생물학적 패스웨이 실시간 확장 시스템

Biological Pathway Real-time Expansion System

최 윤 수, 전 선 희, 서 동 민, 유 석 종
한국과학기술정보연구원

Yunsoo Choi, Sun-hee Jeon, Dongmin Seo,
Seok Jong Yu
Korea Institute of Science & Technology Information

요약

생물학적 패스웨이는 기존 연구 문헌들을 토대로 생의학 분야 전문가들에 의해 수작업으로 구축되기 때문에, 신규 문헌들이 포함하고 있는 최신 정보들이 패스웨이에 적용되는데 많은 시간을 소모하게 된다. 본 논문에서는 텍스트 마이닝 및 정보 추출 기술을 사용하여 구축된 디지털 학술자원인 VSB(Virtual Science Brain) 데이터베이스를 활용하여 실시간으로 생물학적 패스웨이를 확장하는 방법과, 이를 위한 사용자 인터페이스를 소개한다.

Ⅰ. 서론 및 관련연구

생의학 분야의 패스웨이는 기술문헌에 출현한 다양한 전문용어와 그들 간의 의미적 상관 관계를 네트워크 형식으로 표현한 자료구조로서, 생명공학 관점에서는 단백질, 유전자, 세포 등의 생체적 요소 간의 역학관계 혹은 상호작용 등을 세밀하게 기술한 생물학적 심층지식이다[1].

생물학적 패스웨이는 기존 연구 문헌들을 토대로 생의학 분야 전문가들에 의한 수작업으로 구축되기 때문에, 최신 연구 문헌들에 나타난 신규 문헌들이 포함하고 있는 최신 정보가 패스웨이에 적용되는데 많은 시간을 소모하게 된다.

현재 생물학적 패스웨이의 대표로 불리우는 KEGG 패스웨이도 약 4,000여개의 유전자만을 취급하고 있으며, 이는 아직도 밝혀지지 않은 경로가 KEGG내에 존재하고 향후 더 확장될 수 있음을 의미한다[2].

이러한 문제를 해결하고자, 단백질-단백질 상호작용 네트워크와 유전자 온톨로지(GO)를 기반으로 단백질간 상호작용에 신뢰도를 부여하고, 이를 통해 기존 패스웨이를 재구축하는 기존 연구가 있었다[3-4].

기존 연구는 기 구축된 단백질-단백질 상호작용 데이터베이스를 기반으로 패스웨이를 확장하는 방식이기 때문에, 이 또한 최신 연구 문헌들에 나타나는 정보를 신속하게 수용하기 어렵다.

최신 연구 문헌들이 포함하고 있는 정보를 패스웨이에 적용하기 위해서는, 텍스트 문헌을 기계적으로 분석하여 세포명, 화합물명, 질병, 약품, 치료법 등과 같은 핵심 용어들을 자동으로 추출하고, 이들 간의 의미적 연관관계를 문헌내에 기술된 정보를 바탕으로 식별하는 텍스트 마이닝 및 정보 추출 기술을 사용하여 최신 정보에 대한 지식화 작업이 선행되어야 한다.

본 논문에서는 이러한 텍스트 마이닝 및 정보 추출 기술을 사용하여 구축된 디지털 학술자원인 VSB(Virtual Science Brain) 데이터베이스를 활용하여 생물학적 패스웨이를 실시간으로 확장하는 시스템과 사용자 인터페이스를 소개한다.

Ⅱ. 생물학적 패스웨이 확장 시스템

생물학적 패스웨이를 확장하기 위한 기반 자원으로 텍스트 마이닝 및 정보 추출 기술을 사용하여 구축된 디지털 학술자원인 VSB(Virtual Science Brain) 데이터베이스를 활용한다.

VSB는 PubMed 데이터베이스 문헌 중 최근 10년 간 발행된 문헌 18,032,232건으로 부터 구축되었으며, 핵심개체와 개체간 관계로 구성되는 트리플 형식의 자료로 구성된다.

표 1. VSB 데이터베이스 통계 정보

항목	건수
문서(Document)	18,032,232
핵심개체(Entity)	24,336,926
개체간 관계(Relation)	17,925,708
키워드(Keyword)	192,679
동사(Verb)	17,925,550
토큰(Token)	19,703,792

이와 같이 구축된 VSB와 생물학적 패스웨이와의 연동은 핵심개체명을 기반으로 이루어진다. 패스웨이 네트워크 상의 한 노드는 유전자, 단백질 같은 핵심개체로 이루어져 있다. 이 노드의 이름과 동일한 핵심개체명을 VSB 데이터베이스에서 검색을 수행하여, 그 결과로 트리플(개

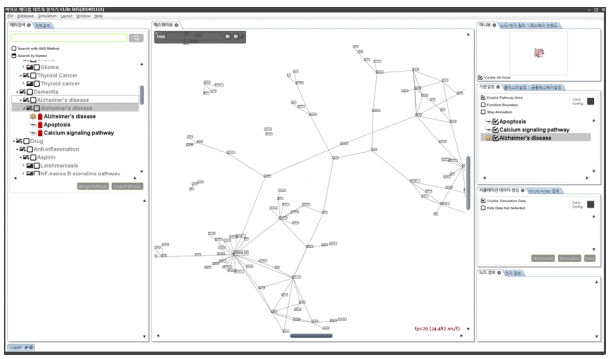
체-관계-개체)을 가져온다. 이 트리플과 생물학적 패스웨이 네트워크를 연결하여, 네트워크를 확장하게 된다.

표 2. 핵심개체 종류

순번	핵심개체 종류	건수
1	compound (복합물)	5,088,700
2	disease (질병)	5,176,030
3	drug (약물)	843,270
4	enzyme (효소)	902,560
5	symptom (증상)	436,280
6	gene_prod (산출물)	4,926,480

본 논문에서 사용하는 핵심개체 중 단백질명은 하나의 단백질명에 대해 여러 개의 동의어를 가지고 있는 경우가 대부분이다. 이러한 동의어들로 인해 검색되고 통합되어야할 트리플이 검색되지 않는 상황을 방지하기 위하여, 동의어 사전을 구축하였다.

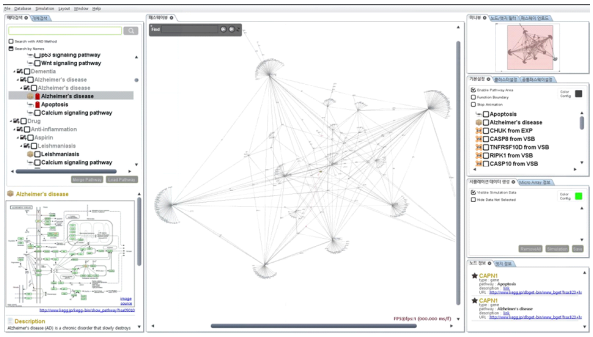
최초 수집되는 패스웨이 대상이 KEGG였기 때문에, KEGG 문서내에 지정되어 있는 동의어를 1차 대상으로 작업하고, 향후 다양한 패스웨이들과의 연계 및 통합을 위해, 인간 유전자, 동의어, 단백질 이름 등에 대한 정보를 담고 있는 HGNC(HUGO Gene Nomenclature Committee)로 부터 동의어를 추출하여 KEGG에서 추출한 동의어와 통합, 확장하여 19,477 건의 동의어 쌍을 확보하였다.



▶▶ 그림 1. 패스웨이 확장 인터페이스

VSB를 사용하여 생물학적 패스웨이에 대한 무조건적인 확장을 수행하면, 매우 복잡하고 거대한 네트워크가 생성되게 된다. 이는 패스웨이를 이용한 분석을 위해 적합하지 않다.

사용자 인터페이스는 연구자들이 실시간으로 관심있는 후보 노드들에서 패스웨이 확장을 수행할 수 있도록 구성되었다. 〈그림 1〉은 패스웨이 확장을 위한 인터페이스로, 대상 패스웨이를 검색하고, 화면에 적재하고, 패스웨이내에 있는 관심 노드를 확장하는 기능을 제공하도록 구성되었다.



▶▶ 그림 2. 확장된 생물학적 패스웨이

〈그림 1〉은 알츠하이머 패스웨이를 최초 적재한 상태이고, 〈그림2〉는 관심 노드들을 선택하여 확장된 패스웨이를 보여준다. 〈그림 1〉과 비교하여 〈그림 2〉는 노드들간의 연결이 더 강화되고, 새로운 연결이 실시간으로 생성되었음을 보여준다.

Ⅲ. 결론 및 향후연구

본 논문에서는 기존에 수작업으로 구축되던 생물학적 패스웨이에 대해, 생의학분야의 최신문헌에 대한 신규정보를 이용하여 실시간으로 확장하기 위한 방법과, 이에 대한 편의성을 위한 사용자 인터페이스를 제안하였다. 본 논문의 결과물인 실시간 패스웨이 자동확장 시스템은 기존 패스웨이에서 발견할 수 없었던 질환을 발생시키는 새로운 기전의 후보 단백질/유전자 발견에 활용이 가능하다.

향후 기반이 되는 VSB에 대한 정확도를 향상시키는 연구와, 확장된 생물학적 패스웨이를 생의학 분야에서 사용하는 다른 분석 도구들과 연계하는 방안에 대한 연구를 수행하여 전체적인 성능과 활용성을 보강할 예정이다.

▮ 참 고 문 헌 ▮

[1] Chun, Hong-Woo, et al. “Pathway construction and extension using natural language processing,” Human Interface and the Management of Information. Information and Interaction for Health, Safety, Mobility and Complex Environments. Springer Berlin Heidelberg, pp. 32-38. 2013.

[2] Kanehisa, Minoru, and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes,” Nucleic acids research Vo. 28 No. 1, pp. 27-30. 2000.

[3] Gurkan Bebek, Jiong Yang, “PathFinder: mining signal transduction pathway segments from protein-protein interaction networks,” BMC Bioinformatics, Vol. 8, 335, 2007.

[4] Arga K. Yalcin, et al., “Understanding signaling in yeast: Insights from network analysis,” Biotechnology and bioengineering, Vol. 97, No. 5, pp. 1246-1258, 2006.