DOI : 10.5392/JKCA.2010.10.12.059

컴포넌트 기반의 질병 및 단백질 분석 시스템의 설계 및 구현

Design and Implementation of an Analysis System for Diseases and Protein Based on Components

박준호*, 여명호**, 이지희*, Li He*, 강광구*, 권현호*, 이진주*, 이효준*,
임종태*, 장용진*, Bao WeiWei*, 김미경*, 류제운****, 강태호***, 김학용****, 유재수*
충북대학교 정보통신공학과*, 국방과학연구소**, ㈜ 매크로임팩트***, 충북대학교 생화학과****

Jun-Ho Park(junhopark@cbnu.ac.kr)*, Myung-Ho Yeo(mhyeo@gmail.com)**,
Ji-Hee Lee(ljhljh82@nate.com)*, Li He(lihe0113@nate.com)*,
Gwang-Goo Kang(joyana@nate.com)*, Hyun-Ho Kwon(outjumper@nate.com)*,
JinJu Lee(jinjulee83@gmail.com)*, Hyo-Joon Lee(reverse999@gmail.com)*,
Jong-Tae Lim(jtlim@cbnu.ac.kr)*, Yong-Jin Jang(yjjang85@gmail.com)*,
WeiWei Bao(pomimi1116@hanmail.net)**, Mi-Kyoung Kim(mini48minwoo@nate.com)*,
Jae-Woon Ryu(jwryu@cbnu.ac.kr)****, Tae-Ho Kang(thkang@netdb.cbnu.ac.kr)***,
Hak-Yong Kim(hkkim@cbnu.ac.kr)****, Jae-Soo Yoo(yjs@cbnu.ac.kr)*

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 질병 및 단백질 분석에 대한 연구를 가속화하고 있다. 하지만 생명 공학 연구자에게 있어서 IT 지식을 기반으로 하는 단백질 분석 도구를 다루는데 많은 어려움이 있다. 이러한 문제를 해결하고자, IT 연구자와 생명 공학 연구자의 협업을 통한 데이터 분석 도구들이 개발되었다. 그러나 기존 데이터 분석 도구들은 확장성 및 여전히 생명과학자들이 사용하기에 어려운 문제가 있다. 본 논문에서는 기존 기법들의 문제점을 해결하는 컴포넌트 기반 질병 및 단백질 분석 시스템을 설계하고 구현한다.

■ 중심어 : | 바이오정보기술 | 생명공학 | 바이오인포매틱스 | 컴포넌트 | 질병 | 단백질 |

Abstract

The research on protein for the diseases analysis and the new medicines development is one of the most important themes in biotechnology. Since the analysis on diseases and protein needs to handle a large scale of data, we don't use the way to approach it by the experiments anymore. In recent, we have accelerated the research on diseases and protein analysis by sharing and connecting the various experimental data by combining the biotechnology with the IT technology. However, many biotechnology researchers have difficulty in handling the protein analysis tools based on the IT knowledge. In order to solve such problems, data analysis tools through the cooperation between IT researchers and biologists have been developed. However, the existing data analysis tools still have the problems that it is very hard for biologists to extend their functions and to use them. In this paper, we design and implement an effective analysis system for diseases and protein based on components that alleviates the problems of the existing data analysis systems.

■ keyword : | BioInformatics | Component | Disease | Protein | DataAnalysis |

* 본 연구는 2010년 교육과학기술부로부터의 지원(지역거점연구단육성사업/충북BIT연구중심대학육성사업단)과 교육과학기술부와 한국산업기술진흥원의 지역혁신인력양성사업으로 수행된 연구결과임.

접수번호 : #100802-003
접수일자 : 2010년 08월 02일
심사완료일 : 2010년 12월 08일
교신저자 : 유재수, e-mail : yjs@chungbuk.ac.kr

I. 서 론

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 생물정보학 - 바이오인포매틱스(BioInformatics)라고 한다[1][2].

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[3]. 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 지원하는 IT 인프라의 부재는 생명 공학 연구의 발전을 저해하는 요소로 작용하기도 한다. 뿐만 아니라, 기존의 도구들은 IT 친화적으로 개발되어, 주로 사용하게 되는 생명 공학 연구자들이 쉽게 사용 할 수 없다는 단점도 가지고 있다. 그러므로 이러한 문제를 해결하기 위해 생명 공학 연구자가 쉽게 접근 할 수 있는 데이터 분석 도구 개발 및 실제 데이터 분석을 지원하는 통합 인프라를 개발하는 것이 시급하다. 이러한 통합 인프라를 이용하여 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 생명 공학 연구자의 협업을 위한 인프라로서 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템을 설계하고 구현한다. 바이오 데이터 정제를 수행하는데 필수적으로 필요로 하는 내장 컴포넌트 외에도 사용자가 추가로 필요로 하는 분석 방법을 정의한 사용자 컴포넌트의 등록 기능을 제공한다. 이를 위해 컴포넌트 라이브러리를 정의하고, 컴포넌트 라이브러리에 따라 사용자 컴포넌트를 개발할 수 있도록 API를 제공하여야 한다. 그리고 컴포넌트를 바탕으로 데이터 분석 모델을 생성하고, 실제 데이터 분석을 수행하기 위한 데이터 분석 엔진을 구성한다. 본 논문에서 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템은 기존 시스템들에 비해 편의성과 효율성을 최대화 한 사용자 인터페이스를 제공하기 위해 Flex3를 활용한 웹 기반의 사용자 인터페이스를 제안한다.

본 논문의 구성은 다음과 같다. 제2장에서는 기존에 제안 된 질병 및 단백질 데이터 분석 시스템에 대한 분석 및 문제점을 기술한다. 제3장에서는 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템의 구조를 기술한다. 제4장에서는 제안하는 시스템의 구현 환경에 대한 설명과 구현 결과 및 시스템 동작 실험 결과를 기술한다. 마지막으로 제5장에서는 결론과 논문의 향후 연구를 기술한다.

II. 관련 연구

현재 생물정보학 분야의 폭발적인 시장 수요 증가가 현실화되고 있으며[4], IT 기술의 발전과 더불어 생물정보 분석과 관련한 다양한 프로그램들이 기하급수적으로 증가하는 추세에 있다. 이러한 생물정보학 분야의 기술 발전에도 불구하고 이를 실제 연구에 활용하기 위해서는 많은 문제점을 가지고 있다. 이러한 문제점들을 해결하고, 생명 공학 연구자들이 보다 손쉽게 연구에 필요한 생명정보 분석 도구들을 효과적으로 활용하기 위한 생명정보 데이터 분석 도구의 개발 및 통합은 현재 생물정보학 분야의 주된 연구 분야 중 하나이다.

유럽의 BT 분야 e-Science 환경 구축에 있어 핵심적인 역할을 수행하고 있는 도구로서 Taverna[5]가 있다. Taverna는 EPSRC(Engineering and Physical Sciences Research Council)에서 추진하는 myGrid 프

로젝트의 한 부분으로서 다양한 생명정보 분석 서비스를 워크플로우 기반으로 통합하여 실행 및 관리를 할 수 있도록 하는 도구로서 [그림 1]에서 보는 바와 같이, 다양한 생명정보 서비스를 구성하기 위한 SCUFL(Simple Conceptual Unified Flow Language) 워크플로우 정의와 특정 SCUFL 워크플로우를 생성하고 편집하기 위한 워크벤치, 그리고 생성 된 워크플로우를 실행하고 단계별 결과를 반환하는 워크플로우 실행 엔진으로 구성된다.

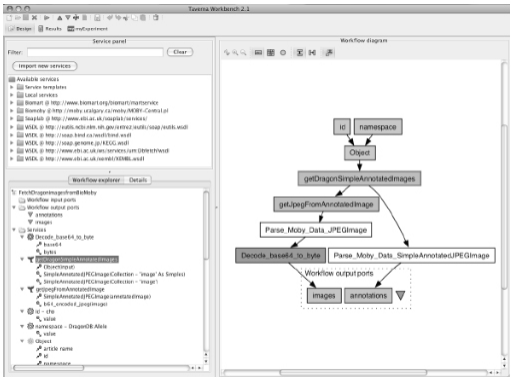


그림 1. Taverna의 수행 화면

미국의 대표적인 도구로서, BioWBI and WEE[6]은 대규모 IT 솔루션 업체인 IBM에서 개발한 Web Services 기반의 생물정보 워크플로우 분석 도구로서, [그림 2] 에서 예시한 바와 같이 워크플로우를 모델링하기 위한 웹 기반의 BioWBI(Bioinformatics Workflow Builder Interface)와 생성된 워크플로우를 실행하고 그 결과물을 반환하고 관리하는 WEE(Workflow Execution Engine)으로 구성된다.

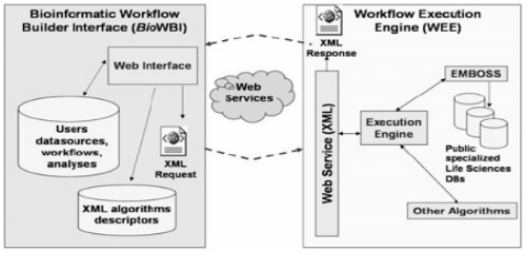


그림 2. BioWBI and WEE 시스템 아키텍처

그 외에도 Wildfire[7]은 2005년도에 발표된 것으로, 클라이언트/서버 구조로 되어 있다. 서버는 기본적으로 배치큐 시스템을 연계한 그리드 컴퓨팅을 통한 EMBOSS 프로그램들을 사용할 수 있다. 이때 사용자는 워크플로우를 클라이언트에서 작성하면 이는 GEL(Grid Execution Language)[8] 언어로 작성되어 서버에서는 GEL 스크립트를 위에서 언급한 스케줄러를 통해 수행하게 된다. Wildfire의 장점은 서버의 다양한 스케줄러를 통해 대량의 생물학 관련 작업을 수행할 수 있다는 점이다. [그림 3]은 GEL을 통한 Wildfire 서버 처리 작업에 대한 개념도이다.

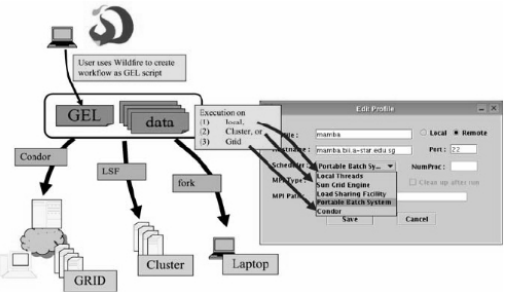


그림 3. GEL을 통한 Wildfire 서버 처리 작업 개념도

앞서 제시한 바와 같이 국외 연구의 경우 워크플로우 기반의 생물정보 서비스 개발 및 통합에 관련된 연구 및 서비스가 매우 활발히 이루어지고 있음을 알 수 있다. 하지만 국내의 경우에는 아직 걸음마 단계에 머물고 있는 현실이다. 한국생명공학연구원 국가생물자원정보관리센터는 2007년 9월 생물학 연구자들이 가장 많이 사용하는 대표적인 생명정보학 분석도구를 통합하여 웹에서 자동적으로 분석 가능한 Biopipe[9] 베타 버전을 공개하였다. BioPipe는 이용자가 필요한 분석도구를 검색하고, 마우스로 분석 파이프라인을 연속적으로 설계할 수 있도록 하여, 실행할 경우 설계된 절차에 따라 자동으로 분석이 가능하도록 하여 보다 손쉽게 연구자들이 원하는 데이터를 분석할 수 있도록 개발되었다. 웹 기반의 소프트웨어로 개발된 BioPipe는 베타 서비스 단계로 웹 인터페이스를 통한 워크플로우 편집은 아직까지는 불편한 상태에 있고, 지속적인 사용자 인터

페이스 개선 작업이 진행 중에 있다.

위에서 언급한 바와 같이, 생명정보 데이터 분석 도구의 개발 및 통합은 활발하게 이루어지고 있지만, 여전히 다음과 같은 문제점을 가지고 있다. 첫째, 기존의 도구들은 확장성이 고려되지 않아서, 생명 공학 연구자들의 요구에 민첩하게 대처하는 것이 불가능하다. 그렇기 때문에 생명 공학 연구자들이 필요로 하는 분석 방법을 적용하기 위해서는 새로운 도구를 개발하여야 하는 부담이 따르게 되고, 이는 질병 및 단백질 데이터에 대한 분석 및 연구에 추가적으로 시간적, 경제적 손실이 발생하게 된다. 뿐만 아니라, 새로운 프로그램을 개발한다고 할지라도 IT 개발자와의 협업의 프레임워크의 부재는 이 마저도 쉽지 않게 만든다. 둘째, 기존의 도구들은 IT 친화적으로 개발이 되었기 때문에, IT 분야, 특히 프로그래밍 분야에 익숙하지 않은 생명과학자들이 사용하기에는 많은 어려움이 있다. 스크립트를 입력하여 분석을 수행하는 프로그램의 경우, 사용 방법의 학습을 위해 수많은 시간이 소요된다. 위에서 언급한 문제들로 인해 생물정보학의 연구는 예상보다 쉽게 발전하지 못하고 있다. 이러한 문제점을 해결하기 위해, 본 논문에서는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템을 제안한다.

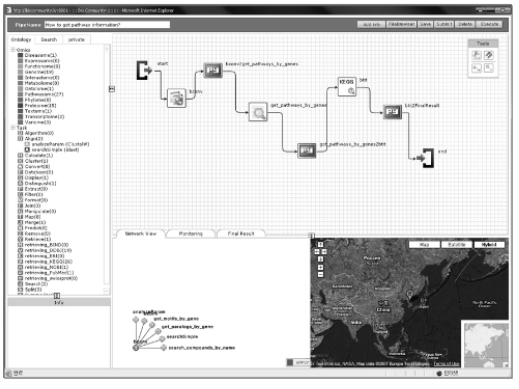


그림 4. BioPipe의 실행 화면

Ⅲ. 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템

본 논문에서 앞서 설명한 기존의 바이오 데이터 분석 도구들의 문제점을 해결하는 새로운 컴포넌트 기반의 질병 및 단백질 데이터 분석시스템을 제안한다. 제안하는 시스템은 기존의 바이오 데이터 분석 도구들과 달리 생명공학자들의 요구 사항을 쉽게 충족시킬 수 있는 기반 인프라를 제공한다. 본 시스템은 생명공학자가 요구하는 다양한 데이터 분석 방법을 구현한 핵심 컴포넌트(자바의 메서드 개념)를 탑재하여, 해당 컴포넌트를 바탕으로 데이터 분석 모델을 생성한 후, 실제 데이터를 정제 및 분석하는 메커니즘을 가진다.

1. 시스템 구성도

본 논문에서 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템은 별도의 완벽한 프로그램을 제공하는 것이 아닌 데이터 정제 및 분석을 수행하는 핵심 컴포넌트만을 개발 및 등록 과정을 거쳐 이를 손쉽게 사용할 수 있는 기반 인프라를 제공한다. [그림 5]는 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템의 구성도를 보여준다. 제안하는 바이오 데이터 분석 시스템의 가장 큰 특징은 바이오 데이터 분석 수행 프로그램을 개발함에 있어서 한 가지 목적이나 분석 방법에 중점을 둔 프로그램을 개발하는 것이 아닌, 핵심 컴포넌트만 손쉽게 개발하여 동작할 수 있도록 하는 기반 인프라를 제공하는 점이다.

제안하는 시스템은 데이터 분석 컴포넌트 라이브러리와 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. 또한 제안하는 시스템은 메인 시스템과 클라이언트 측과의 데이터 통신 기능을 제공하며, SOAP/HTTP 프로토콜을 통해 통신을 수행한다.

데이터 분석 컴포넌트 라이브러리는 데이터 분석을 위해서 필수적으로 필요로 하는 내장 컴포넌트와 사용자가 필요에 의해서 추가로 개발 및 등록을 완료한 컴포넌트에 대한 관리를 수행하며, 컴포넌트 라이브러리에서 관리되는 정보를 기반으로 하여 데이터 분석 엔진은 바이오 데이터 분석 도구를 통하여 사용 가능한 컴

포넌트 정보를 제공한다. 사용자는 바이오 데이터 분석 도구를 이용하여 네트워크 형태의 분석 모델을 생성하고, 컴포넌트 구동 엔진은 생성 된 분석 모델에 따라 실제 데이터 정제 및 분석을 수행하여 결과를 산출한다.

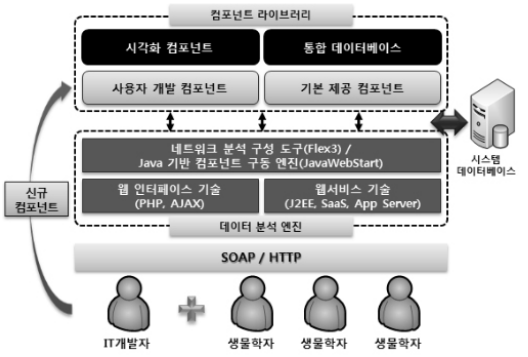


그림 5. 제안하는 시스템의 구성도

2. 컴포넌트 라이브러리

질병 및 단백질 데이터 분석 컴포넌트는 제안하는 시스템에서 기본적으로 제공하는 내장 컴포넌트와 사용자가 필요에 따라 API에 기반을 두어 제작한 사용자 추가 컴포넌트로 구성된다. 제안하는 시스템에서 제공하는 컴포넌트, 즉 내장 컴포넌트는 질병 및 단백질 데이터를 분석하기 위해 필수적으로 필요로 하게 되는 필터링, 파싱, 집계 등의 기능을 수행한다. 질병 및 단백질 분석 시스템은 IT개발자가 쉽게 컴포넌트를 제작하고 운용할 수 있도록 Java 프로그래밍 언어에 기반하며, 기존의 컴포넌트와 사용자 컴포넌트가 유기적으로 연동하여 데이터 분석을 진행하기 위한 API를 제공한다. [표 1]은 사용자에게 제공되는 주요 내장 컴포넌트의 질병 및 단백질 데이터 분석 기능을 보여준다.

컴포넌트 라이브러리는 내장 컴포넌트 및 사용자가 등록한 추가 컴포넌트를 관리하는 기능을 수행한다. 사용자가 추가 컴포넌트를 등록 할 경우, 컴포넌트를 검사하여 컴포넌트에 파라미터, 결과 데이터의 형태 등 API에 기반을 둔 유효한 컴포넌트 인지를 검증한다. 컴포넌트 라이브러리는 검증을 마친 유효한 컴포넌트에 대한 등록 의 수행 및 동적으로 XML 컴포넌트 정보를 생성하여, 이를 데이터 분석 엔진에 제공함으로서 바이오 데이터 분석 도구에서 사용자가 분석 모델을 생성하고 실제 데이터 정제를 수행하기 위한 기반 정보를 제공하는 역할을 수행한다.

표 1. 주요 내장 컴포넌트의 기능

컴포넌트 명	내 용
불러오기	PPI 등의 정보가 담긴 기반 데이터 불러오기
저장하기	각 분석 도구에 따라 분석 된 결과를 저장
데이터 정제	PPI 데이터에서 중복 데이터 정제
가중치 추출	전체 네트워크에서 주어진 가중치에 해당하는 네트워크 추출
데이터 통계	전체 네트워크에서 특정 단백질의 통계 자료 그래프 생성 및 도식화
통계 분석	통계 컴포넌트를 통해서 분석 된 자료를 그래프로 생성
시각화	그래픽 도구를 이용한 단백질 네트워크의 시각화
데이터 검색	단백질 네트워크 검색
통합 데이터베이스	NCBI, UniProt 등의 외부 통합 데이터베이스에서의 데이터 로드

3. 데이터 분석 엔진

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 구현한다.

3.1 바이오 데이터 분석 컴포넌트 구성 도구

바이오 데이터 분석 컴포넌트 구성 도구는 크게 컴포넌트 등록 모듈과, 동적인 인터페이스를 제공하기 위한 드래그 앤 드롭 모듈, XML 데이터를 생성하고 저장하기 위한 모듈로 구성된다. [그림 6]은 제안한 데이터 분석 컴포넌트 구성 도구의 구조를 나타낸다. 컴포넌트 등록 모듈은 컴포넌트 라이브러리에서 제공하는 컴포넌트 정보를 불러온 후, 파싱 된 데이터를 기반으로 컴포넌트 리스트에 각 컴포넌트 위젯을 동적으로 생성하고 등록한다. 드래그 앤 드롭 모듈은 등록 된 각 컴포넌트 위젯에 드래그와 드롭 기능을 제공하여, 캔버스 상에서 자유롭게 배치 및 연결하는 기능을 수행하게 된

다. 마지막으로 XML 로드, 생성 및 저장 모듈은 사용자가 생성한 데이터 분석 모델을 분석하여 동적으로 XML 데이터를 생성하고 데이터베이스에 저장하는 기능을 수행한다. 뿐만 아니라 컴포넌트 라이브러리와 주기적인 통신(HttpService)을 통해 XML 형태의 컴포넌트 정보를 동적으로 불러오기를 수행하여 컴포넌트 등록 모듈이 수행할 수 있는 기반 정보를 제공한다.

그림 6. 데이터 분석 컴포넌트 구성 도구의 구조

3.2 Java 컴포넌트 구동 엔진

사용자가 생성한 네트워크 분석 구성 정보들은 [그림 7]과 같이 XML 데이터로 변환하여 별도의 데이터베이스에 유지한다. 편집 정보 XML에는 편집 정보의 아이디, 생성자, 생성일과 같은 메타데이터 뿐만 아니라, 등록 된 컴포넌트의 정보와 아이디 등의 정보가 기록되어 있다. 이렇게 저장된 XML 데이터는 Java 기반 컴포넌트 구동엔진에서 로드되어 분석을 요청한 사용자의 컴퓨터에서 수행된다. 컴포넌트 구동엔진은 Java기반의 컴포넌트들을 별도의 프로그램의 설치가 없이 웹을 통해 즉시 수행 가능하도록 JavaWebStart[10] 기술을 이용하여 동작한다. 그러므로 데이터 분석 모델은 서버 측에서 생성이 되지만, 실질적인 연산은 클라이언트 측에서 수행하게 되므로, 고차원의 데이터 처리로 인한 서버 측의 과도한 연산 비용이 발생하지 않는 장점을 가지고 있다.

```
<?xml version="1.0" encoding="utf-8" ?>
<Bio_Network_Pipe_IDX>BNP_20091113102424</Bio_Network_Pipe_IDX>
<Bio_Network_Pipe_Owner>arionfit</Bio_Network_Pipe_Owner>
<Bio_Network_Pipe_Subject>바이오네트워크파이프구제</Bio_Network_Pipe_Subject>
<Bio_Network_Pipe_Regdate>20091113 10:24:24</Bio_Network_Pipe_Regdate>
...
<Bio_Component_List>
    <DataLoad>
        <CLevel>1</CLevel>
        <CIDX>C_20091113112458_1</CIDX>
        <Regdate>20091113 11:24:58</Regdate>
        ...
    </DataLoad>
    <Filter_Intersection>
        <CLevel>2</CLevel>
        <CIDX>C_20091113112458_2</CIDX>
        <Regdate>20091113 11:28:13</Regdate>
        ...
    </Filter_Intersection>
    ...
</Bio_Component_List>
```

그림 7. 생성 된 XML 형태의 분석 모델 데이터

Ⅳ. 제안하는 시스템의 구현 및 예제

본 장에서는 제안하는 시스템의 구현 환경과 제안하는 시스템의 구현 결과 및 예제에 대해여 기술한다.

1. 구현 및 동작 환경

제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템은 Windows XP 시스템을 기반으로 한 Apache 2.2.11 서버 환경에서 J2EE 1.4와 J2SDK 1.5를 이용하여 구현하였으며, 데이터베이스 관리 시스템으로 MySQL 5.0을 사용하였다. 또한, 동적인 웹 기반 사용자 인터페이스의 제공하기 위해 데이터 분석 컴포넌트 구성 도구는 Flex3를 이용하여 개발하였고, 웹페이지 상에서 저장하는 데이터를 효율적으로 처리하기 위해 AJAX(Asynchronous Javascript and XML)와 PHP를 이용하여 서비스를 구현하였다.

실제로 데이터 분석을 수행하는 사용자는 Windows XP플랫폼 환경에서 동작 테스트를 진행하였다.

2. 구현 결과 및 예제

[그림 8]은 본 논문에서 제안하는 컴포넌트 기반의 질병 및 단백질 데이터 분석 시스템의 화면 구성도를 보여준다. 서비스를 이용하는 사용자가 IT 분야에 친숙하지 않은 생명공학자라는 점을 고려하여 사용자의 편의성과 효율성을 최대화하기 위해서 컴포넌트 위젯 (데이터 분석을 수행하게 되는 작은 단위)의 동적 배치 기능, 드래그와 드롭 기능, 선 도구를 이용한 컴포넌트 간

의 연결 기능을 제공한다. 이를 통해 직관적인 인터페이스를 제공함으로써 기존의 도구를 이용함에 있어서 발생하였던 어려움을 해결하였다.

그림 8. 데이터 분석 시스템의 화면 구성도

그림에서 ①은 데이터 분석 컴포넌트 구성 도구에서 사용 가능한 컴포넌트 리스트를 보여준다. 이는 컴포넌트 라이브러리와 통신을 수행하여, 데이터 분석에 필수적으로 필요로 하는 내장 컴포넌트, 뿐만 아니라, 필요에 따라 사용자에 의해 구현 및 등록되어 컴포넌트 라이브러리의 검증이 완료 된 사용 가능한 컴포넌트들을 표시한다. ②는 컴포넌트를 이용하여 실제 데이터 분석 도구를 구성을 수행하는 배경이 되는 캔버스이다. 사용자는 컴포넌트 리스트에 사용하고자 하는 컴포넌트 위젯을 드래그와 드롭 방법을 이용하여 캔버스에 위치시키고, 선도구를 이용하여 컴포넌트 간의 연결을 수행하는 것이 가능하다. 선 도구로 연결 된 앞 단의 컴포넌트에서 정제 및 분석 된 결과 데이터는 뒷 단의 컴포넌트에서 입력으로 사용하게 됨을 의미한다. ③은 상태 표시줄로, 수행 중인 내용을 스크립트로 표현함으로써, 사용자는 이를 이용하여 현재 상태에 대한 직관적인 이해가 가능하다. 이러한 도구들을 이용하여 분석 도구를 구성한 후, 확인 버튼을 선택하면, 구성 된 데이터 분석 도구 데이터는 데이터베이스에 저장이 되고, 이를 기반으로 하여 실제 데이터 정제 및 분석을 수행하게 된다.

[그림 9]는 저장 된 데이터 분석 도구 데이터를 기반으로 하여 사용자의 로컬 컴퓨터에서 데이터 정제 및 분석을 수행하는 모습이다. 앞서 저장 된 분석 도구의 데이터는 웹사이트에서 확인이 가능하며, 생성한 데이터 분석 모델을 수행하는 것도 가능하다. 데이터 정제 및 분석을 수행함과 동시에 서버에서는 우선적으로 JavaWebStart 기술을 이용하여 사용자의 컴퓨터가 데이터 분석을 수행하기 위한 기반 환경을 구축을 한 후, 데이터 분석 모델을 확인하여 데이터를 분석하는데 필요한 실제 자바 컴포넌트를 사용자의 컴퓨터로 다운로드하여 분석을 수행 할 준비를 한다. 그 후, 사용자가 정의한 분석 모델에 따라 데이터 분석 컴포넌트가 순차적으로 수행이 되며, 실제 데이터의 정제 및 분석을 수행하게 된다.

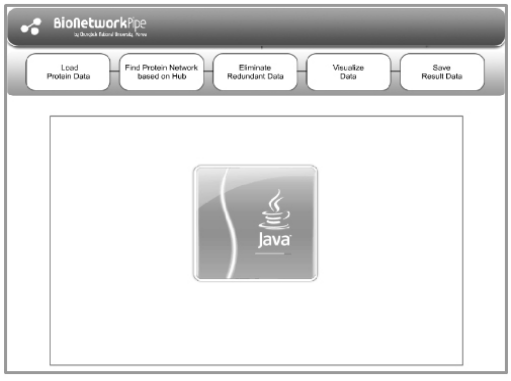


그림 9. 데이터 분석 수행

[그림 10]는 바이오 데이터의 정제 및 수행 결과를 출력한 그림이다. 생성 된 분석 도구 데이터를 기반으로 데이터 분석을 수행한 후, 사용자의 요청에 따라 네트워크의 시각화 및 로컬 컴퓨터에 정제 및 분석 완료 된 데이터를 저장하는 것이 가능하며, 데이터 구성 도구를 구성할 시에, 사용자는 옵션을 설정하므로써 데이터 분석을 수행하는 매 단계 결과를 확인하는 것도 가능하다. 그림에서 보인 데이터 시각화 도구는 분석 결과 데이터를 바탕하는 매 단계 결과를 확인하는 것도 가능하다. 그림에서 보인 데이터 시각화 도구는 분석 결과 데이터를 바탕으로 네트워크의 시각화를 수행하며, 다양한 레이아웃의 변경 및 줌기능, 라벨링 등의 기능을 제공한다.

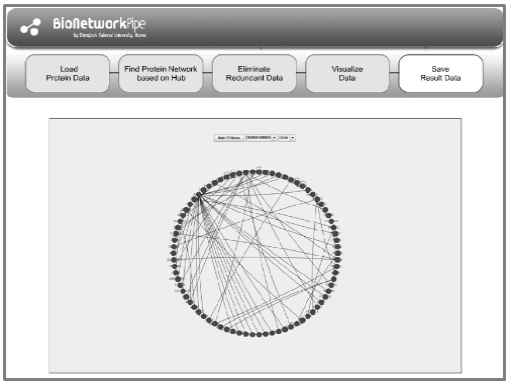


그림 10. 데이터 분석 수행 결과 (시각화)

3. 기존 시스템과의 기능 비교 평가

본 절에서는 제안하는 컴포넌트 기반의 질병 및 단백질 분석 시스템과 기존의 시스템과의 기능 비교 평가를 통해 제안하는 시스템의 우수성을 기술한다. [표 2]는 제안하는 시스템과 기존에 제안 된 시스템의 제공하는 기능의 유/무를 보여준다. 제안하는 시스템은 개발 중인 서버 및 클러스터 기반의 데이터 처리를 제외하고, 외부 웹서비스를 연계한 데이터 처리를 비롯한 선진 시스템에서 제공하는 대부분의 기능을 지원하며, 이에 따른 워크플로우 생성 및 질병 및 단백질 데이터에 대한 처리가 가능하다. 특히 기존의 유사한 기능을 제공하는 선진 시스템에서 제공하지 않았던 통합 데이터베이스 및 처리 결과의 시각화 모듈을 기본 컴포넌트로 제공함으로써 개발 초기의 시스템으로서는 기존의 시스템과 비교하여 수준 높은 기능을 제공한다고 할 수 있다. 뿐만 아니라 기존 대부분의 시스템에서 추가적인 데이터 처리 기법의 확장에 대해서 지원을 하지 않았던 것과 달리 제안하는 시스템에서는 사용자의 요청에 따른 새로운 데이터 분석 컴포넌트의 제작을 지원하기 위한 API를 제공 및 개발 된 컴포넌트를 플러그 인 형태로 사용 할 수 있도록 지원함에 따라 기존의 시스템에 비해 확장성을 높였다. 향후 대규모의 데이터 분석을 수행하고자 하는 연구자를 위하여 개인 PC 단위의 클라이언트에서 수행되던 데이터 분석을 대용량의 서버 및 클러스터에서 수행 할 수 있도록 시스템을 확장할 예정이다.

표 2. 기존 시스템과의 기능 비교 평가

개발 내용	기능	선진 시스템과의 비교			
		Taver -na	Wild fire	Bio pipe	제안하는 시스템
워크플로우 모델링	GUI 기반의 워크플로우 생성	○	○	○	○
	데이터베이스 연계 저장	×	×	×	○
데이터 처리	서버 기반 데이터 처리	○	○	×	×
	클라이언트 기반 데이터 처리	○	○	○	○
	클러스터 기반 데이터 처리 (스케줄링)	△	○	×	×
	외부 웹서비스 제공	○	×	○	○
부가기능	통합 DB 제공	×	×	×	○
	처리 결과 시각화	×	×	×	○
워크플로우 확장	추가 plugin 탑재	○	×	×	○
	확장 API 제공	○	×	×	○
기타	워크플로우 공유	△	×	○	○

V. 결론 및 향후 연구

본 논문에서는 기존 바이오 데이터 분석 도구가 갖는 문제점을 해결한 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 할 수 있을 것으로 기대한다. 향후 연구로는 제한적으로 제공하고 있는 API를 완벽하게 제공하는 것, 뿐만 아니라, 현재 바이오 분야에 국한된 시스템을 다양한 연구 분야로 확장하는 것이다.

참 고 문 헌

[1] A. M. Lesk Introduction to bioinformatics, Oxford Iniversity Press, United Kingdom, 2002

[2] N. M. Luscomebe and G. D, G. M., "What is bioinformatics? A proposed definition and overview of the filed," Methods Inf. Med, Vol.40, pp.346-358, 2001.

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

[4] M. Deng, S. Mehta, F. Sun, and T. Chen, "Inferring domain-domain interactions from protein - protein interactions," Genome Res. Vol.12, No.10, pp.1540-1548, 2002.

[5] http://www.taverna.org.uk/

[6] P. Leo, C. Marinelli, G. Pappadà, G. Scioscia, and L. Zanchetta, "BioWBI: an Integrated Tool for building and executing Bioinformatic Analysis Workflows," Bioinformatics Italian Society Meeting, pp.26-27, 2004.

[7] http://wildfire.bii.a-star.edu.sg/

[8] C. C. Lian, F. Tang, P. Issac, and A. Krishnan, "GEL: Grid Execution Language," Journal of Parallel and Distributed Computing, Vol.65, No.7, pp.857-869, 2005.

[9] http://www.biopipe.net/

[10] http://java.sun.com/javase/technologies/desktop/javawebstart/index.jsp

저 자 소 개

박 준 호(Jun-Ho Park) 정회원

- 2008년 2월 : 충북대학교 정보통신공학과(공학사)
- 2010년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2010년 3월 ~ 현재 : 충북대학교 정보통신공학과 박사과정

<관심분야> : 데이터베이스 시스템, 무선 센서 네트워크, 차세대웹, LMS, LCMS

여 명 호(Myung-Ho Yeo) 정회원

- 2004년 2월 : 충북대학교 정보통신공학과(공학사)
- 2006년 2월 : 충북대학교 정보통신공학과(공학석사)
- 2010년 2월 : 충북대학교 정보통신공학과(공학박사)
- 2010년 2월 ~ 현재 : 국방과학연구소 연구원

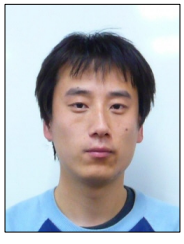
<관심분야> : 메인 메모리 기반 데이터베이스, 시공간 데이터베이스, 무선 센서 네트워크

이 지 희(Ji-Hee Lee) 준회원

- 2006년 2월 : 청주대학교 정보통신공학과(공학사)
- 2009년 2월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 무선 센서 네트워크, 유비쿼터스 컴퓨팅

이 하(Li He) 준회원



- 2008년 9월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 무선 센서 네트워크, 위치 기반 서비스, 데이터베이스

강 광 구(Gwang-Goo Kang) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 2월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 무선 센서 네트워크, 위치 기반 서비스, 데이터베이스

권 현 호(Hyun-Ho Kwon) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 2월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : 데이터베이스, 무선 센서 네트워크, 모바일 디바이스

이 진 주(JinJu Lee) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 2월 ~ 현재 : 충북대학교 정보통신공학과(공학석사)

<관심분야> : 시공간 데이터베이스 시스템, 이동 객체, 무선 센서 네트워크

이 효 준 (Hyo-Joon Lee) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

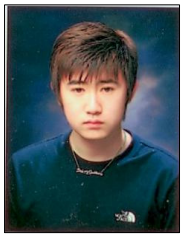
<관심분야> : DB 시스템, 센서 네트워크, 저장시스템, 파일시스템

임 종 태(Jong-Tae Lim) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : DB 시스템, 센서 네트워크, 위치기반서비스

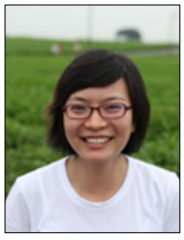
장 용 진(Yong-Jin Jang) 준회원



- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 3월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : DB 시스템, 센서 네트워크, 저장시스템, 파일시스템

포 미 미(WeiWei Bao) 준회원



- 2009년 9월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : DB 시스템, 센서 네트워크, 저장시스템, 파일시스템

김 미 경 (Mi-Kyoung Kim) 준회원

- 2009년 2월 : 충북대학교 정보통신공학과(공학사)
- 2009년 9월 ~ 현재 : 충북대학교 정보통신공학과 석사과정

<관심분야> : DB 시스템, 센서 네트워크, 저장시스템, 파일시스템

류 제 운(Jea-Woon Ryu) 정회원

- 2006년 2월 : 충북대학교 생화학과(이학학사)
- 2008년 2월 : 충북대학교 생화학과(이학석사)
- 2010년 3월 ~ 현재 : 충북대학교 생화학과 박사과정

<관심분야> : 생물정보학, 단백질 상호작용, 신호전이, 시스템생물학

강 태 호(Tae-Ho Kang) 정회원

- 1999년 2월 : 호원대학교 정보통신공학과(공학사)
- 2002년 8월 : 충북대학교 정보산업공학과(공학석사)
- 2007년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2010년 2월 : 충북대학교 전기전자컴퓨터공학부 Post-doc.
- 2010년 3월 ~ 현재 : ㈜ 매크로임팩트 연구원

<관심분야> : 생물정보학, 단백질 상호작용, 신호전이, 시스템생물학

김 학 용(Hak-Yong Kim) 종신회원

- 1985년 2월 : 충북대학교 농화학과(농학사)
- 1987년 2월 : 충북대학교 화학과(이학석사)
- 1994년 5월 : 미국 코네티컷대학교, 분자세포생물학과(이학박사)
- 1998년 3월 ~ 현재 : 충북대학교 생화학과 교수

<관심분야> : 시스템생물학, 신호 전이, 단백질 네트워크, 생체동역학

유 재 수(Jae-Soo Yoo) 종신회원

- 1989년 2월 : 전북대학교컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술원 전산학과(공학석사)
- 1995년 2월 : 한국과학기술원 전산학과(공학박사)
- 1995년 3월 ~ 1996년 8월 : 목포대학교 전산통계학과(전임강사)
- 1996년 8월 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 및 컴퓨터정보통신연구소 교수

<관심분야> : 데이터베이스시스템 정보검색 센서네트워크 및 RFID, 멀티미디어데이터베이스, 분산객체 컴퓨팅