서에 나타나는 단백질 이름에 Gene Ontology (GO) 클래스를 부여하는 방법을 제안한다. 이를 위해서 문서에서 단백질 이름을 인식하고, GO의 용어들을 의미적 특성으로 분류하여 용어내 단어들 간의 의미 관계를 파악하고, 인식된 단백질 이름의 주변 문맥에서 매칭되는 GO 용어의 단어들을 조합하여 가장 적절한 GO 용어를 선택하는 방법을 제안한다.

단백질 이름 인식

본 논문에서 유전자 및 단백질 이름 리스트는 대용량 단백질 서열 데이터베이스인 Swiss-Prot/TrEMBL[3]의 DE (description)와 GN (gene name) 항목을 추출하여 사용하였다. 위 리스트를 이용하여 문서에서 단백질 이름을 인식할 때, 표 1에 나타나는 생물학적인 지식을 기반한 용어의 다양한 표기방법도 고려하였다.

표 1. 다양한 용어 표현 예제

약어	설명
Sla2p	SLA2 gene product
UBA(2)	UBA2
Mcm2-7	Mcm2, Mcm3, ..., Mcm7
hMMS19	Human homolog of MMS19
ScSls1p	Saccharomyces cerevisiae homolog of SLS1 gene product

표 1에 나타난 약어들은 그 설명이 문서에 명확히 나타나지 않는 예들인 반면, 단백질 이름이지만 단백질 서열 데이터베이스에 아직 등록되지 않았거나 알려진 단백질이지만 등록된 약어 또는 설명과는 다른 표현으로 기술된 경우에 두문자어(acronym)나 동격어(구)(appositive)로 표현되는 경우가 많으므로, 본 논문에서는 두문자어 및 동격어구를 추

[2] BioCreAtIvE(http://www.pdg.cnb.uam.es/BioLINK)에서 언급한 예문

[3] http://kr.expasy.org/sprot/

출하여 사용하였다. 두문자어는 1,2-bis(o-aminophenoxy)ethane-N,N,N'N'-tetraacetic acid tetra(acetoxy-methyl) ester (BAPTA-AM)에서와 같이 숫자표현이 약어에 포함되지 않는 경우를 주의하여서 중간에 한 단어를 건너 뛰는 것을 허용한 상태로 단어의 첫 문자 또는 중간 문자가 괄호안의 단어에 포함되는 조건으로 추출하였다. 동격어구는 문장내에서 일정한 패턴으로 기술되는 경우가 많으므로 다음의 패턴 등을 이용하여 추출하였다.

- 약어, 설명.
- 약어 is 설명
- 설명, designated 약어,
- 설명 also known as 약어
- 설명 (... has been termed 약어)

Gene Ontology 용어 분류 및 논리적 표현

Gene Ontology의 용어들은 크게 단백질의 기능을 분류한 molecular function, 일반적인 생물현상을 분류한 biological process, 그리고 세포내 구조를 분류한 cellular component로 나눠져있다.[4] Molecular function의 용어들은 마지막 단어에 따라 크게 'activity' (7,971개), 'binding' (374개), 기타(6개)로 나눠진다. 'activity' 바로 앞에는 주로 단백질이나 효소 이름이 나오거나, '-er/-or'로 끝나는 단어들이 나와서, 이 단어가 가리키는 단백질 집합을 의미하고, binding으로 끝나는 GO 용어는 binding 앞에 나오는 단백질과 결합하는 단백질들의 집합을 의미한다. Cellular component의 용어들은 이들이 가리키는 세포 구조에 위치하는 단백질들과 연계된다. Biological process의 용어들은 마지막 단어에 따라 크게 'biosynthesis' (8,662개), 'catabolism' (8,781개), 'metabolism' (2,797개), 'regulation' (7,304개), 기타(8,186개)로 나눠지는데, 각 클래스에 포함된 용어는 주로 마지막 단어 앞에 나타나는 단백질의 합성, 분해, 대사 및 조절을 의미한다.

본 논문에서는 GO 용어들을 마지막 단어에 따라 위의 클래스들로 분류하고, 각 용어를 agent 또는 (predicate, agent)로 표현하였다. 이때, agent는 주로 단백질이나 생명현상을 가리키고, predicate은 bind나 regulate와 같은 단백질이 주체가 되는 작용을 의미한다. 본 논문에서는 의미표현의 간략화를 위해 용어의 마지막 단어가 activity나 pathway와 같이 용어의 나머지 단어로 유추가능한 경우는 제외하였다. 예를 들어, 예제 1의 calmodulin binding activity는 (bind, calmodulin)으로 표현된다. 그런데 bind의 경우, 영어 표현으로 associate나 interact와 같은 의미로 사용되고, regulate의 경우 activate나 inhibit이 하위 의미로 사용되므로, molecular function의 binding 클래스와 biological process의 regulation 클래스는 표2와 같은 유사 혹은 하위 의미의 단어 리스트를 작성하여 사용하였다. 그 외에 GO 용어 'cell movement'가 문서에서 'IFN-gamma selectively enhanced *transmigration of Th1-type cells*'와 같이 유사 단어로 표현되는 경우도 있고, GO 용어 'integral to plasma membrane'이 문서에서 'to release it from the plasma membrane'와 같이 추론과정이 필요한 경우도 있지만, 유사단어쌍을 자동으로 찾아내거나 추론하는 시스템을 본 논문에서는 적용하지 않았다.

표 2. 클래스별 유사/하위 단어 리스트

클래스	유사 혹은 하위 단어
binding	associate, interact

[4] http://www.geneontology.org 본 논문에서는 2003년 9월 데이터를 사용하였다.

Regulation	accelerate, activate, augment, block, decrease, down-regulate, inactivate, increase, induce, inhibit, prevent, stimulate, upregulate

GO 용어를 (predicate, agent) 쌍으로 표현하는 이유는 GO 용어 그대로가 문서에 나타나는 경우보다는 predicate과 agent로 구분되어 문장 내에서 나타나고, 또한 둘 사이에 문법적인 의존관계가 있을 때가 많기 때문이다. 예를 들어, 예제 1에서 (bind, calmodulin) 쌍은 동사-주어 관계로 나타나고, 'rRNA modification'은 (modify, rRNA)로 표현되고 "the modification of rRNA"와 같이 전치사구로 수식하는 구조로 나타나고, 'inhibition of caspase activation'은 (inhibit, caspase activation)으로 표현될 수 있는데 예제2에서 동사-목적어 관계로 나타난다.

예제2) Protein: RIP3

GO:0001719 – inhibition of caspase activation

PMID: 10339433

Text: Overexpression of a dominant-negative mutant of *RIP3* strongly inhibited the caspase activation but not the NFkappaB activation induced by TNFalpha.

이보다 훨씬 복잡한 문법 구조를 가질 때도 많은데, 예를 들어, 예제3에서 GO 용어 'regulation of cell growth'의 각 단어는 멀리 떨어져 있지만 문법적인 의존관계, 즉 동사-목적어 관계와 명사구-전치사구 관계로 연결되어 있다.

예제3) Protein: mAbs

GO:0001558 – regulation of cell growth

PMID: 10727470

Text: To identify molecules regulating this interaction, we generated *FDC-staining monoclonal antibodies (mAbs)* and screened them for **their** ability to block FDC-mediated costimulation of growth and differentiation of CD40-stimulated B cells.

예제4에서처럼 GO 용어 'SH3/SH2 adaptor protein activity'가 여러 문장에 걸쳐 나타나는 경우도 있지만, 이 문제를 일반적으로 해결하기 위해서는 추론 과정이 필요하므로 본 논문에서는 다루지 않는다.

예제4) Protein: Grap-2

GO:0005070 – SH3/SH2 adaptor protein activity

PMID: 9878555

Text: In this study, we report the molecular cloning of a novel adaptor-like protein, *Grap-2* (Grb-2 related adaptor protein 2), using the multisubstrate docking protein Gab-1 as bait in the yeast two-hybrid system. Sequence analysis revealed that Grap-2 **contains** a SH3-SH2-SH3 structure that has a high degree of sequence homology to those of the Grb-2 and Grap adaptor molecules.

문맥에 적절한 Gene Ontology 용어 매칭

지금까지 문장에서 단백질 이름과 GO 용어가 문서에서 어떤 형태 및 구조로 나타나는지를 살펴보았다. 본 절에서는 이들이 어떤 구조로 문서내에서 같이 나타나고, 어떤 특징이 이들과 연관되어 있음을 알려주는지를 살펴본다.

단백질 이름과 GO 용어의 연관성을 가장 쉽게 알 수 있는 경우는 GO 용어가 단백질 이름와 동일한 경우이다. 예를 들어, 단백질

이름 'Xylulokinase'는 GO 용어 'Xylulokinase activity'로 바로 연계시킬 수 있다. 그러나 문서에서 나타나는 용어는 GO 용어와 의미적으로는 동일하지만 기호 사용이나 표기 방법에서 서로 다른 경우가 대부분이다. 예를 들어, GO 용어 'Deoxyribonuclease II activity'와 'Interleukin-15 receptor activity'는 약어를 이용하여 문서에서 'DNase II'와 'IL-15 receptor'로 표현된다. 그 다음 간단한 구조는 단백질 이름과 GO 용어가 동사로 연결된 경우로 다음은 그 예들이다.[5]

예제5) Protein: Rad21 subfamily, Rec8 subfamily
GO:0007062 - sister chromatid cohesion
PMID: 10207075
Text: Our work and that of others defined mitosis-specific (*Rad21 subfamily*) and meiosis-specific (*Rec8 subfamily*) proteins **involved** in sister chromatid cohesion in several eukaryotes, including humans.

예제6) Protein: h-warts/LATS1
GO:0004674 – protein serine/threonine kinase activity
PMID: 10207075
Text: A human homologue of the Drosophila warts tumor suppressor, *h-warts/LATS1*, **is** an evolutionarily conserved serine/threonine kinase and a dynamic component of the mitotic apparatus.

예제7) Protein: ORC
GO:0006260 – DNA replication
PMID: 10438470
Text: *The origin recognition complex (ORC)* **is** an initiator protein for DNA replication, but also effects transcriptional silencing in Saccharomyces cerevisiae and heterochromatin function in Drosophila.

그외에는 예제8에서처럼 같은 문장내에 멀리 떨어져 있지만 문법적 의존관계를 가지는 경우도 있고, 예제9에서처럼 대명사를 통하여 여러 문장에 걸쳐 표현되는 경우도 있다.

예제8) Protein: Mms19
GO:0006350 – transcription
PMID: 11071939
Text: An intriguing example is *the Saccharomycecerevisiae Mms19 protein* that **has** an unknown dual **function** in NER and RNA polymerase II transcription.

예제9) Protein: LDLR
GO:0006629 – lipid metabolism
PMID: 10049586
Text: *The low-density lipoprotein receptor (LDLR) family* is a group of related glycoprotein receptors that are bound to by a diverse array of ligands. **These receptors** play critical roles in the endocytotic processes of plasma apolipoproteins and therefore regulate cholesterol homeostasis and lipid metabolism (Krieger and Herz, 1994).

본 논문에서는 같은 문장내에 문법적 의존관계를 가지는 단백질 이름과 GO 용어를 연계시키기 위해서 문장내 문법적 의존구조를 파악하여 문법적 의존관계를 가지는 단백질 이름과 GO 용어를 연계시키고, 여러 문장에 걸쳐 표현되는 연관관계를 파악하기 위해서 조응현상을 해결하는 방법을 제안한

[5] 단백질 이름은 이탤릭체로, GO 용어는 밑줄로, 연결하는 동사는 굵게 표시되어 있다.

다.

문장내 문법적 의존구조를 파악하기 위해서 결합범주문법(Steedman 2000)을 이용하여 각 단어에 부여하는 범주에 적합한 의존구조를 부여하였다. 예를 들어, 동사 'inhibit'의 경우 문법적 범주로 '(s\np)/np'를 부여하는데,[6] 이에 적합한 의존구조를 추가하여 'inhibit'에 '(s:W(X,Y)\np:Y)/np:X'를 부여하였다.[7]

여러 문장에 걸친 연관관계를 추출하기 위해서는 조응현상을 해결해야 하는데, 본 논문에서는 단백질 이름과 GO 용어의 연관관계를 파악하는 시스템으로까지 통합되지는 않은 상태이지만, 기본적으로 중심화 이론(Centering Theory, Grosz *et al.* 1995)에서 제안하는 것과 같이 바로 전 문장의 주어와 목적어를 우선 선행사 후보로 간주하고, 조응현상의 대상이 단백질 이름이므로 선행사 후보를 단백질 관련 표현으로 제한하는 방법으로 구현하였다.

Experiment and Results

본 논문에서는 GoA 프로젝트에서 구축한 Swiss-Prot/TrEMBL에 포함된 단백질과 GO 용어를 연계시킨 데이터 중 303개 연관관계를 학습데이터로 사용하였다.[8] GoA 데이터에서 연계된 GO 용어를 GO의 최상위 카테고리별로 분류하면 표3과 같이 나타난다.

표 3. GoA 코퍼스에 나타난 카테고리별 용어 빈도

카테고리	학습데이터에서 사용된 횟수 (번)
Molecular function	105
Biological process	140
Cellular component	58

단백질 이름과 GO 용어를 연계시키기 위해 본 논문에서 제안하는 방법의 타당성을 검증하기 위해 다음과 같은 방법들을 실험하였다.

방법1) 같은 문장내 단백질 이름과 GO 용어가 같이 나타나면 연계시킴[9]

방법2) 같은 문장내 단백질 이름과 GO 용어의 모든 단어들이 같이 나타나면 연계시킴

GO 용어가 예를 들어 'protein', 'cell', 'binding'과 같이 한 단어로 이루어진 경우, 단백질 이름을 설명하기 위해 사용되는 경우가 드물므로, 본 실험에서는 두 단어 이상으로 이루어진 GO 용어에 대해서만 실험하였다. 방법1과 방법2를 44개의 논문 초록에 적용한 결과는 표4와 같다.

표 4. 방법1과 방법2에 대한 실험 결과

	방법1	방법2
인식한 GO 용어 수	853	1,087
두 단어 이상으로 이루어진 GO 용어 수	202	436
같은 문장내 단백질 이름과 GO용어가 연계된 횟수	106	174
잘못 연계된 횟수	41	235
정확율	72.1%	42.5%

[6] 괄호 밖의 np(목적어 명사구)를 오른쪽(/)에서 받고 나서 괄호 안의 np(주어 명사구)를 왼쪽(\)에서 받은 다음 s(문장)을 출력하는 의미의 범주이다.

[7] W는 단어 자체를 가리키고 W(X,Y)는 X와 Y가 W에 의존한다는 의미이다.

[8] http://www.ebi.ac.uk/GOA/

[9] GO 용어가 문서에서 다양한 표현으로 나타나는 것을 해결하기 위해 단어의 어간만을 매칭하였고, 용어 내에 하나 이하의 단어가 포함되는 것을 허용하였다.

표4에서 알 수 있듯이 방법1이 방법2보다 정확율이 높지만 재현율은 훨씬 낮은 것을 확인할 수 있다. 문장의 문법적인 구조를 보지 않고 공기정보만을 사용하였을 때 생기는 이러한 문제를 보완하기 위해 본 논문에서 제안한 문장의 의존구조를 이용할 경우, 예제10과 같은 표현에서 정확한 연관관계를 추출할 수 있다.

예제10) Protein: E47
GO:0045595 – regulation of cell differentiation
PMID: 10781029
Text: The E2A protein *E47* is known to be involved in the regulation of tissue-specific gene expression and cell differentiation.

Discussion

생물의료분야 문서에서 나타나는 단백질 이름을 Gene Ontology의 용어와 연계시키기 위해 Chiang *et al.* 2003과 Raychaudhuri *et al.* 2002에서는 공기정보만을 기계학습에 이용하였지만, 본 논문에서는 문맥정보를 이용하는 방법을 제안하였다. 온톨로지 용어가 문서에서 다양한 형태로 발견되기 때문에 사용하는 문맥정보로 어휘내 문자 혹은 기호 수준의 다양성, 유사 혹은 하위 단어 수준의 어휘적 다양성, 의미적 구조에 따른 의존구조 안에서의 다양성 등을 고려하였다. 그리고 단순히 용어의 공기정보만을 이용하는 방법을 실험하여 제안방법으로 이러한 방법을 보완할 수 있음을 보였다.

References

[1] J.C. Park, H.S. Kim, and J.J. Kim, Bidirectional incremental parsing for automatic pathway identification with Combinatory Categorial Grammar, Proc. *Pacific Symposium on Biocomputing*, 6, 2001, 396-407

[2] J.C. Park. Using Combinatory Categorial Grammar to extract biomedical information, *IEEE Intelligent Systems in Biology*, *16*(6), 2001, 62-67

[3] L. Hirschman, J.C. Park, J.-I. Tsujii, L. Wong, and C. Wu. Accomplishments and challenges in literature data mining for biology, *Bioinformatics*, *18*(12), 2002, 1553-1561

[4] J.H. Chiang and H.C. Yu, MeKE: discovering the functions of gene products from biomedical literature via sentence alignment, *Bioinformatics*, *19*(11), 2003, 1417-1422

[5] Barbara Grosz, Aravind K. Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse, *Computational Linguistics*, 1995, 203-225

[6] Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii, Extracting the Names of Genes and Gene Products with a Hidden Markov Model, Proc. *Int'l Conf. on Computational Linguistics*, 2000

[7] Mark Steedman, The syntactic process, *MIT Press*, 2000

[8] Torgeir R. Hvisten, Astrid Laegreid, and Jan Komorowski, Learning rule-based models of biological process from gene expression time profiles using Gene Ontology, *Bioinformatics*, *19*(9), 2003, 1116-1123

[9] S. Raychaudhuri, J.T. Chang, P.D. Sutphin, R.B. Altman, Associating Genes with Gene Ontology Codes Using a Maximum Entropy Analysis of Biomedical Literature, *Genome Research*, *12*(1), 2002, 203-214