서지분석을 통한 노화 관련 유전자 정보 데이터베이스 구축

Construction of the Aging Related Gene Database using Text-mining

유 석 종*, 박 준 호, 유 재 수
한국과학기술정보연구원 국가슈퍼컴퓨팅연구소*,
충북대학교 정보통신공학부

Seok Jong Yu*, Junho Park, Jaesoo Yoo
KISTI National Institute of Supercomputing and Networking*, Chungbuk National University

요약

최근 노령화가 급속히 진행되면서 노화에 대한 연구가 활발히 진행되고 있다. 하지만 노화현상은 광범위한 표현형을 지니고 있는 생명현상으로 이에 대한 체계적인 연구를 지원하기 위한 웹포털 사이트가 필요한 실정이다. 특히 노화에 따른 질병과의 연관성 및 관련 유전자에 대한 정보를 수집하고 이를 체계적으로 분석할 수 있는 통합정보시스템은 향후 노화연구를 지원하기 위한 가장 핵심적인 요소라고 할 수 있다. 본 연구에서는 기존 노화와 관련된 461개의 유전자를 기반으로 관련된 질병과의 연관성을 OMIM 데이터베이스를 활용하여 분석하였다. 또한 관련 단백질의 기능을 GO데이터베이스 분석을 통해 유전자의 기능을 분석하였다. Pubmed에서 제공하는 노화관련 논문들의 MeSH 정보 분석을 통해서 노화와 관련된 용어를 분석하였다. 노화와 관련된 64개의 유전자를 키워드로 NCBI의 pubmed 데이터베이스로부터 관련논문을 다운로드 받아 생물학적 상호작용 정보를 추출했다. 생물학적 상호작용은 NCBI에서 제공하는 Metamap 데이터베이스를 기반으로 각각의 생물학적 용어를 정의했다. 현재 노화 유전자 64개에 대해 128,729개의 생물학적 상호작용 정보를 추출했고, 8대 노인성만성질환에 대해 301,176개의 생물학적 상호작용 정보를 추출하였다.

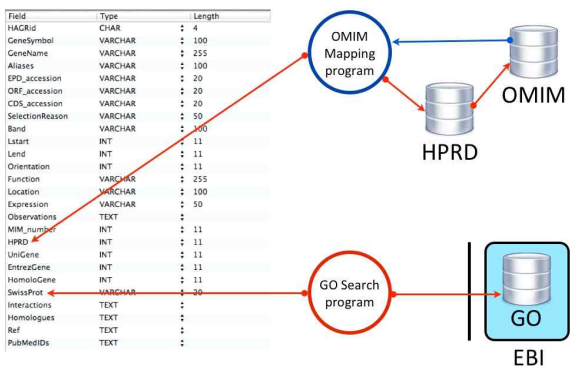
Ⅰ. 서론

이를 위해서는 근본적으로 정보공학기법을 활용해 문헌의 서지데이터에서 다양한 수준의 생물학적 정보를 추출하는 것이 핵심적인 기술이 될 것이며, 기존 문헌에서 다양한 단백질-단백질 상호작용, 단백질-유전자 상호작용에 대한 정보를 추출하여 분석하는 것이 핵심적인 정보가 될것이다. 현재 국제적으로는 GenAge 데이터베이스[1]가 가장 많은 노화관련 유전자정보를 제공하고 있으며, 분자 수준, 생리학적 수준, 병리학적 수준의 노화 관련 데이터를 통합하여 노화의 전반적인 형질에의 정보는 Digital Ageing Atlas(DAA)[2]가 제공하고 있다. 본 연구에서는 이러한 다차원의 생물학적 네트워크를 생성하기 위해 서지분석을 위한 텍스트마이닝기법 개발과 이를 기반으로 생물학적 네트워크의 데이터베이스화 및 이를 연구자가 손쉽게 열람하고 그 생물학적 의미를 파악할 수 있도록 하였다.

Ⅱ. 방법

1. 노화 관련 유전자 정보분석

GenAge[1]의 461개유전자를 기반으로 OMIM[3]정보와 GO[4]정보를 추가할 수 있도록 데이터베이스를 설계했다. 설계된 데이터베이스의 정보입력을 위해서 OMIM과 GO를 매핑할 매핑용 프로그램을 직접 제작했다. 기존의 GenAge데이터 필드에 HPRD ID값을 이용해 HPRD데이터베이스를 검색하고 여기에서 OMIM정보를 수집해 GenAge와 OMIM간의 상관관계를 매핑했다. 또한 GO정보는 EBI에서 제공하는 GO API를 이용해 직접 검색할 수 있는 검색 및 매핑용 프로그램을 제작했다. MeSH용어분석은 노화 관련용어를 기준으로 Pubmed의 MeSH정보를 XML로 다운로드받아 용어의 빈도를 통계분석 하였다.



▶▶ 그림 1. 노화 관련 데이터베이스 설계 및 외부 데이터 통합

본 연구는 KISTI의 국가슈퍼컴퓨팅서비스 개발 및 기술 연구 과제(K-13-L01-C02) 및 질병관리본부 학술용역과제(2013E6200100)의 지원을 받아 수행되었습니다.

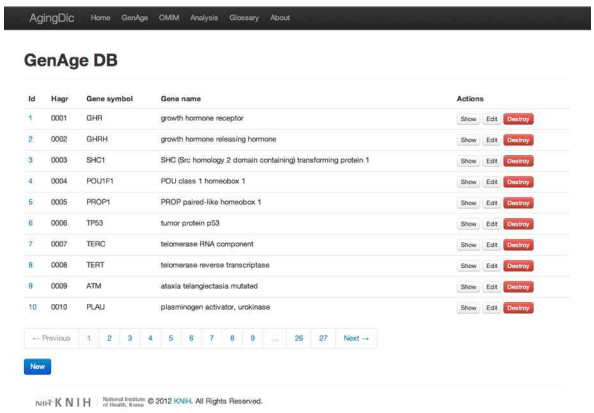
2. 노화 관련 유전자 서지분석

기존에 노화관련 유전자 정보를 활용하여 64개의 유전자를 선정하였으며, 각각의 유전자이름을 질의어로 Pubmed의 관련 논문을 다운로드 받았다. 서지분석은 Stanford파서와 Metamap정보를 활용하여 수행하였으며, 추출된 단백질 혹은 유전자간의 상호작용정보를 추출하여 생물학적 네트워크를 구축하였다.

Ⅲ. 결과

1. 노화 관련 유전자 데이터베이스 구축

노화와 관련된 중요 유전자는 이미 GenAge데이터베이스를 통해 보고되었다. 본 연구에서는 향후 서지분석을 통해 찾아진 유전자를 지속적으로 관리할 수 있도록 AgingDic 데이터베이스 내에 유전자 정보 및 관련 정보를 수록하고자 했다. 특히 서지분석을 통해 찾아진 유전자와 기존 보고된 유전자간의 상관성과 향후 관련된 유전자들이 관여하는 신호전달 네트워크 등을 지속적으로 확장할 필요성이 있기 때문에 데이터베이스 설계시 확장성을 고려해 설계했다. 또한 개발된 GO와 OMIM정보 추출 소프트웨어를 활용해 AgingDic데이터베이스에 자동으로 관련 정보를 입력할 수 있도록 함으로써 향후 자동화된 데이터베이스 구축시스템을 구축했으며, 해당유전자의 정보뿐만 아니라 관련 논문, GO, OMIM등을 열람할 수 있도록 개발했다.



▶▶ 그림 2. 노화 관련 유전자 데이터베이스

2. 서지분석을 통한 노화 관련 생물학적 네트워크 구축

노화에 대해 추출한 64개의 유전자에 대해 세부적인 텍스트마이닝 작업을 수행했다. 분석방법은 앞장에서 설명한 서지분석 알고리즘을 이용했으며 가장 특징적인 부분은 MetaMap을 활용하여 다양한 생물학적 용어를 검색할 수 있도록 했다. 기존의 텍스트마이닝 도구들은 주로 단백질-단백질 상호작용 정보를 수집했는데, 그 외의 정보들은 분석하기 힘들었다. 본 연구에서는 Metamap의 약 100여개 온톨로지를 이용하여 문헌내의 용어를 찾아주기 때문에 보다 폭넓은 분석이 가능하다. 텍스트마이닝 결과 전체 128,729개의 상호작용정보를 추출했다. 가장 상호작용 정보를 많이 가지고 있는 유전자는 JUN, FAS, FOS유전자였으며, MT-CO1등 연구가 미미한 유전자도 존재했다. 노인성 만성 8대 질환에 대해서도 노화 유전자에 대한 분석과 동일한 기법으로 텍스트마이닝 작업을 수행했으며 각각의 질병에 대한 키워드를 이용하여 관련된 문헌 내에 다양한 생물학적 상호작용 정보를 수집했다. 분석결과 총 301,176개의 생물학적 상호작용을 추출하였으며, 질병별로는 고혈압에 대한 상호작용정보가 가장 많았고 백내장에 대한 정보는 상대적으로 적었다.

Ⅳ. 결론 및 토의

노화에 대한 기존 연구를 이를 다양한 관점에서 분석하는 것은 향후 노화 연구에 대한 시작점으로 현재 이를 제공하기위한 종합적인 분석은 이루어지고 있지 못했다. 본 연구에서는 노화에 관여하는 유전자에 대해 질병 및 단백질의 기능분석을 추가하여 연구자가 노화 관련 유전자의 다양한 정보를 열람할 수 있도록 데이터베이스를 구축하였다. 또한 노화에 관여하는 유전자에 대해 질병 및 단백질의 기능분석을 추가하여 연구자가 노화 관련 유전자의 다양한 정보를 열람할 수 있도록 데이터베이스를 구축하였으며, 서지분석을 통해 기존 문헌에서 노화와 관련된 생물학적 상호작용정보를 추출하여 생물학적 조절기작에 대한 네트워크를 구축함으로써 향후 노화관련 연구에 활용할 수 있도록 하였다.

▌참 고 문 헌▐

[1] Tacutu, R., Craig, T., Budovsky, A., Wuttke, D., Lehmann, G., Taranukha, D., Costa, J., Fraifeld, V.E., de Magalhães, J.P., 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Research 41, D1027-33

[2] Craig, T., Smelick, C. and de Magalhaes, J. P. 2010, The Digital Ageing Atlas: http://ageing-map.org

[3] Amberger, J., Bocchini, C.A., Scott, A.F., Hamosh, A., 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Research 37, D793-6.

[4] Gene Ontology Consortium, 2013. Gene Ontology annotations and resources. Nucleic Acids Research 41, D530-5.