단백질 상호작용 추출을 위한 확장성을 가진 텍스트 마이닝 기법

이현철[0], 여은주[1], 강희영[2], 조완섭[1], 김학용[2], 유재수[3]
충북대학교 정보산업공학과[0]*, 경영정보학과**, 생명과학부***, 전기전자 컴퓨터공학부
enigma7@korea.com[0], triumph2, nannaingir@daum.net, wscho, hykim, ysj@cbnu.ac.kr

An Extensible Text Mining Technique for the Extraction of Protein-Protein Interaction

Hyunchul Lee[0], Eunju Yeo[1], Heeyoung Kang[2], Wansup Cho[1], Hakyoung Kim[2], Jaesoo Yoo[3]
Dept. of Information Industrial Engineering, Chungbuk National University[0]
Dept. of Management Information System, Chungbuk National University[1]
Division of Lifesciences, Chungbuk National University[2]
School of Electrical and Computer Engineering, Chungbuk National University[3]

요 약

단백질 간의 상호작용에 대한 연구는 생물학적 프로세스를 이해하기 위해 중요한 부분이다. 이러한 단백질 간의 상호작용에 대한 정보는 주로 생명과학 관련 연구논문에 존재하지만 컴퓨터로 자동으로 처리하여 상호작용에 관한 정보를 추출할 수 있기 위해서는 텍스트 마이닝 기술이 적용되어야 한다. 바이오 텍스트 마이닝에서 대두되고 있는 중요한 쟁점은 대용량의 연구논문에서 필요한 정보를 어떻게 효율적으로 정확하게 추출할 것인가에 대한 내용이다. 또한, 관심이 있는 단백질의 종류나 관련성을 표시하는 문장내 패턴의 다양성을 수용하기 위하여 개발하는 시스템의 확장성을 높이는 것도 소프트웨어 공학적인 측면에서 중요한 이슈이다. 이 논문의 목적은 생물학적 내용을 담고 있는 연구논문으로부터 단백질간의 상호작용을 추출하는 확장성을 가진 텍스트 마이닝 기법을 제안하는데 있다.

1. 서 론

생명공학의 고속 발달로 인하여 그와 관련된 다양한 형태의 바이오 데이터가 대량으로 생성되고 있다. 서열 및 구조 데이터의 정보도 여러 공개 데이터베이스에 축적되고 있으며, 이와 관련된 각종 연구보고서 및 논문들의 증가 추세도 이전의 패턴을 넘어서고 있다. 이러한 대량의 바이오 데이터는 자연언어로 되어있기 때문에 자동으로 데이터를 처리할 수 있는 방법이 필요하게 되었다. 몇 십 년 전부터 생물의학에서 자연언어처리(NLP)는 환자의 데이터를 자동으로 처리하는 부분에서 많이 사용되어져 왔지만 이 데이터는 항상 동일한 형식을 가지고 있었다. 하지만 바이오 데이터는 각기 다른 형식으로 저장되기 때문에 기존의 방법과는 다른 방법이 요구되어 졌다[11].

본 연구의 목적은 생물학적 내용을 담고 있는 연구 논문의 초록으로부터 자연언어처리와 마이닝 기술을 이용하여 관심 있는 단백질간의 상호작용을 파악하고, 이를 바탕으로 단백질 기능과 세포 반응의 분석을 위한 기반을 조성하고자 한다.

유전자 서열이 밝혀진 뒤, 여러 공개 데이터베이스에 공개되었는데[12], 이 자료를 통하여 각각의 유전자 기능 및 서열에서의 역할을 분석하게 되었다. 유전자 이외에 단백질 수준에서 구조와 기능을 분석하는 연구의 중요성이 대두되었는데, 유전자에 대한 정보에 비해 단백질에 대한 정보는 비교적 한정되어 있다. 따라서 단백질 데이터베이스의 구축이 필요하며, 이를 바탕으로 단백질간의 상호작용 분석은 미지의 단백질 기능 및 단백질 네트워크의 분석으로 세포내 작용을 밝히고 생물학적 프로세스를 이해하기 위해 반드시 필요한 부분이다.[1]

본 논문은 2004년 과학기술부 국책생물정보학 연구개발 사업의 2차년도 지원을 받았음.

본 논문에서 제안한 기법은 형태소 분석을 위하여 Brill POS tagger를 사용하였고, 단어의 접미사 변형을 막기 위해 Porter Stemming Algorithm을 적용하였다. 그리고 기존의 바이오 텍스트 마이닝 시스템과는 확장성에 차별화를 두어 설계하였다. 확장성이란, 연구자에 의해 단백질 사전의 내용을 바꿈으로써, 상호작용 추출의 범위의 수정이 가능하도록 하고 상호작용에 사용되는 패턴 또한 연구자에 의해 수정, 추가가 가능함을 말한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구를 간단하게 소개한다. 3장에서는 시스템에 대한 전체적인 구조에 대하여 설명한다. 4장에서는 결론 및 향후 연구방향으로 본 논문을 맺는다.

2. 관 련 연 구

바이오 인포메틱스에서 자연언어처리를 이용한 텍스트 마이닝의 분야는 아래와 같이 여러 가지 예로 활용되어 지고 있다.

- 단백질간의 상호작용 검색 [1,2,3]
- 단백질-유전자 상호작용 검색 [4]
- 대사경로 분석 [5]
- 사전 구축 [6,7]

생물학적 데이터의 분석은 연구에 따라 다양한 목적을 가지고 있기 때문에 각 목적에 맞도록 국내에서도 많은 프로그램들이 여러 연구자들에 의해 만들어 지고 있다[8]. 하지만 관심 있는 단백질간의 상호작용을 추출하는 프로그램은 개발되지 않고 있다. 물론 AngioDB[13]에서 상호작용 추출에 대한 내용을 언급하고는 있지만, 확장성에 대해서는 논의가 되지 않고 있다. 본 논문은 생물학 문서에서 자동화된 단백질 상호작용 추출[1]의 내용을 기반으로 본 연구 목적에 맞도록 수정, 설계 되었다.

3. 시스템 동작 절차

시스템의 동작 절차는 그림 1과 같고, 그림 2는 입력된 문장이 처리되는 과정을 예를 통해서 나타낸 것이다. 그림 1의 첫 번째 단계에서는 Brill POS tagger package[9]를 이용하여 입력된 문장의 형태소를 분석한다 그림 2의 . 두 번째 단계에서는 분석된 형태소에서 동사 추출 및 단어의 접미사의 변형을 막기 위하여 Porter Stemming Algorithm[10]을 적용한다. 세 번째 단계에서는 상호작용 추출이 가능한 키워드와 비교한 뒤, 적절한 문장을 선택한다. 네 번째 단계에서는 추출된 문장을 단백질 사전과 비교를 통하여 단백질 이름을 인식한다. 그리고 마지막 단계에서는 패턴 비교를 통하여 단백질간의 상호작용을 추출한다. 각 단계의 상세한 설명은 다음 절에서 하겠다.

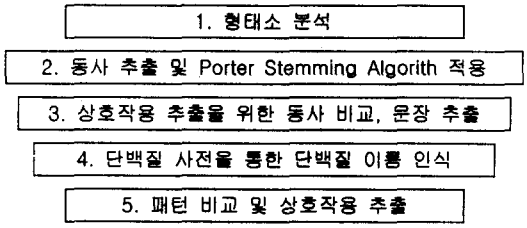

그림1. 바이오 텍스트 마이닝 시스템 동적 절차

3.1 형태소 분석

입력된 파일에 포함된 각 단어의 형태소를 분석하기 위해서 Brill POS tagger package[9]를 사용하였다. 분석에 사용되는 태거는 약 50가지로 나누어지는데 대표적인 것은 표1과 같다. 표 1은 형태소 분석에 사용되는 태거와 그 의미를 나타낸 것이다. 입력된 파일의 문장들은 이 과정을 거쳐 그림2 단계 ①의 문장이 단계 ②와 같이 각 단어의 형태소가 분석된다. 형태소의 분석은 문장에서 동사를 검색하고, 패턴 비교 기법을 적용하기 위해서 이루어진다.

표1. 태거의 종류와 의미

태 거	의 미
CC	coordinating confunction
CD	cardinal number
DT	determiner
IN	preposition or subordinating conjugation
JJ	adjective
NN	noun, singular or mass
NNP	noun, plural
VB	verb
VBN	verb, part participle

3.2 동사 추출 및 Porter Stemming Algorithm 적용

상호작용을 추출할 수 있는 키워드와 비교를 위해 전 단계에서 분석된 형태소를 이용하여 동사를 추출한다. 추출된 동사의 품사변화에 따라 변화되는 접미사를 없애기 위해 Porter Stemming Algorithm[10]을 적용하였다. 이는 다음단계에서 처리될 단백질 상호작용 추출 키워드와의 단순한 비교를 위하여 행해진다. 예를 들어, interact라는 단어는 -tion, -ted, -ting과 같이 여러 종류의 접미사가 붙을 수 있는데, 이 과정을 거친 후에는 모두 'interact' 라는 하나의 단어가 된다. 따라서 키워드의 비교는 'interact'라는 단어를 사용하여 처리된다. 그림2의 단계 ②의 밑줄친 부분이 동사가 추출된 부분이며, 굵게 표시된 부분이 알고리즘을 적용한 결과이다.

3.3 상호작용 추출을 위한 동사 비교, 문장 추출

전 단계에서 Porter Stemming Algorithm을 적용시킨 동사와 상호작용 추출에 사용되는 키워드를 비교 한 뒤, 일차적으로 상호작용을 포함하고 있을만한 문장을 추출한다. 확장성을 위하여 키워드는 추가가 가능하도록 설계되었다.

상호작용 추출에 사용되는 키워드는 'interact', 'bind', associate', 'complex', 'substrate', 'couple', 'phosphorylate'의 7개 동사를 사용하였다. 이는 Medline에서 'human, protein'이라는 검색어를 사용하여 얻어진 연구논문 초록 가운데 1000여개를 대상으로 하여 상호작용에 대한 내용을 갖는 문장에서 2번 이상 나타난 동사를 대상으로 한 것이다. 그림 2의 단계②에서 추출키워드와 같은 interaction이 추출되고, 이 키워드가 포함된 문장이 상호작용 추출에 선택된다.

표2. 단백질 상호작용 추출을 위한 패턴.A,B는 단백질이름

키워드	패 턴
Interact	interact between A and B
	interact of A with B
	interact of A and B
	interact with A and B
	A interact with B
	A-B interact
	A/B interact
Associate	associ between A and B
	associ of A with B
	A associ with B
Bind	bind of A to B
	A bind to B
	A bind B
	A bound to B
Complex	A-B complex
	A/B complex
Phosphorylate	A phosphorylat to B
	A is phosphorylat by B
Substrate	A substrat of B
Couple	A coupl to B

3.4 단백질 사전을 통한 단백질 이름 인식

추출된 문장에 포함되어 있고 상호작용에 관여하는 단백질 이름을 인식하는 과정이다. 상호작용의 추출을 위해서는 문장에 포함된 단백질 이름이 반드시 2개 이상이여야 한다.

본 연구는 초록에 나타난 모든 단백질 이름을 대상으로 하는 것이 아니라, 연구자가 관심이 있는 단백질 이름만을 대상으로 하기 때문에 기존의 연구[8]와는 다른 특정 범위를 가지게 된다. 단백질 이름의 인식은 수동으로 구축된 사전을 이용하여 문장에 나타난 단어와의 패턴 비교 기법을 사용하여 그림2의 단계 ③의 굵게 표시된 부분과 같이 단백질 이름이 인식을 하게 된다.

단백질 사전은 데이터베이스로 만들어졌기 때문에 연구자에 의

단계①	These results indicate that pCTLA4-Ig may be a useful reagent to define the precise nature of the interaction between B7 and Cytotoxic T-lymphocyte antigen 4 (CTLA4).
단계②	These/DT results/NNS indicate/VB that/IN pCTLA4-Ig/JJ may/MD be/VB a/DT useful/JJ reagent/NN to/TO define/VB the/DT precise/JJ nature/NN of/IN the/DT interaction/NN between/IN B7/CD and/CC Cytotoxic/NNP T-lymphocyte/JJ antigen/NN 4/CD (CTLA4)./JJ
단계③	These/DT results/NNS indicate/VB that/IN pCTLA4-Ig/JJ may/MD be/VB a/DT useful/JJ reagent/NN to/TO define/VB the/DT precise/JJ nature/NN of/IN the/DT interaction/NN between/IN B7/CD and/CC Cytotoxic/NNP T-lymphocyte/JJ antigen/NN 4/CD (CTLA4)./JJ
단계④	interaction between B7 and Cytotoxic T-lymphocyte antigen 4 (CTLA4).
단계⑤	interaction between B7 and CTLA4

그림 2. 정보 추출 과정의 예

해 관심 있는 단백질 리스트로 사전을 변경할 수 있다. 본 연구에 사용된 단백질 사전은 인간의 건강과 질병에 관련된 기능과 연관성을 갖는 단백질을 도메인으로 하는 Human Protein Reference Database (www.hprd.org)의 데이터를 이용하였다. 구축된 사전에는 동의어를 포함하여 약 7000개 단백질 이름을 포함되어 있다.

단백질 이름의 인식을 위한 문자 비교의 범위를 형태소 분석을 통하여 줄이고자 하였다. 하지만, 단백질 이름이 워낙 다양하여 모든 형태소로 분석이 되었다. 따라서 모든 형태소에서 단백질 이름을 비교 하였다. 그리고 한 문장 내에서 단백질 이름 동의어인 경우에 잘못된 상호작용 관계를 추출할 수 있기 때문에 이를 방지하기 위하여 한 문장내의 동의어에 대한 중복 처리를 하였다. 그림2의 단계 ③에서 추출된 단백질 이름은 B7, Cytotoxic T-lymphocyte antigen 4, CTLA4 3가지이지만 Cytotoxic T-lymphocyte antigen 4와 CTLA4는 동의어 이므로 단계 ④와 같이 하나의 단백질로 처리된다.

3.5 패턴 비교 및 상호작용 추출

이전 단계에서 상호작용 추출 키워드를 포함하고, 단백질 이름이 2개 이상인 문장을 추출하였다. 마지막 단계에서는 추출된 문장을 패턴비교 알고리즘을 사용하여 단백질간의 상호작용을 추출하게 된다. 상호작용 추출을 위한 패턴은 표2와 같이 간단한 규칙과의 비교를 통하여 적용한다. 패턴은 데이터베이스화하여 연구자에 의해 추가/삭제가 가능하도록 설계를 하였다. 그리고 부정적인 내용을 갖는 문장이나 복합 문장에 관한 처리[1]를 하게 된다. 그림2의 단계 ⑤와 같이 단백질의 상호작용에 대한 정보가 추출된다.

4. 결론 및 향후 연구계획

본 논문에서는 생물학적 내용을 담고 있는 연구 논문의 초록에서 텍스트 마이닝 기술을 적용하여 확장성을 가진 단백질간의 상호작용을 추출하는 방법을 제안하였다. 이 방식을 사용하여 시스템을 구성할 경우, 연구자가 관심 있는 범위의 단백질간의 상호작용을 파악할 수 있고, 특정 단백질의 반응을 쉽게 찾아낼 수 있어 기존의 모든 유전자와 단백질을 대상으로 처리되던 바이오 텍스트 마이닝과는 차별화를 보인다. 이러한 연구 방법으로 시스템을 구축할 경우 생물학 관련 분야뿐만 아니라 의학 분야에서도 도움이 될 것으로 기대된다.

또한 이 시스템을 구축하고 정확도와 회수율을 측정해 기존의 시스템과의 비교를 향후과제로 진행할 계획이다.

참 고 문 헌

[1] Ono T, Hishigaki H, Tanigami A, Takagi T : Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics 2001 Feb;17(2):155-61

[2] Blaschke C, Andrade MA, Ouzounis C, Valencia A : Automated extraction of biological information from scientific text : protein-protein interactions. Proc Int Conf Intell Syst Mol Biol 1999;30A(2): 60-7

[3] Thomas J, Milward D, Ouzounis C, Pulman S, Carroll M : Automatic extraction of protein interactions from scientific abstracts. Pac Symp Biocompute 2000:541-52

[4] Sekimizu T, Park HS, Tsujii J : Identifying the interaction between genes and gene products based on frequently Seen Verbs in Medline Abstracts. Genome Inform Ser Workshop Genome Inform 1998;9:62-71

[5] Friedman C, Kra P, Yu H, Krauthammer M, Rzhetsky A : GENIES : a natural-language processing system for the extraction : identifying protein names from biological papers. Pac Symp Biocompute 1998;33(2):707-18

[6] Ohta Y, Yamamoto Y, Okazaki T, Uchiyama I, Takagi T : Automatic construction of knowledge base from biological papers. Proc Int Conf Intell Syst Mol Biol 1997;5:(2)218-25

[7] Rindflesch TC, Hunter L, Aronson AR : Mining molecular binding terminology from biomedical text. Proc AMIA Symp 1999;34(12):127-31

[8] Bio Text Miner : http://bi.snu.ac.kr/~jheom/cgi-bin-perm/CrazyWWWBoard.cgi?db=BioTextMiner

[9] Brill, E. : Some advances in transformation-based part of speech tagging. In Proc of the Twelfth National Conf on Artificial Intelligence. AAAI Press 1994;

[10] Porter, M.F An algorithm for suffix stripping. Program, 1980;14;127-30

[11] 김태현, 이현숙, 박수준, 박선희 : 바이오 텍스트 마이닝 기술동향 ETRI IITA IT정보, 2003,9,3(www.itfind.or.kr).

[12] Kyoto Encyclopedia of Genes and Genomes[KEGG] available http://www.kegg.com/

[13] Yong S. Choi, Jaehyuk Cha : AngioDB : 혈관신생 인자에 대한 데이터베이스 구축 및 활용 연구, 대한 생화학• 분자생물학회 OMICS 연수강좌 자료집, 2003년 8월.