온톨로지를 이용한 웹서비스 기반 바이오 정보 시스템의 설계 및 구현

박용일*, 박성수*, 이종근*, 홍동완**, 윤지희*
*한림대학교 컴퓨터공학과
**춘천정보대학 인터넷미디어과
e-mail:taz102@hallym.ac.kr

Design and Implementation of a Web-Service based Bio-Informatics System using GO knowledge-base

Yong-Il Park*, Sung-Su Park*, Jong-Keun Lee*,
Dong-Wan Hong**, Jee-Hee Yoon*
*Dept. of Computer Engineering, Hallym University
**Dept. of Internet-media, ChunCheon Information College

요 약

최근 국, 내외에서 생물정보학 데이터베이스 구축이 활발히 진행되어 왔고, 각 바이오 정보 시스템의 데이터 통합 연구가 진행 중이다. 대표적인 바이오 데이터베이스 시스템인 GenBank, DDBJ, EBML 등은 같은 의미의 데이터라 하더라도 각 시스템의 내부 데이터 구조 및 데이터 표현 형식이 상이하여 통합에 어려움이 따른다. 이를 해결하기 위해 통합 데이터 형식을 지원하는 웹 서비스 기반 데이터 통합 방식이 제안되고 있다. 현재 국내 웹 서비스 기반의 바이오 정보 제공 사이트들은 SOAP을 이용한 단순 메시지 전달 기법으로 초보적인 단계라 할 수 있다. 본 논문에서는 SOAP을 이용한 단순한 메시지 전달 기법만이 아닌 레지스트리 서버 검색을 통해 서비스 제공자를 찾고, WSDL문서를 분석한 후 사용자에게 검색 메소드를 제공함으로써 빠르고 정확한 서비스를 제공하여 기존에 구축된 시스템의 단점을 보완한다. 또한 상이한 스키마로 이루어진 데이터들을 효과적으로 통합하기 위해 온톨로지를 이용한 웹 서비스 기반 바이오 정보 시스템을 제안하고 구현한다.

1. 서론

최근 생물정보학에 대한 관심 증가와 활발한 연구로 인하여 생물학 데이터가 기하급수적으로 증가하였으며, 생물학, 전산학 전문가들이 공동으로 이를 분석, 관리하기 위한 바이오 정보 시스템 개발에 많은 시간과 노력을 투자하고 있다[1]. 현재 개발된 바이오 정보 시스템의 대표적인 시스템으로 미국 국립생물정보센터의 GenBank[2], 일본 국립유전학연구소의 DDBJ(DNA Data Bank of Japan)[3][4], 유럽 생물정보학연구소의 EMBL(European Molecular Biology Laboratory)[5], 스위스의 생물정보학연구소 Swiss-PROT[6]등이 있다. 이들은 시스템 목적에 맞게 구축되어, 사용 방법이 상이하므로 생물학 전문가는 시스템 사용이 숙련되기까지 많은 시간과 노력을 투자해야 한다. 또한 생물학 전문가는 각 시스템의 연관된 데이터를 통합, 검색하여 보기를 원하는데 여러 시스템의 사용방법을 제대로 습득하지 못하여 통합 바이오 데이터를 획득하기가 쉽지 않다.

이와 같은 문제점을 해결하기 위하여 연관성 있는 통합 데이터를 제공하는 바이오 정보 시스템의 개발이 필요하다. 대표적인 데이터 통합 기술로 링크 기반 통합, 뷰 기반 통합, 데이터웨어하우징 방법 등이 있다[7]. 하지만 이와 같은 통합 방법을 사용할 경우, 기존에 구축된 바이오 정보 시스템의 내부 데이터 구조가 상이하여 시스템 간 통합 데이터 모델 설계가 어렵고, 내부 데이터 구조 변경과 바이오 데이터 업데이트에 대한 실시간 반영이 수월하지 않다. 이러한 문제점을 해결하기 위한 방안으로 웹 서비스 기반 개발방식이 제안되었다[7]. 정보통신부 등의 국가정책기관에서 웹 서비스를 데이터 공유 방법과 컴포넌트 기반 서비스 개발의 표준으로 재정하여 현재 사업, 군사, 의료 시스템 등의 개발에 활용되고 있다. 그러나, 웹 서비스의 핵심 기술인 UDDI(Universal Discovery Description & Integration) 레지스트리는 IBM 의 UDDI business 레지스트리와 SAP의 sap@uddi.business 레지스트리 등과 같이 사업(e-business, enterprise) 분야 위주로 개발되어 있어, 바이오 정보 시스템 분야에 적합한 UDDI 레지스트리 개발이 필요하다. 또한 의미적 데이터 통합을 지원하기 위하여 각 유전체 데이터베이스의 데이터 연관성을 고려한 온톨로지[8,9,10]를 지원해야 한다. 본 논문에서는 레지스트리 서버 검색을 통하여 서비스 제공자를 찾고 WSDL(Web Se-rvices Description Language)를 분석, 변환하여 생물학 전문가에게 등록, 검색 작업의 편의성을 제공하고 유전자 데이터들의 의미론적 통합을 위하여 온톨로지를 이용한 웹 서비스 기반 바이오 정보 시스템을 제안 및 구현한다.

2. 관련연구

2.1 국내바이오정보시스템

국내 바이오 정보 시스템은 생물체의 기본 구성단위인

DNA의 서열 및 구조 정보, 온라인 문헌정보(online publication ; abstract DB) 등의 자료를 제공하고 있으며, 대표적인 데이터베이스 시스템으로 한국 유전자 데이터베이스(Kor-ean Sequence Database; KSDB)[11]와 "KRISTAL 2002"[12] 등이 있다. 한국 유전자 데이터베이스 진누리(Gene Nuri)는 유전자 데이터 등록처(Gene In)를 통한 등록, E-mail 접수, 논문 검색 등을 통하여 유전자 데이터베이스를 구축하고 있다. "KRISTAL 2002"는 웹 기반 GenBank 검색 서비스로, dbEST, dbGSS, dbSTS 등을 제공한다.

2.2 Bio-MOBY

Bio-MOBY[13,14,15]는 2001년 이질적인 바이오 데이터베이스들을 통합하기 위한 목적으로 개발된 시스템으로, 유전자 데이터의 검색 방법이 단순하고, 소스가 공개되어 확장이 용이하다. Bio-MOBY는 웹 서비스를 제공하는 MOBY-S(MOBY-Service)와 시멘틱 웹 서비스를 제공하는 S-MOBY(Semantic MOBY)가 동시에 구축, 연구되고 있다. 클라이언트와 서버 사이에서 주고받는 구조화된 데이터를 말하는 MOBY Obj-ect, 레지스트리 역할을 하는 MOBY Central, 객체 및 서비스 요소들로 구성되어 있다. Bio-MOBY 시스템 구조는 그림 1.과 같다.

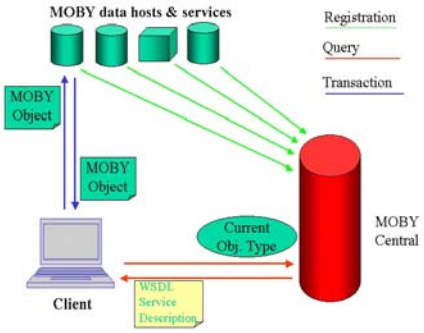

그림 1. Bio-MOBY 시스템 구조

2.3 온톨로지(Ontology)

생물정보학에서는 데이터의 특성과 데이터베이스의 목적에 따라 다양한 생물학 데이터베이스가 구축되어 있고 연구 진행 중이다. 이러한 데이터베이스들은 서로 독립적인 것이 아니라 내부 데이터의 의미가 같은 연관관계이어서, 유기적 데이터 결합이 필요하다. 임의의 시스템의 내부 데이터와 각 시스템 간 정보 공유와 교환을 위해 의미적 연관성 제공을 온톨로지가 한다. 온톨로지는 어휘, 개념, 관계 등을 포함하는 특정 분야 지식의 의미 모델로 기존 지식의 추출 및 신뢰성이 보장되는 검증 과정을 거쳐 확장 가능하다. 대표적인 유전자 정보 관련 온톨로지로 슐츠클라이머 온톨로지(Schulze Kremer's Ontology)[16], UMLS(unified medical lang-uage system)[17], Gene Ontology[8] 등이 있다.

3. 웹서비스 기반 바이오정보시스템

3.1 개발의의

본 시스템의 개발은 다음의 의의를 가진다. 첫째, 생물학 전문가들이 발견한 실험 유전자 데이터를 유전자 데이터베이스에 쉽게 등록, 검색하는 사이트를 구축하여 개선된 생물정보학 서비스를 제공한다. 둘째, 기존 데이터 통합 방식에서의 문제점들을 해결하는 표준화된 웹 서비스 기반 바이오 시스템을 구축한다. 이는 분산된 데이터를 통합하여 생물학 전문가에게 제공함으로써 최신 생물 정보 데이터 연구를 할 수 있도록 도와준다. 셋째, 유전자 온톨로지(G-ene Ontology)를 이용하여 연관된 데이터를 통합하여 생물학 전문가에게 지원, 개선된 서비스를 제공한다.

3.2 웹서비스

웹 서비스[18]는 그림 2.와 같이 서비스 제공자(Service Provider), 서비스 요청자(Service Re-questor), 서비스 대리자(Service Agency)로 구성된다. 서비스 제공자는 웹 서비스를 구현하여 서비스를 제공하는 측면의 소프트웨어 시스템이고, 서비스 요청자는 서비스를 받기 위해 웹 서비스를 호출하는 소프트웨어 시스템이다. 그리고 서비스 대리자는 웹 서비스를 카테고리 별로 분류하고 웹 서비스의 내용을 저장함으로써 서비스 요청자가 자신이 필요한 웹 서비스를 쉽게 찾을 수 있도록 디렉토리 서비스를 제공하는 레지스트리이다. 관련 기술은 다음과 같다.

- UDDI[19] : 웹 서비스에 관한 정보를 등록, 탐색하기 위한 분산형 웹 기반 등록기이다.
- WSDL[20] : 웹 서비스의 IDL(Interface Def-inition Language)버전이며 특정 웹 서비스의 방법과 프로토콜, 데이터 포맷들을 더욱 상세하게 정의하는 표준 스크립트이다. WSDL은 XML 포맷으로 구성되고 HTTP를 통해서 전달되며 인터페이스를 정의하는 IDL에 해당한다.
- SOAP(Simple Object Access Protocol)[21]: 클라이언트의 작업 요청과 시스템의 응답을 XML 문자열로 구성하고 전송하는 표준 프로토콜이다. SOAP은 HTTP와 XML의 결합으로 분산 환경에서 정보의 상호교환을 하는 프로토콜이다.

그림 2. 웹 서비스 구성

웹 서비스를 제공하지 않는 사이트들은 자체 WSDL을 레지스트리에 등록하고 그것을 기준으로 검색 메소들을 사용자에 제공하여 바이오 정보 검색 서비스를 제공할 수 있다. 본 시스템은 메이저급 레지스트리와 연계하여 국, 내외의 생물학 전문가에게 서비스 제공을 목표로 하였다. 본 시스템은 사설 UDDI를 개발하여 Bio-MOBY 레지스트리와 연계하였다. 구현한 자체 사설 UDDI 레지스트리를 Bio-MOBY SOAP을 이용하여 Bio-MOBY의 레지스트리 서버를 검색하고 서비스 제공자의 WSDL을 분석, 변환하여 사용자는 빠른 검색과 확장이 가능하다. 그림 3.은 레지스트리 서버를 검색하고 생물학 전문가에 메소드를 제공하는 과정이다.

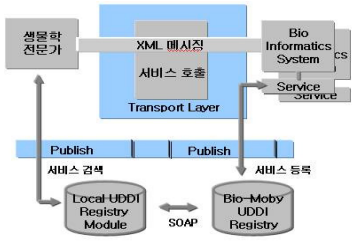

그림 3. 시스템 개념도

3.3 온톨로지

여러 바이오 정보 시스템에서 다루는 데이터들은 서로 독립적인 것이 아니라 의미상 연관(association)이 있어 관련 자료들을 함께 보일 수 있는 서비스가 필요하다. 생명 유지에 필요한 단백질이나 유전자 등은 하나의 종(species)에만 국한된 것이 아니라 여러 종에 공통적으로 존재하므로 모든 종에서 서로 연관된 데이터에 대한 용어의 확립이 필요하다. 연관 정보는 연관된 바이오 데이터의 연결에 필요한 주석(annotati-on)정보를 담고 있는 것으로, 여러 바이오 정보 시스템의 데이터 간 정보 공유와 교환에 사용된다. 이와 같이 바이오 정보 통합에 필요한 지식 정보를 온톨로지라 한다. 바이오 정보 시스템에서의 온톨로지는 단백질 간의 상호작용(protein-protein interaction)을 표현한 유전자 네트워크상에서 그 위치 등의 정보를 데이터베이스에 저장한 지식정보로 연관 데이터 간의 공유에 사용된다. 본 연구에서는 대표적인 온톨로지 컨소시움인 "Gene Ontology"의 지식 정보를 이용하여 지식 관리 시스템(Knowledge-base Managem-ent System)을 구축한다. 본 시스템은 주로 "Homo Sapiens"와 "Rattus norvegious" 위주로 검색하는데, MGI(Mouse Genome Database)의 온톨로지를 GenBank의 GeneID와 Swiss-PROT의 sp와 연계하여 통합, 구축하였다. 검색 분야에 사용되는 "Gene Ontology"는 다음 세 가지 범주(category)로 구성된다.

- Cellular component : 셀(cell)의 구성 요소(component)로 "anatomical structure"상의 네이밍이 주된 정보이다.
- Molecular Function : gene product가 무엇을 할 수 있는지에 대한 정보를 포함하고 있다.
- Biological Process : Molecular Function이 순서대로 조합하여 이루어진 것이다. GO에서는 10개 이상의 구별되는 단계(Molecular Function step)가 있어야 "Biological Process"로 인정한다.

표 1. Gene Ontology 적용 분야[8]

	SGD	FlyBase	MGI
Celluar Component	1,670	1,415	2,832
Molecular Function	1,698	6,604	3,769
Biological Process	1,720	1,084	2,859

표 1.은 유전자에 대한 Gene Ontology 적용 분야를 나타낸 것이다. 예를 들어 MGI에 등록된 유전자 중 2,859개가 Biological Process의 범주에 해당하는 세부 기능과 관련되어 있다.

Gene Ontology 데이터는 그림 4.의 DTD 구조를 갖는 XML 문서로 표현된다. Component, Function, Process 등은 DAG(Directed Acyclic Graph)형태로 구성되는데 XML 문서의 포함(nesting) 구조의 제한성을 극복하기 위하여 'rdf link' 방식을 활용한다. Ontology는 "term" 태그 내에 구성되는데 구성 정보는 Gene ID인 'go:accession', Gene Product name인 'go:name', 유사한 유전자 정보(homology)를 표현하기 위한 'go:synonym', 유전자의 설명이 포함된 'go:def-inition', 유전자 인스턴스 간 관계를 나타내는 '(go:part-of | go:isa)'와 외부 데이터와의 연계 정보를 담고 있는 'go:dbxref', 주석 정보를 담고 있는 'go:association', 유전자 데이터 발견자와 발견 시기를 포함하고 있는 'go:history' 태그로 구성된다.

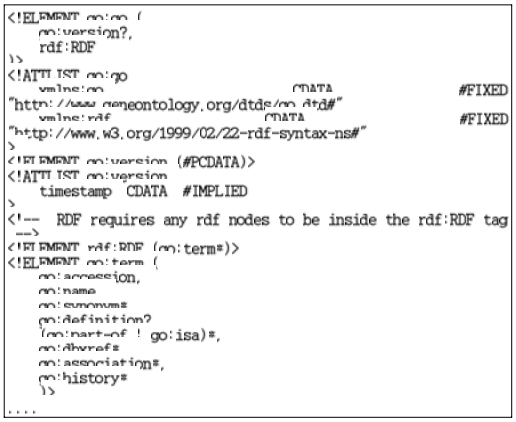

그림 4. GO DTD

본 시스템의 Gene Ontology를 확장하는 방법으로 Gene Ontology에서 제공하는 정보를 관계형 데이터베이스 My SQL 4에 구축하였고, 정보 추출 시스템에서 얻어진 단백질 상호작용 정보와 조건 정보를 데이터베이스에 통합 동의어(synonym)/약어(acronym) 정보로 추출하였다, 또한 용어들의 형태소 분석, 구조 분석 등을 통하여 용어 간의 연결성을 높였다.

4. 사용자 인터페이스

본 시스템의 사용자 인터페이스는 생물학 전문가에게 편의성을 제공한다. 생물학 전문가가 유전자 데이터 등록 시 기존에 등록되어 있는 연관된 서비스 리스트를 보여줌으로써 생물학 전문가에게 편의성을 보장한다.

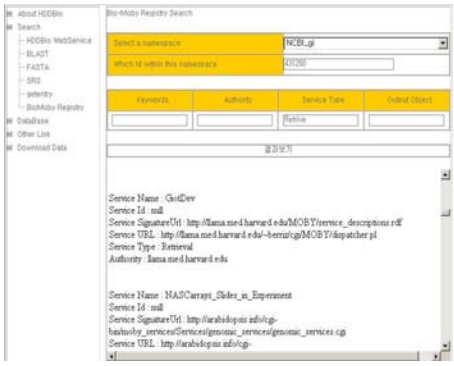

그림 5. 레지스트리 등록화면

그림 5.는 레지스트리 등록 시 'find' 명령을 이용하여 기존에 등록된 연관성이 있는 바이오 데이터 리스트를 보여주는 화면이다. GenBank 포맷의 인간 유전자 중 유전자 ID가 '25246'인 질의를 입력 받고, 입력 받은 웹 서비스 클라이언트는 SOAP 메시지 처리 모듈에 서비스 요청을 한다. 서비스 요청을 받은 SOAP 메시지 처리 모듈은 UDDI 레지스트리 모듈과 온톨로지 관리 모듈에게 서비스 검색을 요구한다. 온톨로지 모듈은 질의를 분석하고 연관 정보를 추출하여 결과들을 반환한다. UDDI 레지스트리 모듈은 반환 결과를 이용하여 서비스 목록 검색 시 연관된 서비스 제공자의 WSDL 문서들을 가져오고, 모듈 내에서 하나의 WSDL 문서로 자동 통합, 변환한다. 변환된 WSDL 문서는 SOAP 메시지 처리 모듈로 전달된다. 메소드를 받은 SOAP 메시지 처리 모듈은 WSDL 문서 내에 기술된 메소드를 이용하여 검색한 결과를 사용자에게 보여준다. 사용자 인터페이스에는 온톨로지 관리 모듈에서 분석한 'RGD: 2220'이 같이 보여진다. 'RGD: 2220'은

RatID가 2220이라는 것을 의미한다. 생물학 전문가는 자신이 등록하려는 유전자 ID '25246'이 인간 유전자에서도 발견이 되고, 쥐의 유전자에도 발견이 되는 유전자임을 알 수 있다.

5. 시스템 구조도

본 시스템의 구조도는 그림 6.과 같다. 시스템은 5부분의 모듈로 구성되어 있다. 각 모듈의 기능은 다음과 같다.

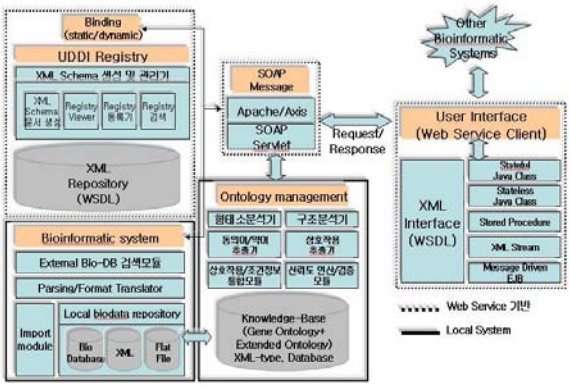

그림 6. 시스템 구조도

・웹 서비스 클라이언트(사용자 인터페이스)모듈: 검색 시 사용자는 사용자 인터페이스를 통하여 질의하고, 웹 서비스 클라이언트는 SOAP 메시지 처리 모듈에 질의를 요청하며 메소드를 받아 사용자와 서비스 제공자를 연결한다. 서비스 등록 시 SOAP메시지 처리 모듈에 서비스 리스트를 요청하고 서비스 등록을 도와준다.
・SOAP메시지 처리 모듈: 사용자로부터 서비스 요청이 있을 시 레지스트리 서버에 검색, 등록을 요청하고, 온톨로지 관리 모듈에 온톨로지 분석을 요청한다.
・UDDI 레지스트리 모듈: 외부 레지스트리 서버를 검색하고 검색한 WSDL 문서를 분석, 변환하여 메소드를 제공한다. 서비스 등록 요청 시 외부 레지스트리에 서비스를 등록한다. Bio-MOBY 레지스트리와 연결할 수 있다.
・로컬 바이오 정보 시스템 모듈: 자체 로컬 데이터베이스를 구축하여 검색이 빈번하거나, 온톨로지 연관이 있는 데이터를 저장하여 빠른 결과를 제시한다.
・온톨로지 관리 모듈: 형태소 분석기와 구조 분석기를 통하여 동의어/약의어, 상호작용/조건정보를 추출하고, 신뢰도 연산, 검증 등을 거쳐 신뢰성 있는 온톨로지를 제공한다.

본 시스템의 플랫폼은 로컬 시스템으로 마이크로소프트 윈도우즈 2000서버에 SQL Server 2000을 이용하여 구축하였다. 사설 UDDI와 온톨로지 관리 시스템은 한컴 리눅스 4에 Apache Tomcat 5.0, Axis 1.2 RC2, UDDI 4j, JUDDI 0.9 RC3, MySQL4.0을 이용하여 구현하였다. 개발 소프트웨어는 JDK 1.4.2, Bio-JAVA 1.3, JDBC Type 4 Driver, Xerces-j 2.5.0 등을 활용하였다.

6. 결론 및 향 후 연구과제

본 논문에서 구현한 시스템은 생물학 전문가들이 발견한 실험 유전자 데이터를 유전자 데이터베이스에 효율적으로 등록, 공유할 수 있도록 온톨로지를 활용한 지능형 웹 서비스 기반 바이오 정보 시스템이다. 본 시스템은 생물학 전문가에게 편리한 사용자 인터페이스를 제공하여 여러 바이오 데이터베이스에 저장된 연관된 데이터의 통합, 검색이 가능하다. 본 논문의 온톨로지 관리 시스템은 현재 로컬 시스템으로 구축되어 있는데, 기존에 구축된 다른 온톨로지 시스템과의 공유를 위하여 웹 서비스 기반으로 확장할 계획이다.

7. 참고문헌

[1] Stein, L: Creating a bioinformatics nation. Nature417, 119-129(2002)
[2] GenBank : http://www.ncbi.nlm.nih.gov
[3] DDBJ Project: http://wwwddbj.nig.ac.jp
[4] H.Sugawara, S.Miyazaki: Biological SOAP Server and Web Services provided by the public sequence data bank. Nucleic Acids Research.Vol31, 3836-3839 (2003)
[5] EMBL Project: http://www.ebi.ac.uk
[6] Swiss-PROT Project: http://www.expasy.ch
[7] Stein, L. Integrating Biological Database. Nature Reviews-Genetics. Vol4,337-345 (2003)
[8] Gene Ontology: http://www.geneontology.org
[9] Gene Ontology Consortium. Gene Ontolgy:tool for the unification of biology. Nature Genet. Vol25, 25-29 (2000)
[10] Stevens,R.,Goble,C.A.& Bechhofer,S. Ontology-based knowledge representation for bioinformatics. Briefings In Bioinformatics. Vol1, 398-414 (2000)
[11] http://bric.postech.ac.kr
[12] http://www.ccbb.re.kr
[13] http://www.bioMOBY.org
[14] Mark D, Wilkinson and Matthew Links. BioMOBY: An open source biological web
services proposal. Briefings In Bioinformatics. Vol3, 331-341 (2002)
[15] Wilkinson MD, Gessler D, Farmer A, Stein L. The BioMOBY Project Explores Open
-Source, Simple, Extensible Protocols For Enabling Biological Database Interoperability. Proc Virt Conf Genome and Bioinf. Vol3, 17-27 (2003)
[16] http://igd.rz-berlin.mpg.de/~www/oe/mbo.html
[17] http://umlsks.nlm.nih.gov/
[18] Web Services Architecture: http://www.w3.org/TR/2002/WD-ws-arch-20021114/
[19] UDDI: http://www.uddi.org
[20] WSDL: http://www/w3.org/TR/2001/NOTE-wsdl-20010315
[21] SOAP: http://www/w3.org/TR/2000/NOTE-SOAP-2000508