

질병 및 단백질 상호 작용 분석을 위한 가시화 시스템의 설계 및 구현

Design and Implementation of a Visualization System for Analyzing Disease and Protein Interaction

박준호, 유석종, 임종태, 이지희, 포미미, 김미경, 김현주, 조미림, 류제운*, 김학용*, 유재수
 충북대학교 정보통신공학과, 충북대학교 생화학과*

Junho Park, SeokJong Yu, JongTae Lim, JiHee Lee, Bao WeiWei, MiKyoung Kim, HyunJu Kim, MiRim Jo, JaeWoon Ryu, HakYong Kim, JaeSoo Yoo
 Chungbuk National University

요약

최근 웹 서비스 기술을 이용하여 질병이나 단백질과 같은 바이오 데이터의 분류 및 제공하는 것과 관련된 연구들이 활발하게 진행되고 있다. 웹 서비스 기반 바이오 데이터 서비스에 대한 연구는 데이터베이스의 구축을 중심으로 보고되고 있으나 이를 기반으로 한 분석 응용 도구에 대한 연구는 미미한 실정이다. 이에 따라 본 논문에서는 통합 바이오 데이터베이스를 구축하고 이를 활용한 가시화 시스템을 설계하고 구현한다. 본 시스템을 이용하여 특정 질병에 관련된 단백질 정보를 신호 전달 경로 네트워크상에 가시화하여 연구자에게 직관적으로 주요 단백질에 대한 네트워크내의 위치 정보를 제공한다.

I. 서론

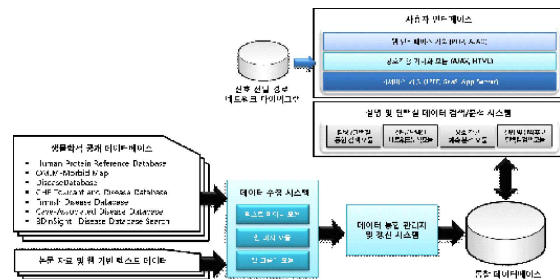
생명 현상에 대한 연구가 활발해짐에 따라 그 실험 결과의 분석을 위하여 다양한 접근 방법이 적용되고 있다. 생명 공학 분야에서의 데이터 분석 방법은 가정을 세우고 많은 시간과 노력으로 실험하여 분석했던 과거의 가정 중심 방법으로부터, 정보처리 기술과 데이터 분석 방법의 발전으로 인해 많은 시간과 노력을 줄일 수 있는 데이터 중심 방법으로 옮겨갔다. 이러한 생물학 실험과 컴퓨터 정보 처리를 융합한 모든 연구 분야를 일컫어 바이오인포매틱스(Bioinformatics)라고 한다[1, 2].

질병에 대한 연구는 많은 과학자들의 중요 연구테마이자, 일반인을 포함한 모든 사람들의 큰 관심사이기도 하다. 현재 질병 및 단백질 데이터를 분류하고, 웹을 통해 제공하는 다양한 서비스가 존재한다. 하지만, 데이터베이스의 기본 정보에 대한 탐색 기능만 제공하고 이에 대한 분석 정보를 제공하지 않기 때문에, 해당 질병 및 단백질의 특성을 파악하기 위해 많은 시간이 소요된다. 따라서 효율적인 연구를 위해서 데이터베이스의 구축뿐만 아니라 이러한 데이터를 바탕으로 다양한 분석을 제공하는 시스템이 요구된다.

본 연구에서는 다양한 질병에 대해 관련이 있는 중요 단백질 및 신호 전달 네트워크 정보를 제공하기 위한 통합 바이오 데이터베이스를 구축하고, 이를 활용하여 질병 관련 단백질에 대한 신호 전달 경로 네트워크를 가시화하는 시스템을 설계하고 구현한다.

II. 제안하는 시스템

제안하는 시스템은 질병 및 단백질 정보를 효율적으로 제공하기 위한 통합 데이터베이스와 가시화 시스템으로 구성이 되어 있다. [그림 1]은 제안하는 시스템의 구조를 나타낸다. 기존의 바이오 데이터베이스들은 자신들의 연구 목적에 따르는 데이터 위주로 이루어진 서비스이기 때문에 다양한 연구에 있어서 필요 한 정보의 상호 연관성을 연결하여 정보를 제공하고 분석을 수행하기에 어려움이 있다. 그러므로 데이터의 통합 제공 및 분석을 수행하기 위해서 각 서비스로부터 데이터를 추출해야하며, 제안하는 시스템에서는 [그림 1]과 같이 공개 데이터베이스에 최적화 된 웹 파서와 텍스트마이너를 구현하여 통합 데이터베이스[3]를 구축하였다.



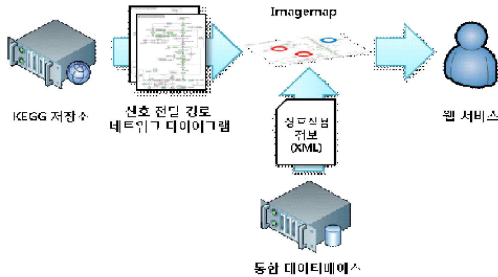
▶▶ 그림 1. 제안하는 시스템의 구조

추출된 데이터는 통합 관리자를 통하여 유일한 식별자를 부여하고 이를 기준으로 통합 데이터베이스를 구성함으로써 데이터를 체계적으로 분류 및 관리한다. 또한, 특정 질병 및 단백질 정보에 대한 여러 정보를 특성에 따

* 이 논문은 2011년 교육과학기술부의 지원(지역거점연구단 육성사업/충북BIT연구중심대학육성사업단)과 교육과학기술부와 한국산업기술재단의 지역혁신인력양성 사업으로 수행된 연구결과임.

라 테이블 단위로 관리하기 때문에 데이터 중복에 따른 불필요한 공간의 낭비 및 검색 시간을 줄였다. 통합 데이터베이스는 XML 형태로 데이터를 제공하여 각 속성별 내용 파악이 용이하도록 하였다.

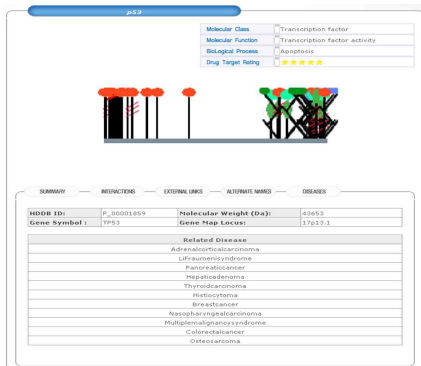
제안하는 가시화 시스템은 질병 관련 단백질의 상호 작용 정보를 신호 전달 경로 네트워크상에 가시화하는 기능을 제공한다. 제안하는 시스템은 KEGG[4]의 분자 또는 유전자 간의 상호 작용하는 대사과정 정보를 나타내기 위한 그래프 기반의 신호 전달 네트워크 다이어그램을 탑재하고, 이를 기반으로 가시화를 수행한다. 통합 데이터베이스에서는 각 질병 및 단백질의 상호 작용 정보를 제공하므로, 질의 요청 시 데이터를 검색하여 상호 작용 정보와 연결 관계에 있는 단백질이 위치하는 신호 전달 네트워크 다이어그램을 로드 및 바인딩 한다. 상호 작용 정보는 다이어그램 상에 HTML 이미지 맵 태깅을 사용하여 표현하므로 별도의 이미지 처리 기술을 활용하지 않아 서버 측의 과도한 연산 비용이 발생하지 않는 장점을 가지고 있다.



▶▶ 그림 2. 상호 작용 정보 가시화 시스템의 동작 구조

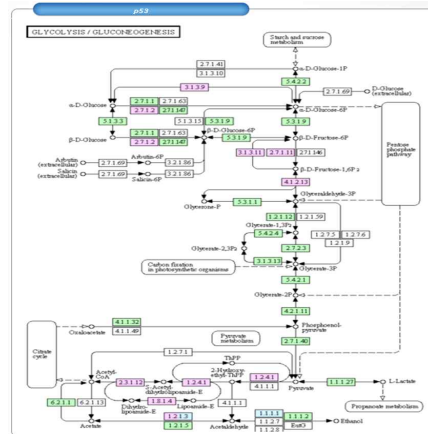
III. 구현 및 예제

본 절에서는 예제를 통해 제안하는 시스템을 유용성을 기술한다. [그림 3]은 통합 데이터베이스 검색 결과를 제공하는 화면을 나타낸 것이다. 상단의 검색 폼에 입력한 검색어에 대한 통합 데이터베이스 검색 결과를 제공하며, 질병 및 단백질에 대한 정보와 상호 작용 정보에 대한 열람이 가능하다. 사용자에게 의해 요청 된 질의는 통합 데이터베이스로부터 XML 문서를 결과로 반환한다. 서비스 페이지는 XML 데이터를 처리하여 동적으로 페이지를 생성한다.



▶▶ 그림 3. 통합 데이터베이스 검색 결과

[그림 4]는 암과 관련된 단백질 p53에 대한 가시화 분석을 수행한 결과이다. 제안하는 시스템은 사용자가 검색한 질병과 관련된 단백질에 대해, 연관된 신호 전달 네트워크를 표시하며 이중에 질병 관련 중요 단백질이 어디에 위치하는지를 직관적으로 가시화 할 수 있는 시스템이다. 이를 통해 연구자는 질병과 관련 단백질의 신호 전달과 정을 가시화함으로써 질병의 기전에 대한 원인 단백질 작용과 관련 단백질들의 조절 기작을 이해할 수 있을 것으로 기대된다.



▶▶ 그림 4. 질병관련 단백질의 상호 작용 가시화

IV. 결론

본 연구에서는 통합 바이오 데이터베이스를 구축하고 이를 활용한 가시화 시스템을 설계하고 구현하였다. 이를 위해, 기존 바이오 데이터베이스를 분석하여 다양한 속성을 XML 데이터 구조로 정리하고, 웹퍼서와 텍스트 마이닝 도구를 이용하여 통합 데이터베이스를 구축하였으며, 이를 이용하여 특정 질병이나 단백질의 상호 작용 정보를 신호 전달 경로 네트워크상에 가시화함으로써 질병 및 신약 연구의 중요한 단서 정보를 제공한다. 향후 연구로는 구축된 통합 데이터베이스를 활용하여 질병에 작용하는 기전, 단백질, 상호작용 등의 다양한 정보들을 분석하고 제공하여 질병 단위의 분자 네트워크 연구나 질병 사이의 연계성 분석 등에 활용할 예정이다.

■ 참고 문헌 ■

- [1] N. M. Luscombe, D. Greenbaum and M. Gerstein, "What is Bioinformatics? A proposed definition and overview of the field," *Methods of Information in Medicine*, vol.4(4), pp.346-358, 2001.
- [2] A. M. Lesk, "Introduction to Bioinformatics," Oxford University Press, United Kingdom, pp.2-20, 2002.
- [3] M. Yeo, Y. Lee, K. Roh, H. Park, H. Kim, T. Kang, H. Kim and J. Yoo, "Design and Implementation of a Directory System for Disease Services," *International Journal of Contents*, vol.6(1), pp.59-64, 2010.
- [4] KEGG: Kyoto Encyclopedia of Genes and Genomes, <http://www.genome.jp/kegg/>