웹 서비스 기반 바이오 정보 통합 분석 도구

*최요한 ○ *유성준 **김민경 ***박현석
*세종대학교 컴퓨터공학부 ** 이화여자대학교 공학연구소 *** 이화여자대학교 컴퓨터학과
skyhani@hotmail.com sjyoo@sejong.ac.kr hspark@macrogen.co.kr minkykim@ewha.ac.kr

WSBAT: Web Services based Biodata Analysis Tool

*Yohan Choi○ , *Seongjoon Yoo, **Minkyung Kim, ***Hyun Seok Park
*School of Computer Engineering, Sejong University, **Center for Engineering Research, Ewha University, ***Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 웹 서비스 기술을 이용하여 바이오 데이터 및 데이터 메소드를 제공하는 것과 관련된 연구 들이 진행되고 있다. 웹 서비스 기반 바이오 데이터 서비스에 대한 연구 자료는 시스템 구조 및 API 등을 중심으로 보고되고 있으나 이를 기반으로 한 통합 응용 도구 개발 관련 연구는 미미한 실정이다. 이에 따라 이 논문에서는 웹 서비스 API 등을 이용하여 바이오인포매틱스 연구자들이 이용할 수 있는 데이터 통합, 검색, 브라우징 기능을 제공하는 분석 도구를 개발하였다. 사용자는 이 도구를 이용하여 바이오 데이터 간의 상호연관성을 보다 쉽게 발견할 수 있으며 보다 다양한 검색 결과를 여러 가지 형태로 볼 수 있게 될 것이다.

1. 서 론

바이오인포매틱스 연구에 있어 가장 필요하고 중요하게 여겨지는 자원은 바이오 데이터베이스와 데이터 분석 도구이다. 현재 바이오 데이터베이스를 구축해 놓은 곳으로는 미국 국립생물정보센터의 GenBank[1], 일본의 DDBJ[2](DNA Data Bank fo Japan)와 유럽의 바이오인포매틱스 연구소인 EMBL (European Molecular Biology Laboratory) [3,4] 등이 있다. GenBank, DDBJ, EMBL에서는 각각 350억 개의 뉴클레오타이드 시퀀스 데이터베이스와 3,000만개의 엔트리들을 보유하고 있으며, 그 내용이 계속 늘어나고 있다[2].

이들의 경우, 서열 데이터를 중심으로 매일 상호 교환되고 있기 때문에 데이터 양의 차이는 많지가 않다[5]. 그러나, 서열 데이터베이스를 제외한 다른 주제의 데이터베이스들의 경우엔 데이터베이스의 데이터 교환이 일어나지 않는다. 이런 이유로 해서 생물학 연구를 하기 위해 서열 데이터 이외의 바이오 데이터를 종합적으로 얻는 데는 많은 어려움이 있다.

서로 다른 스키마의 데이터베이스를 통합하는 방법으로는 federated database 방식이나 데이터웨어하우징 방법 등이 사용되고 있는데 바이오 데이터베이스 통합에 적용하는 데는 각각 문제점이 존재한다. 즉 federated database 방식의 경우에 있어서는 완벽한 연결성을 보장하지 못하며, 데이터웨어하우스의 경우에는 유지하는 데에 많은 비용이 든다는 것이다.

이 때문에 일부 바이오 데이터 관련 기관들은 자신들의 데이터베이스 및 데이터 분석 툴들에 대하여 웹 서비스를 적용하여 웹 서비스 시스템을 구축하고 서비스를 제공하고 있으며, 향후 여러 관련 단체들도 웹 서비스를 제공할 예정이다. 현재 웹 서비스 기반으로 서비스를 제공하는 바이오 연구기관 및 단체는 아직 초기 단계이긴 하나 DDBJ, KEGG[6], BioDAS[7], EMBL 등이 있으며 각각 자신들의 연구 특성에 맞는 컴포넌트 형식의 웹 서비스 기반 데이터 서비스를 제공하고 있다.

하지만 이러한 바이오 정보를 제공하는 기관들의 서비스는 자신들의 연구 목적에 따르는 데이터베이스 위주로 이루어진 서비스이기 때문에 다양한 연구에 있어서 필요한 정보의 상호 연관성을 연결하여 분석하고 검색하기엔 어려움이 있다. 이러한 문제를 해결하기 위하여 웹 서비스를 기반으로 연구 기관 및 단체들의 서비스를 통합하여 데이터들의 상호 연관성을 찾아 서비스를 제공하고 보완하는 필요성이 제기되고 있다.

2. 웹 서비스 기반 바이오 관련 연구 동향

현재 바이오 데이터베이스를 제공하는 주요 기관들 중에서 웹 서비스 기술을 이용하는 곳은 일본의 DDBJ와 KEGG, 유럽의 EMBL 등이 있으며 유전자의 주석(Annotation) 정보를 제공해주는 BioDas(Bio Distributed Annotation System) 등이 있다. 또한 여러 바이오 데이터베이스 관련 단체들도 웹 서비스를 접목시킨 바이오 데이터 제공을 준비중이다. 아래에 웹 서비스 기반으로 서비스하는 주요 데이터베이스에 대해 간단히 살펴본다.

2.1 KEGG(Kyoto Encyclopedia of Genes and Genomes

KEGG는 교토대학의 Kanehisa 박사팀에 의해 운영되

는 유전자 Pathway 관련 지놈 데이터베이스 연구소 이며 유전자의 Pathway에 관련하여 지놈 프로젝트의 결과물들과 잘 조합하여 제공중이다. KEGG에서는 SSDB (Sequence Similarity Database), PATHWAY, GENES,DBGET/LinkDB등을 웹 서비스로 제공 중이다.

2.2 EMBL(European Molecular Biology Laboratory)

1980 년에 창설된 EMBL 은 유럽분자생물학 연구소이다. 현재 유럽의 16 개국이 가입하고 공동으로 연구하고 있다. 세계 최초로 뉴클레오타이드 시퀀스 데이터베이스를 구축했으며 DNA 와 RNA 시퀀스, 유전자에 대한 주석(Annotation) 및 특징들 뿐만 아니라 단백질 시퀀스와 구조의 정보도 가지고 있다[4]. EMBL 에서는 현재 웹 서비스를 이용하여 유전자에 대한 여러 정보를 XML 형식의 BSML 과 SCIOBJ 형태로 제공해 주고 있으며 BIOML, GAME 등 다양한 형태로도 지원할 예정이다[3].

2.3 BioDAS(Bio Distributed Annotation System)

BioDAS는 분산형 유전자 주석 모델로서 유전자들의 주석 정보들을 상호 교환하기 위한 오픈 시스템으로 여러 데이터베이스들로부터 주석 정보를 모아 서비스를 제공해 줄 수 있다. 웹 서비스로 제공 되는 서비스로는 바이오 데이터에 관련된 주석 정보를 서비스 하고 있다.

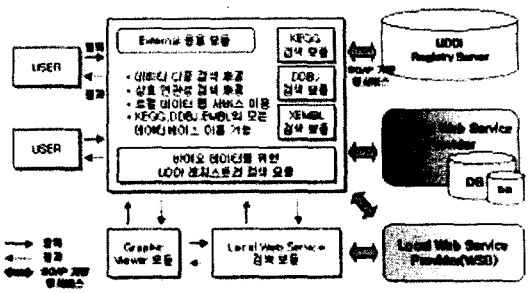
2.4 BioMOBY

BioMOBY[8]는 자체 데이터베이스를 구축해 놓고 있진 않지만 웹 서비스를 통하여 바이오 데이터를 발견하고 배포하기 위해 구조 형성을 목표로 하는 Open Source Research 프로젝트이다[9]. BioMOBY는 분산된 바이오 데이터를 웹 서비스 기술을 통하여 분산된 데이터베이스에 쉽게 접근 가능 하도록 하고 있다.

이상에서 현재 웹 서비스를 기반으로 바이오 정보 데이터 및 데이터 분석 툴들에 대한 각각의 서비스에 대하여 살펴보았다. 다음 장에서는 웹 서비스 API 등을 이용하여 바이오인포매틱스 연구자들이 이용할 수 있는 데이터 통합, 검색, 브라우징 기능을 지원하는 도구의 기능 및 구조를 제시하고 이의 구현에 대해 기술한다.

3. 웹 서비스 기반 바이오 데이터 통합 응용 시스템 설계

[그림 1]은 웹 서비스 기반 바이오 데이터 통합 응용 시스템(WSBAT: Web Services based Biodata Analysis Tool)에 대한 전체 구조도이다.
WSBAT은 [그림 1]에서와 같이 크게 세 개의 구성 요소를 갖는다. 첫번 째 구성요소는 웹 서비스 기반으로 바이오 데이터베이스를 중심으로 한 외부 응용 통합 모듈이다. 다음은 BioJAVA[10]를 이용하여 웹 서비스를 적용한 로컬 웹 서비스 검색 모듈(WSB:Web Service using BioJAVA), 그리고 JAVA 그래픽 모듈을 사용하여 유전자들간의 상호작용 (Interaction) 관계를 그래픽으로도 볼 수 있도록 하는 모듈로 구성된다.



[그림 1] 시스템 구조도

3.1 외부 웹 서비스 기반 서열 데이터 통합 모듈 및 상호작용 데이터 웹 서비스 제공 모듈

KEGG, DDBJ, XEMBL, BioDAS의 데이터 정보 서비스를 제공받는 외부 통합 검색 모듈과 BioJAVA 를 웹 서비스에 접목시킨 데이터 제공 모듈로 구성되어 있다. 가능한 서비스를 보면 KEGG에서는 SSDB, PATHWAY, GENES, DBGET/LinkDB의 4가지 서비스와 XEMBL의 뉴클레오타이드 정보 서비스 그리고 DDBJ의 Blast, ClustalW, Fasta, GetEntry, SRS, TxSearch, ExClustalW, Gtop, PML, Gib등 10가지에 대한 검색 서비스를 통합 구현하였다.

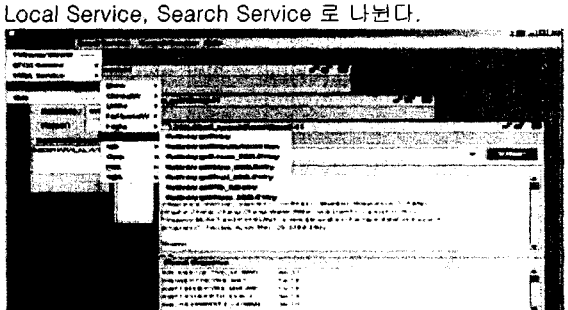
3.2 데이터 상호작용 그래픽 뷰어 모듈

SOAP을 통해서 웹 서비스로 제공되는 바이오 데이터들은 모두 텍스트 형식이기 때문에 데이터의 표현에 있어서는 많은 제한이 따른다. 특히 단백질 같은 바이오 데이터들의 상호작용이나 PATHWAY 관계등을 보여 주기 위해선 그래픽적인 요소가 필요하다.

이를 위해서 데이터들의 관계를 그래픽으로 볼 수 있는 자바 그래픽 모듈을 구현하여 텍스트 기반의 데이터 정보 뿐만 아니라 단백질등의 상호작용을 그래픽으로 볼 수 있도록 하였다.

3.3 웹 서비스 유저 인터페이스

[그림 2]에서 보는 바와 같이 각각의 서비스를 다중으로 검색 가능하도록 데스크탑 형식의 클라이언트 인터페이스를 구현하였으며 메뉴를 보면 크게 External Service, Local Service, Search Service 로 나뉜다.



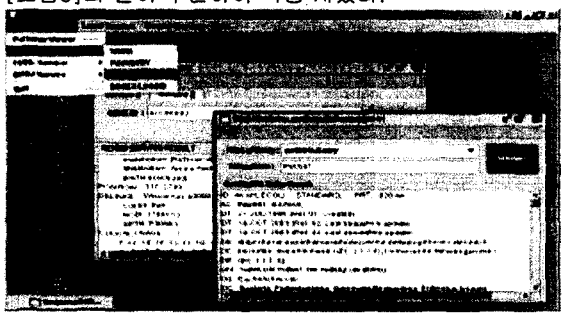
[그림 2] 어플리케이션 인터페이스

External Service는 KEGG, XEMBL, DDBJ 에서 제공되는 서비스 메뉴이며 Local Service는 BioJAVA라이브러리를

이용하여 웹 서비스로 구현한 정보를 제공하는 메뉴이다. Search Service는 UDDI[11]를 이용한 또 다른 외부 바이오 웹 서비스를 검색하여 WSDL 문서를 가져 올 수 있는 메뉴도 함께 제공하고 있다.

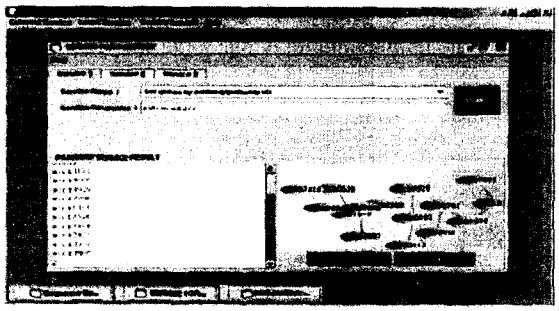
4. WSBAT의 동작

어플리케이션을 바탕으로 웹 서비스 기술을 이용한 데이터 정보 및 데이터들의 상호 연관성에 대한 서비스를 [그림3]과 같이 구현하여 적용 시켰다.



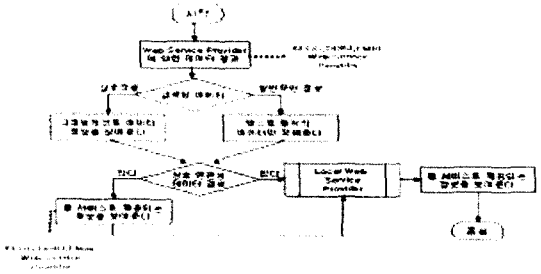
[그림 3] 웹 서비스 기반 데이터 서비스

메뉴에서 KEGG Service를 선택하면 KEGG 검색 모듈을 이용하여 GENES 서비스에 의해 데이터를 제공받고 상호연관성 있는 데이터를 DDBJ Service를 이용하여 보여주고 있다.



[그림 4] 웹 서비스 기반 그래픽 뷰어

그림[4]는 KEGG에서 서비스를 받아 상호 작용하는 유전자들에 대한 데이터 정보를 텍스트와 그래픽으로 동시에 볼 수 있으며 그래픽을 이용한 여러 서비스도 제공할 예정이다.



[그림 5] 서비스 흐름도

[그림 5]는 서비스 흐름을 보여주고 있다.
사용자의 질의에 대한 웹 서비스 기반 서비스를 SOAP [12] 을 통해 결과 데이터를 처리한 뒤 텍스트 기반 및 시각적인 뷰어를 제공하고 데이터의 상호 연관성을 검색하여 또 다른 웹 서비스 제공자로 인한 서비스를 검색 할 수가 있다.

5. 결론 및 향후 과제

오늘날의 바이오 데이터는 하루가 다르게 데이터의 양이 증가하고 있다. 또한 바이오에 대한 여러 연구들로 인해 이질적으로 분산된 바이오 데이터들의 처리와 그 분석에 있어서 많은 어려움을 겪고 있는 실정이다.

본 논문에서는 이러한 데이터들에 대하여 통합된 어플리케이션으로부터 서로간의 상호 연관된 정보를 찾고 보다 효율적으로 데이터 검색을 할 수 있도록 하기 위한 연구 중 SOAP을 통한 컴포넌트 기반 XML형식의 자유로운 데이터 및 메소드 교환 등 여러 장점을 가지고 있는 웹 서비스 기반 기술을 접목 시켜 서비스를 구현하였다. 또한 공개 UDDI 레지스트리에서 바이오 데이터 서비스 검색을 위한 UDDI 프로그래밍도 구현중이다.

데이터 제공자들이 데이터를 제공함에 있어 웹 서비스 방식을 사용함으로써 데이터의 활용도를 높임은 물론 그로부터 발생한 2차적인 데이터베이스 혹은 분석 도구활용도를 높일 수 있을 것으로 기대된다. 하지만 아직까지 바이오 정보에 대한 웹 서비스는 초기 단계이며 여러 바이오 데이터베이스 관련 연구 단체들에 의한 웹 서비스 기술이 갖추어 지고 있지 않아 많은 연구가 필요하다. 또한 웹 서비스에서 가장 취약한 보안문제도 함께 연구 되어야 할 과제이다.

참 고 문 헌

[1] GenBank : http://www.ncbi.nlm.nih.gov/genbank
[2] XML Central of DDBJ : http://www.xml.nig.ac.jp
[3] EMBL Project : http://www.ebi.ac.uk/xembl
[4] Catherine Brooksbank*, Evelyn Camon, Midori A. Harris, Michele Magrane, Maria Jesus Martin, Nicola Mulder, Claire O'Donovan, Helen Parkinson, Mary Ann Tuli, Rolf Apweiler, Ewan Birney, Alvis Brazma, Kim Henrick, Rodrigo Lopez, Guenter Stoesser, Peter Stoehr and Graham Cameron, The European Bioinformatics Institute's data resource, Nucleic Acids Research, Vol. 31. No. 1. 2003
[5] 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. 2003
[6] KEGG Project : http://www.genome.ad.jp/kegg
[7] BioDAS : http://www.biodas.org
[8] BioMOBY project : httpd://www.biomoby.org
[9]Mark D. Wilkinson and Matthew Links, BioMOBY: An open source biological web services proposal, VOL 3. NO 4 331.341. DECEMBER 2002
[10] BioJAVA : http://www.biojava.org/tutorials
[11] UDDI : http://www.uddi.org
[12] SOAP : http://www.w3.org/TR/soap12