

# 대사경로 재구축을 위한 텍스트 마이닝 기법<sup>†</sup>

## (Text-mining Techniques for Metabolic Pathway Reconstruction)

권혁렬\*, 나종화\*\*, 유재수\*\*\*, 조완섭\*\*\*\*

(Hyuk-Ryul Kwon, Jonghwa Na, Jae Soo Yoo, Wan-Sup Cho)

**요약** 대사 공학의 발전과 함께 생물체에 유전자 재조합기술과 관련 분자생물학 및 화학공학적 기술을 이용하여 새로운 대사회로를 도입하거나 기존의 대사회로를 제거·증폭·변경시켜 세포나 균주의 대사 특성을 조절하는(directed modification) 일련의 기술들이 가능해지고 있다. 하지만 이러한 대사회로를 조절하기 위해서는 많은 선행 연구에 대한 고찰이 필요하며, 일선 연구자들은 방대한 선행 자료를 검색하고 일일이 읽으면서 자신에게 필요한 정보를 수집하고 있다. 따라서 효율적으로 대사 모델을 구축하고, 방대한 대사관련 연구논문으로부터 대사흐름 관련 정보를 자동으로 추출하는 기술의 개발이 중요한 이슈로 부각되고 있다. 본 논문에서는 대사경로 재구축을 위한 서열과 패턴 기반의 텍스트 마이닝 기법을 제안한다. 제안된 기법은 웹 로봇을 이용하여 최신의 논문을 반자동적으로 수집하고 이를 이용하여 최신의 논문을 로컬 데이터베이스로 구축한다. 또한 생물학 개체명의 인식율을 높이기 위해 유전자 온토로지를 이용하며, NCBI에서 제공하는 Tokenizer 라이브러리를 이용하여 개체명의 파괴 없이 인식할 수 있게 하였다. 본 연구에서 제안한 텍스트 마이닝 기법에서는 패턴을 이용하여 논문으로부터 대사경로 지식을 추출하게 되므로 올바른 패턴을 확보하는 것이 중요한 문제이다. 논문에서는 패턴의 수집을 위하여 대표적인 대사경로 전문 사이트인 일본의 KEGG 경로 데이터베이스에서 추출한 Glycosphingolipid 종에 대한 20,000 여건의 논문에서 66개의 패턴을 추출하였다. 제안된 기법의 유효성을 입증하기 위하여 Glycosphingolipid 종의 GLS 대사경로 19개 개체명을 이용하여 시스템을 평가하였다. 그 결과 논문 125,907 건에 대하여 정확도 96.3%, 재현율 95.1%, 처리시간 15초의 성능을 보였다. 본 논문에서 제안된 시스템은 대사 경로 재구축에 유용하게 활용될 수 있을 것으로 기대된다.

**핵심주제어** : 대사경로, 텍스트 마이닝, 대사공학, 대사경로 재구축

**Abstract** Metabolic pathway is a series of chemical reactions occurring within a cell and can be used for drug development and understanding of life phenomenon. Many biologists are trying to extract metabolic pathway information from huge literatures for their metabolic-circuit regulation study. We propose a text-mining technique based on the keyword and pattern. Proposed technique utilizes a web robot to collect huge abstract papers and stores them into a local database. We use gene ontology to increase compound recognition rate and NCBI Tokenizer library to recognize useful information without compound destruction. Furthermore, we obtain useful sentence patterns representing metabolic pathway from papers and KEGG database. We have extracted 66 patterns in

\* 이 논문은 2007년 교육인적자원부의 재원으로 한국학술진  
홍재단의 지원을 받아 수행된 연구임 (지방연구중심대학  
육성사업/충북BIT연구중심대학육성사업단)

\*\* 충북대학교 바이오정보기술 석사과정

\*\*\* 충북대학교 정보통계학과 교수

\*\*\*\* 충북대학교 전기전자컴퓨터공학부 교수

교신저자 : 충북대학교 경영정보학과/BK21 u-Biz팀 부  
교수 (wscho@cbnu.ac.kr)

20,000 documents for Glycosphingolipid species from KEGG, a representative metabolic database. We verify our system for nineteen compounds in Glycosphingolipid species. The result shows that the recall is 95.1%, the precision 96.3%, and the processing time 15 seconds. Proposed text mining system is expected to be used for metabolic pathway reconstruction.

**Key Words** : Metabolic Pathway, Text Mining, Metabolic Engineering, Metabolic Pathway Reconstruction

## 1. 서 론

대사(metabolism)란 생물체 내에서 일어나는 물질의 분해나 합성과 같은 모든 물질적 변화를 의미 한다. 생물체 내에서 에너지 생산을 위해서는 여러 가지 대사 과정을 거치게 되며, 이 과정에서 다양한 대사산물들이 발생하게 된다[1].

대사 공학이란 이러한 생물체에 유전자 재조합 기술과 관련 분자생물학 및 화학공학적 기술을 이용하여 새로운 대사회로를 도입하거나 기존의 대사회로를 제거·증폭·변경시켜 세포나 균주의 대사 특성을 우리가 원하는 방향으로 바꾸는(directed modification) 일련의 기술을 말한다[2]. 이러한 방식으로 미생물로 하여금 토양의 유기 공해물질 등 환경 오염물질을 분해하도록 하거나, 신약물질 등 유용한 화학 물질을 미생물 세포 내에서 생산하도록 하는 것이 가능하게 된다[1]. 이런 방법에 의해 원하는 물질들의 과량 및 대량생산이 일부 가능해진 반면, 이를 유전자의 도입으로 인해 균주의 성장능력저하, 부산물의 과량생산 등의 예상치 못한 부작용도 발견되고 있다[2]. 이에 따라 대사회로의 인위적인 변경에 따른 결과를 미리 예측하고 해당 미생물의 대사 체계를 디자인하여 조작하기 위해서, 특정 대사 과정 일부가 아닌 전체에 대한 이해와 분석과정이 필수적이다[1]. 또한, 최근 일부 미생물의 모든 유전자 염기서열이 밝혀지고 앞으로 더욱 많은 데이터가 쏟아져 나올 것으로 예상되고 있다. 이는 앞으로 거의 모든 대사회로의 조작이 가능해지는 것을 의미한다. 그러나 단순한 대장균의 경우만 해도 약 2,000여 개의 반응이 조절을 받으며 복잡하게 일어나고 있는데 어디를 조작해야 하는가를 알아내기란 쉬운 일이 아니다[2].

이처럼 대사회로를 조작하기 위해서는 많은 양의 정보가 요구된다. 이를 위해 일선 연구자들은

그러한 정보를 수집하기 위하여 많은 대사 작용 관련 논문을 읽고 정보를 수집, 이러한 정보를 이용하여 대사회로를 조작하는 방식으로 연구를 진행하고 있다. 이러한 정보 수집 과정에서 많은 시간과 인력이 소비되고 있다.

따라서 대사 모델 구축에 필요한 시간과 인력의 낭비를 줄여, 효율적으로 대사 모델을 구축하고 이를 논문에 적용하기 위해서는, 대사관련 연구 논문으로부터 대사호름 관련 정보의 추출을 자동화하기 위한 시스템 개발이 필요하게 되었다.

국내에서 개발된 대사호름 정보를 추출하기 위한 시스템으로는 대사 작용 정보 자동 추출 시스템[1]이 존재한다. 이 시스템은 생물학 문헌으로부터 대사호름 정보를 추출하기 위해 생물학 문헌의 문법 정보를 제공해주는 EngCG[3]라는 라이브러리와 컴파운드 이름에서 접미사의 특징을 정의해 생물학 문헌 속에 존재하는 컴파운드를 인식하고 대사호름 정보를 추출하고 있다. 이 시스템은 컴파운드 이름의 변형이나 신조어를 인식하는데 탁월한 효과를 가지고 있으나, 생물학 문헌 속에 컴파운드가 아닌 것을 컴파운드로 인식하는 오류 가지고 있다. 이로 인해 낮은 정확도를 가지고 있다. 본 논문은 사용자가 원하는 정보를 효율적이고 정확한 정보를 제공하기 위해 유전자 온톨로지와, 생물학 논문으로부터 수동으로 추출한 패턴, NCBI에서 제공하는 Tokenizer 등을 이용한, 대사경로 재구축을 위한 텍스트 마이닝 기법을 제안하다.

본 논문에서는 대사 공학 관련 논문들로부터 추출한 패턴(66개)과 사용자가 입력한 개체명을 이용하여 대사 작용에 관여하는 물질에 대한 설명이 들어있는 Abstract 논문을 자동 추출하는 텍스트 마이닝 시스템을 설계하고 구현한다. 또한 자동 추출된 Abstract 논문이 실제 존재하는지 문헌상에서 직접 확인하고, 잘못 인식된 내용을 수정 할 수

있도록 한다. 그리고 본 논문에서 제시하는 시스템을 평가하기 위해서 KEGG 경로 데이터베이스에 대사 경로가 구축된 Glycosphingolipid 종에 대한 논문 82건을 수집하여 시스템 평가를 실시하였다. 그 결과 논문 125,907 건에 대하여 정확도 96.3%, 재현율 95.1%, 처리시간 15초의 성능을 보였다. 본 논문에서 제안된 시스템은 대사 경로 재구축에 유용하게 활용될 수 있을 것으로 기대된다.

본 논문의 구성은 다음과 같다. 2 장에서는 바이오 텍스트 마이닝 관련 기법들과 시스템들에 대한 사례들을 소개한다. 3 장에서는 본 논문에서 제안하는 대사경로 재구축을 위해 설계된 시스템의 구조와 검색 알고리즘에 대해 소개한다. 4 장에서는 Glycosphingolipid (GLS)의 대사 체계에 실험 및 평가 결과를 소개하고, 마지막으로 5 장에서 본 논문의 결과가 가지는 의의와 향후 연구과제 제시하며 결론을 맺는다.

## 2. 관련 논문

본 장에서는 생물학 분야 논문 검색을 위한 기존에 개발되어진 텍스트 마이닝 기법에 대해 알아본다.

### 2.1 텍스트 마이닝 기법

대량의 생명공학 관련 데이터가 쏟아져 나오면서 데이터 마이닝의 중요성이 증대되고 있으며, 여러 분야에서 텍스트 마이닝을 개발하고 있다. 기존의 텍스트 마이닝 시스템들은 서로 다른 방법론과 접근방식으로 지식발견의 문제를 해결해 나가고 있다.

MedStract[4]의 경우는 바이오 텍스트에 특징적으로 나타나는 단어들을 고려해 UMLS[3] 시소러스에 수반된 사전을 사용하여 개체명과 품사를 인식하게 하고 있다[5, 6, 7]. 그리고 독립된 몇 가지 오토마타를 단계적으로 적용해 명사구와 동사구 등을 인식하고, 정의한 패턴에 따라 상호작용 정보를 추출한다. 그리고 웹 기반의 사용자 인터페이스를 제공하며, 개체명 검색 결과를 테이블과 그래프 형태로 제공한다[8]. 이처럼 시소러스에 수반된 사

전을 이용하여 개체명을 인식할 경우 개체명의 변이에 대한 인식률이 저하된다는 단점을 수반하고 있다.

GENIES[8]의 경우에는 단백질 또는 유전자 명칭을 확인하는 Term tagger와 문장, 단어, 구를 결정하는 Preprocessor, 그리고 제약 규칙과 의미적 패턴으로 되어 있는 문법을 사용해 적절한 상호작용관계를 확인하는 Parser, 구문분석의 오류를 여러 가지 휴리스틱을 사용해 처리하는 Error recovery 모듈로 구성되어 있다. 이 시스템은 추출된 개체간의 상호작용 정보를 이용해 pathway를 구성하게 된다[8].

BIOBIBLIOMETRICS[9]는 유전자 이름을 이용해 생물학 관련 문헌 DB에서 정보를 검색한 후 가시화한다[10]. BIOBIBLIOMETRICS 시스템은 문헌 DB에서 두 유전자가 공존하는 정도로부터 유사도를 구하고, 특정 임계 값 이상이면 관련이 있다고 판단한다. 실제로 생물학 관련 연구자가 특정 유전자를 검색하면 그 유전자와 관련된 다른 유전자들이 검색되고 이것을 가시화해 보여주는데, 이 결과로부터 유전자와 유전자 사이의 관계를 연구하는데 도움을 준다. 이러한 방법은 사용자가 쉽게 사용할 수 있는다는 장점이 있지만, 사용자가 임의로 임계값을 높였을 경우 사용자가 원하는 결과 값을 얻지 못할 수 있다. 또한 임계값을 너무 낮추었을 경우 광범위한 결과 값을 사용자에게 주어 정교한 검색을 하기에는 어려움이 따른다.

[11]은 코퍼스를 이용하여 개체명을 인식하고, 이를 학습 알고리즘에 적용하여 인식하는 방법이다. 코퍼스란 분석의 대상이 되는 문헌들을 수집해 놓은 집합을 뜻한다[12]. [11]은 생물학 분야 개체명 인식을 위한 많은 논문들에서 학습 데이터로 사용되고 있다. 이러한 코퍼스를 이용한 학습 기반법의 장점은, 시스템이 개체명을 인식하기 어려운 개체명도 인식이 가능하다는 점이다. 하지만 목적에 맞는 학습에 필요한 대용량의 코퍼스를 구축하기 위해서는 많은 시간과 비용이 소모된다는 단점을 가진다. 또한 학습 기반 방식에서 개체명 인식의 정확도는 학습에 사용된 코퍼스의 영향을 받게 되므로, 코퍼스 자체의 정확도 및 신뢰도 역시 중요하다고 할 수 있다[1].

### 3. 대사경로 재구축을 위한 텍스트 마이닝 시스템

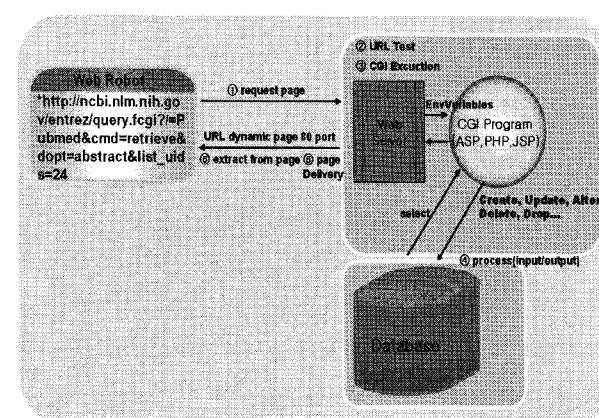
본 논문에서 제안하는 텍스트 마이닝 시스템은 (그림 1)과 같이 웹 로봇, 파서, 데이터베이스, 검색 시스템, OLS[13]등 5단계의 컴포넌트로 나누어 볼 수 있다. 본 절에서는 각각의 기능에 대하여 살펴보도록 한다.



(그림 1) 텍스트 마이닝 시스템 5단계 컴포넌트

#### 3.1 웹 로봇

웹 로봇은 자동적으로 웹의 하이퍼텍스트 구조를 따라 다니며 문서를 추출하고, 재귀적으로 문서에서 참조되는 다른 문서들을 추출하는 방식으로 동작하는 프로그램이다[4]. 이 프로그램을 통해 바이오 관련 1700만 건 해당되는 논문들을 수집하게 된다. 또한, 이를 이용해 최신의 논문을 수집할 수 있어 사용자에게 최신의 논문을 제공할 수 있

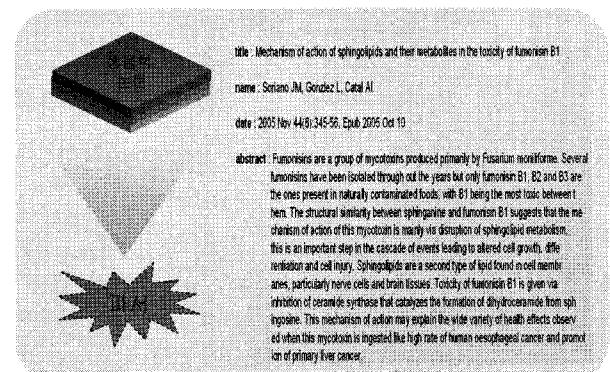


(그림 2) 웹 로봇

게 된다. PubMed[13]에서 부여한 PMID는 각 문서의 고유 아이디로 순차적으로 부여된 것을 확인하였다. 따라서 PMID를 이용한 중복 추출 제거와 최신성을 유지하며 생물학 논문을 수집 할 수 있었다. (그림 2)는 웹 로봇의 동작 모습을 보여주고 있다.

#### 3.2 파서

웹 로봇을 이용하여 수집된 생물학 논문들은 처음에 정제되지 않아 웹 로봇에 의해 수집된 그대로 사용되기에는 무리가 있다. 이를 해결하기 위해 본 논문에서는 파서를 이용하여 수집된 생물학 논문을 정제하여 사용하고 있다. (그림 3)는 파서에 의해서 정제되는 모습을 보여주고 있다.

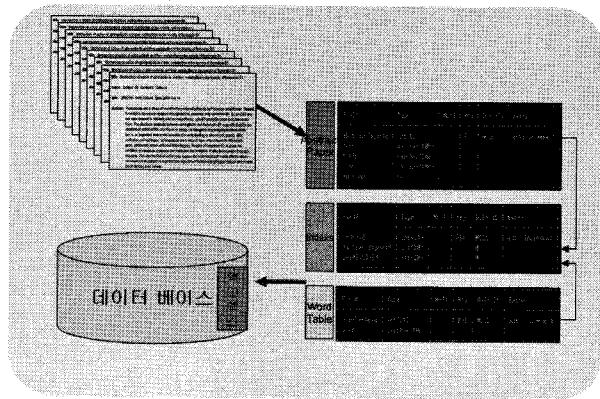


(그림 3) 파서

#### 3.3 데이터베이스

대사경로 재구축을 하기 위해서는 대량의 논문에 접근하여 사용자가 원하는 논문을 찾는 과정이 필요하다. 현재 Pubmed에 구축된 Abstract 논문은 1700만건에 달하고 있다. 이를 웹으로 접근하였을 경우에는 많은 시간적 비용이 들어간다. 이를 해결하기위해 웹 로봇을 이용하여 정기/비정기적인 Pubmed에 접근하여 최신 문헌을 로컬 데이터베이스에 업데이트 하고 있다. 또한 데이터베이스 구축시 역 인덱스를 이용하여 보다 빠른 검색을 할 수 있게 하였다. 현재 로컬 데이터베이스에는 실험 데이터로 125,907건의 Abstract 논문을 저장하고 있으며 지속적으로 늘려 나갈 것이다. (그림

4)는 데이터베이스 구축 과정을 보여 주고 있다.



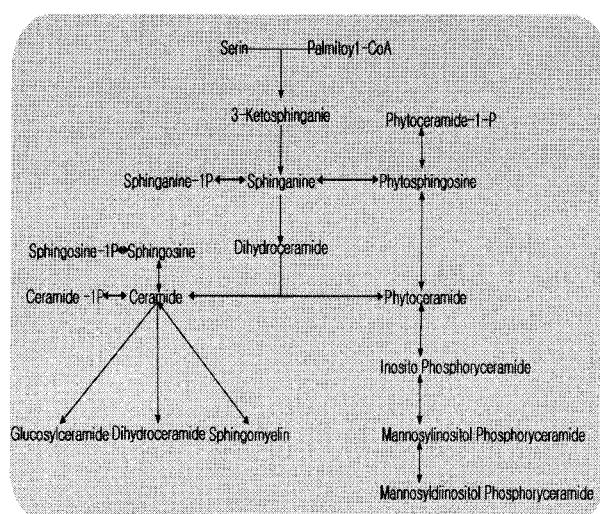
(그림 4) 데이터베이스

### 3.4 검색 시스템 및 검색기법

검색 시스템은 웹 로봇과 파서를 통해 얻어진 데이터를 가지고 사용자가 입력한 개체명과 패턴이 존재하는 Abstract 논문 검색 기능을 수행하며, 본 논문에서 제시하는 시스템의 중심적인 엔진 역할을 담당하고 있다. 이 절에서는 검색시스템에 대해 알아보기로 한다.

#### 3.4.1 패턴 추출

본 논문에서는 KEGG 경로 데이터베이스에 생합성 경로가 이미 밝혀진 Glycosphingolipid(GLS)의 19개의 컴파운드를 이용하여 Pubmed에서 검색한 결과 약 20,000 여건이 검색되었다. 약 20,000여 건에 대해서 생명공학 전문가가 수작업으로 검사한 결과 개체명 한 문장 안에 개체명 2개 이상과 패턴이 존재하는 논문은 82건과 패턴 66개를 찾을 수 있었다. 여기서 사용된 대사 흐름 논문 GLS는 eukaryotic cell에서 세포의 성장, 증식 및 사멸에 관여하고, 피부의 보습 유지와 항암효과, 세포의 apoptosis 등의 작용을 담당하고 있는 개체명이다[17]. GLS는 (그림 5)과 같은 구조로 생합성 경로를 가지고 있다.

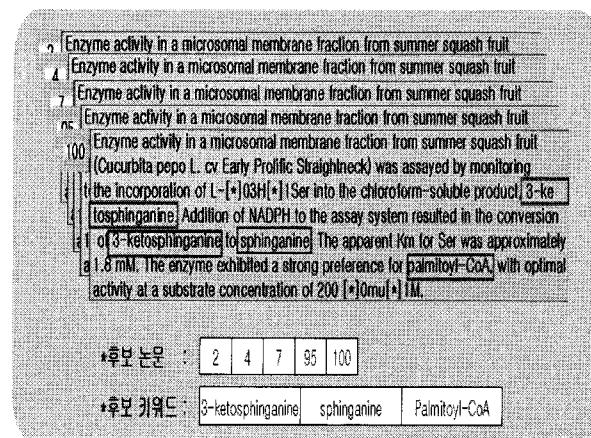


(그림 5) Glycosphingolipid(GLS) 생합성 경로

#### 3.4.2 개체명 검색 기법

개체명 검색 기법은 사용자가 입력한 2개 이상의 개체명을 이용하여 하나의 논문에 2개 이상 개체명이 존재할 경우 이 Abstract 논문을 후보 논문으로 등록한다. 또한 후보 논문에 출현한 개체명에 대해서는 후보 개체명으로 등록함으로서 1차적인 필터링을 통해 보다 빠른 검색을 제공하고 있다.

이때 중복으로 나타난 개체명과 Abstract 논문은 중복되어 저장되지 않는다. 대사경로에서 하나 이상의 개체명은 다른 효소 하나 이상과 반응하여 다른 개체명을 생산해내기 때문이다. (그림 6는 GLS에 포함된 개체명인 3-ketosphing-anine,



(그림 6) 개체명 검색 기법

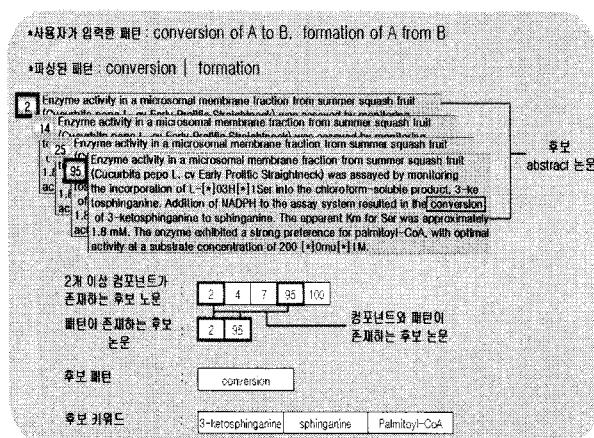
sphinganine, palmitoyl-CoA 3개를 이용하여 하나의 Abstract 논문에 개체명 2개 이상이 존재하는 Abstract 논문을 후보 논문으로 그리고 후보논문에 출현한 개체명을 후보 개체명으로 등록한 모습이다.

### 3.4.3 패턴 검색 기법

패턴 검색 기법은 사용자가 입력한 패턴을 이용하여 개체명이 2개 이상 존재하고 사용자가 입력한 패턴이 1개 이상 존재하는 Abstract 논문을 후보 논문으로 등록한다. 이때 대상 Abstract 논문은 개체명 검색 기법에 의해 검색된 후보 논문만을 대상으로 삼는다. 또한 사용자가 다음과 같은 패턴 conversion of A to B, formation of A from B를 입력하였을 경우 빈번하게 사용되어지는 전치사나, 접속사등과 같은 of, to, from과 같은 단어들은 생략하고 패턴으로서 의미를 가지는 conversion, formation과 같은 단어들만 검색하여 불필요한 단어의 검색을 생략하였다. (그림 5)는 사용자가 입력한 패턴을 이용하여 개체명이 2개 이상 존재하는 Abstract 논문과 Abstract 논문에 출현한 개체명과 패턴을 후보 논문과 개체명 패턴으로 등록하는 모습이다.

### 3.4.4 개체명 패턴 조합

개체명 검색 기법과 패턴 검색 기법을 이용하여



(그림 7) 패턴 검색 기법

추출된 각각의 후보 논문 리스트에는 후보 논문에서 추출된 개체명과 패턴에 대한 정보를 함께 가지고 있다. 이는 개체명과 패턴 조합시 불필요한 조합방지를 막기 위해서이다.

(그림 7) 서와 같이 2번 후보 Abstract 논문이 추출되면 이에 따르는 후보 패턴과 후보 개체명이 같이 저장되게 된다. 이러한 정보를 이용하여 개체명 패턴 조합을 실시하게 된다. 만약 2번 후보 Abstract 논문에 후보 개체명으로 3-ketosphinganine, sphinganine가 등록되어져 있고 후보 패턴으로 conversion이 등록되어져 있다면, 개체명 패턴 조합을 실시하기에 앞서 사용자가 처음 입력한 패턴 정보를 이용하여 conversion of A to B라는 패턴 원형을 찾게 된다. 이때 원형 패턴에 존재하는 A | B | C라는 구분자를 이용하여 후보 논문에 존재하는 개체명을 대입시켜 다음과 같은 개체명 패턴 조합을 실시하게 된다.

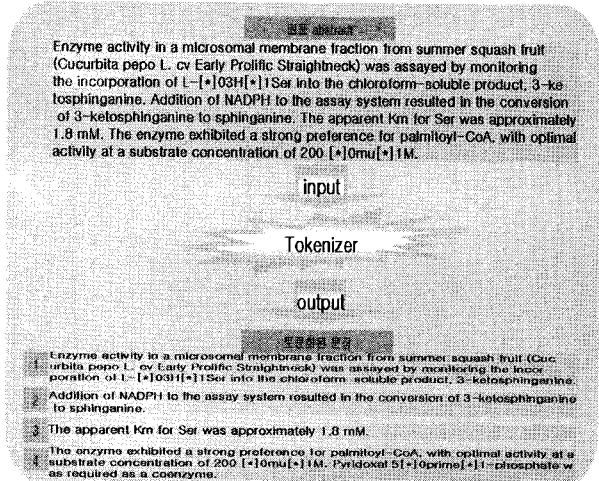
- ① conversion of 3-ketosphinganine to sphinganine,
- ② conversion of sphinganine to 3-ketosphinganine

이 때 “conversion of 3-ketosphinganine to 3-ketosphinganine”와 같은 개체명의 조합은 제외시키게 되는데 이는 하나 이상의 개체명은 다른 효소와 만나 다른 개체명을 생산해낼 수 있기 때문이다.

### 3.4.5 논문 검색 기법

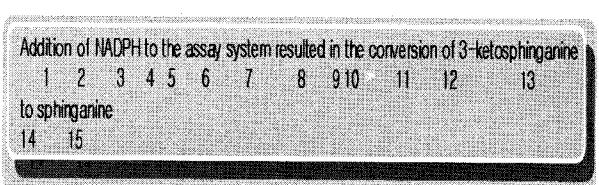
논문검색 기법은 위에서 언급한 개체명 검색 기법, 패턴검색 기법, 개체명 패턴 조합으로 얻어진 정보를 가지고 한 문장 안에 존재하는 Abstract 논문을 추출하는 방식이다. 이는 하나의 문장 안에 두 개 이상의 개체명과 패턴이 같이 존재한다면 그 논문은 대사경로를 나타내는 부분일 확률이 높기 때문이다. 이러한 방식을 적용하기 위해서 선행되어져야 할 일은 하나의 Abstract 논문을 문장 단위로 토큰화하는 일이 이루어져야 한다. 이때 중요한 것은 개체명 이름의 손실 없이 문장단위로 토큰화하는 것이 매우 중요하다. 이는 일선 연구자들이 일정한 규칙을 정해 그에 따르는 개체명을 사용하기 위한 시도가 이루어지고 있으나 제대로

지켜지지 않는 경우가 대부분이고 또한 개체명을 표시할 때 숫자, 기호 등이 포함되거나 여러 단어들이 이어져 하나의 이름을 구성하는 경우가 대부분이기 때문에 잘못된 토큰화는 개체명의 손실을 가져와 정확한 논문을 찾는데 걸림돌이 되고 있다. 이를 보안하기 위해 본 논문에서는 NCBI에서 제공하고 있는 Tokenizer[15]라는 라이브러리를 이용하여 개체명 손실 없이 Abstract 논문을 문장단위로 토큰화하여 이용하고 있다. Tokenizer를 이용하여 토큰화된 문장은 (그림 8)과 같다.



(그림 8) Tokenizer

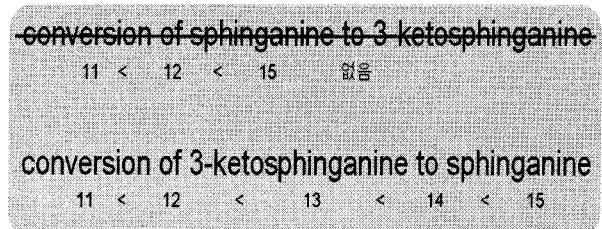
이와 같이 Abstract 논문이 Tokenizer를 통해 문장단위로 토큰화가 이루어지면 후보 Abstract 논문에서 존재하는 개체명과 패턴 조합을 가지고 문장단위 검사를 실시하게 된다. 이때 한 문장 속에 존재하는 단어는 고유한 위치정보를 가지며, 이 위치정보를 이용하여 개체명과 패턴 조합과 함께 비교하게 된다. (그림 8)는 Tokenizer를 통해 토큰화된 문장에 대한 위치 정보를 나타내고 있다.



(그림 9) 단어 위치 정보

본 논문에서 제시한 시스템은 이러한 단어위치

정보를 이용하여 개체명 패턴 조합을 비교하게 된다. 만약 이 후보 Abstract 논문에 다음과 같은 개체명 패턴 조합이 존재한다면 시스템은 다음과 같이 비교를 실시하게 된다. (그림 9)은 단어 위치 정보를 이용하여 개체명 패턴 조합 비교를 실시한 모습이다.



(그림 10) 개체명 패턴 조합 비교

(그림 10)과 같이 개체명 패턴 조합에 있는 첫 번째 단어를 이용하여 토큰화된 문장에서 단어위치를 찾게 된다. (그림 10)과 같이 처음 conversion이라는 단어의 위치정보를 찾고 이 위치에 대한 정보를 개체명 패턴 조합에 있는 conversion은 토큰화된 문장에서 찾은 conversion과 같은 위치 정보를 저장하게 된다. 이후 다음 단어인 of를 찾을 경우 (그림 9)에서와 같이 위치정보 2와 12라는 of 단어를 만나게 된다. 이때 시스템은 앞에서 찾은 conversion이라는 위치 정보를 이용하여 11이라는 위치 정보보다 더 큰 위치정보를 가지는 of를 찾게 된다. 한 문장 안에 사용자가 입력한 개체명 2개와 패턴이 존재한다 하더라도 위치 정보에 맞지 않으면 이는 대사 작용을 나타내는 문장이라 보지 않고 다음 문장과 다른 후보 논문을 검색하게 된다. 이러한 방식을 통해 사용자가 입력한 개체명과 패턴이 모두 존재하고 사용자가 의도한 문장을 도출해내게 된다. 이러한 결과물이 있을 경우 이 후보 Abstract 논문은 사용자에게 결과물로 보여지게 되고 또한 사용자의 편의를 위해 개체명과 패턴이 존재하는 문장에 대해서는 검은색과 개체명 패턴 조합 부분은 파란색으로 사용자에게 보여지게 된다.

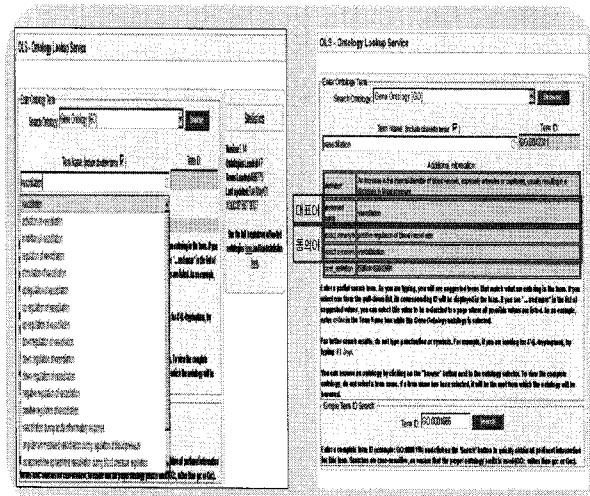
#### 3.4.6 OLS(Ontology Lookup Service)

OLS[13]는 Ontology Lookup Service의 약자로

사용자가 개체명을 입력하면 개체명에 대한 대표어와 동의어를 검색하여 사용자에게 서비스해주는 시스템이다. 개체명 표기에 대해 일정한 형식으로 개체명을 표기하도록 표준을 만들어 놓았으나, 제대로 지켜지지 않는 경우가 대부분이다. 이에 본 논문에서 개체명을 입력하면 개체명에 대한 대표어와 동의어를 찾아주는 OLS 웹 서비스를 이용하여 개체명 인식률을 높이는데 사용했다. (그림 11)는 사용자가 개체명을 입력했을 때 사용자에게 개체명에 대해 대표어와 동의어를 보여주는 모습이다.

#### 4. 시스템 평가

본 장에서는 제안된 시스템의 유효성을 입증하기 위하여 실제 인터넷상에 공개된 데이터베이스를 가지고 수행한 실험결과를 설명한다.



(그림 11) OLS(Ontology Lookup Service)

##### 4.1 시스템 실험조건 및 실험환경

본 장에서는 KEGG 경로 데이터베이스로부터 대사경로가 밝혀진 Glycosphingolipid(GLS)인 19개의 개체명을 이용하여 실험을 하였다. GLS는 eukaryotic cell에서 세포의 성장, 증식 및 사멸에 관여하고, 피부의 보습 유지와 항암효과, 세포의 apoptosis 등의 작용을 담당하고 있는 개체명이다 [17]. 이러한 GLS 19종에 대하여 개체명이 2개 이

상 존재하고 패턴이 존재하는 논문을 pubmed에서 검색한 결과 약 20,000 여건의 Abstract 논문이 검사되었고 이중, 개체명과 패턴이 모두 존재하는 논문은 82건이고 이중 발견한 패턴은 66건으로 이를 바탕으로 실험을 실시하였다. <표 1>은 본 논문에서 실험할 때 사용한 실험 조건이다.

<표 1> 실험 조건

개체명	패턴	개체명과 패턴이 존재하는 논문 수	총 논문
19개	66개	82건	125,907건

실험 환경은 다음 <표 2>와 같은 운영체제 Windows 2000 Server와 CPU2.8GHz, 램 1024kb, 데이터베이스는 MySQL 환경에서 실험평가가 이루어졌다.

<표 2> 실험 환경

운영체제	CPU	램	데이터베이스
Windows 2000 Server	2.8GHz	1024kb	My-sql

##### 4.2 시스템 평가

위의 <표 1>를 보면 실험 대상으로 한 문헌 정보 수는 125,906건이다. 이중 개체명 19개와 패턴 66개를 포함하고 있는 Abstract 논문은 82건이다. 본 실험 평가에서 주안점을 둔 것은 텍스트 마이닝 시스템의 처리 시간과, 정확도, 마지막으로 재현율에 대해서 중점으로 실험을 실시하였다. 정확도와 재현율은 다음과 같은 수식으로 계산 하였다 [14].

$$\text{정확도} = \frac{\text{TP}}{\text{TP} + \text{TN}}$$

$$\text{재현율} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

위 수식에서 사용되는 TP, TP + TN, TP + FP 정의는 아래와 같다.

TP = 시스템에 의해 적합하게 검색된 문장의 개수

TP+TN = 개체명과 패턴을 포함하는 총 문장 수

TP+FP = 시스템에 의해 검색된 총 문장 수

<표 3> 정확도와 재현율 평가

TP	TP+TN	TP + FP	정확도	재현율	처리시간
79	82	83	96.3%	95.1%	15초

실험 데이터에 대한 평가 결과는 <표 3>과 같이 정확도 96.3%와 재현율 95.1%를 가지고 있다. 이러한 결과는 개체명과 패턴을 가지고 있는 Abstract 논문과 적은수의 데이터를 가지고 테스트를 하였기 때문에 정확도가 높게 나타났다고 보여지며, 향후 다양한 데이터에 대한 검증이 추가로 필요하다. 그러나 정확도의 검증은 결국 전문가의 수작업을 통한 과정을 거쳐야 하므로 다양한 대규모 데이터에 대하여 실험하는 것은 현실적으로 쉽지는 않다. 이러한 평가의 어려움을 감안하여 앞으로 KEGG 경로 데이터베이스에서 밝혀진 다른 유전자를 이용하여 시스템 평가를 추가로 시행할 예정이다.

(그림 12) 제안된 시스템 인터페이스

## 5. 결 론

본 논문에서는 대사경로 재구축 과정에서 문헌정보로부터 필요한 대사경로 지식을 추출하는 패턴 기반의 텍스트 마이닝 기법을 제안하였다. 제안된 텍스트 마이닝 시스템에서는 웹 로봇을 이용하여 최신의 논문을 반자동적으로 수집하고, 이를 이용하여 최신의 논문을 로컬 데이터베이스로 구축

하였다. 또한 생물 개체명의 인식율을 높이기 위해 유전자 온토로지를 이용하여, NCBI에서 제공하는 Tokenizer 라이브러리를 이용하여 개체명의 파괴 없이 인식할 수 있게 하였다. 지식 발견에서 중요한 역할을 담당하는 대사경로 표현에 관한 패턴들을 확보하기 위하여 일본의 KEGG 경로 데이터베이스를 활용하였으며, 여기서 추출한 66개의 패턴을 사용하였다. 물론 패턴은 사용자가 추가할 수 있도록 하였다. 공개된 생명공학 데이터인 Glycosphingolipid 종의 GLS 대사경로에 대하여 제안된 시스템을 실험하고, 평가한 결과 논문 125,907 건에 대하여 정확도 96.3%, 재현율 95.1%, 처리시간 15초의 성능을 보였다. 제안된 시스템을 사용하여 문헌정보로부터 추출한 모든 지식이 대사 작용을 정확히 표현한다고 볼 수는 없으나 전문가들에게 수작업으로 인한 시간을 줄여주는데 큰 도움이 될 것이다. 앞으로 다양한 종에 대한 대사경로를 추가로 검색하여 패턴의 완성도를 높임으로써 발견된 지식의 정확도는 높이는 연구가 필요하다. 또한 시스템의 처리 성능을 높이기 위한 알고리즘 개선도 필요하다.

## 참 고 문 헌

- [1] 최은정외, “대사 작용 정보 자동 추출을 위한 텍스트 마이닝시스템 설계 및 구현”, 석사학위 청구논문, 2005.
- [2] 이상엽, 네이처 바이오테크놀로지 10월호 초청논문.
- [3] <http://www2.lingsoft.fi/cgi-bin/engcg>
- [4] 여은주외, “대용량 생물정보 문헌으로부터 단백질 상호작용분석을 위한 바이오 테이터 마이닝 시스템,” KCSE 한국 소프트웨어공학 학술대회, 2005.2.
- [5] D.Hindle, "Deterministic parsing of syntactic non-fluencies," In Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics, 1983.
- [6] D.McDonald "Robust partial parsing through incremental multialgorithm processing," In P.Jacobs, editor, Text-based Intelligent Systems,

1992.

- [7] J. Pustejovsky et al., "Semantic indexing and typed hyperlingking," In AAAI Symposium on Language and the Web, Stanford, CA, 1997.
- [8] 임해창외, "바이오 Text-Mining 시스템 개발" *한국정보과학회지*, pp.60~68, 2003.6.
- [9] 김태경 "Etherboot 기반의 CGRID 구축과 서열분석에의 적용", 컴퓨터 정보학회, pp. 195~207, 2005.12.
- [10] B.Stapley and G.Benoit, "BIOBIBLIOMETRICS: Information Retrieval and Visualization from Co-occurrences of Gene Name in MEDLINE Abstracts" Pacific Symposium on BioComputing, pp.526~537, PSB 2000.
- [11] Lorraine Tanabe and W. John Wilbur, "Tagging gene and protein names in the biomedical text", J. of Bioinformatics, 18(8), 1124~1132. 2002.
- [12] Jun'ichi Kazma, Takaki Makino, Yoshihiro Ohta, Jun'ichi Tsujii, "Tuning Support Vector Machines for Biomedical Named Entity Recognition", Genome Informatics. 2002.
- [13] <http://www.ebi.ac.uk/ontology-lookup/>
- [14] [http://www.ncbi.nlm.nih.gov/sites/entrez?db\\_PubMed](http://www.ncbi.nlm.nih.gov/sites/entrez?db_PubMed)
- [15] <http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/textTools/current/Usages/Tokenerizer.html>
- [16] 이현철외 "텍스트 마이닝 기법을 이용한 단백질 상호작용 추출," 충북대학교 컴퓨터정보통신연구지, 2003.
- [17] 권해룡외, "스팡고질 생합성 경로의 재구축," 석사학위 청구논문.



권 혁 렐 (Hyuk-Ryul Kwon)

- 충북대학교 경영정보 학과 (경영학사 졸업)
- 충북대학교 바이오인포메틱스 학과 (공학석사 재학중)

- 관심분야 : Text-Mining, 바이오 인포메틱스
- Email : khl80@naver.com



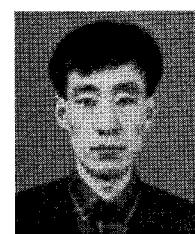
나 종 화 (Jonghwa Na)

- 서울대학교 계산통계학과 이학박사
- 공군사관학교 전산학과 전임강사
- 서울대 통계연구소 특별연구원
- 현 충북대학교 교수(06-현재), 기업정보화지원센터 소장
- 펜실베니아주립대학(PSU) 방문교수
- 관심분야 : 전산통계, 데이터마이닝, 수리통계



유 재 수 (Jae Soo Yoo)

- KAIST 전산학과 (공학박사)
- 현 충북대학교 BIT 연구중심 대학육성사업단 부단장
- 현 충북대학교 전전컴 교수
- 관심분야 : Database, BIT, 정보검색, 멀티미디어 분산객체 시스템



조 완 섭(Wan-Sup Cho)

- KAIST 전산학과 (공학 박사)  
미국 U. of Florida Post. Doc.
- 현 충북대학교 경영정보학과 교수
- 한국전자통신연구소 연구원 &Post. Doc.
- 현 충북대학교 경영정보학과 부교수
- 관심분야 : Data Warehouse & OLAP, CRM, DB, BIT
- Email : wscho@chungbuk.ac.kr