BIOLOGY ORIENTED TARGET SPECIFIC LITERATURE MINING FOR GPCR PATHWAY EXTRACTION

GPCR 경로 추출을 위한 생물학 기반의 목적지향 텍스트 마이닝 시스템

Eunju Kim[1], Seolkyoung Jung[1], Eunji Yi[1], Gary Geunbae Lee[1], Soo-jun Park[2]

[1] Natural Language Processing Lab, Department of CSE, Pohang University of Science and Technology(POSTECH), Pohang, Korea
[2] Bioinformatics Research Team, Computer and Software Research Lab, ETRI, Taejon, Korea

***E-mail: { hosuabi, snow, juicy, gblee}@postech.ac.kr, psj@etri.re.kr**

Abstract

Electronically available biological literature has been accumulated exponentially in the course of time. So, researches on automatically acquiring knowledge from these tremendous data by text mining technology become more and more prosperous. However, most of the previous researches are technology oriented and are not well focused in practical extraction target, hence result in low performance and inconvenience for the bio-researchers to actually use. In this paper, we propose a more biology oriented target domain specific text mining system, that is, POSTECH bio-text mining system (POSBIOTM), for signal transduction pathway extraction, especially for G protein-coupled receptor (GPCR) pathway. To reflect more domain knowledge, we specify the concrete target for pathway extraction and define the minimal pathway domain ontology. Under this conceptual model, POSBIOTM extracts interactions and entities of pathways from the full biological articles using a machine learning oriented extraction method and visualizes the pathways using JDesigner module provided in the system biology workbench (SBW) [14]

Introduction

1990년대 후반, 생물학 관련 문서가 급격히 증가함에 따라 자연어 처리 기술을 이용하여 생물학적 문서로부터 정보를 추출하고자 하는 연구가 활발히 진행되기 시작하였다. 연구 초기의 주요 관심사는 MEDLINE 등의

본 연구는 ETRI 위탁연구과제 지원으로 수행되었음.

논문 데이터베이스로부터 검색된 초록(abstract)들을 대상으로 molecular interaction 정보를 추출하는 것이었다. 이러한 연구 결과로서 단백질-단백질과 유전자-단백질 상호작용, 분자적 결합 관계, 유전자나 단백질과 약품간의 상호작용 등의 정보 추출에 미리 정의된 유의미한 단어 사전에 기반한 문장 내 명사 인식을 이용하는 방법들이 제시되었다[1-5]. 또한 이벤트 추출을 위해 문장의 파싱(parsing)을 이용하는 방법론들도 제시되었다[6-8].

개개의 상호작용 추출에 중점을 두고 있는 연구들과는 달리 전체적인 네트워크나 경로에 중점을 둔 연구들도 있다. Human과 Saccharomyces cerevisiae에서 유전자의 동시 발생 네트워크를 구축하는 데 명사의 공기(collocation) 정보를 이용하는 방법에 대한 연구가 이에 해당한다[9,10].

논문 초록을 대상으로 biomedical 또는 세포내 경로에 대한 정보를 자동으로 추출하고 시각화하는 시스템의 prototype으로서 Ng와 Wong[11]은 간단한 규칙 기반 pattern matching을 이용한 시스템을 제시하였다. 또한 Leroy와 Chen[12]은 추출한 정보로부터 전치사 기반의 템플릿(template)을 구성하는 의학적 파서(medical parser)를 제안하였으며, Friedman[13] 등은 세포내 경로에 대한 정보를 추출하기 위한 시스템인 GENIES를 제시하였다.

본 논문에서는 전체적인 생물학적 경로의 추출, 특히 GPCR에 관련된 경로 추출에 초점을 두는 시스템으로서 POSTECH bio-text mining system (POSBIOTM)을 제안한다. POSBIOTM은 신호 전달 경로의 추출에 초점을 둔 GENIES 등의 다른 시스템들과 다음과 같은 점에서 차별점을 지닌다. 첫째, 신호 전달 경로 추출 과정에 도메인 지식을 이용하기 위하여 생물학적으로 의미 있는 도메인 기반의 추출 대상(target)을 정의하였다. 둘째, 경로의 참여자(participant) 및 각 참여자들의 특성(property) 추출을 위해 정보 추출(information extraction, IE) 기반의 방법론을 채택하였으며, 자동으로 규칙을 찾기 위하여 교사 기계 학습(supervised machine learning) 방법을 사용하였다. 이 때 파서 대신 자동 개체명 추출 모듈을 사용함으로써 구조적 변형에 안정적인 모델을 구축하였다. 마지막으로, 추출한 정보의 자동 시각화를 위하여 SBW(System Biology Workbench)[14]의 시각화 모듈인 JDesigner를 이용하여 표준화와 모듈 재사용에 주력하였다.

본 논문은 다음과 같이 구성된다. 먼저 GPCR 관련의 생물학적 경로 추출을 위하여 target-specific한 자료 모델을 정의한다. 다음으로 POSBIOTM에서 정의하는 경로 추출 대상에 대해 설명한 뒤 POSBIOTM의 경로 추출 과정 및 시스템 구조를 설명한다. 마지막으로 실험 결과를 보이고 결론을 제시한다.

GPCR 관련 경로 추출 자료 모델

본 논문에서는 GPCR 관련 생물학적 경로 추출을 위한 해당 도메인 지식을 포함하는 자료 모델(그림1)을 다음과 같이 정의하였다. 생물학적 경로는 반응(reaction)을 기본 구성자로 가지며, 각 경로는 하나 이상의 기본 구성자로 이루어지는 것으로 정의한다. 이 때 각 경로는 전체 경로의 기능이나 특성을 결정하게 되는 조직, 기관 등에서의 생물학적 위치 등 환경에 관한 정보를 특성으로 가질 수 있다.

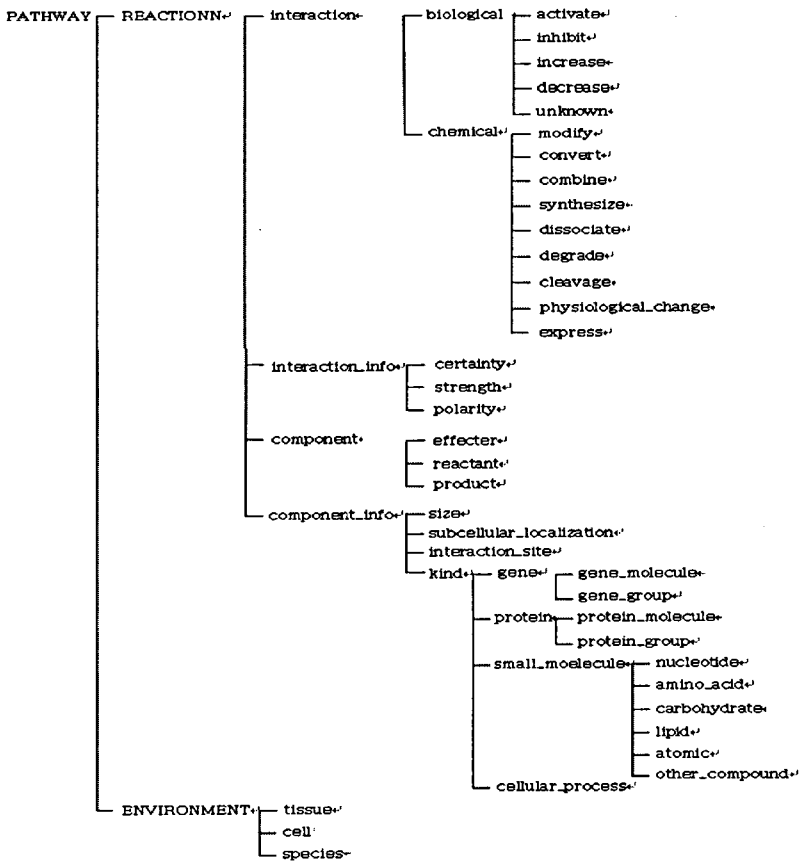

그림 1 GPCR 관련 경로 추출 자료 모델

반응은 하나의 상호작용과 여러 개의 반응 성분(component)으로 구성되고 그에 따른 추가 정보를 가지도록 정의한다.

상호작용은 크게 생물학적(기능적) 상호작용과 화학적(분자적) 상호작용으로 나눈다. 생물학적 상호작용은 하나의 반응 성분이 다른 성분의 상태에 생물학적으로 어떻게 영향을 미치는가에 대한 정보를 포함한다. 이 때 'unknown'은 2개의 성분이 서로 관계는 있으나 그 내용이 구체적으로 명시되어 있지 않은 경우를 나타낸다. 화학적 상호작용은 상호작용 유형에 따라 group으로 나누었다. 예를 들어 'modify' 그룹의 경우 'phosphorylate', 'acetylate', 'ubiquinate', 'prenylate' 등의 상호작용을 포함한다. 또한 상호작용의 세기, 확신도, 긍정/부정 정보를 특성으로 가지도록 하여 여러 문서로부터 추출된 정보들 간의 모순을 해결하는 데 이용하도록 한다.

반응 성분은 반응에서의 생물학적 역할에 따라 효과자, 반응자, 산출물로 나눈다. 각각의 성분은 크기, 세포 내 국지화(localization), 상호작용이 일어나는 종(species) 및 기관(organ) 등을 하위 특성으로 가질 수 있다. 특히 생물학적 반응의 성분으로는 단백질, 유전자, 소분자, 세포내 과정 등이 있다. 이 때 소분자의 경우 다른 화학제품, 독소 등과 구별하기 위해 세포나 기관 내에 본래 존재하는 소분자로 범위를 제한하여 다른 화학제품이나 독소 등과 구

별되도록 한다. 또한 세포적 과정은 질병과 관련된 과정을 제외한 생물학적 경로의 결과로서 나타나지는 세포의 기능을 의미한다. 이와 같은 정의된 자료 모델을 기반으로 개체명, 상호작용 관계, 경로 추출 대상 등을 태깅(tagging)하여 말뭉치(corpus)를 구성하였으며, 이 말뭉치는 POSBIOTM의 개체명(named entity) 인식 및 관계 규칙 학습(relation rule learning) 모듈에서 이용된다.

GPCR 관련 경로 추출 대상

일반적인 IE 태스크의 경우 템플릿 요소(template element), 템플릿 관계(template relation), 시나리오 템플릿(scenario template)으로 추출 대상을 분류할 수 있다. POSBIOTM은 이러한 일반적 추출 대상, 특히 MUC-7에서 정의된 대상을 경로 추출에 적합하도록 수정하여 템플릿 요소, 템플릿 관계, 이벤트(event)를 추출 대상으로 정의한다.

템플릿 요소(Template Element)

경로 추출에 있어서 전체 신호 전달 경로는 프레임(frame) 집합으로 간주되며, 이 때 대상 슬롯들은 경로의 참여자 혹은 참여자의 특성이 된다. 템플릿 요소는 프레임에서 슬롯을 채우는 내용으로, 객체(entity)와 객체간의 상호작용(interaction)으로 구성되며 객체와 상호작용은 각각 특성을 가질 수 있다. POSBIOTM에서 객체로 정의되는 대상은 단백질(protein), 유전자(gene), 소분자(small molecule), 세포 과정(cellular process) 등이며 이들 객체는 각기 종(species), 기관(organ), 분자 크기 등의 특성을 가질 수 있다. 상호작용은 bind, activate 등과 같이 객체들 간의 관계를 나타내는 키워드로, 확신도(certainty), 세기(strength), 긍정/부정(polarity) 등의 특성을 가질 수 있다.

템플릿 관계(Template Relation)

템플릿 관계는 참여자와 참여자의 특성간의 관계를 나타내는 것으로 각 객체와 그 하위 특성과의 연결을 나타낸다. POSBIOTM에서는 템플릿 관계로서 species_of, organ_of, size_of, certainty_of, strength_of, polarity_of 등을 정의하고 있다.

species_of, organ_of는 각기 참여자의 종과 기관에 대한 관계로서, 같은 단백질이라 하더라도 다른 종이나 기관에서는 다르게 발현되거나 그 기능이 다른 경우가 있음을 고려한 관계이다. 참여자 객체의 크기에 대한 관계인 size_of는 실험실에서 유용하게 쓰일 수 있는 정보이다. 이러한 관계들 외에 확신도, 세기, 긍정/부정 등에 대한 관계들은 같은 경로나 문서 상에서 추출된 상호작용 사이의 모순을 해결하는 데 매우 중요하다.

이벤트(Event)

이벤트 추출은 하나의 프레임에 대해 적절한 슬롯들을 각 객체들과 상호작용으로 채우는 것을 말한다. 이벤트는 경로 상의 한 반응(reaction)으로서, 참여자로는 상호작용(interaction), 효과자(effecter), 반응자(reactant)가 있으며 경우에 따라 산출물(product)이 포함될 수 있다. 이 때 효과자는 상호작용을 시작하거나 자극하는 하나의 템플릿 요소나 이벤트이고, 반응자는 효과자에 반응하는 템플릿 요소나 이벤트, 산출물은 상호작용에 의해 생성되는 물질을 나타낸다.

특히 산출물 슬롯의 경우, 실제 모든 생물학적 상호작용은 특정한 결과물을 만들게

되지만 'phosphorylate'와 같이 상호작용명으로부터 결과물(반응자에 인산염을 붙인 것)을 바로 예상할 수 있는 경우에는 산출물을 생략하도록 한다. 그러나 'dissociate'와 같은 상호작용의 경우 명칭만으로는 반응자가 2개 혹은 그 이상의 결과물로 나눠진다는 것만을 알 수 있고 실제 결과물을 바로 예상할 수 없으므로, 산출물 슬롯을 해당 결과물 개수의 템플릿 요소들로 채워 결과물을 명시한다.

GPCR 관련 경로 추출 과정

POSBIOTM의 경로 추출은 다음과 같은 과정을 통해 이루어진다.

먼저 경로 추출 대상 문서를 수집하기 위해 GPCR 관련 단어를 키워드로 MEDLINE에 질의를 던져 검색된 논문의 요약문(abstract)에 연결된 전문(full article)을 수집한다. 이때 검색된 전문들이 PDF 등 plain text 형식이 아닐 경우가 많으므로 문서 형식을 plain text로 전환하는 전처리(pre-processing) 과정을 거친다.

수집 후 전처리가 완료된 각 대상 문서는 먼저 문장 단위로 분할되고, 각 문장은 구문분석 과정으로 품사 태깅(tagging), 명사구 분리(noun phrase chunking), 개체명 태깅 과정을 거친다.

구문분석 과정에서 얻어진 여러 가지 정보들로 태깅된 각 문장들은 어휘적, 구문적, 개념적 수준에서의 지식 부호화 형식인 LSP(lexico-semantic patterns)[16] 표현으로 변환되고, 이렇게 변환된 문장을 대상으로 학습 방법에 의해 자동으로 학습된 경로 추출 규칙을 이용하여 관계와 이벤트를 추출한다.

마지막으로 추출된 경로 정보를 SBML (Systems Biology Markup Language) 파일 형식으로 변환한 후 시각화 모듈에 넘겨 경로를 시각화한다.

POSBIOTM 시스템 구조

POSBIOTM은 크게 개체명(named entity, NE) 인식, 관계/이벤트 추출 규칙 학습, 관계/이벤트 추출, 시각화 모듈의 네 부분으로 구성된다(그림2).

개체명 인식

생물학 분야의 개체명 인식의 경우 문서마다 같은 개체명이라도 다양한 철자 형태로 기술되는 경우가 많아 개체명 인식이 어려워지는 문제점이 있다. 이러한 철자이형태(spelling-variant form) 문제를 해결하기 위해 HMM(hidden-Markov model) 기반의 방법론에 edit-distance를 결합한 모델을 개체명 인식에 이용하였다[17].

관계(relation)/이벤트(event) 추출 규칙 학습

기존의 관계 추출 시스템들은 해당 대상인 도메인에 맞추어 전문가가 만들어낸 추출 규칙에 의존하는 경우가 많았다. 그러나 이러한 규칙 생성 방법은 많은 시간과 노력을 요구할 뿐만 아니라 이식성도 부족하여 도메인이 변경될 경우 대부분의 규칙을 새로 작성해야 하는 등의 문제가 있다.

POSBIOTM은 교사 학습 방법을 통해 자동으로 규칙을 추출, 생성하여 이러한 문제점을 극복하였다. 학습 알고리즘으로는 LSP와 유사한 문맥 기반(context-based) 정규 형식의 규칙을 학습, 생성하는 알고리즘인 WHISK 알고리즘[19]을 채택하였다.

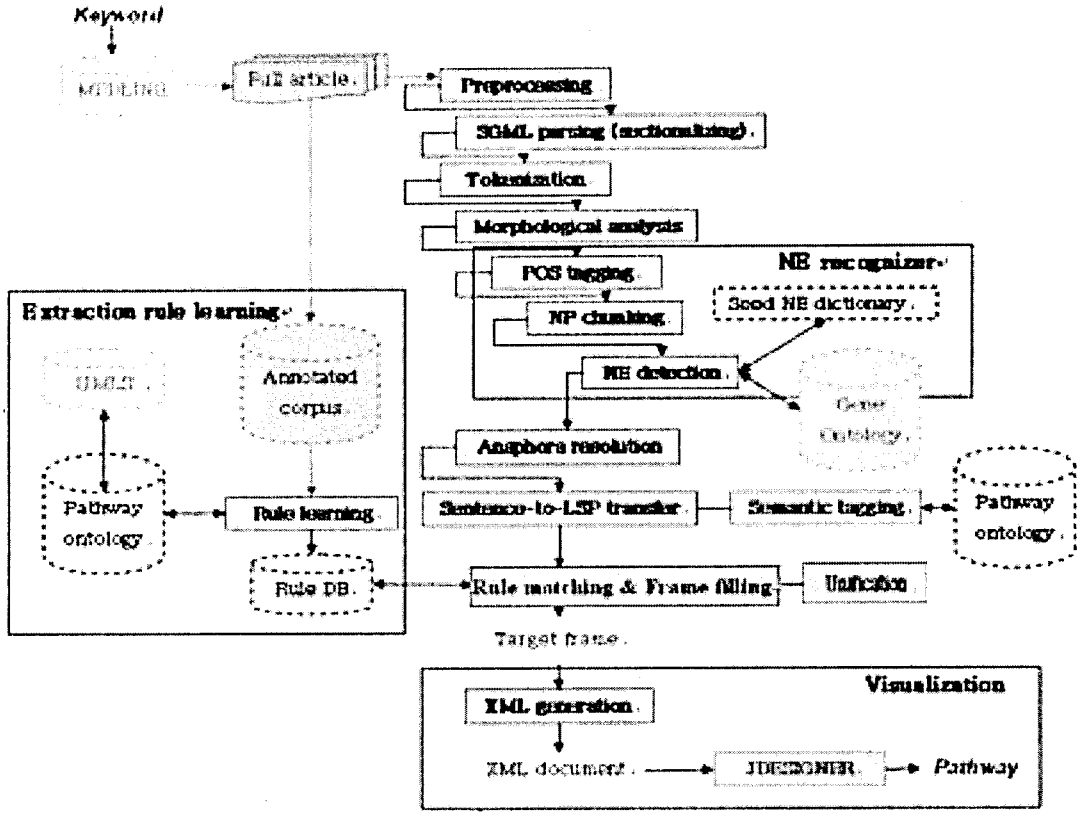

그림 2 POSBIOTM 전체 구조

관계(relation)/이벤트(event) 추출

관계/이벤트의 추출에는 앞서 학습된 추출 규칙을 이용한다. 규칙은 추출 패턴과 추출 결과물의 두 부분으로 나뉘어 작성되어 있다.

Pattern:*(P)*[be]*(BI)*for*(CP)*
Output: Biological Interaction:
{effector $1} {interaction $2} {reactant $3}

그림 2 GPCR 경로추출을 위한 규칙 예

그림3은 GPCR 경로를 추출하기 위해 적용되는 규칙의 한 예이다. 규칙 패턴에서 *는 다음 필드가 패턴에 등장할 때까지 지나가라는 의미이고 '안의 token은 문서상의 실제 문자에 해당한다. 또한 []안의 token은 품사 태깅 정보를 나타내며, ()안은 추출 대상 개체들이다.

대상 문서에 대해 각 규칙 패턴을 만족하는 내용을 만나면 해당 패턴에 대한 규칙의 결과물을 해당 규칙의 추출 결과물로 내놓는다.

이와 같이 추출된 반응들은 중복되는 내용을 삭제하고 서로 대치되는 내용을 잘 처리하여 통합하는 과정을 거친다.

JDesigner를 통한 시각화

POSBIOTM은 논문 전문으로부터 추출된 경로 및 객체에 대한 정보들을 생물학자들이 쉽게 확인하고 수정할 수 있도록 하기 위한 시각화 모듈을 포함한다.

생물학자들이 직접 그려서 이용하는 형태가

대부분이던 기존의 경로 시각화 모듈에서 한 발 나아가, POSBIOTM은 문서로부터 추출된 정보를 시각화용 markup language로 변환하여 시각화 모듈에 넘겨 주므로 추출된 경로 정보들을 바로 시각적으로 확인할 수 있다.

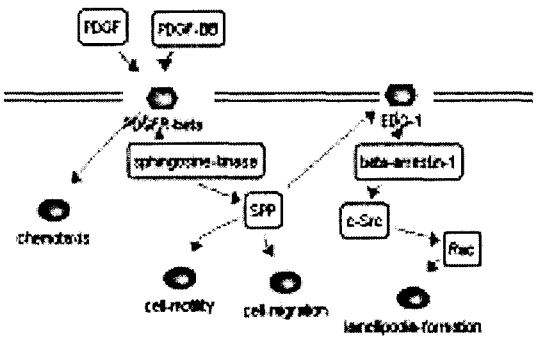

그림 3 JDesigner를 통한 시각화

또한 POSBIOTM에서 시각화 모듈로 채택한 JDesigner의 경우 SBW(System Biology Workbench)에 등록되어 있으므로 SBW에 등록된 다양한 시뮬레이션 관련 프로그램들과 연동할 수 있다는 장점도 가진다.

더불어 보다 나은 시각화를 위해 기본적인 JDesigner에 다음과 같은 수정을 가하였다. 객체와 관계의 종류 및 그에 대한 세부 정보는 경로를 표현하는 데 있어 생물학적으로 중요하다. 그러나 JDesigner에는 이러한 특성을 자세히 표시하는 기능이 없기 때문에 이러한 정보들을 표현할 수 있도록 하였다. 또한 경로의 이해에 매우 중요한 세포 내 위치 정보를 제공하기 위해 세포막 등의 경계 표시를 강화하여 보다 쉽게 경로를 확인할 수 있도록 하였다.

Experiments and Results

POSBIOTM은 현재 개발중인 새로운 시스템이므로, 여기에선 전체 구조의 기본적인 모듈들에 대한 예비 단계로서의 실험적 결과를 제시한다

개체명 인식

경로 추출을 위한 대상 템플릿 요소들은 다양한 범주들을 포함해야 하지만, 개체명 인식에 대한 접근 방식이 기계 학습 기반이므로 주석된 말뭉치만 있으면 다른 범주로 쉽게 확장할 수 있다. 본 논문에서는 이러한 점을 고려하여 개체명 인식 알고리즘의 성능을 평가하기 위하여 인식 대상 개체명의 80% 가량을 차지하는 단백질명에 대해 GENIA 말뭉치[15] 대상으로 개체명 인식 실험을 우선적으로 수행하였다.

실험 결과, 기본적인 HMM 기반 인식 모델의 경우에는 F-score로 55.69% (정확도: 72.25%, 재현률: 45.35%)의 성능을 보였으며 HMM에 edit-distance를 도입한 인식 모델의 경우에는 F-score로 58.47% (정확도: 67.43%, 재현률: 51.62%)의 성능을 보였다. 이를 통해 생물학 분야의 개체명, 특히 단백질명의 인식에 있어서 POSBIOTM이 채택한 edit-distance 기반 HMM 방법론이 기존 HMM 모델 기반 방법에 비해 나은 성능을 보임을 확인할 수 있었다.

관계/이벤트 추출

POSBIOTM의 관계/이벤트 추출 성능을 평가하기 위한 pilot study로서 기본적인 이벤트 추출에 초점을 맞추어 실험을 수행하였다. 2,305 단어 분량의 논문 전문을 대상으로 하였으며, 이벤트 추출만의 성능을 보다 정확히 평가하기 위하여 대상 문서에 대해 수동으로 개체명을 태깅한 뒤 실험을 수행하였다.

대상 문서로부터 WHISK 알고리즘에 의해 학습될 수 있는 32개의 이벤트에 대한 규칙을 생성하고 이를 추출에 이용한 결과 POSBIOTM에 의해 추출된 이벤트는 총 39개였으며 평가 결과 재현률 65.6%, 정확도 53.8%의 성능을 보였다. 추출된 이벤트 중 많은 부분이 중복된 것을 감안하여 실험에서는 단지 유일한 이벤트만 카운트에 고려하였다.

Discussion

보다 나은 경로 추출을 위해서는 구분리기(phrase chunker), 품사 태거, 개체명 인식기 등의 자연어 분석 프로그램들의 성능을 향상시키는 것이 필요하다. 또한 학습 말뭉치 양을 증가시켜 규칙 학습의 효율을 높이는 작업도 필요하다. 현재 자동으로 양질의 학습 말뭉치 양을 증가시켜 시스템의 성능을 올리는 연구가 진행되고 있다.[18]

Conclusion

본 논문에서는 생물학적 추출 타겟을 기반으로 문서로부터 자동으로 GPCR 경로 정보를 추출하고 시각화하는 생물학적 텍스트 마이닝 시스템을 제안한다. POSBIOTM은 기존에 전산 기술 위주였던 경로 추출 시스템과 비교하여 좀더 실용적이고 생물학적으로 의미있는 시스템이 될 것이다. 하지만, 아직 POSBIOTM은 개발 중이기 때문에 앞으로 좀더 면밀한 분석과 실험이 요구된다.

Acknowledgements

본 연구의 생물학적 조언을 아끼지 않은 포항공대 생명과학과 류성호 교수와 김윤동 연구원에게 감사의 말씀을 전합니다.

References

1. Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A., Automatic extraction of biological information from scientific text: protein-protein interactions. *Intelligent Systems for Molecular Biology* 60-67 (1999)
2. Thomas, J., Milward, D., Ousounis, C., Pulman, S., Carroll, M., Automatic extraction of protein interactions from scientific abstracts. *Pac. Symp. Biocomput.* **5**, 541-52 (2000)
3. Sekimizu, T., Park, H.S, Tsujii, J., Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome informatics Workshop* 62-71 (1998)
4. Rindflesch, T.C., Rayan, J.V., Hunter, L., Extracting molecular binding relationships from biomedical text. *Applied Natural Language processing and the North American Chapter of the Association for Computational Linguistics* 188-95 (2000)
5. Rindflesch, T.C., Tanabe, L., Weinstein, J.N., Hunter, L., EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac. Symp. Biocomput.* **5**, 517-28 (2000)
6. Yakushiji A., Tateisi, Y., Miyao Y., Event extraction from biomedical papers using a full parser. *Pac. Symp. Biocomput.* (2001)
7. Park J.C., Kim H.S., Kim J.J., Bidirectional incremental parsing for automatic pathway

identification with combinatory categorical grammar. *Pac. Symp. Biocomput.* (2001)

8. Pustejovsky, J., Castaño, J., Zhang J., Kotecki, M., Cochran B., Robust relational parsing over Biomedical Literature: Extracting inhibit relations. *Pac. Symp. Biocomput.* 362-73 (2002)
9. Jenssen, T.-K., Komorowski, J., Laegreid, A., Hovig, E. Pubgen, Discovering and visualizing gene-gene relations. *Currents in Computational Molecular Biology* 48-9 (2000)
10. Stapley, B. J., Benoit, G., Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. *Pac. Symp. Biocomput.* **5**, 529-40 (2000)
11. Ng S.K., Wong M., Toward routine automatic pathway discovery from on-line scientific text abstracts. *Genome Inform Ser Workshop Genome Inform.* **10**, 104-12 (1999)
12. Leroy G., Chen H., Filling preposition-based templates to capture information from medical abstracts. *Pac. Symp. Biocomput. 350-61 (2002)*
13. Friedman C., Kra P., Yu H., Krauthammer M., Rzhetsky A., GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics.* **17,** Suppl 1:S74-82 (2001)
14. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle J, Kitano H., The ERATO Systems Biology Workbench: enabling interaction and exchange between software tools for computational biology. *Pac Symp Biocomput.* 450-61 (2002)
15. Tateishi, Y., Ohta, T, Collier, N., Nobata, C., and Tsujii, J., Building annotated corpus in the molecular-biology domain. *Proc. COLING 2000 Workshop on Semantic Annotation and Intelligent Content,* 28–34 (2000)
16. Lee G., Seo J., Lee S., Jung H., Cho B., Lee C., Kwak B., Cha J., Kim D., Ahn J., Kim H., Kim K., SiteQ: Engineering high performance QA system using lexico-semantic pattern matching and shallow NLP, *Proc. of the 10th text retrieval conference* (2001)
17. Yi E., Lee G.G., Park S., HMM-based protein name recognition with edit-distance using automatically annotated corpus. *Proc. of the workshop on BioLINK text data mining SIG: Biology literature, information and knowledge. ISMB03* (2003).
18. An J., Lee S., Lee G.G., Automatic acquisition of named entity tagged corpus from World Wide Web. *Proc. of interactive posters/demonstrations. ACL03*, 165 (2003)
19. Soderland S., Learning information extraction rules for semi-structured and free text. *Machine Learning*, **34**, 233-72 (1999)