컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템의 설계 및 구현

Design and Implementation of a Data Analysis System for Diseases and Protein based on Components

박준호, 여명호*, 이지희, Li He, 강광구, 권현호, 이진주, 이효준, 임종태, 장용진, Bao WeiWei, 김미경, 강태호**, 김학용, 유재수
충북대학교, 국방과학연구소*, 매크로임팩트(주)**

Park, Jun-Ho, Yeo, Myung-Ho*, Lee, JiHee, Li He, Kang, GwangGoo, Kwon, Hyun-Ho, Lee, JinJu. Lee HyoJoon, Lim, JongTae, Jang, Yong-Jin, Bao WeiWei, Kim, MiKyoung, Kang, TaeHo, Kim, HakYong, Yoo, JaeSoo
ChungBuk National Univ.,
Agency for Defense Development*,
MacroImpact Inc.**

요약

최근 질병 분석 및 신약을 개발하기 위한 단백질에 대한 연구는 생명 공학의 큰 테마 중 하나이다. 질병 및 단백질 데이터를 분석하기 위한 연구는 대용량의 데이터 처리를 요구하기 때문에 과거 실험을 통해 접근하던 방식에서 벗어나 최근 IT 기술의 결합을 통해 다양한 실험 데이터를 공유하고, 연계함으로써 연구를 가속화하고 있다. 하지만 생명 공학 연구자에게 있어서 IT 지식을 기반을 둔 단백질 분석 도구를 다루는데 많은 어려움이 있다. 이러한 문제를 해결하고자, IT 연구자와 생명 공학 연구자의 협업을 통한 데이터 분석 도구를 개발이 폭넓게 시도되고 있지만, 연구자 간의 협업을 도울 수 있는 통합 인프라는 전무한 실정이다. 본 논문에서는 IT 연구자와 생명 공학 연구자의 협업을 위한 인프라로서 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템을 설계하고 구현한다.

I. 서론

생명 현상에 대한 연구가 활발해 짐에 따라 그 실험 결과의 분석을 위하여 다양한 접근 방법이 적용되고 있다. 생명 공학 분야에서의 데이터 분석 방법은 가정을 세우고 많은 시간과 노력으로 실험하여 분석 했던 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 일컬어 바이오인포매틱스(BioInformatics)라고 한다[1].

인간의 질병 및 단백질에 대한 연구는 많은 과학자들의 중요 연구테마이자, 일반인을 포함한 모든 사람들의 큰 관심사이기도 하다. 현재 질병 데이터 및 단백질 데이터의 분석 기능을 제공하는 다양한 서비스가 존재한다. 하지만, 서로 다른 목적을 위해서 개발 된 데이터 분석 서비스이기 때문에 모든 생명 공학 연구자들의 요구 사항을 충족시키는 것은 불가능하다. 뿐만 아니라, 연구 목적에 부합하는 새로운 데이터 분석 도구를 개발하기 위한 IT 연구자와 생명 공학 연구자의 협업을 지

* 이 논문은 2010년 교육과학기술부의 지원(지역거점연구단육성사업/충북BIT연구중심대학육성사업단)과 교육과학기술부와 한국산업기술재단의 지역혁신인력양성 사업으로 수행된 연구결과임.

원하는 IT 인프라의 부재는 생명 공학 연구의 발전을 저해하는 요소로 작용하기도 한다.

따라서 연구자 간의 협업을 통해 데이터 분석 도구 개발을 지원하는 통합 인프라가 필요하다. 이러한 통합 인프라를 이용해 연구자들의 요구 사항을 충족시키는 데이터 분석 도구의 개발을 함으로써 질병 및 단백질 데이터에 대한 고차원적인 분석과 연구에 드는 시간적, 경제적 손실을 막아 질병의 연구에 있어서 큰 성과를 얻는데 도움이 된다.

본 논문에서는 IT 연구자와 생물학자의 협업을 위한 인프라로서 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템을 설계하고 구현한다.

Ⅱ. 제안하는 시스템

제안하는 시스템은 Java 기반의 데이터 분석 컴포넌트라이브러리와 네트워크 분석 구성 도구 및 컴포넌트 구동 엔진을 포함한 데이터 분석 엔진으로 구성되어 있으며, 사용자가 필요로 하는 분석 컴포넌트 개발을 지원하기 위해 필요한 API를 제공한다. 제안하는 시스템은 각 서브시스템과 클라이언트 측과의 데이터 통신 기능을 제공하며 SOAP/HTTP 프로토콜을 통해 통신을 수행한다.

제안하는 시스템의 컴포넌트 라이브러리는 데이터 분석 도구에서 제공하는 컴포넌트의 관리를 수행한다. 기본적으로 컴포넌트는 Java를 기반으로 하여 개발되며, 컴포넌트들은 바이오 데이터를 실질적으로 정제 및 분석하는 역할을 수행한다. 컴포넌트 개발자는 자신이 개발 완료 한 컴포넌트를 웹 사이트를 통해 시스템에 등록하는 것이 가능하다. 컴포넌트 라이브러리에서는 기본 제공 컴포넌트뿐만 아니라 개발자에 의해 새롭게 시스템에 등록 된 컴포넌트들의 목록을 XML 데이터로 제공하며, 데이터 분석 엔진의 네트워크 분석 구성 도구에서는 등록 된 컴포넌트 리스트를 로드 및 바인딩 하여 사용자들이 즉시 사용 할 수 있도록 한다.

데이터 분석 엔진은 사용자에 의해 데이터 분석 모델을 생성하고, 모델에 따라 컴포넌트를 구동함으로써 데이터를 분석하는 기능을 수행한다. 생명 공학자들이 기존의 분석 도구를 다루는데 어려움을 가지고 있는 것을 고려하여, 그림을 그리듯이 네트워크 분석 모델을 설계할 수 있도록 웹 기반 그래픽 인터페이스를 제공한다. 사용자는 자신이 분석을 하고자 하는 데이터 업로드 컴포넌트를 비롯하여 바이오 데이터 분석 및 정제 컴포넌트, 결과 데이터 저장 및 시각화 컴포넌트 등을 캔버스에 배치하고, 라인 툴 기능을 이용해 컴포넌트 간의 파이프라인 구조로 연결하여 분석 모델을 작성한다. 뿐만 아니라 컴포넌트에서 데이터 분석 시 필요한 매개변수 및 각 분석 단계별 완료 결과 확인 여부를 설정하는 것이 가능하다. 작성 완료 된 분석 모델은 네트워크 분석 도구에서 XML데이터로 생성 되어 데이터베이스 상에 저장된다. 사용자가 데이터 분석을 실행 할 경우, 컴포넌트 구동 엔진이 작동하여 데이터 분석을 수행한다. 컴포넌트 구동 엔진은 Java 기반의 컴포넌트를 별도의 프로그램 설치가 없이 웹상에서 즉시 수행이 가능하도록 JavaWebStart 기술을 이용하여 동작한다. 그러므로 데이터 분석 모델은 서버 측에서 생성하지만, 실질적인 연산은 클라이언트 측에서 수행하게 되므로, 고차원의 데이터 처리로 인한 서버 측의 과도한 연산 비용이 발생하지 않는 장점을 가지고 있다.

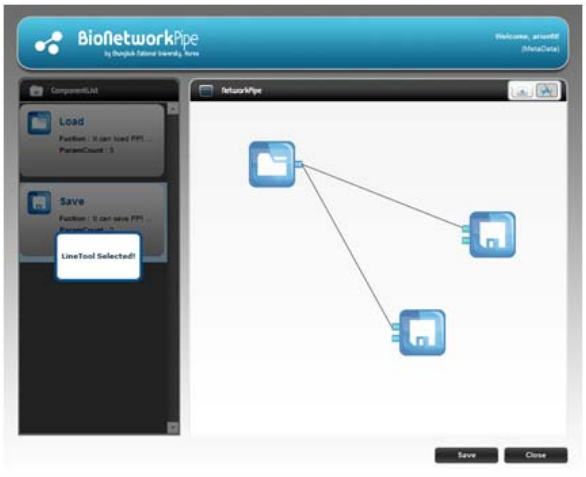
▶▶ 그림 1. 제안하는 시스템의 구조

Ⅲ. 구현 및 예제

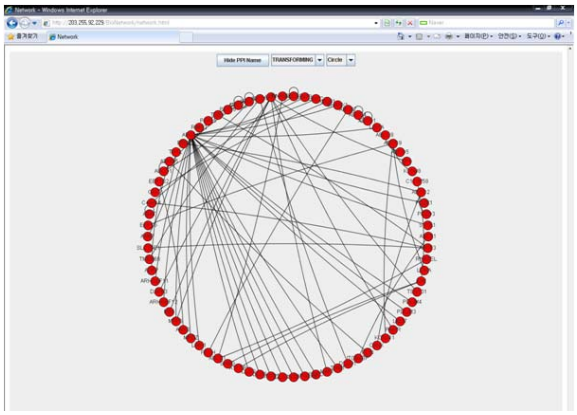
본 절에서는 구현 예제를 통해 제안하는 시스템을 유용성을 기술한다. [그림 2]는 바이오 데이터 분석도구를 이용하여 데이터 분석 모델을 생성하는 과정을 나타낸 것이다. 컴포넌트 라이브러리에서 제공하는 등록 된 컴포넌트 리스트 정보를 이용하여 컴포넌트 리스트에는 사용가능한 컴포넌트의 리스트가 생성된다. 컴포넌트를 드래그 앤 드롭 방식을 이용하여 캔버스에 배치한 후, 데이터의 분석 및 분석 결과 간의 다중 분석에 대한 요구 사항을 라인 툴을 이용하여 정의한다. 구성 된 분석

모델은 최종적으로 XML 데이터로 생성되고, 컴포넌트 구동 엔진을 통해 Java 컴포넌트를 이용하여 바이오 데이터 정제 및 분석을 수행한다.

[그림 3]은 바이오 데이터 분석 시스템을 이용하여 분석 완료 된 데이터를 시각화한 결과이다. 제안하는 시스템은 Java 기반 컴포넌트 구동 엔진을 이용하여 네트워크 분석 및 분석 결과 데이터 파일의 다운로드와 시각화 기능을 별도의 프로그램 설치 없이 웹상에서 즉시 수행 하는 것이 가능하다. 웹 기반 데이터 시각화 도구는 분석 결과 데이터를 바탕으로 네트워크의 시각화를 수행하며, 다양한 레이아웃의 변경 및 줌기능, 라벨링 등의 기능을 제공한다.



▶▶ 그림 2. 컴포넌트 기반의 데이터 분석도구



▶▶ 그림 3. 바이오 데이터 분석 결과를 이용한 웹 기반 데이터 시각화 기능

Ⅳ. 결론

본 논문에서는 IT 연구자와 생물학자의 협업을 위한 인프라로서 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템을 설계하고 구현하였다. 제안하는 시스템은 Java 기반의 데이터 분석 컴포넌트 라이브러리와 데이터 분석 엔진으로 구성되어 있으며, 제공하는 컴포넌트 외에도 추가로 사용자가 필요로 하는 컴포넌트의 탑재를 위해 필요한 API를 제공한다. 네트워크 분석 구성 도구는 데이터 분석 컴포넌트를 웹상에서 그래픽 사용자 인터페이스를 이용하여 분석 모델을 작성한다. 컴포넌트 구동 엔진은 XML 데이터로 생성 된 분석 모델을 이용하여 고차원 데이터 처리 및 분석 기능을 수행하여 결과를 제공한다. 제안하는 시스템은 IT기술자와 생물학자들의 협업을 위한 통합 인프라를 제공함으로써, 생물정보학 분야의 연구를 가속화 시킬 수 있다.

▌참 고 문 헌▐

[1] Lesk A. M. "Introduction to bioinformatics", pp.2-20, Oxford Iniversity Press, United Kingdom, 2002

[2] Sugawara, H., Miyazaki, S., "Biological SOAP servers and web services provided by the public sequence data bank, Nucleic Acids Research," Vol.31, No.13, pp. 3836-3839, 2003