바이오 데이터 분석을 위한 웹 콘텐츠 개발에 관한 연구

송 영 옥*, 최 승 권**, 신 승 수***, 조 용 환****
우송대학교*, (주)애니솔루션**, (주)사이젠텍***, 충북대학교****

A Study on Web Contents Development for Bio Data Analysis

Song Young-Ok*, Choi Seung-Kwon**, Shin Seung-Soo***,
Cho Youn-hwan****
Woosong Univ.*, ANY Solution Co., Ltd**, Cyzentech Co.,Ltd. Lab***, Chungbuk National Univ.****
E-mail : yysong@lion.woosong.ac.kr, skchoi@anysol.com,
shinss@chungbuk.ac.kr, yhcho@cbucc.chungbuk.ac.kr

요 약

현재 많은 분야에서 IT 기반 기술의 급속한 발전을 이용하고 있다. 그 중 계속적인 발전을 보이고 있는 생명공학 분야의 기술과 정보기술의 한 형태로 바이오인포메틱스 분야에 많은 관심이 증가되고 있다. 본 논문에서는 이와 같은 바이오인포메틱스의 분야에서 많은 사용자들이 이용해야하는 생명공학 기술과 데이터베이스 등을 이용하여 웹을 기반으로 하는 콘텐츠를 개발하는데 있어 핵심 기술들을 논하고자 한다. 이와 같은 BIT에 관련된 콘텐츠 개발로 인해 현재 제한되어 있는 연구 환경을 개선하고 비용절감의 효과를 보일 수 있도록 한다.

Abstract

Recently, rapid progress in IT based technology is used in many development areas. We are interested in the bioinformatic part which is the combination of fields of biotechnology and information technology. this thesis discuss core technology for development of web based contents using biotechnology, database, and so on. this development of contents connected with the BIT is useful in improving many research environment and saving working cost.

I. 서론

최근 들어 정보 기술의 발달로 다양한 분야에서 정보기술을 기반으로 하여 콘텐츠 개발 및 제공되고 있다. 그 중 인터넷의 웹 서비스 기술을 기반으로 개발된 콘텐츠 분야를 많이 접할 수 있으며 이로 인해 많은 사용자들이 자료의 공유와 다양한 정보를 쉽게 접함으로서 각자의 연구 활동에 많은 영향을 받고 있다.

본 연구에서 웹 기반으로 바이오 콘텐츠 개발을 하고자 한다. 바이오인포메틱스는 최근 들어 새로이 생긴 학술 분

야로 앞으로도 많은 발전이 이루어 질 수 있는 중요한 분야로 대두되고 있다. 이러한 바이오 데이터를 분석하기 위해 기존의 연구 방법에서 탈피하여 웹을 기반으로 하는 데이터 분석 환경을 구성함으로서 많은 바이오 데이터 연구자들에게 보다 편리한 연구 환경과 다양한 정보를 쉽게 찾을 수 있는 개선된 연구 환경의 필요성이 요구된다.

본 논문에서는 이러한 연구 환경의 개선과 보다 효율적인 분석 시스템의 필요성에 따라 다중 사용자가 이용할 수 있는 바이오 콘텐츠를 개발하는데 있어 필요한 각 요소와 기술에 관하여 연구하고자 한다.

II. 이론적 배경

웹 콘텐츠란 웹 사이트라는 용기에 들어 있는 어떠한 내용물 전체를 지칭하는 것으로 단순하게 사용자들에게 DB형태로 제공되는 텍스트 정보뿐만 아니라 특정 사이트가 가지는 모든 구성요소를 콘텐츠라 할 수 있다.

이미 바이오 데이터를 분석하기 위하여 대표적인 몇 개의 웹 사이트들이 제공되고 있다[4]. 기존에 제공 중인 사이트들은 바이오 데이터를 분석하는데 있어 필요한 데이터베이스를 구성하여 놓고 있으며 이러한 자료를 전 세계적으로 이용할 수 있도록 함이 주요 사이트 제공 이유라 할 수 있다.

표 1. 데이터베이스 제공 URL [1][2]

구분	DB	URL
DNA DB	GenBank	http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html
	EMBL	http://www.ebi.ac.uk/embl
	DDBJ	http://www.ddbj.nig.ac.jp
Protein DB	nr(ExPasy)	ftp://www.expasy.ch/databases/sp_tr_nrdb/
	nr(NCBI)	ftp://ncbi.nlm.nih.gov/blast/db/
	SwissProt	http://www.expasy.ch/sprot/sprot-top.html
	TrEmbl	http://expasy.ch/sprot/sprot-top.html
	PIR	http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml
	GenPept	http://www.ncbi.nlm.nih.gov/Entrez/protein.html
	PDB	http://www.rcsb.org/pdb/index.html
	Genomes	http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html

위의 표 1에서는 주로 연구 활동에 이용되는 DNA DB와 Protein DB에 대해 제공하고 있는 URL을 정리한 것이다[2][3][5]. 이와 같은 사이트에서는 바이오 데이터를 분석하기 위한 전체 콘텐츠 제공을 목적으로 하기보다는 데이터베이스 자료를 활용할 수 있는 텍스트 기반 도구들을 제공하고 있다. 또한 개개인의 필요에 따라 데이터 분석 자료를 보관할 수 있는 방법이 로컬 컴퓨터에 파일 형태로 저장하거나 필요할 때마다 재 검색을 요구하고 있다. 이러한 작업 때문에 분석의 연속성은 기대하기 힘든 실정이다.

따라서 본 논문에서는 데이터 분석을 위한 여러 가지 요소를 컴포넌트 기반으로 개발하고 전체 콘텐츠를 제공할 필요성에 따라 바이오 분야의 콘텐츠를 개발하고자 이에 필요한 각 구성 요소들을 설계하고자 한다.

III. 바이오 콘텐츠 개발

1. 바이오 콘텐츠의 전체 시스템 구조

개발하고자 하는 바이오 콘텐츠의 전체 시스템 구조를 구성하기 위하여 다음 그림 1과 같이 구성할 수 있다.

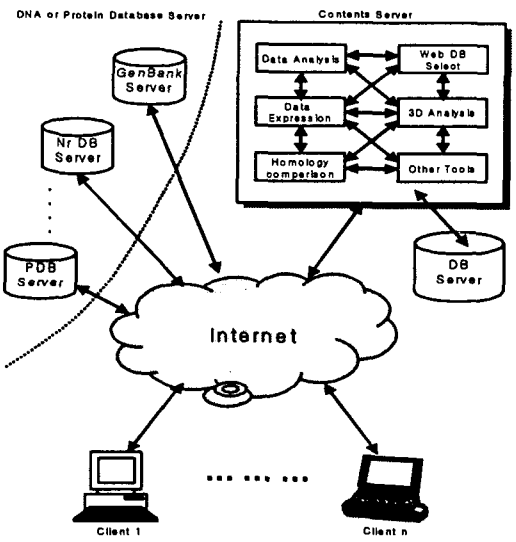

그림 1. 바이오 콘텐츠 제공 시스템 구조

분석에 필요한 DNA 또는 Protein 데이터베이스 서버와 바이오 콘텐츠 제공을 위한 콘텐츠 서버 및 개별 데이터 관리를 위한 데이터베이스 서버를 포함하고 이를 이용하는 클라이언트로 구성된다.

2. 바이오 콘텐츠에 필요한 컴포넌트 구성

기존의 바이오 분야에 활용할 수 있도록 제공되는 사이트들에서 주로 데이터베이스 자료의 활용에 초점을 맞추어 제공한 것에 반해 본 논문에서 바이오 데이터를 분석하기 위하여 연구하고자하는 바이오 웹 콘텐츠에서는 데이터 분

석 설계에 필요한 각종 컴포넌트를 포함하여 다양한 종류의 개발 환경을 제공함으로써 바이오 연구 분야에서 쉽고 편리하게 활용할 수 있는 콘텐츠를 제공하고자 한다.

다음 그림 2는 이와 같은 웹 기반 바이오 콘텐츠를 설계하기 위하여 필요한 여러 가지 컴포넌트들을 표현한 것이다.

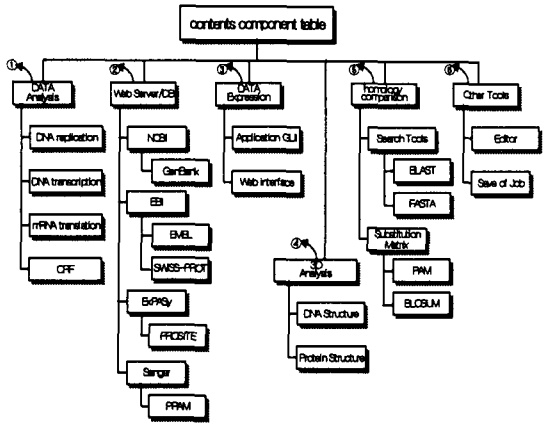

그림 2. 바이오 콘텐츠를 위한 컴포넌트 구성

전체 컴포넌트들은 다음과 같이 세부 요소로 구성되어있다.

① 데이터 기본 분석 컴포넌트 : DNA 데이터를 이용하여 RNA 전사, DNA 번역, ORF 검색 등을 할 수 있는 기본 요소로 텍스트 기반 분석 도구를

② DB 서버 구성 : 데이터 기본 분석 컴포넌트를 통한 데이터를 이용하여 검색하고자 하는 여러 나라와 지역에서 제공하는 데이터베이스를 원활하게 이용할 수 있도록 구성하며 검색 조건 또한 다양하게 주어질 수 있도록 구성되어야하는 컴포넌트

③ 데이터 표현 컴포넌트 : 데이터베이스를 통한 검색이나 3D 표현된 데이터의 결과 등을 시각적이고 판단이 편리하도록 표현하는 방법을 이용하는 컴포넌트

④ 3D 분석 컴포넌트 : DNA, Protein등의 데이터를 일반 텍스트 분석뿐만 아니라 3D 시뮬레이션을 통해 분석할 수 있도록 구성하는 컴포넌트

⑤ 유사성 비교 컴포넌트 : 각종 데이터베이스 서버를 통해 유전자의 유사성을 비교 검색할 수 있는 컴포넌트로 가장 적절한 알고리즘이 적용되어야 하는 컴포넌트

⑥ 기타 컴포넌트 : 데이터를 분석하는데 편리성을 제공하기 위해 데이터 편집기 등을 포함하는 컴포넌트

위와 같은 몇 가지의 컴포넌트 뿐만 아니라 보다 다양하고 유용한 기능을 구성하는 컴포넌트 개발이 계속되어야 한다.

3. 다중 사용자를 위한 웹 데이터베이스 설계

본 논문에서 연구하는 바이오 콘텐츠의 가장 기본 요소라 할 수 있는 것은 보다 다양한 데이터베이스 활용이다. 이러한 데이터베이스 서버들은 아직은 국내에서는 미약한 부분이지만 전 세계적으로 이용되는 몇 개의 데이터베이스 서버를 활용하여 데이터 연구에 활용하는 것이 현재로서는 경제적이며 데이터 분석시간을 단축할 수 있다. 그러나 데이터 분석 결과에 대한 저장 보관의 필요성에 따라 기관내 또는 개인 데이터베이스 구축이 본 콘텐츠에서도 필수로 포함되고 있다.

그리고 개인을 위한 단일 데이터베이스가 아닌 다중 사용자가 활용하면서 동시에 개별적인 데이터베이스 구축을 필요로 하는 경우를 고려하여야 한다. 이러한 데이터베이스는 본 콘텐츠의 이용을 위한 회원등록(①) 후 회원인증(②)을 통하여 전체 바이오 콘텐츠를 이용할 수 있는 권한을 획득한다(③). 그리고 DNA 또는 Protein 데이터베이스 서버를 이용하여 데이터를 검색하고(④) 검색 결과의 HTML 페이지에서 이용 가능한 텍스트 필드를 선택(⑤)하여 이것을 바탕으로 개인별 데이터베이스 구축에 필요한 스키마를 구성(⑥)할 수 있는데, 필드명에 따르는 필드속성을 지정(⑦)하여 개인별 테이블이 생성(⑧)된다. 이렇게 생성된 데이터베이스의 테이블을 이용하여 데이터 분석한 각 결과들을 보관할 수 있다. 이러한 전체 과정을 다음 그림3과 같이 나타낼 수 있다.

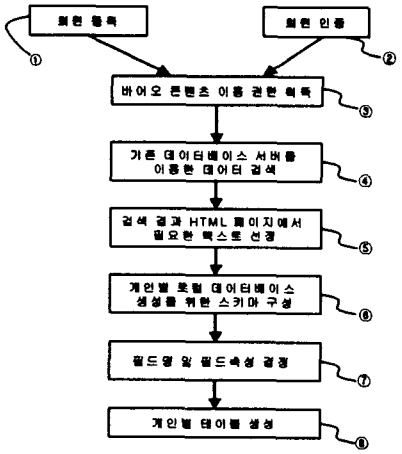

그림 3. 개인 데이터베이스 테이블 생성 단계

4. 개발 언어 선택

기존에 바이오 분야에 대한 분석 시스템들에서 이용하던 Perl 등의 언어를 비롯하여 여러 가지 언어들이 이용 가능하며, 본 논문에서 바이오 콘텐츠 개발을 위해 가장 적절한 언어를 선택하는 것이 커다란 과제로 남아있다. 현재까지 자료 파싱을 위해 주로 이용되고 있던 Perl 또는 Python등은 바이오 분야를 위한 패키지 제공으로 실제 개발 시간을 어느 정도 단축할 수 있다는 장점을 가지고 있는 반면 Java언어는 콘텐츠 개발이 용이한 언어라 할 수 있고 바이오 패키지가 제공되고 있지만 Perl 언어에 비해 아직까지는 미약하다 할 수 있다[2].

바이오 콘텐츠 개발을 위해 이용될 수 있는 언어로는 다음 표2와 같은 종류와 용도로 구분할 수 있다.

표 2. 바이오 콘텐츠 개발을 위한 언어

업무분류	세부항목	언어종류
인터페이스 구성	웹 인터페이스	HTML, JAVA, EJB JS, PHP4, etc
	어플리케이션	VC++, JAVA
자료 파싱	PHP4, Perl, Python, C, C++, JAVA, XML etc.	
데이터베이스 관리시스템	MySQL, ORACLE etc.	

Ⅳ. 결론 및 향후 연구과제

본 논문에서 웹 기반 바이오 콘텐츠를 개발하기 위하여 필요한 구성요소들을 살펴보았다. 이러한 구성요소를 컴포넌트 기반으로 개발하여 제공한다면 관련 분야에 대하여 연구하는 사용자들에게 보다 다양한 분석 시스템을 제공할 수 있을 것이며, 이러한 분석 결과들을 손쉽게 보관하고 재이용할 수 있으므로 효율성을 높일 수 있을 것이다. 이에 따라 앞으로 지속적인 개발로 데이터 분석 시간을 단축할 수 있는 방안과 각종 데이터베이스 서버의 부하를 줄이고 늘어나는 네트워크 사용에도 불구하고 효율적으로 이용할 수 있는 방법이 개발되어야 한다.

참 고 문 헌

[1] 송영옥 "인터넷 기반 바이오 데이터 분석 시스템", 충북대학교, 2003.

[2] Cynthia Gibas & Per Jambeck, "Developing bioinformatics Computer Skills", 2001.

[3] James D. Tisdall, "Beginning Perl for Bioinformatics", O'REILLY, 2002.

[4] http://www.ncbi.nlm.nih.gov

[5] ETRI(IT-InformationCenter), "bioinformatics : Technoogy & Market Analysis", 2001.