웹 서비스를 이용한 바이오 서열 정보 데이터베이스 및 통합 검색 시스템 개발

(Development of Integrated Retrieval System of the Biology Sequence Database Using Web Service)

이 수 정(Sujung Lee), 용 환 승(Hwan-Seung Yong)

이화여자대학교 컴퓨터학과

요 약

최근, 바이오 관련 장비, 기술들이 발전함에 따라, 바이오 관련 데이터나 그것을 제공하는 호스트들이 급속하게 증가하고 있다. 또한, 이러한 데이터들은 개발 커뮤니티들의 수만큼, 분산되고 이질적인 면을 가지고 있어서, 바이오 관련 데이터베이스의 통합과 연동기능의 제공이 중요한 문제가 되고 있다. 그러나, 현재까지 진행되고 있는 많은 통합 연구 시스템의 대부분이 링크기반, 데이터웨어하우징 구축 기반으로 하고 있어서, 데이터 스키마나 데이터의 변경시, 실시간 업데이트와 같은 문제점을 보인다. 이러한 비효율적인 면을 개선시키고자, 플랫폼, 스키마의 변화에 구애 받지 않고 서비스를 가능하게 하는 웹 서비스 기술을 이용한 통합 시스템이 제안되고 있다. 본 논문에서도 이러한 흐름에 맞추어, 웹 서비스를 이용한 바이오 서열 데이터의 데이터베이스와, 통합 검색 시스템을 개발하였다. 개발된 시스템은 BSML을 포함한 다양한 포맷의 데이터로 서열정보를 제공하며, 또한 외부 데이터베이스의 검색을 병렬로 처리하여, 검색 성능을 향상시키도록 하였다.

키워드 : 바이오인포메틱스, 데이터베이스, 웹 서비스, SOAP, UDDI, WSDL

ABSTRACT

Recently, the rapid development of biotechnology brings the explosion of biological data and biological data host. Moreover, these data are highly distributed and heterogeneous, reflecting the distribution and heterogeneity of the Molecular Biology research community. As a consequence, the

integration and interoperability of molecular biology databases are issue of considerable importance. But, up to now, most of the integrated systems such as link based system, data warehouse based system have many problems which are keeping the data up to date when the schema and data of the data source are changed. For this reason, the integrated system using web service technology that allow biological data to be fully exploited have been proposed. In this paper, we built the integrated system of the bio sequence information based on the web service technology. The developed system allows users to get data with many format such as BSML, GenBank, Fasta to traverse disparate data resources. Also, it has better retrieval performance because the retrieval modules of the external database proceed in parallel.

Keywords : BioInformatics, Database, Web Services, SOAP, UDDI, WSDL

1. 서 론

80년대 중반 Human Genome 프로젝트가 시작된 이후로, 분자 생물학 분야의 데이터가 기하 급수적으로 증가하고 있으며, 그 정보들은 개발 커뮤니티의 특징에 따라 다수의 분산되고 이질적인 데이터 저장소에 저장된다. 이로 인해, 연구자들은 의미 있는 바이오 정보를 얻기 위하여, 여러 독립된 데이터베이스를 검색하고, 그 결과를 수작업으로 분석, 변형하며 통합한다. 그러나 현재와 같이 다양한 데이터베이스가 독립적으로 존재하는 경우, 통합적인 정보검색과 새로운 지식 추출은 기술적으로 어려운 일이다. 이러한 이유로 상호 이질적인 데이터베이스 간의 정보 교환을 위한 데이터 표준화 작업과, 바이오 관련 데이터베이스의 통합과 연동기능을 제공하는 시스템 개발에 대한 연구가 무엇보다 필요하다.

현재까지 개발된 통합 데이터 검색 시스템을 살펴보면, 자체적으로 운영하는 데이터베이스를 여러 개 두어 링크로 연결하거나, 생물학적인 개념을 포괄적으로 표현하기 위하여 거대한 온톨로지 기반으로 데이터를 모델링 하는 방법이 있다. 이러한 방법들은 링크 기반, 데이터웨어하우징 구축 기반의, 공개된 대부분의 생물학 데이터베이스에서 많이 사용하고 있는 방법으로써, 소스 데이터의 변경이나, 새로운 데이터베이스를 추가할 경우, 구축, 유지 면에서 비효율적인 단점이 있다.[1,2,3]

최근에 들어, 화두가 되고 있는 XML[4]과 웹 서비스 기술[5,6,7]은, 위에서 설명한 바이오 데이터 교환과 통합 문제를 큰 어려움 없이 해결해줄 수 있는 실마리를 제공하고 있으며, 세계적으로 몇몇 연구 기관에서도 이를 이용한 연구를 진행하고 있다.

본 논문에서도, 이러한 흐름에 맞추어 XML과 웹 서비스 기술을 이용한 유전체 서열 데이터베이스를 구축하고, 외부 데이터베이스들을 통합 검색하는 시스템을 개발하였다. 우선, 생물 종을 벼와 돼지로 한정한, 핵산, 단백질 서열 정보 데이터베이스를 구축하였다. 그리고 WSDL, SOAP, UDDI와 같은 웹 서비스 기술을 이용하여 구축한 데이터베이스의 검색 모듈을 웹 서비스화하여 공개하고, 외부 데이터베이스를 웹 서비스 병렬 처리를 통해 통합 검색 할 수 있도록 하였다. 즉, 개발된 시스템은 사용자의 질의 시 얻고자 하는 정보가 구축된

데이터베이스에 존재하지 않을 경우, 외부 데이터 베이스를 검색을 하고 그 결과를 자동적으로 구축된 데이터베이스에 저장되도록 한다. 또한 기존의 많은 생물학 데이터들은 매우 다양한 형식으로 만들어져 있기 때문에, 데이터베이스 상호간 용어의 호환성을 높이고, 데이터 교환을 편하게 하기 위하여, 데이터를 BSML[8]을 포함한 GenBank[9,10], Fasta 등 다양한 형태로 제공하도록 하였다.

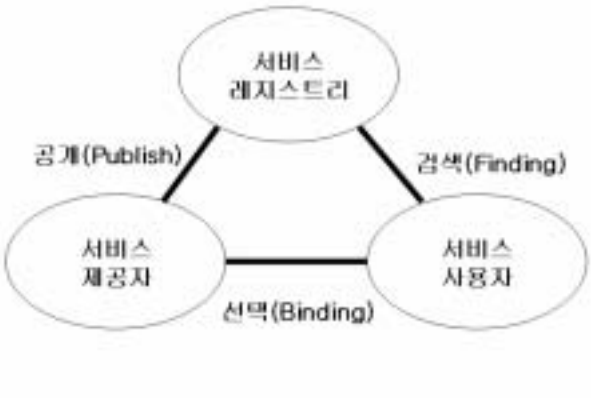
본 논문의 구성은 다음과 같다. 2 장에서는 우선, 본 논문의 핵심 기술이라 할 수 있는 웹 서비스 기술이 무엇인지 소개하고, 현재까지 바이오 정보 시스템을 통합하려는 다양한 시도들을 살펴본다. 3 절은 기존의 통합 검색 시스템을 간단히 살펴보고, 개발할 바이오 데이터 통합 검색 시스템 모델을 설계하고 데이터를 어떻게 전처리 하여 데이터베이스화 했는지를 테이블 스키마와 함께 제시하며, 웹 서비스에 기반한 통합 검색 모듈을 웹 서비스 구조에 비추어 설명한다. 4 절에서는 구현된 시스템의 결과로 사용자 인터페이스 중심으로 살펴본다. 마지막으로 5 절에서는 본 연구의 결론으로 연구의 의의 및 향후 연구 과제를 제시한다.

2. 관련 연구

이번 장에서는 본 연구의 핵심이라고 할 수 있는 웹 서비스 기술에 대해서 간단히 소개를 하고, 현재까지의 바이오 데이터 통합 시도 사례를 살펴보고자 한다.

2.1. 웹 서비스

웹 서비스는 XML 메세징을 통하여 인터넷을 통하여 접근 가능한 네트워크 명령들의 집합을 기술한 인터페이스로서, 단일 또는 서로 다른 플랫폼과의 상관 관계없이 상호운영(Interoperating)이 가능하도록 해주는 표준화된 기술이다. 웹 서비스는 (그림 1)에 나타난 것과 같이 크게 세 가지로 구성된다.



(그림 1) 웹 서비스 구조

· 서비스 제공자 (Service Provider): 웹 서비스를 개발하는 역할을 함. 공개적으로 접근할 수

있는 중앙 레지스트리에 웹 서비스의 세부사항을 공개(Publish)함.

· 서비스 사용자 (Service Consumer: Client): 웹 서비스를 사용하는 애플리케이션을 의미함. 레지스트리를 검색(Find)하여 서비스의 세부사항을 알아내고. 그 후 원하는 서비스에 binding하여 실제로 해당서비스의 기능을 이용함.

· 서비스 레지스트리 (Service Registry): 서비스 제공자가 자사의 웹 서비스 상세내역을 올려두고 Client가 발견할 수 있도록 하는 중앙 저장소를 의미함.

(그림 1)에 보이는 세 가지 오퍼레이션[검색, 선택, 공개]에 관한 표준안은 WSDL (Web Service Description Language), SOAP(Simple Object Access Protocol), UDDI(Universal Description, Discovery, and Integration)로 구성된다. 각각에 대해서 살펴보면 다음과 같다.

· WSDL : 웹 서비스 기술 (description)에 관한 표준안

WSDL은 서비스에 대한 자세한 정보를 명시하는 데 사용하는 언어로 애플리케이션이 서비스를 어떻게 호출하는지와 어떤 매개변수들이 필요한지를 알게 해준다. 실제 WSDL 문서는 복잡하며 import, type, schema, message, port, binding, service와 같은 엘리먼트들을 포함하고 있다.

· SOAP : 웹 서비스 선택(Binding)에 관한 프로토콜

SOAP은 XML포맷을 사용하여 애플리케이션이 웹서비스의 메소드를 호출할 수 있도록 하는 인터넷 프로토콜로서, 서비스가 요구하는 정보를 포함하고 잇는 문서를 애플리케이션에서 서비스로 건네주고 응답을 받는다. SOAP역시 XML에 기반하기 때문에 애플리케이션이 동작하는 운영체제, 언어 그리고 객체 모델에 관계없이 애플리케이션 사이의 통신을 할 수 있게 된다.

· UDDI : 웹 서비스 공개(publish)와 검색(find)에 관한 표준안

UDDI는 클라이언트가 동적으로 웹 서비스를 찾을 수 있게 하는 방법을 제공한다. UDDI 인터페이스를 이용하면 비즈니스는 외부 비즈니스 파트너가 제공하는 서비스에 동적으로 연결 할 수 있다. UDDI 레지스트리는 두 가지 종류의 클라이언트를 가지고 있다. 첫째는 서비스를 공개(Publish)하려는 비즈니스이고, 둘째는 서비스를 얻으려는 클라이언트이다.

2.2 바이오 데이터 통합 시도 연구 동향

Lincoln D. Stain 은 바이오 데이터베이스의 통합을 위한 기존의 접근 방법을 링크기반의 통합, 뷰의 통합, 데이터웨어하우징(DW) 구축을 통한 통합 접근, 이 세 가지로 구분하고 있다[1]. 이들 세 가지 방법의 특징을 살펴보면 다음 <표 1>과 같다. 최근에 들어서, 바이오 데이터 통합 문제를 다루는 많은 개발 커뮤니티들이, 앞에서 살펴본 기존의 통합방법에서

보이는 설계, 구축, 유지 보수의 효율성 면에서 문제점을 인식하고, 웹 서비스란 새로운 방법을 사용하여 바이오 데이터의 통합을 시도하고 있다. 웹 서비스는 상이한 플랫폼에 상관없이 인터넷이 가능한 곳이라면 쉽게 시스템에 접근 할 수 있도록 하는 기술로, 바이오 데이터베이스 분야에서도 DB상호간의 연동성을 가지게 하여 의미 있는 정보 추출의 효율성을 높일 수 있다. 근 한 4~5년 전부터 바이오 분야에서도 그 영향이 확대되고 있으며, 현재 대표적인 웹 서비스 기반의 연구로는 BioMoby[11], BioDAS[12], XML Central of DDBJ[13], XEMBL[14,15] 등이 있다.

(1) BioMoby

BioMoby는 웹 서비스를 통하여 바이오 데이터를 발견하고 배포하는 구조 개발을 목표로 하고 있는 Open Source Research 프로젝트로써 시스템 내부의 포맷이나 스키마에 상관없이 클라이언트가 바이오 데이터를 가지고 있는 많은 수의 Source들과 상호 작용할 수 있도록 해주는 시스템을 개발하고 있다.

기존의 많은 수의 바이오 정보나 분석 툴을 서비스 하고 있는 호스트들은 대부분이 이질적인 CGI 기술기반의 인터페이스를 통하여 사용자들에게 서비스를 제공하고 있으나, BioMoby는 웹 서비스를 통하여 사용자들에게 이질적으로 분산된 데이터베이스에 손쉽게 접근 가능토록 해주고 있는 점에서 의의가 있다.

(2) BioDAS

BioDAS(Bio Distributed Annotation System)는 유전체 서열 데이터의 주석 정보를 상호교환하기 위한 오픈 소스 시스템으로, 지리적으로 분산된 여러 주석 데이터 베이스 들에서부터 정보를 모아서 하나의 통합된 관점으로 데이터를 표현할 수 있도록 한다. 현재, TIGR, Wormbase, Ensembl, Flybase서버가 DAS/2 프로토콜로 데이터를 제공하고 있다.

(3) XML Central of DDBJ

DDBJ(DNA Data Bank of Japan)는 동양에서 가장 활발한 활동을 하고 있는 DNA 유전자 은행으로써 1986년 일본 National Institute of Genetics(NIG)에서 시작하였다. DDBJ는 10여 년간을 GenBank/NCBI, EMBL/EBI 와 함께 사용자들에게 Primary database를 제공하기 위해서 협력해왔고, 그 결과 데이터의 양은 기하 급수적으로 늘어났다. 그리하여 사용자들은 이러한 데이터들을 로컬 사이트에 다운로드 받기가 어려워 짐에따라, Center for Information Biology and DDBJ에서는 사용자에게 데이터들을 직접 다운로드 받지 않고 데이터에 접근 할 수 있는 컴퓨터 환경을 만들고자 하였는데 이것이 바로 웹 서비스를 이용한 시스템이다. DDBJ SOAP 서버와 WSDL의 조합으로, 사용자로 하여금 데이터베이스나 툴의 개발 없이도 자신만의 Computational 환경을 구축 할 수 있도록 하였다.

(4) XEMBL

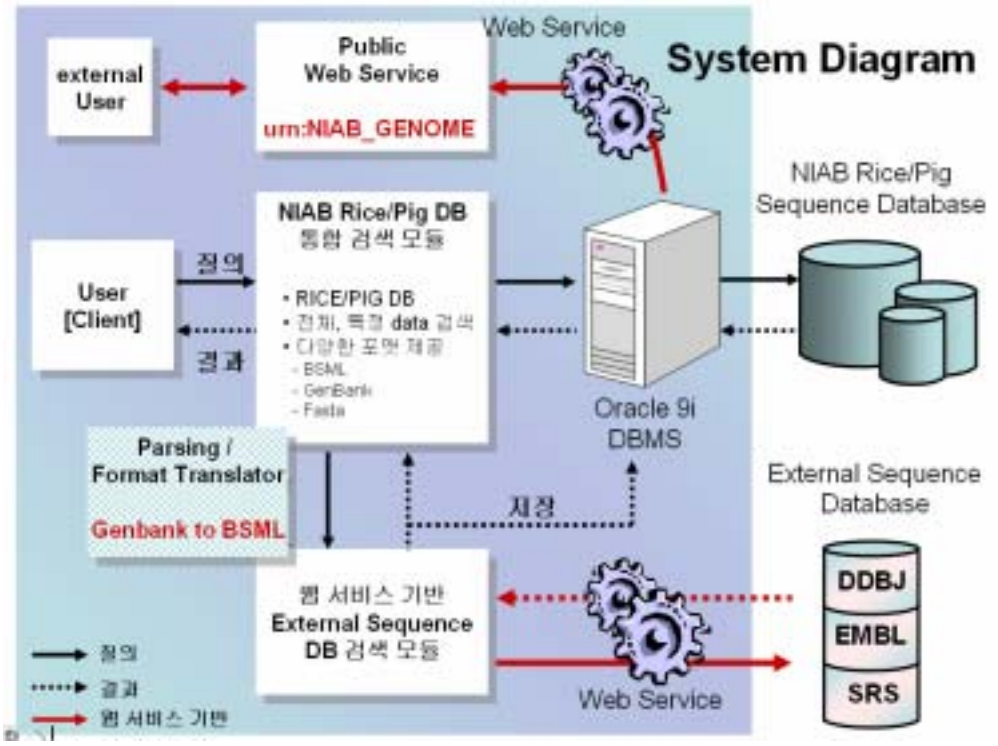
영국의 EBI(European Bioinformatics Institute)의 XEMBL서비스는 EBI의 EMBL 핵산 서열 데이터베이스를 다양한 XML형태로 제공하고자 하는 목적으로 만들어진 것으로, 웹 서비스 기술을 이용하여 EMBL데이터의 XML 형태의 접근을 쉽고 간단하게 해줄 수 있는 인터페이스를 제공한다. 개발자는 XEMBL사이트에서 XEMBL서비스의 WSDL 파일을(http://www.ebi.ac.uk/xembl/XEMBL.wsdl) 다운로드 받은 후, 웹 서비스 툴 킷을 이용하여 Stub과 Skelleton 파일을 생성함으로써, XEMBL사이트의 내부 시스템의 플랫폼을 몰라도 EBI의 EMBL데이터베이스에 접근 가능한 독자적인 애플리케이션을 작성할 수 있다. XEMBL이 현재 지원하고 있는 XML표준으로는 BSML과 AGAVE 두 가지이며, 앞으로 BIOML과 GAME을 지원하겠다고 말하고 있다.

구분	설명	
링크 기반의 통합 (link Integration)	특징	□ 가장 친숙하고 접근하기 쉬운 방법 □ 웹 기반의 하이퍼링크를 통해 데이터 접근
	단점	□ 상호 링크된 데이터베이스 간, 데이터의 모호성 □ 잘못된 링크를 따라가기 쉬움 - 업데이트 된 페이지에서 그렇지 않은 페이지로의 이동 □ 통합과 해석의 문제가 연구자의 책임
	예	□ SRS, Entrez 대부분의 통합 검색 시스템에서 채택
뷰 기반 통합 (View Integraion)	특징	□ 정보는 소스데이터베이스에 두면서, 하나의 큰 데이터베이스 환경을 구축하여 각각의 시스템을 전체 시스템의 한 부분으로 구성하는 방법 □ 쿼리 언어 분석하여 서브 쿼리로 나눔 - 얻고자 하는 정보를 가지고 있는 소스 데이터베이스로 서브 쿼리를 보냄 - 결과를 변환, 통합하는 소프트웨어 필요
	단점	□ 쿼리 처리 기능이, 가장 늦은 데이터 소스에 의해 제한됨
	예	□ Klesli/K2 – 펜실베니아 대학
데이터웨어하우징 (DW) 구축 (Data warehousing)	특징	□ 단일 데이터베이스, 즉 DW로 구축 □ 소스 데이터베이스들의 데이터 의미를 통합된 데이터 모델 사용 □ 다양한 SW 작성 - 소스 데이터베이스로부터 데이터를 가지고 오는 기능 - 가지고 온 데이터를 통합 데이터 모델로 변환 시키는 기능 - 통합 모델로 변환된 데이터를 DW로 로딩하는 기능
	단점	□ 소스데이터베이스에 새롭게 추가된 데이터의 실시간 업로드 문제 □ 소스 데이터베이스의 스키마 변경시 DW의 업그레이드 문제 - 소프트웨어의 기능 추가(버전 업 된 새로운 SW 개발 필요
	예	IGD Project(Integrated Genome Database)

<표 1> 기존의 바이오 데이터 통합 방법 비교

3. 바이오 서열 정보 통합 검색 시스템 설계

본 장에서는 바이오 서열 정보에 대한 통합 검색 시스템을 설계하고자 한다. 우선 (그림 2)은 웹 서비스를 이용한 유전체 서열 DB 통합 검색 시스템의 전체 구성도이다. 사용자의 입장에서 검색 할 수 있는 데이터베이스는 크게 NIAB(개발한 유전체 서열 DB을 이하 NIAB 라 칭함) 데이터베이스와 외부 서열 데이터베이스 등 두 부분으로 나눌 수 있다. 우선 사용자는 NIAB DB를 검색하고, 만약 NIAB DB에 존재하지 않는 아이템에 대한 검색을 할 경우, 웹 서비스를 통한 다른 외부 서열 데이터베이스로의 검색을 가능하도록 한다. 또한 NIAB 데이터베이스 검색 모듈을 웹 서비스 제공자의 입장에서 공개한다. 이에 관련된 내용은 아래에서 자세히 설명을 하겠다.



(그림 2) 시스템 구성도

3.1. 바이오 서열 정보 데이터베이스 구축

현재 NCBI[16] GenBank 데이터베이스에는 핵산 서열이 벼 유전자 595567개, 돼지의 유전자 169416개가 등록되어 있으며, 단백질 서열이 벼 64937개, 돼지 9335개가 등록되어 있다. 이중에서 핵산 서열은 EST서열만을 우선적으로 Bulk로 다운 받았고, 단백질 서열은

벼, 돼지 두 가지 모두 등록되어 있는 것을 모두 다운로드 받았다.

다운로드 받은 데이터는 사람이 읽고 해석하기 편한 텍스트파일의 형태로 ASCII 텍스트 파일이며, 그 크기는 모두 합쳐 1GB가 훨씬 넘는다. 이렇게 큰 텍스트 형태의 데이터를 데이터베이스로 구축하기 위해서는 데이터 전처리 과정이 필요하다. 다운로드 받은 데이터의 형식은 크게 GenBank 포맷과 FASTA 포맷 두 가지이고, 이러한 형태의 파일을 처리하기 위해 BioJava라이브러리[17]를 이용하여 자바 전처리 프로그램을 작성하여 구동하였다. BioJava는 http://www.biojava.org/download/binaries에서 구할 수 있으며, 바이너리 파일을 다운 받으면 된다. 현재 BioJava는 버전 1.3이 배포되고 있으며. 본 연구에서는 1.2.2 버전을 사용하였다.

3.1.1. BioJava 라이브러리를 이용한 데이터 전처리

NCBI에서 다운로드 받은 서열 데이터의 형태는 *.gbk란 이름의 텍스트 파일이며, 데이터베이스를 구축하기 위해서는 이러한 텍스트 파일의 데이터를 테이블의 형식에 맞게 잘라서 넣어야 한다. 본 연구에서는 텍스트 파일을 분석해 파싱한 다음 그 결과로 나온 데이터를 자동으로 입력하는 프로그램을 작성하였다. GenBank 텍스트 파일의 데이터 포맷은 키워드와 내용으로 구성되어 있다. 주어진 키워드에 대해 관련 내용이 기술되는 형식을 취하고 있다. 키워드는 반복되어 나타나는 키워드와 반복이 없는 키워드로 나누어진다. 반복 없는 키워드는 어떠한 GenBank 파일에서도 오직 한 번씩만 나타나는 것으로서, 이는 BioJava 라이브러리 내에서 제공되는 GenBank 텍스트 파일 지원 클래스를 이용하여 파싱할 수 있다. 그러나 반복되는 키워드의 경우는 각 GenBank 파일마다 각기 다른 형태를 가지며, 반복되는 횟수에도 차이가 있다. 반복되는 키워드 부분은 GenBank데이터의 상당 부분을 차지하고 있으며, 이로 인해 반복되는 키워드가 너무 많은 GenBank 파일의 경우는 파싱 작업 시 메모리 과부하가 일어난다. 이 경우 BioJava 라이브러리만으로는 파싱이 불가능하며, 따라서 이를 위해 클래스를 수정, 개선하였다. 본 논문에서는 새롭게 작성한 함수를 이용하여 반복 없는 키워드와 마찬가지로 반복되는 키워드에 해당하는 내용을 파싱할 수 있도록 하였다. FASTA 포맷의 파일은 구성자체가 서열 Description부분 한 줄과 나머지 서열로써 구성되어 있어 파싱 작업을 GenBank 파싱에 비해 훨씬 수월하게 할 수 있었다.

3.1.2 데이터베이스 스키마

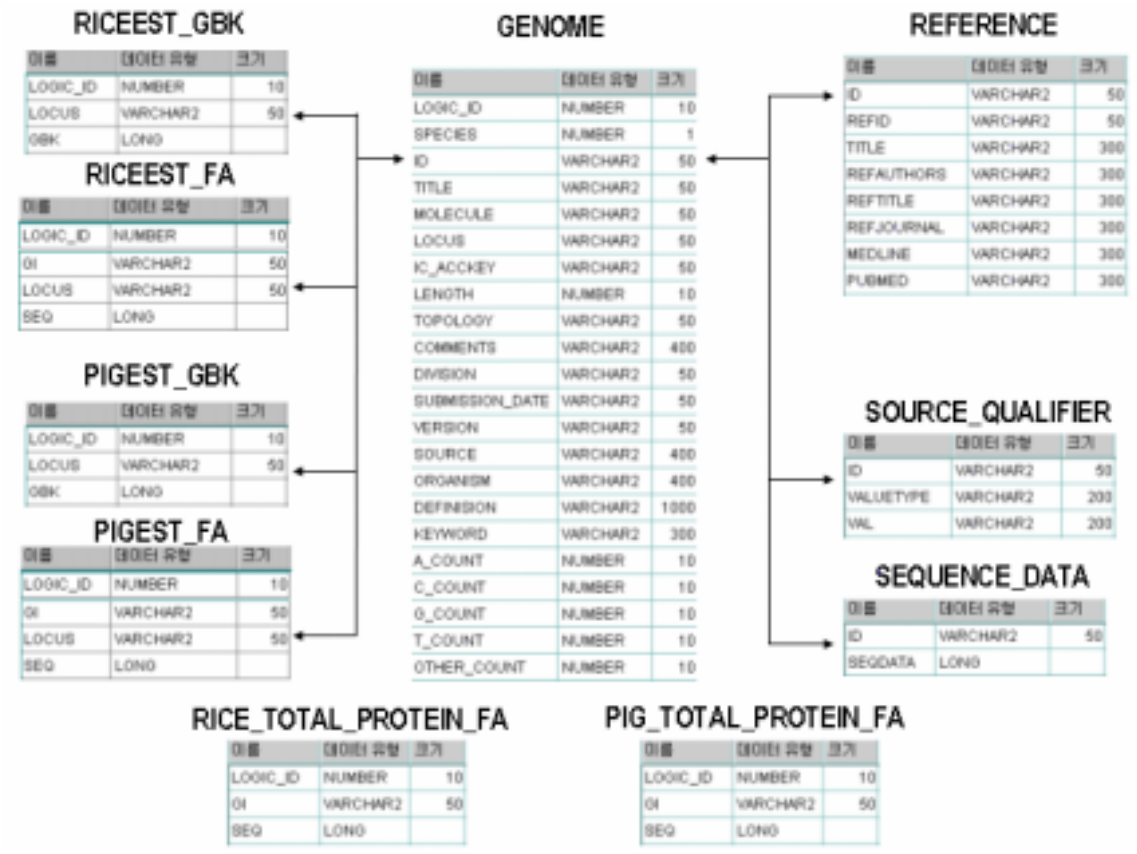
아래 <표 2>은 데이터베이스 테이블 설명이며, (그림 3)는 테이블과의 관계를 나타낸 스키마이다.

Table name	Table 설명
GENOME	해당되는 생물체의 서열이 언제 밝혀졌는지, 이 서열이 어떤 기능을 하고, sequence의 핵산의 구성이 어떻게 되는지에 대한 일반적 정보를 저장한다.

REFERENCE	관련된 논문에 관한 내용, 논문의 저자, 제목, 저널의 이름과 같은 정보저장.
SOURCE_QUALIFIER	서열의 길이, source organism의 이름, Taxon ID 번호 와 같은 각 레코드에 있는 Feature의 설명
SEQUENCE_DATA	Sequence 서열을 저장함.
RICEEST_GBK	종이 RICE이고 GenBank형식의 유전자 서열파일을 저장함
RICEEST_FA	종이 RICE이고 FASTA형식의 유전자 서열 파일을 저장함
RICE_TOTAL_PROTEIN_FA	종이 RICE이고 FASTA형식의 단백질 서열을 저장함
PIGEST_GBK	종이 PIG이고 GenBank형식의 유전자 서열파일을 저장함
PIGEST_FA	종이 PIG이고 FASTA형식의 유전자 서열 파일을 저장함
PIG_TOTAL_PROTEIN_FA	종이 PIG이고 FASTA형식의 단백질 서열을 저장함

<표 2> 데이터베이스 테이블 설명

데이터베이스 내에 있는 서열 정보를 BSML포맷에 맞추어 제공할 수 있어야 한다는 부분이 스키마를 설계하는데 있어 가장 중요하게 고려되었던 점이다. 서열 정보의 일반적인 정보를 담은 GENOME 테이블을 포함한, 모두 10개의 테이블로 핵산과 단백질 서열 정보를 충분히 표현할 수 있도록 했다.



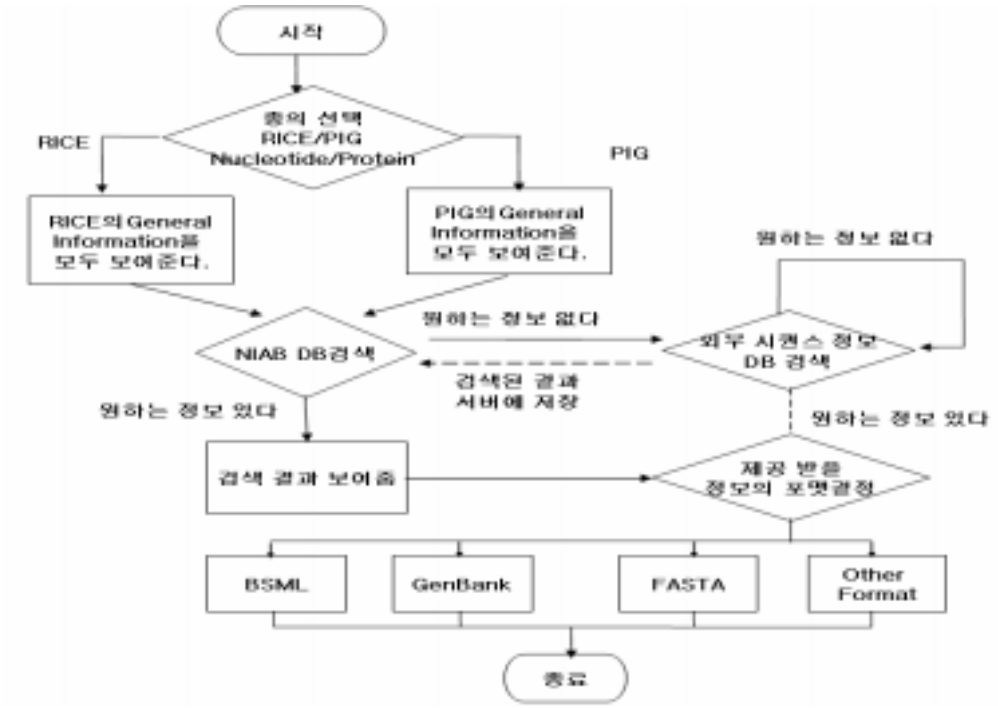
(그림 3) 데이터베이스 스키마

3.2. 웹 서비스 기반 DB 통합 검색 및 검색 모듈 공개

3.2.1. 웹 서비스 기반 바이오 서열 DB 통합 검색

메타 검색이란 여러 데이터베이스를 일일이 찾아 다니면서 검색하지 않고, 한 번에 검색 결과를 확인할 수 있는 장점이 있는 검색 방법을 말한다. 본 연구에서도 이러한 메타 검색 개념을 적용하여 NIAB 데이터베이스에 없는 정보에 대해서는 웹 서비스를 통한 외부 서열 데이터베이스로 검색을 할 수 있도록 했다. 사용자가 입력한 키워드로 검색된 결과가 벼나 돼지에 관련된 정보라면 사용자에게 검색 결과를 보여주는 동시에 내부적으로 NIAB 데이터베이스에 파싱되어 저장되도록 하였고, 그렇지 않다면 사용자에게 검색된 결과만을 보여주게 하였다. 다시 말해서 이 시스템은 시간이 지남에 따라, 벼나 돼지에 관련된 사용자의 질의가 많아질수록, 그 만큼의 데이터가 데이터베이스에 저장되어 정보가 축적되게 된다.

아래 (그림 4) 사용자가 질의를 입력했을 때의 검색 프로세스를 나타낸 것으로, 사용자는 우선 얻고자 하는 정보가 핵산 서열인지, 단백질 서열인지를 선택을 하고, 그 후에 벼와 돼지와 같은 종을 선택한다. 선택 파라티터 값에 의하여 우선적으로 NIAB 데이터베이스에 있는 전체 서열 아이템의 일반적인 정보를 보여주게 되며, 사용자가 Accession Number와 같은 서열의 키워드 정보를 추가로 입력함으로써 해당 아이템의 정보가 NIAB데이터베이스에 존재하는지 확인할 수 있다. 키워드로 입력한 정보에 해당하는 아이템이 NIAB 데이터베이스에 존재 하는 경우에는 검색 결과를 사용자 인터페이스를 통해 보여주고 다양한 포맷으로 데이터를 저장할 수 있도록 사용자에게 물으며, 만약 얻고자 하는 정보가 NIAB데이터베이스에 존재 하지 않을 경우에는 자동적으로 외부 서열 데이터베이스를 검색할 수 있도록 사용자 인터페이스를 구성한다. 사용자는 그 인터페이스를 통하여 외부의 데이터베이스를 선택적으로 접근 할 수 있다.



(그림 4) 통합 검색 프로세스 흐름도

(그림 4)에서 보듯이, 본 연구에서는 뒤에서 설명할 서비스를 제공자로써 구축된 데이터베이스의 검색 모듈을 공개함은 물론, 외부 시스템을 접근 하는 클라이언트, 즉 서비스 사용자로써의 역할도 할 수 있도록 시스템을 설계하였다. 외부 서열 데이터베이스가 공개하는 해당 서비스의 WSDL(Web Service Description Language)파일을 다운로드 받은 후, 웹 서비스 툴킷에 포함되어 있는 JAVA2WSDL 툴을 사용하여 클라이언트 측 스텁 코드를 생성했다.

클라이언트-사이드 스텁은 마치 메소드가 로컬에 있는 것처럼 애플리케이션이 원격 웹 서비스의 메소드를 호출할 수 있게 해준다. 아래 <표 3>은 본 시스템에서 접근하는 외부 서열 데이터베이스의 목록이다. 서버는 크게 일본의 XML Central of DDBJ와 유럽의 EBI 두 개이며, 서비스 개수는 모두 GetEntry, SRS, DDBJ, XEMBL 등 모두 4개이다.

서비스이름	제공기관	함수 개수	역 할
GetEntry	XML Central for DDBJ	27개	Accession Number 등등의 서열 해당 정보로 결과 추출
SRS	XML Central for DDBJ	2개	원하는 데이터베이스에 SRS쿼리 폼으로 질의 가능
DDBJ	XML Central for	6개	Locus, Gene, Project 정보를 입력으로 하여

	DDBJ		DDBJ 엔트리 추출
XEMBL	EBI	1개	Accession Number를 받아 BSML, Agave형태의 XML포맷으로 데이터 추출

<표 3.2> 외부 서열 데이터베이스 검색 서비스

3.2.2. 웹 서비스 제공자로서 DB 검색 모듈 공개

앞에서는 유전체 서열 정보를 담아 구축한 데이터베이스의 검색과 외부 데이터베이스를 어떻게 통합하여 검색할 것인가 하는 문제에 대해서 설명을 하였다. 이번 절에서는 구축한 NIAB 유전체 서열 DB와 함께 시스템이 외부 사용자에 의해 접근 가능토록 즉, 서비스 제공자의 역할도 할 수 있도록 앞에서 개발한 다양한 검색 서비스를 웹 서비스로 공개하였다. 서비스 이름은 urn:NIAB_GENOME이라 하였으며 그 서비스 내에 이용할 수 있는 함수는 모두 8가지로 구성하여 개발하였다 <표 4>.

함수 명	함수 설명
getNIAB_GENOME_Info	서열 데이터의 일반적인 정보 검색
getNIAB_GENOME_Sequence	서열 데이터만을 검색
getNIAB_REFERENCE_List	서열에 관련된 논문에 관한 내용, 저자 등 검색
getNIAB_GENBANKEntry	해당서열의 GenBank 파일 검색
getNIAB_GI_Number	해당 서열의 GI Number 검색
getNIAB_Nuc_FASTAEntry	해당 서열의 FASTA 파일 검색
getNIAB_Protein_FASTAEntry	해당 단백질 서열의 FASTA 파일 검색
getNIAB_BSMLEntry	핵산 서열의 BSML파일 검색

<표 4> 8개의 NIAB 데이터베이스 검색 모듈

3.2.3. 웹 서비스 레지스트리 검색 모듈

웹 서비스 기술은 사용자의 직접적인 제약 없이, 서비스 개발자가 서비스를 가능한 많은 클라이언트에게 제공할 수 있어야 하고, 클라이언트 역시 많은 서비스 제공자들로부터 원하는 서비스를 찾을 수 있어야 한다. 이러한 문제를 다루고 있는 기술이 UDDI로써, UDDI는 비즈니스와 해당 서비스에 대한 정보를 구조화된 방법으로 수용하기 위해 디자인된 공용 레지스트리로, UDDI를 통해 개발 커뮤니티와 커뮤니티가 제공하는 웹 서비스에 대한 정보를 게시하고 검색할 수 있다. 현재 제품화 되어 서비스 중인 공개 UDDI레지스트리는

세가지로 나누어 볼 수 있다. (IBM 레지스트리, Microsoft 레지스트리, HP레지스트리, 기타 사설 레지스트리). 본 논문에서는 바이오 분야의 공개된 사설 레지스트리인 BioMoby 레지스트리를 접근하는 애플리케이션을 개발했다.

3.3. 다중 질의문의 병렬 웹 서비스 검색

본 논문에서는 바이오 서열 정보 검색시 질의를 효율적으로 처리하기 위하여 다중 질의문의 병렬 웹 서비스 검색 기능을 제안하고자 한다. 최근에 웹 서비스 기술을 통하여 바이오 정보를 통합하고 분석하려는 환경을 제공하려는 많은 연구가 진행되고 있으나, 동일 서비스를 하는 서버는 그리 많지 않다. 그래서 우선 구현할 시스템에서는 동일한 파라미터 값을 갖는 여러 질의문을 동시에 각각 서버에 보낼 경우를 고려하여, 자바 쓰레드 기법을 통한, 웹 서비스 검색을 병렬화 할 수 있도록 했다. 이에 대한 성능 검증은 4장에서 설명하겠다. 본 연구를 통하여 얻은 결과는 향후 동일 서비스를 하는 서버가 여러 개 존재할 경우, 질의문에 대한 서버별 검색을 병렬화 하여 검색기능의 향상을 기대해 볼 수 있다.

4. 구현 및 결과

이 절에서는 3절에서 밝힌 바이오 데이터 통합 검색 시스템 설계를 토대로 구현한 결과를 제시한다.

4.1 구현 환경

운영체제	Windows 2000 Advanced Server
서블릿 엔진(웹 서버)	Tomcat v4.0.6
DBMS	Oracle 9i
웹 서비스 개발 도구	IBM WSTK v3.3.2, Axis v1.1, IBM UDDI Registry
개발 언어	JAVA, JSP, JAVA Bean
라이브러리	BioJava library v1.2.1 , Xerces v2.2.0(XML Parser)
데이터 연결	ODBC:JDBC

<표 5> 시스템 개발환경

본 논문에서 시스템을 구현하고 테스트한 환경은 <표 5>와 같다. DBMS는 오라클 9i release2를 사용했고, 데이터 처리를 위해서 BioJava 라이브러리를 이용하였다. 웹 인터페이스로 검색 서비스를 구현하기 위해서 IBM 웹 서비스 툴 킷을 이용하여 생성한 각

서비스의 스텁 코드는 JAVA Bean으로 서블릿 엔진에 설치하고, 그를 이용하기 위한 GUI는 JSP로 작성하였다.

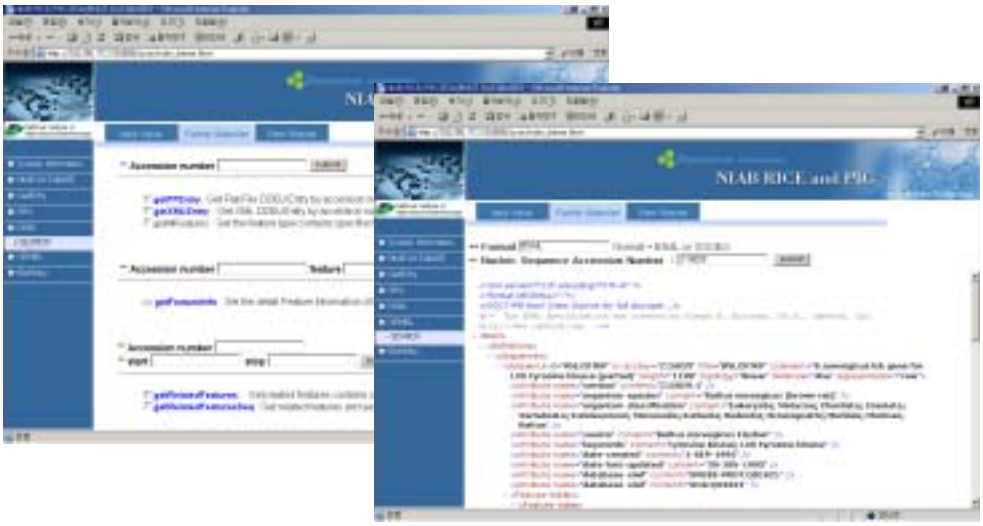
4.2 구현 결과

4.2.1 서비스 사용자측 모듈 – 통합 검색 클라이언트

본 시스템에서 접근하고자 하는 외부 서열 데이터베이스는 크게 일본의 XML Central of DDBJ, 유럽의 EBI서버 두 가지로 나눌 수 있다. 이 두 서버로부터, 유전체 서열 분야에 해당하는 서비스의 종류를 4가지(GetEntry, SRS, DDBJ, XEMBL)로 한정하여 통합 검색 클라이언트를 개발하였다. 접근하는 방법은 다음과 같이 두 가지로 설정하여 사용자 인터페이스의 메뉴를 구성하였다.

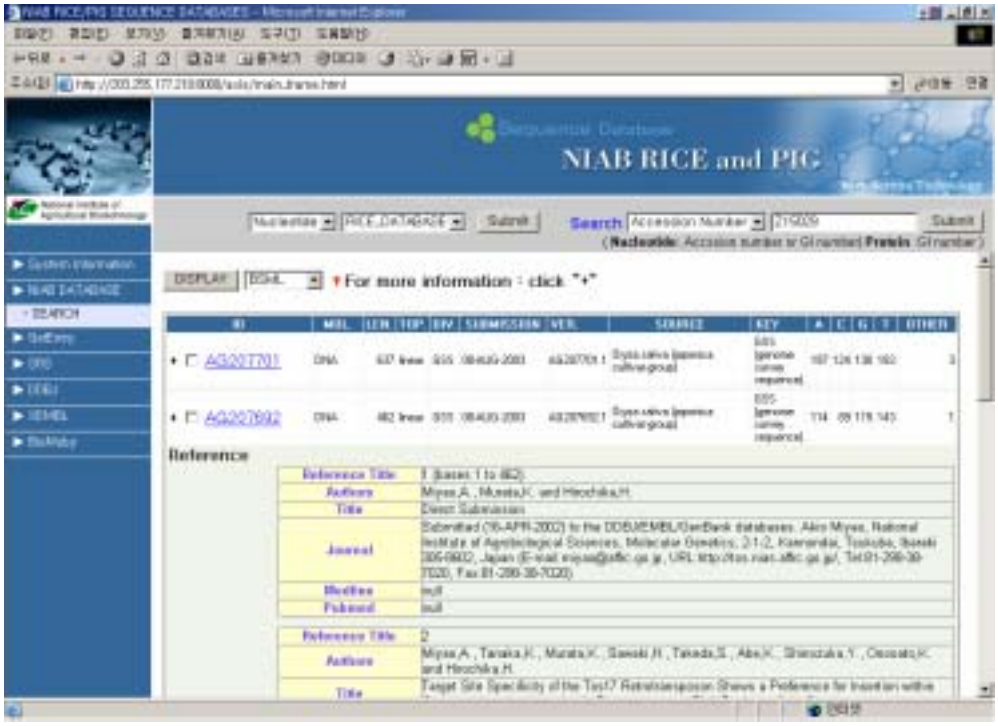
- ✓ 외부 데이터베이스의 직접 접근
- ✓ 구축한 NIAB 데이터베이스 검색 후, 검색 결과가 없을 시에 자동 접근

(그림 5)의 (가)는 DDBJ서비스의 함수 선택 인터페이스이며, (나)는 XEMBL서비스 인터페이스로 Accession Number Z15029의 BSML형식의 정보를 검색한 화면이다. (그림 6)은 벼의 핵산 서열을 NIAB데이터베이스에서 검색하여 조회하는 인터페이스이다. Accession Number앞의 '+' 기호를 클릭하면 해당 아이템에 대한 자세한 서열 정보를 펼쳐 볼 수 있도록 구성하였다.



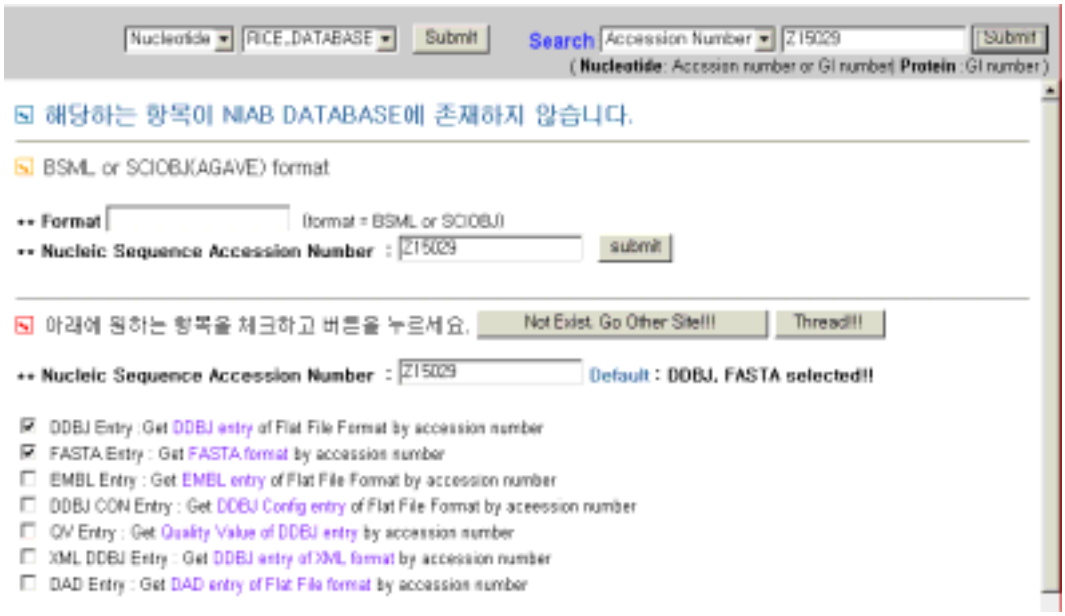
(가) (나)

(그림 5) 외부데이터베이스로의 직접 접근 인터페이스



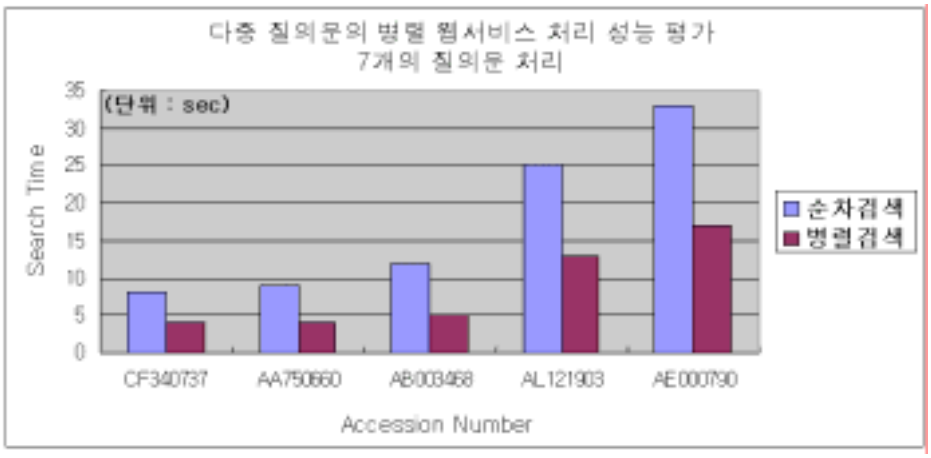
(그림 6) 벼의 핵산 정보 조회

전체 목록을 조회함은 물론, 사용자가 원하는 정보에 대한 키워드를 입력하여 특정 서열에 대한 정보도 검색할 수 있도록 하였다. 아래 (그림 7)은 예로 Accession Number가 Z15029인 서열을 검색했을 때 NIAB 데이터베이스에 없을 경우의 사용자 인터페이스로, 선택적으로 외부 데이터베이스를 검색할 수 있도록 했다.



(그림 7) 검색 결과가 없을 시, 외부데이터베이스 접근 인터페이스

이때, 동일한 파라미터 값에 대한 사용자의 다중 질의문을 어떻게 처리할 것인가 하는 문제가 생긴다. 다중 질의문의 처리를 순차 검색과 쓰레드로 처리한 병렬 검색 두 가지로 성능평가를 해 보았다.



(그림 8) 다중질의문의 병렬 웹 서비스 처리 성능 평가

(그림 8)은 동일한 파라미터를 가지고 DDBJ, FASTA, EMBL, XML포맷 엔트리 외에 서열의 Config정보, 텍스트 파일 포맷의 DAD엔트리 정보를 추출하는 7개의 질의문을 하나의 서버에 각각 병렬로 보낼 경우, 전체 7개의 질의를 모두 처리하는데 걸린 시간을 측정한 것이다. 서열 데이터 정보의 크기가 다른 5개의 서열 Accession number를 파라미터 값으로 실험을 해본 결과 병렬 검색이 순차 검색에 비해 두 배 정도의 성능이 우수하다는 것을 알 수 있었다.

이는 앞으로 동일 서비스를 하는 서버가 여러 개 존재 할 경우, 서버 별 웹 서비스 검색을 병렬화 하여, 서버 별 검색 시간을 고려한 제어를 통하여 검색의 효율을 높일 수 있을 것으로 기대한다.

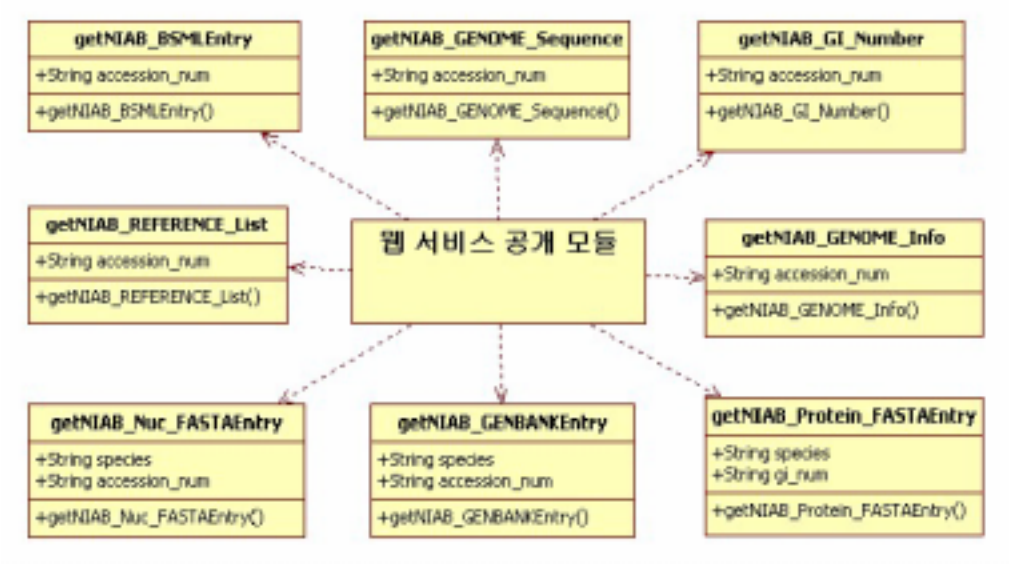
4.2.2 서비스 제공자측 모듈 – 검색 서비스 공개

(가) 데이터베이스 구축

우선, 생물 종을 벼와 돼지로 한정한, 서열 정보 데이터, 즉 대용량의 GenBank 텍스트 파일을 NCBI 사이트에서 다운로드 받는다. 다운로드 받은 파일을 BioJava 라이브러리를 이용해서 파싱하고, 3절에서 제시한 스키마에 알맞게 정보를 추출하여 데이터베이스 테이블에 저장한다.

(나) 데이터베이스 검색 서비스 공개

서비스 제공자라 함은 웹 서비스 기술을 이용하여, 파라미터 입 출력 값만으로도 검색과 같은 서비스를 외부 사용자에게 공개할 수 있는 시스템을 말한다. 본 연구에서는 앞에서 구축한 데이터베이스를 다양한 입 출력 값으로 검색하는 모듈을 개발하고, 이를 공개한다. 아래 (그림 9)는 검색 모듈의 파라미터를 살펴볼 수 있도록 클래스 다이어그램으로 표현해 본 것이다.



(그림 9) 웹 서비스로 공개할 검색 모듈의 클래스 다이어그램

서비스 이름은 urn:NIAB_GENOME이라 하고, AXIS서버에 공개하기 위한 WSDD 파일을 작성한 후, 커맨드 창에서 %java org.apache.axis.client.AdminClient deploy.wsdd 라 명령을 주면 정상적으로 등록되었다는 메시지와 함께 완료된다.

(그림 10) 정상적으로 서비스 공개 후 확인

AXIS 서버에 등록이 잘되었는지 확인은 익스플로러에서 서버 홈 밑의 axis/servlet/AxisServlet 경로로 들어갔을 때, (그림 10)과 같이 서비스 이름과 개발된 함수 명이 제대로 보여지면 정상적으로 등록이 된 것이다

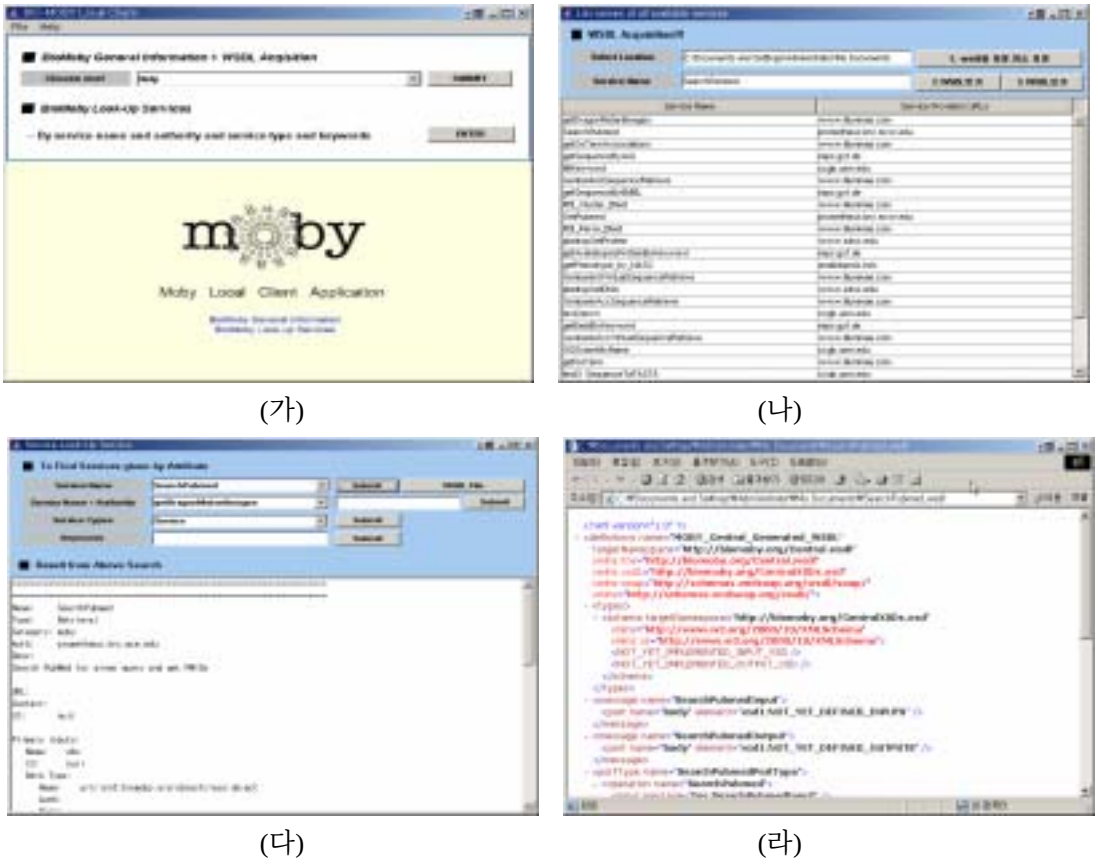
이렇게 웹 서비스 서버로 등록이 완료 되면, 이제 외부 사용자 누구나, 아래 (그림 11)의 개발된 데이터베이스로의 검색 기능이 명시된, 공개된 WSDD의 파일을 가지고, 본 연구에서 개발한 데이터베이스를 검색하는 자체적인 애플리케이션을 개발할 수 있다.

(그림 11) 웹 서비스로 공개하기 위한 WSDL파일

4.2.3 서비스 레지스트리 검색 모듈

본 절에서는 바이오 분야의 공개된 사설 레지스트리인 BioMoby 레지스트리를 접근하는 애플리케이션을 개발했으며 이를 공개한다. BioMoby 레지스트리 검색 애플리케이션을 JAVA WEB START(따로 설치 없이 바로 다운로드 받아 사용 가능)를 사용하여 바로 실행시킬 수 있다. BioMoby 레지스트리 검색 클라이언트 애플리케이션을 이용하여

BioMoby 레지스트리에 공개되어 있는 서비스들의 목록을 조회하고, 해당서비스의 WSDL파일을 생성시켜 사용자의 PC에 저장 할 수 있도록 프로그래밍 하였다.



(가) (나)

(다) (라)

(그림 12) BioMoby 레지스트리 검색 클라이언트

(그림 12) 의 (가)는 BioMoby 레지스트리 검색 클라이언트의 초기 화면이며, (나)는 BioMoby 레지스트리에 공개되어 있는 모든 서비스의 목록을 표로 나타내어 조회 가능토록 했으며 , (다)는 레지스트리에 공개되어 있는 서비스들을 여러 파라미터를 주어 검색 가능하도록 한 화면이다. (나) (다) 에서 WSDL생성 버튼을 누르면, 사용자의 컴퓨터에 WSDL파일이 저장되며, 필요할 경우 익스플로러로 WSDL파일을 열어볼 수 있다.

5. 결론 및 향후 과제

본 논문에서는 기존의 바이오 정보 통합 시스템의 특징을 간단히 살펴본 후, 최근 세계적인 많은 바이오 정보 시스템에서 채택하고 있는 웹 서비스 기술을 기반으로 한 새로운 정보 검색 시스템을 제시하였다. 우선 대용량 텍스트 파일로 받은 벼와 돼지의 핵산

서열과 단백질 서열 데이터를 바이오 자바를 이용한 전처리 과정을 통하여 데이터베이스로 구축하였으며, 이를 검색 할 수 있는 검색 모듈을 자바언어를 이용하여 개발하였다. 데이터베이스에 없는 정보에 대해서는 웹 서비스를 통한 여러 외부 서열 데이터베이스로의 검색을 메타 검색 개념을 적용하여 병렬로 처리 할 수 있도록 하였다. 또한 검색으로만 끝나는 것이 아니라, 검색한 정보가 벼와 돼지분야의 것이면, 그 데이터를 데이터 파싱 작업을 통해 새롭게 구축된 데이터베이스에 축적시킨다. 또한 다양한 검색 기능을 통해, 다양한 포맷으로 서열 정보를 얻을 수 있도록 한 본 시스템이 웹 서비스 아키텍쳐의 서비스 제공자로써의 역할을 할 수 있도록 검색 기능을 웹 서비스로 공개하고, UDDI 레지스트리에 공개하여 사용자로 하여금 접근 가능토록 했다.

웹 서비스 기술을 통하면 큰 어려움 없이, 외부 시스템의 어플리케이션을 이용할 수 있기 때문에, 최근 바이오 분야에서도 웹 서비스가 각광을 받았다. 그러나 웹 서비스 기술 자체가 보안, 프라이버시의 문제, 효율적인 메세징, 라우팅 기술이 요구된다는 문제점도 가지고 있다. 본 연구에서 개발한 시스템도 예외가 아니다. 또한 내용적인 면을 살펴보면, 지금 상태의 시스템은 바이오 데이터의 일차적인 정보만을 제공하는 단순 데이터 검색이라 할 수 있다. 이는 향후 데이터 정보간의 상호 연결성을 고려하여 보다 의미 있는 정보를 제공할 수 있도록 하는 연구가 필요할 것으로 생각된다.

참고문헌

[1] Stein, L., 'Integrating Biological Databases', Nature Reviews-Genetics, Vol4, 337-345, 2003

[2] 유성준, 김용국, 박성호, 박성희, '웹 서비스 기반 바이오 정보 통합 기술 동향', 데이터베이스 연구회지, 19권 1호, 2003

[3] Frishman D, Heumann K, Lesk A, Mewes HW. Comprehensive, comprehensible, distributed and intelligent databases: current status. Bioinformatics.;14(7):551-61. Review, 1998

[4] Achard, F., G. Vaysseix and E. Barillot, 'XML, Bioinformatics and Data Integration', Bioinformatics Review 17(2), pp. 115-125, 2001

[5] Stein, L. 'Creating a Bioinformatics Nation', Nature 417, pp. 119-120, 9 May 2002

[6] 정보문화사 Professional Java Web Services

[7] Oracle9 i Application Server 웹 서비스 기술 백서
http://otn.oracle.co.kr/tech/webservices/pdf/webservices_twp.pdf

[8] Labbook, Inc., BSML DTD and Genomic XML Viewer. http://www.labbook.com

[9] GenBank : http://www.ncbi.nlm.nih.gov/Genbank/GenbankOverview.html

[10] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids

Res.;31(1):23-7, 2003 Jan 1

[11] Wilkinson MD, Links M. BioMOBY: 'an open source biological web services proposal.', Brief Bioinform.;3(4):331-41, 2002 Dec

[12] http://www.biodas.org/, Brian King, A Web Services Description of DAS, 2002.

[13] http://xml.nig.ac.jp/index.html. DDBJ Project

[14] Wang L, Riethoven JJ, Robinson A. XEMBL: distributing EMBL data in XML format. Bioinformatics. (8):1147-8, 2002 Aug;18

[15] Wang, L., P. Rodriguez-Tome, N. Redaschi, P. McNeil, A.J. Robinson, and P. Lijnzaad., Accessing and distributing EMBL data using CORBA (common object request broker architecture), Genome Biology 1(5), 2002 Aug;18

[16] Entrez online documentation : http://www.ncbi.nlm.nih.gov/Database/index.html

[17] BioJava Tutorial : http://www.biojava.org/tutorials/index.html

<저자약력>

이수정

E-Mail: echocrystal@hanmail.net

2002년 이화여자대학교 공과대학 컴퓨터학과 졸업 (학사)

2004년 이화여자대학교 과학기술대학원 컴퓨터학과 (공학석사)

2004년 - 현재 삼성전자 정보통신총괄 근무중

용환승

<논문지 제 11-D권 제2호 참조>