임상 및 바이오 정보 통합 데이터 처리를 위한 범용적 임포팅 시스템의 설계 및 구현

Design and Implementation of a General Purpose Importing System for Integrated Data Processing on Clinical and Bio Information

임종태[1], 류은경[1], 김기연[1], 김천중[1], 윤수용[1], 박선용[1], 노연우[1], 육미선[1], 정지원[1], 최기태[1], 유석종[2], 유재수[1]*
[1]충북대학교 정보통신공학과,
[2]한국과학기술정보연구원

Jongtae Lim[1], Eunkyung Ryu[1], Kiyeon Kim[1], Cheonjung Kim[1], Sooyong Yoon[1], Sunyong Park[1], Yeonwoo Noh[1], Miseon Yuk[1], Jiwon Jeong[1], Kitae Choi[1], Seokjong Yu[2], Jaesoo Yoo[1]*
[1]Chungbuk National University,
[2]Korea Institute of Science and Technology Information

요약

질병의 발병기작과 같이 생명현상을 이해하기 위해서는 질병관점에서 생명현상을 분석하는 연구가 필요하며, 이를 지원하기 위한 질병과 관련된 임상 및 바이오 정보의 통합된 데이터베이스가 필요하다. 통합된 데이터베이스를 구축하기 위해서는 다양한 출처로부터 수집되는 다양한 임상 및 바이오 정보를 하나로 통합하여 저장 및 관리하는 기법이 필요하다. 따라서 본 논문에서는 임상 및 바이오 정보 통합 데이터베이스를 위한 범용적 임포팅 시스템을 설계하고 구현한다. 제안하는 시스템을 통해 사용자는 다양한 출처로부터 수집되는 임상 및 바이오 정보를 데이터 형태에 상관없이 하나의 통합 데이터베이스에 입력할 수 있다.

Ⅰ. 서론

생명과학분야에서 생명현상의 근본적인 기작을 이해하고 이를 활용하여 생명현상을 모델링하고 실험적으로 확인하는 것은 필수적이다. 또한 암과 같은 중요 질병에 대한 연구가 활발해지면서 질병의 원인 기작을 분석하고 이를 기반으로 신약 개발 등 다양한 응용 연구가 진행되고 있다. 최근 빅데이터가 이슈가 되면서 하둡과 하둡 에코시스템을 이용하여 임상 및 바이오 정보를 분석하는 연구가 중요하게 연구되고 있다[1]. 특히 시스템 생물학 분야에서는 질병과 관련된 임상 정보와 신호전달 정보를 통합하여 질병관점에서 생명현상을 분석한다. 하지만 임상 및 바이오 정보는 다양한 데이터베이스에 산재해있고, 각 정보는 상이한 데이터 형태를 가지고 있다. 따라서 임상 및 바이오 정보를 활용하기 위해서는 각각 다른 분석 시스템이 필요하며, 다른 형태의 데이터를 연관 분석한다는 것은 불가능하다. 질병의 발병기작과 같이 생명현상을 이해하기 위해서는 질병관점에서 생명현상을 분석하는 연구가 필요하며, 이를 지원하기 위한 질병과 관련된 임상 및 바이오 정보의 통합된 데이터베이스가 필요하다.

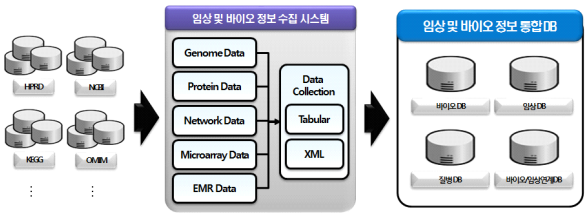
통합된 데이터베이스를 구축하기 위해서는 다양한 출처로부터 수집되는 다양한 임상 및 바이오 정보를 하나로 통합하여 저장 및 관리하고, 이를 처리할 수 있는 새로운 접근 방법이 필요하다. 따라서 본 논문에서는 임상 및 바이오 정보 통합 데이터베이스를 위한 범용적 임포팅 시스템을 제안한다. 제안하는 시스템을 통해 사용자는 다양한 출처로부터 수집되는 임상 및 바이오 정보를 데이터 형태에 상관없이 하나의 통합 데이터베이스에 입력할 수 있다. 사용자는 간편하게 자신이 설계한 데이터베이스 스키마를 입력하고, 임포팅 시스템은 데이터베이스 스키마를 기반으로 정보를 저장한다.

Ⅱ. 임포팅 시스템 설계

본 논문에서는 임상 및 바이오 정보 통합 데이터베이스를 위한 범용적 임포팅 시스템을 설계한다. 설계한 시스템은 입력되는 데이터에 상관없이 통합 데이터베이스에 정보를 저장할 수 있다. 그림 1은 제안하는 시스템의 전체 시스템 구조를 보여준다. 제안하는 시스템은 HPRD[2], NCBI[3], KEGG[4], OpenEMR[5] 등으로부터 수집되는 유전체, 단백체, 발현체, 신호전달, 임상 정보를 데이터 포맷, 형태와 상관없이 통합 데이터베이스에 저

* 교신저자 : yjs@chungbuk.ac.kr
본 연구는 미래창조과학부 및 정보통신산업진흥원의 대학 IT연구센터육성 지원사업(NIPA-2014-H0301-14-1022), 한국과학기술정보연구원의 「고성능 컴퓨팅 기반 빅데이터 기술 개발 (K-14-L06_C04-S01)」사업, 그리고 미래창조과학부의 방송통신정책연구센터운영지원사업(KCA-2013-003)으로부터 지원받아 수행된 연구임.

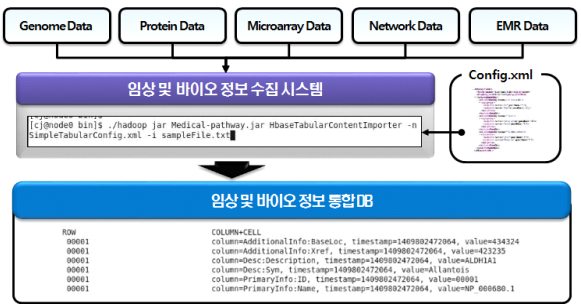
장한다. 제안하는 시스템은 데이터가 Tab으로 구분된 Tabular 파일과 XML 파일을 지원한다. 제안하는 임포팅 시스템을 통해 수집된 데이터는 바이오 데이터베이스, 임상 데이터베이스, 질병 데이터베이스, 바이오 임상 연계 데이터베이스 형태로 저장되어 빅데이터 분석에 활용된다.



▶▶ 그림 1. 제안하는 시스템 구조

Ⅲ. 임포팅 시스템 구현

그림 2는 제안하는 시스템의 프로세스 흐름을 나타낸 것이다. 제안하는 임상 및 바이오 정보 수집 시스템에는 데이터스키마 정보를 저장하고 있는 config 파일과 실제 임상 및 바이오 정보를 저장하고 있는 데이터 파일이 입력된다. 제안하는 임상 및 바이오 정보 수집 시스템은 입력된 config 파일을 참조하여 데이터베이스 스키마를 생성하고 데이터 파일로부터 실제 데이터를 파싱(Parsing)하여 통합 데이터베이스에 저장한다. 빅데이터 분석을 위해 하둡과 하둡 에코시스템과 연동하여 동작할 수 있도록 Hbase에 통합 데이터베이스를 구축했다.



▶▶ 그림 2. 제안하는 시스템의 프로세스 흐름도

그림 3과 4는 제안하는 임상 및 바이오 정보 수집 시스템에 입력되는 config 파일을 보여준다. config 파일은 입력되는 정보의 데이터스키마 정보를 가진다. 실제 임상 및 바이오 정보 수집 시스템에 입력되는 config 파일은 그림 4와 같이 XML 포맷으로 구성된다. 하지만 사용자 중에는 XML 문서를 직접 작성하기 어려운 사용자가 있을 수 있다. 따라서 제안하는 시스템은 사용자가 쉽게 XML 문서를 작성할 수 있도록 쉬운 문법으로 config 파일을 작성할 수 있는 config 생성 모듈을 지원한다. 사용자는 그림 3과 같이 쉬운 문법을 가진 변환 파일을 작성하고 config 생성 모듈을 동작시켜 그림 4와 같은 config 파일을 생성할 수 있다.

```
TestHbaseTable
##5
!!PrimaryInfo
@@ID,5
@@Name,3
!!Desc
@@Description,6
@@Sym,1
!!AdditionalInfo
@@Xref,4
@@BaseLoc,2
```

▶▶ 그림 3. Config 변환 파일

```
-<hbase-table>
   <table-name>TestHbaseTable</table-name>
   <rowkey-position>5</rowkey-position>
  -<columnfamilies>
   -<columnfamily name="PrimaryInfo">
    -<columns>
       <column name="ID" position="5"/>
       <column name="Name" position="3"/>
     </columns>
    </columnfamily>
   -<columnfamily name="Desc">
    -<columns>
       <column name="Description" position="6"/>
       <column name="Sym" position="1"/>
     </columns>
    </columnfamily>
   -<columnfamily name="AdditionalInfo">
    -<columns>
       <column name="Xref" position="4"/>
       <column name="BaseLoc" position="2"/>
     </columns>
    </columnfamily>
  </columnfamilies>
</hbase-table>
```

▶▶ 그림 4. Config 파일

Ⅳ. 결론

본 논문에서는 임상 및 바이오 정보 통합 데이터베이스를 위한 범용적 임포팅 시스템을 설계하고 구현하였다. 제안하는 시스템을 통해 향후 다양한 형태의 임상 및 바이오 분야 데이터를 손쉽게 통합할 수 있는 방법을 제시하였다. 향후 연구로는 제안하는 시스템을 이용하여 통합 데이터베이스를 구축하고 Mahout이나 R과 같은 빅데이터 분석 도구와 연동하여 질병 중심의 임상 및 바이오 정보 연관 분석을 수행하는 빅데이터 분석 시스템을 구현할 것이다.

■ 참 고 문 헌 ■

[1] D. Howe, M. Costanzo, P. Fey, T. Gojobori, L. Hannick, W. Hide, D. P. Hill, R. Kania, M. Schaeffer, S. St Pierre, S. Twigger, O. White, S.. Y. Rhee, "Big data: The Future of Biocuration", Nature, Vol.455, pp.47-50, 2008.
[2] http://www.hprd.org/
[3] http://www.ncbi.nlm.nih.gov/
[4] http://www.genome.jp/kegg/
[5] http://www.open-emr.org/