바이오인포매틱스 분야 회색문헌 및 백색문헌의 연구 동향 비교 분석

Analyzing Research Trends in Bioinformatics based on Comparison between Grey and White Bioinformatics Literatures

김예은, 연세대학교 문헌정보학과, arcoiris@yonsei.ac.kr
김정주, 연세대학교 문헌정보학과, kjj198@yonsei.ac.kr
송민, 연세대학교 문헌정보학과, min.song@yonsei.ac.kr
Ye Eun Kim, Dept. of Library and Information Science, Graduate School of Yonsei University
Jung Ju Kim, Dept. of Library and Information Science, Graduate School of Yonsei University
Min Song, Dept. of Library and Information Science, Yonsei University

본 연구의 목적은 바이오인포매틱스 분야의 회색문헌과 백색문헌의 초록을 대상으로 단어 동시출현(word co-occurrence)네트워크 분석을 통해 해당 분야의 연구 동향을 비교 분석하고자 하였다. 이를 위해 2010년부터 2012년까지 발표된 회색문헌인 회의자료(proceeding)와 백색문헌인 학술논문(journal article)의 초록을 SCOPUS, IEEEXplore, Microsoft academic search에서 수집하였다. 단어 동시출현 네트워크를 분석한 결과 회색문헌의 주요 연구는 분석도구 및 방법으로, 백색문헌의 주요 연구는 바이오인포매틱스의 주요 연구대상인 유전자 발현, 단백질 서열 및 구조 등으로 나타났다.

1. 서 론

생물정보학 또는 생명정보학으로 번역되는 바이오인포매틱스(Bioinformatics)는 생물학을 의미하는 bio와 정보학을 뜻하는 informatics의 합성어로(이식 2003), 컴퓨터를 활용하여 생물학적 데이터를 수집, 관리, 저장, 평가 및 분석하는 기술이다. 바이오인포매틱스는 인간게놈 프로젝트의 완성으로 유전정보의 양이 증가함에 따라 그 중요성이 점차 높아지고 있다. 이에 따라 바이오인포매틱스 분야의 새로운 분석 기술과 분석 도구, 프레임워크 등을 다루는 수많은 연구 자료가 생산되고 있으며, 바이오인포매틱스 분야의 연구 동향에 대한 분석은 계속해서 늘어나고 있다.

바이오인포매틱스와 같은 과학 기술 분야에서는 자료의 최신성이 중요한 역할을 담당한다. 출판사를 통한 인쇄, 편집, 발행 및 배포의 과정을 거치지 않는 회색문헌(Grey Literature)은 백색문헌(White Literature) 보다 최신성을 지니며, 이를 통해 연구의 선취권을 획득할 수 있다는 점에서 그 어떤 자료보다도 핵심적으로 이용될 수 있다(이지연, 이지연 2007).

이에 본 연구에서는 바이오인포매틱스 분야의 회색문헌과 백색문헌에서 자연언어 분석 기반의 텍스트 마이닝을 통해 단어를 추출하고 추출된 단어들의 동시출현 네트워크를 분석함으로써 바이오인포매틱스 분야의 연구 동향을 파악하고 비교해보고자 한다.

2. 관련 연구

2.1 바이오인포매틱스 연구 동향

바이오인포매틱스의 중요성이 강조됨에 따라 연구 동향을 분석한 다양한 연구들이 진행되고 있다. Molidor et al.(2003)은 바이오인포매틱스는 포스트 게놈 시대를 이끌어가고 있으며, 이는 앞으로의 분자 생명 과학의 핵심 부분이 될 것이라 예상하였고, 임달혁 외(2004)는 바이오인포매틱스의 패러다임이 생산된 데이터를 거슬러 올라 컴퓨터와 분석도구들을 활용하여 유전자 서열의 감추어진 발현성 및 표현형을 찾아가는 것으로 연구흐름의 방식이 진행되고 있다고 분석하였다. Perez-Iratxeta et al.(2006)은 MEDLINE 데이터베이스의 바이오메디컬 문헌의 초록을 이용하여 바이오인포매틱스의 발전 및 토픽의 최근 동향에 대해 연구하였으며, 해당 분야가 빠르게 성장하고 있음을 밝혀냈다.

2.2 회색문헌의 정의 및 유형

회색문헌의 정의와 유형은 하나로 통일되어 있지 않고 국가나 주제 분야 또는 학자에 따라 여러 측면에서 다양하게 나타나고 있다. 남영준(2002)은 서지정보원을 통해 공개적으로 드러나지 않아 확인과 접근 및 이용이 불투명한 정보자료를 통칭하는 것이라 정의하였고, 정현이(2000)는 최신정보원으로서 가치가 있지만 공식적인 출판 경로를 통해 생산, 배포되지 않는 자료로서 이용하기가 쉽지 않은 자료라 정의하며 회색문헌의 유통상의 특성을 강조하고 있다. 또한 Farace(1997)는 생산 주체에 초점을 맞춰 모든 정부, 대학, 기업 등에서 인쇄 및 전자형태로 발간되는 자료로 상업 출판의 통제를 받지 않는 것이라 정의하였다.

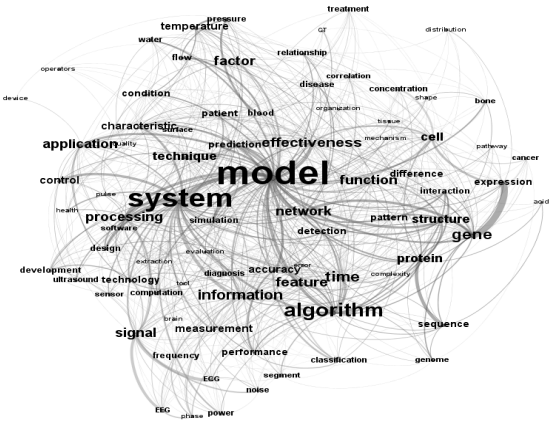
유형에 있어서 회색문헌과 백색문헌을 정확하게 구분하는 것은 어렵지만 일반적으로 회색문헌은 각종 보고서(연구보고서, 기술보고서, 정책보고서), 회의록, 회의자료, 학위논문, 정부간행물, 통계자료, 뉴스레터 등을 포함한다.

3. 연구대상 및 방법

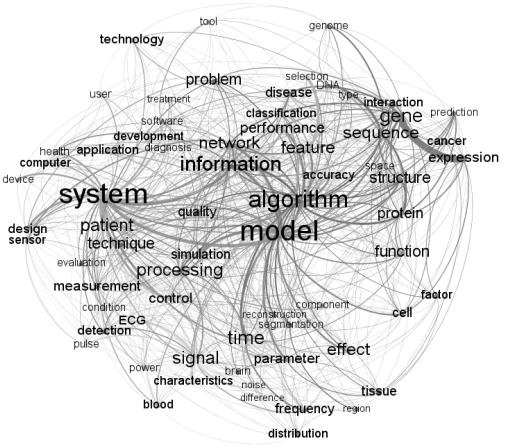
본 연구에서는 회색문헌으로 바이오인포매틱스 분야의 주요 학회 7곳에서 최근 3년간(2010-2012) 발표된 회의자료(proceeding)와 백색문헌으로 주요 학술저널인 BMC Bioinformatics와 Bioinformatics의 최근 3년간(2010-2012) 출판된 학술논문의 초록을 연구대상으로 선정하였다. SCOPUS, Microsoft Academic Search, IEEEXplore를 통해 4,009편의 학술논문 초록과 2,794편의 회의자료 초록을 수집하였다. 수집한 자료의 분석은 콘텐츠 분석 툴킷인 Apache Tika parser를 이용하여 자연어 처리를 위한 품사 태깅을 통해 단어 추출을 한 후 네트워크 분석 및 시각화 도구인 Gephi 0.8.2를 이용하여 단어 동시 출현 네트워크를 분석하였다.

4. 분석결과

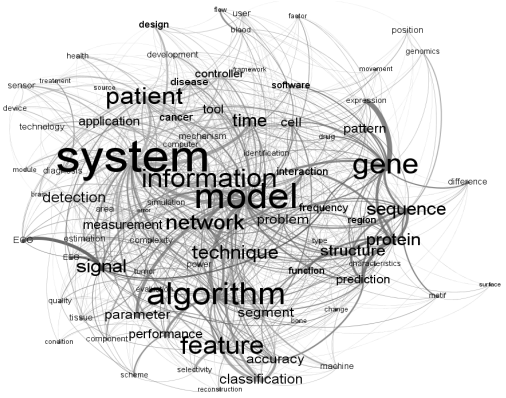
4.1 회색문헌 단어 동시출현 네트워크



<그림 1> 2010년도 회색문헌 단어 동시출현 네트워크



<그림 2> 2011년도 회색문헌 단어 동시출현 네트워크

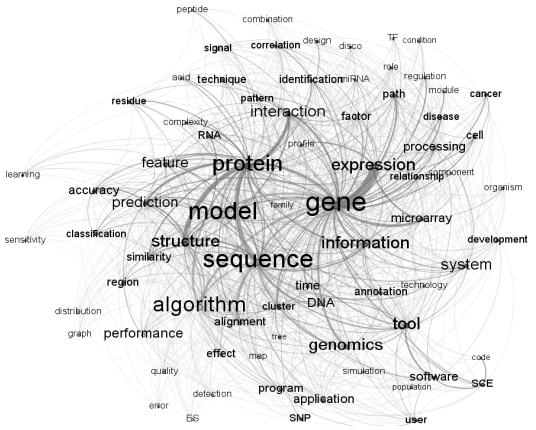


<그림 3> 2012년도 회색문헌 단어 동시출현 네트워크

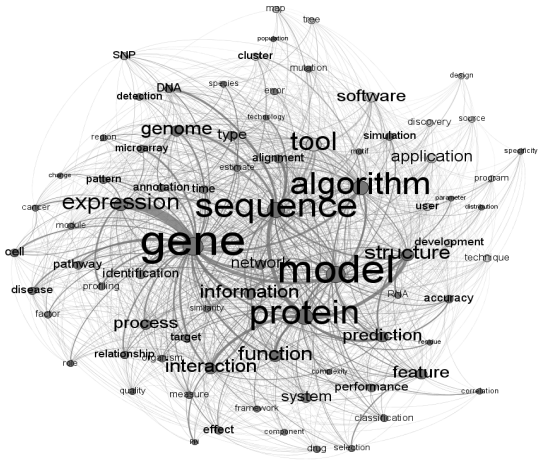
회색문헌의 최근 3년간(2010-2012) 단어 동시출현 네크워크를 분석한 결과 system, model, algorithm 등의 유전자 및 단백질 구조를 분석하기 위한 도구와 방법에 관한 단어의 동시출현 빈도가 높았고, 각각의 주요 단어들을 중심으로 커뮤니티를 형성하고 있었다. 그리고 시간이 지남에 따라 바이오인포매틱스의 주요 연구 대상인 gene, gene expression, protein sequence, genome 등의 단어 동시출현 빈도 및 커뮤니티의 크기가 증가하고 있음을 알 수 있었다.

4.2 백색문헌 단어 동시출현 네트워크

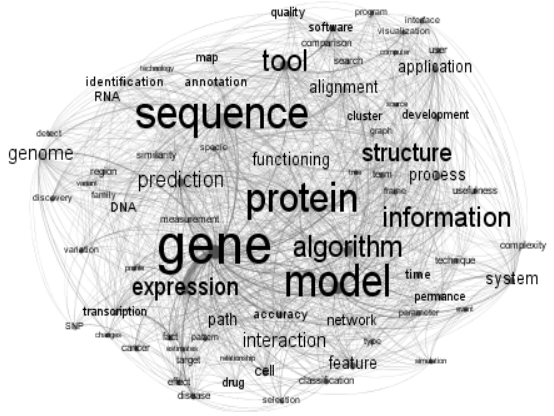
백색문헌의 최근 3년간(2010-2012) 단어 동시출현 네트워크를 분석한 결과 유전자 및 단백질 구조 등을 분석하기 위한 도구와 방법에 집중되었던 회색문헌의 결과와는 다르게 algorithm, model과 더불어 gene, gene expression, protein, protein sequence 등의 단어 동시출현 빈도가 높게 나타났고, 이들을 중심으로 한 커뮤니티의 크기가 크게 형성되고 있음을 알 수 있었다.



<그림 4> 2010년도 백색문헌 단어 동시출현 네트워크



<그림 5> 2011년도 백색문헌 단어 동시출현 네트워크



<그림 6> 2012년도 백색문헌 단어 동시출현 네트워크

또한 분석결과를 통해 백색문헌의 경우 2010년부터 시간이 경과함에 따라 유전자 분석도구와 모델, 알고리즘에 대한 관심이 점차 바이오인포매틱스 분야 연구자들에게 있어서 증가하고 있음을 확인할 수 있었다.

5. 결 론

본 연구는 바이오인포매틱스 분야의 회색문헌과 백색문헌에 출현하는 단어의 동시출현 네트워크를 분석함으로써 연구 동향을 비교하고자 하였다. 그 결과 회색문헌인 회의자료는 백색문헌인 학술논문보다 분석도구 및 방법을, 백색문헌은 유전자 발현과 단백질 서열 및 구조를 연구 주제로 하며, 이를 중심으로 연구를 진행하고 있음을 추측할 수 있다.

마지막으로 본 연구를 기반으로 명사구(noun phrase) 동시출현 네트워크 분석을 통한 회색문헌과 백색문헌의 연구 동향 비교 연구를 후속연구로 제언하고자 한다. 이는 보다 유용한 단어의 추출이 가능할 것이기에 바이오인포매틱스 분야의 회색문헌 및 백색문헌의 유형과 출판 및 발표연도의 범위를 확대하여 분석을 실시한다면 보다 의미 있는 연구 결과를 얻을 수 있을 것이라 기대한다.

참고문헌

남영준. 2002. 디지털 시대의 회색문헌 이용 활성화에 관한 연구. 『정보관리학회지』, 19(4): 233-256.

이식. 2003. Bio-Informatics 개론 및 국내외 동향. 『전자공학회지』, 30(10): 17-28.

이지연, 이지연. 2007. 국내 과학기술분야 회색문헌의 효율적 관리방안에 관한 연구. 『정보관리연구』, 38(2): 25-57.

임달혁. 2006. 바이오데이터베이스와 도구를 활용한 바이오인포매틱스의 동향. 『약제학회지』, 34(1): 73-79.

정현이. 2000. 회색문헌의 이용실태에 관한 연구. 석사학위논문, 중앙대학교 대학원 문헌정보학과.

조현양. 2008. 인용 분석을 통한 학문간 회색문헌의 활용도 비교 연구. 『한국문헌정보학회지』, 42(1): 273-294.

Molidor, R. et al. 2003. “New trends in bioinformatics: from genome sequence to personalized medicine.” *Experimental Gerontology,* 38(10): 1031-1036.

Perez-Iratxeta, C. et al. 2006. “Evolving research trends in bioinformatics.” *Briefings in Bioinformatics,* 8(2): 88-95.

Farace, D. J. 1997. “Rise of the phoenix: A review of new forms and exploitations of grey literature.” *Publishing Research Quarterly,* 13(2): 69-76.

Patra, S. K. and Mishra, S. 2006. “Bibliometric study of bioinformatics literature.” *Scientometrics,* 67(3): 477-489.