

웹 서비스 기반 유전자 주석정보 통합검색 시스템 구축

이 회 전†, 용 환 승

이화여자대학교 컴퓨터학과 데이터베이스 연구실

Development of Integrated Retrieval System

Based on Web Service for Gene Annotation Database

Hee-Jeon Lee, Hwan-Seung Yong

Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 바이오인포매틱스 분야에서는 유전자 주석정보 데이터들의 통합 방안에 대한 논의가 활발하게 진행 중에 있다. 본 논문에서는 BioDAS의 웹 서비스 개념을 이용, 분산된 주석 데이터 서버들간의 통합검색 시스템을 구축함으로써 메타검색 시스템을 구현하였다. 본 시스템은 사용자에게 메타검색 기능 및 결과 저장기능을 제공해 주며 외부 사용자에게 웹 서비스를 제공한다.

한 면을 지니고 있다.

본 연구에서는 유전자 서열 주석정보들의 통합 검색을 위하여 웹 서비스 기술을 기반으로 분산된 데이터베이스 중, 특히 유전자 주석 데이터와 관련된 기존 데이터베이스들로부터 통합검색 시스템을 설계하고 구축하였다. 데이터베이스 스키마는 BioPerl[1]의 GFF(General Feature Format)[2] 스키마를 기반으로 하였다. 구현한 시스템은 메타 검색 기능 및 저장기능이 가능하기 때문에 사용자는 여러 데이터베이스에 접속하여 자신이 원하는 정보를 개별적으로 찾는 번거로운 검색 작업을 하지 않아도 된다. 또한 시스템은 웹 서비스 기반의 저장 기능을 기반으로 하기 때문에 쉽게 통합, 확장이 가능할 것으로 기대된다.

본문의 구성은 다음과 같다. 제 1장 서론에 이어 제 2장에서는 연구와 관련해 웹 서비스와 BioDAS[3]에 대해 소개하였다. 제 3장에서는 통합

1. 서 론

오늘날 주석(annotation) 정보 데이터베이스 관련 연구자들은 계산적인 방법과 실험적인 방법 등을 통해 유전자 서열 데이터의 주석들을 통합해 나가고 있다. 그러나 이들은 유전자 주석 데이터들을 위해 편리한 원스톱 소스를 제공해 주는 것이 아니라, 어떤 특정 영역에 대한 정보를 위해 다양한 웹사이트들을 체크하여 주거나 여러 다른 형태의 데이터들을 FTP를 통해 다운로드 받거나 혹은 전체 그림을 얻기 위해 메뉴얼하게 통합을 수행하는 수준이다. 이러한 방법들은 중앙 데이터베이스의 사용자 제한과 같은 정책상의 문제나 실시간 통합이 용이하지 않은 등의 기술적 문제로 인한 한계점을 가지고 있기 때문에 유전자 주석정보의 통합 시스템으로 부족

검색 연동시스템의 시스템 디아이어그램 및 아키텍처에 관해 논의하고 제 4장에서는 시스템 구현에 대해 설명하였다. 제 5장에서는 결론 및 향후 과제에 대해 논하였다.

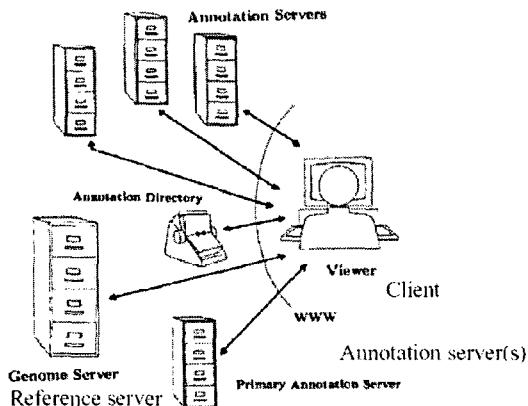
본 시스템은

<http://203.255.177.216:8000/axis/body.html>에서 사용이 가능하다.

2. 관련기술 및 연구 동향

본 연구와 관련, 통합검색 시스템 구축을 위해 사용한 웹 서비스 기술과 BioDAS 개념에 대해 살펴본다.

웹 서비스란 인터넷이나 네트워크로 다른 객체에 RPC(Remote Procedure Calls)를 수행하는 기술로써, 플랫폼 종립적인 표준인 HTTP나 XML을 사용함으로써 클라이언트에게 전체 시스템 구현을 숨길 수 있는 장점을 가지고 있다[4]. 웹 서비스 아키텍처는 발행(publish), 검색(find), 바인드(bind)라는 세 가지 기본적인 오퍼레이션을 지원한다. 서비스 제공자는 서비스 중개자에게 서비스를 발행하고 서비스 수요자는 서비스 중개자를 이용해서 서비스를 검색하며, 이때 서비스 수요자와 서비스 제공자 사이에 바인딩이 일어난다[5].



[그림 1] DAS 아키텍처

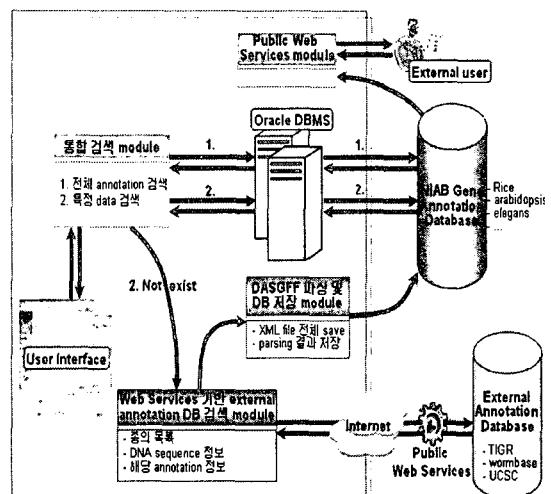
그림 1은 DAS의 구성도이다. BioDAS(Distributed Sequence Annotation System)는 유전자 서열 데이터의 통합을 위한 방안으로, 하

나의 클라이언트가 다양한 서버들로부터 정보들을 통합하는 client-server 시스템이다. 이는 HTTP를 통해 클라이언트가 요청한 주석처리된 정보를 XML 문서로 보여준다[6].

시스템에서 하나의 레퍼런스 서버는 source와 주석 데이터 관련 서버들을 검색해 준다. 하나 혹은 그 이상으로 구성되는 주석 데이터 관련 서버는 요청한 서열 정보를 GFF 형식으로 보여준다. GFF 형식은 Sanger Institute에서 제안한 General Feature Format 또는 Gene Finding Format이라고도 불린다. GFF 형식은 실험결과를 표현하기 용이하고 유전자 파싱의 통합적인 정보를 얻을 수 있다. 뿐만 아니라, 예측을 보고하기에도 용이하며 가시화를 쉽게 할 수 있는 등의 장점을 지니고 있다. 현재 DAS 시스템에 참여하고 있는 서버로는 TIGR, wormbase, UCSC, Ensembl, flybase 등이 있다.

3. 유전자 주석정보 통합검색 시스템 구성

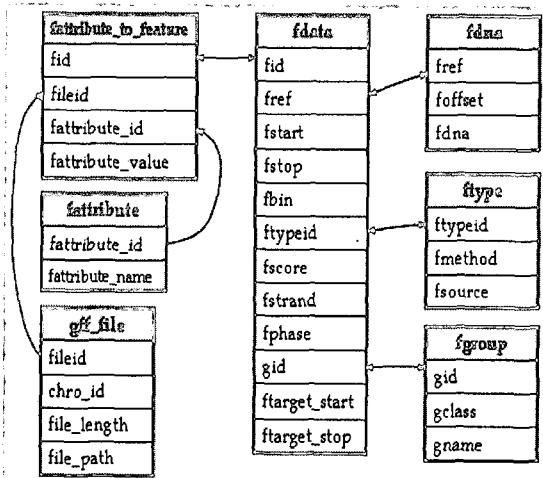
다음 그림 2는 통합검색 시스템의 시스템 디아이어그램이다.



[그림 2] 시스템 디아이어그램

그림 2의 로컬 데이터베이스(NIAB Gene annotation Database)에는 기본적으로 DAS 서버들로부터 웹 서비스를 통해 서비스 받은 *Arabidopsis*, *Rice*, *Elegans* 등에 관한 주석 데이터들이 로드되어

있다. 데이터베이스 스키마는 BioPerl의 GFF 스키마를 기반으로 하였다. 다음 표 1은 데이터베이스 스키마이다.



[표 1] 데이터베이스 스키마

웹 서비스 기반 DAS 클라이언트가 되기 위해 OmniGene 프레임워크[7]을 사용하였다. OmniGene은 웹 서비스와 J2EE 기술을 통해 생물학 관련 데이터들을 교환하는 표준화 작업 중의 하나로 데이터베이스에 접근하기 위한 미들웨어 역할을 하는 프레임워크이다. 웹 서비스 기술은 SOAP의 버전 3인 AXIS[8](Apache eXensible Interaction System) 버전 1.1을 사용하였다. 다음은 시스템의 각 기능들에 대한 설명이다.

① 통합검색 기능

사용자가 필요한 주석정보 데이터가 로컬 데이터베이스에 존재할 경우, 시스템은 오라클 DBMS를 거쳐 결과 데이터를 리턴해 준다. 그러나 원하는 정보 데이터가 로컬 데이터베이스에 존재하지 않을 경우, 시스템은 웹 서비스를 통해 그림 2의 외부 annotation 서버들로부터 주석정보 데이터를 서비스 받아 사용자에게 리턴해 준다. 이를 통해 사용자는 필요한 정보를 찾기 위해 여러 사이트에 접속할 필요 없이 UI를 통해 정보를 찾을 수 있는 메타 검색

이 가능하다.

② 웹 서비스 기반 외부 주석정보 데이터베이스 검색 기능

TIGR, Wormbase, UCSC 등 웹 서비스를 제공하는 서버들로부터 생물학 관련 정보 데이터들의 XML 형식 데이터를 받는 기능이다. 웹 서비스 기반 외부 데이터베이스 검색 기능을 통해 서비스 받을 수 있는 정보의 종류로는 서버로부터 데이터를 받을 수 있는 종의 목록을 리턴 해주는 DASDSN XML 데이터, 사용자가 특정 영역에 대해 질의한 DNA 서열 정보를 리턴 해주는 DASDNA XML 데이터, 마지막으로 해당 영역에 대한 주석정보를 리턴 해주는 DASGFF XML 파일이 있다.

③ DASGFF 파싱 및 데이터베이스 저장 기능

사용자의 요청 혹은 로컬 데이터베이스 관리자의 필요에 의해 서비스 받은 DASGFF XML 파일을 데이터베이스에 로드하는 기능이다. 이를 통해 서비스 받은 XML 파일 전체를 저장할 수 있고 SAX 파서를 통해 파싱 과정을 거쳐 각 항목을 데이터베이스 내 테이블에 저장한다.

④ public 웹 서비스 기능

통합검색 시스템 구축을 위해 생성한 자바 메소드들을 웹 서비스를 이용하기 원하는 외부 사용자들을 위해 배포하는 기능이다. 이를 위해 WSDD(Web Service Deployment Descriptor)[9] 파일을 만들고 웹 서비스를 제공하였다.

4. 통합검색 시스템 구현

본 시스템의 개발환경은 Windows XP professional이며 DBMS로는 Oracle 9i release 2를 사용하였다. 개발 언어는 Java이며 웹 서버는 Apache 버전 4.06을 이용하였다. 검색의 종류는 다

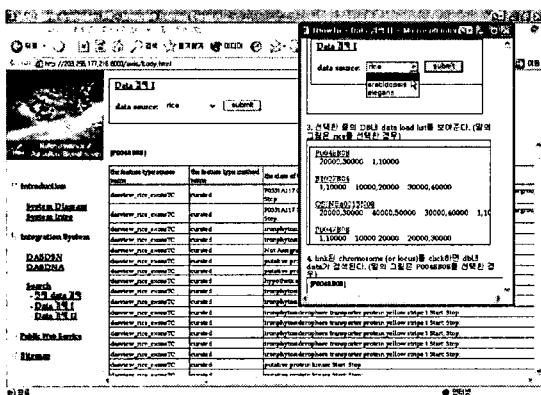
음과 같다.

① DASDSN, DASDNA 검색

사용자가 UI를 통해 주석정보 데이터베이스 서버를 선택하면 선택된 서버에서 가지고 있는 종의 목록인 DASDSN XML 파일을 웹 서비스를 통해 받을 수 있으며 특정 영역에 대한 유전자 서열 정보는 DASDNA XML 파일로 받을 수 있다.

② 통합검색

사용자는 로컬 데이터베이스에 저장되어 있는 전체 주석정보 데이터의 목록을 검색할 수 있다. 다음 그림 3은 검색결과에 대한 GUI이다.



[그림 3] 검색 GUI

검색 결과는 각각에 대해 로컬 데이터베이스 내에 저장되어 있는 해당 DASGFF 파일이 링크돼 있다. 또한 각 종에 대해 로컬 데이터베이스에 로드 돼 있는 locus 혹은 염색체 데이터의 전체 테이블이 검색 가능하다. 뿐만 아니라 사용자는 특정 영역에 대한 주석정보를 요청할 수도 있다. 만약 요청한 유전자 주석정보 데이터의 영역에 대해 로컬 데이터베이스가 정확한 값을 가지고 있지 않을 경우, 시스템은 검색을 요청한 영역을 포함하는 더 큰 영역의 데이터를 찾는다. 만약 더 큰 영역도 없다면 시스템은 웹 서비스를 통해 사용자가 요청한 영역의 데이터를 외부 데이터베이스 서버로부터 서비스 받아 사용자에게 보여준다. 사용자는 요청한 데이터의 결과를 보고 그 유용성을 판단해 로컬 데이터베이스 내에 주석정

보 데이터를 저장할 수 있다.

5. 결론 및 향후 과제

본 논문에서는 BioDAS와 OmniGene의 웹 서비스기술을 이용해 분산된 주석정보 데이터베이스 서버들을 이용한 통합검색 시스템을 개발하였다. 시스템은 사용자에게 유전정보에 대한 메타검색 기능을 제공해줄 뿐만 아니라, 저장 기능을 통해 시스템 확장의 용이성도 갖추고 있다. 이러한 시스템 개발을 통해 앞으로 방대한 양의 유전 정보를 좀더 체계적이며 조직적으로 관리할 수 있을 것으로 기대한다. 또한 웹 서비스를 배포함으로써 외부 사용자는 본 연구를 통해 구현한 시스템과 유사한 시스템을 로컬 시스템에서 쉽게 구현할 수 있다.

향후 과제로써 사용자에게 검색 결과를 좀더 사용자 친화적으로 보일 수 있게 하기 위해 사용자 인터페이스에 대한 연구와 함께 웹 서비스 기반 통합검색 시스템을 통해 얻은 주석정보 데이터를 효율적으로 이용하는 방안에 대한 연구가 필요할 것으로 기대된다.

[참고 문헌]

- [1] <http://www.bioperl.org/>
- [2] <http://www.sanger.ac.uk/>
- [3] <http://www.biodas.org/>
- [4] 정지훈, “웹 서비스”, 한빛미디어, 2002
- [5] S.Jeelani Basha, “Professional Java Web Service”, wrox, 2002
- [6] Robin D. Dowell and Lincoln Stein, “The Distributed Annotation System”, Research article, BMC Bioinformatics, 2001
- [7] <http://omnigene.sourceforge.net/>
- [8] <http://203.255.177.216:8000/axis/body.html/>, “AXIS User’s Guide”
- [9] Jurgen Kaljuvee, “Bioinformatics Data Integration Approaches via SOAP Web Services and XML”, O’REILLY Bioinformatics Technology Conference, February, 2003