웹기반의 GenBank 특허 데이터 검색 시스템의 설계 및 구현

양진옥o, 김상수
한국생명공학연구원
joy@joy.kribb.re.kr, sskimb@mail.kribb.re.kr

The Design and Implementation of a Web-Based Search Engine for GenBank Patent Data

Jin Ok Yang[0], Sangsoo Kim
Korea Research Institute of Bioscience and Biotechnology (KRIBB)

요 약

NCBI 의 GenBank 데이터베이스는 전세계에서 수집된 염기 서열 데이터들의 집합이며, 그 중 특허로 등록되어 있는 데이터들을 GenBank 특허 데이터라 부른다. 본 논문에서는 한국생명공학연구원의 유전체 사업단에서 개발해 오고 있는 웹기반 GenBank 특허 데이터 검색 시스템의 설계와 구현에 대해서 설명한다. 본 시스템은 일반 속성(attribute)을 저장하고 검색하기 위해 DBMS 를 사용하고, DNA 시퀀스 검색을 위해 BLAST 를 사용한 약결합 아키텍처(loosely-coupled architecture)를 채택하고 있다. 즉, 일반 속성으로 저장될 수 있는 데이터들은 데이터베이스의 테이블들의 컬럼 값으로 저장하고 SQL 언어를 통해 검색할 수 있도록 하였으며, DNA 시퀀스 검색을 위해서는 BLAST 에서 제공하는 인덱스를 구축하고 BLAST 명령어를 사용하여 검색할 수 있도록 하였다. 또한, 검색 결과들이 기존의 외부 특허 시스템과 연동하도록 하기 위해, 결과 분석 모듈을 구현하여 검색 결과들이 다른 웹 사이트의 데이터를 가리키도록 하였다. 마지막으로, 이러한 DNA 검색 시스템을 구현할 때에 고려해 되야 되는 이슈들을 설명한다.

1. 서 론

NCBI(National Center for Biotechnology Information) [NCBI01]는 NIH(National Institute of Health) 산하의 기관으로 생명 정보 기술을 다루기 위해 1988 년에 설립되었다. NCBI 는 염기 서열 데이터베이스인 GenBank 를 운영하고, 데이터 분석, 검색 등의 소프트웨어와 자료들을 제공하고 있다. 또한, NCBI 에서는 짧은 염기 서열인 EST, genomic 서열, OMIM 에서 제공하는 phenotypic description 등의 데이터들을 제공한다. 그리고, BLAST, Entrez 등의 강력한 검색 엔진을 제공하여 사용자들이 빠르게 원하는 정보를 찾아 볼 수 있도록 하고 있다.

GenBank[GenBank01]는 염기 서열 데이터베이스로서, 2001 년 6 월까지 12,244,000 개의 시퀀스를 가지고 있다. GenBank 데이터베이스는 그 크기가 지수적으로 증가하고 있으며, 따라서 이러한 데이터들을 웹에서 검색할 수 있는 시스템은 필수적이다.

본 논문에서는 한국생명공학연구원의 유전체 사업단에서 개발해 오고 있는 웹기반 GenBank 특허 데이터 검색 시스템 [JOY01]의 설계와 구현에 대해서 설명한다. 본 시스템은 일반 속성(attribute)을 저장하고 검색하기 위해 DBMS 를 사용하고, 염기 서열 검색을 위해 BLAST 를 사용한 다. 이를 위해, 일반 속성으로 저장될 수 있는 데이터들은 데이터베이스의 테이블들의 컬럼 값으로 저장하고, SQL 언어를 통해 검색할 수 있도록 하였다. 그리고, 염기 서열 유사성 검색을 위해서는 BLAST 에서 제공하는 인덱스를 구축하고 BLAST 명령어를 사용하여 검색할 수 있도록 하였다.

본 논문의 구성은 다음과 같다. 제 2 장에서는 본 논문에서 제안하는 웹기반의 GenBank 특허 데이터 검색 시스템의 아키텍처에 대해서 설명한다. 제 3 장에서는 GenBank 특허 데이터의 구조에 대해서 설명한다. 제 4 장에서는 본 시스템의 구성 모듈에 대해서 자세히 설명한다.. 제 5 장에서는 이러한 검색 시스템을 구축할 때 발생하는 여러 가지 이슈들에 대해서 토론하며, 제 6 장에서 결론을 맺도록 한다.

2. 시스템 아키텍처

본 장에서는 제안하는 시스템의 아키텍처에 대해서 간단하게 설명한다. 구성 모듈에 대한 자세한 설명은 제 4 장에서 설명하도록 한다.

그림 1 은 본 시스템의 아키텍처를 도식화하고 있다. 본 시스템은 크게 두 가지 기능을 제안한다. 첫번째 기능은 GenBank 특허 데이터 파일로부터 데이터베이스에 로딩될 수 있는 형태의 데이터로 변환하고, 이를 DBMS 의 데이터베이스와 BLAST 의 인덱스에 저장하는 기능이다. 이 기능을 위해 **'GenBank 데이터 변환 모듈'**을 구현하였다. 두번째 기능은 웹을 통한 특허 자료 검색 기능이다. 이 기능을 위해 **'GenBank 검색 모듈'**을 구현하였다. 이 모듈에서는 웹 서버로부터 받은 질의를 분석하여 데이터베이스의 질의인 경우에는 PostgreSQL[POSTGRES]로 변환하여 결과를 얻어내고, DNA 시퀀스 질의인 경우에는 BLAST 프로그램을 사용하여 유사한 시퀀스들의 집합을 얻어낸다. 그리고, **'질의 결과 분석 모듈'**에서는 DBMS 와 BLAST 검색 엔진의 질의 결과를 HTML 페이지로 변환하는 작업을 수행한다.

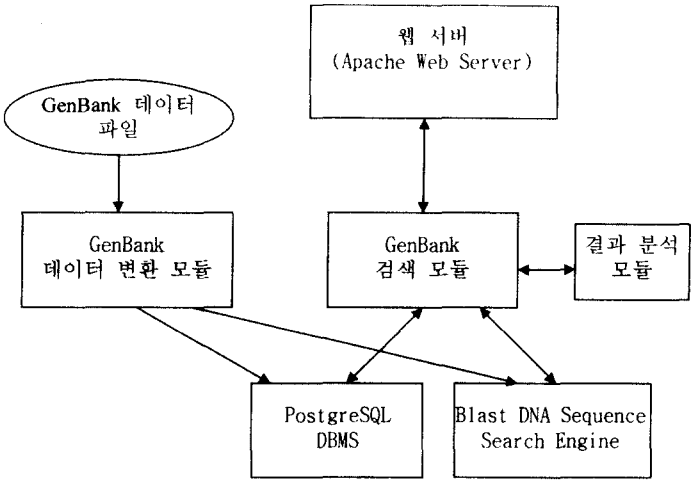

그림 1: GenBank 특허 데이터 검색 시스템의 아키텍쳐.

3. GenBank 데이터의 구조

GenBank 데이터의 자세한 구조는 참조문헌 [GenBank02] [OWK01]를 참조하고, 여기서는 간단하게 설명하도록 한다. GenBank 데이터는 다음의 구성 요소로 이루어져 있다.

- **Locus:** LocusName, SequenceLength, Molecule Type 등으로 구성
- **Definition:** 시퀀스에 대한 간단한 설명
- **Accession:** 시퀀스 레코드에 대한 유일 식별자
- **Version:** Nucleotide 시퀀스 식별자 번호
- **Source:** 조직 이름(organism name)을 포함한 자유 포맷 형태의 정보
- **Reference:** 이 시퀀스의 저자들에 의한 출판 정보
- **Features:** 유전자, 유전자 생성물(gene products)등에 대한 정보
- **Protein ID:** 단백질 시퀀스 식별자 번호
- **GI:** 단백질 변환(protein translation)에 대한 "GeneInfo Identifier" 시퀀스 식별자 번호
- **Translation:** nucleotide coding sequence(CDS)에 대응되는 amino acid translation
- **Base Count:** 이 시퀀스내의 A, C, G, T 의 개수
- **Origin:** 시퀀스 데이터

4. 구성 모듈

■ GenBank 데이터 변환 모듈

본 모듈에서는 GenBank 파일을 파싱(parsing)하여 다음의 두 가지 데이터로 변환한다.

1) PostgreSQL 의 벌크 로딩(bulk loading) 포맷으로 변환한다.
2) BLAST 의 인덱싱될 수 있는 포맷으로 변환하는 기능을 수행한다.

그리고, 변환된 벌크 로딩 데이터들은 PostgreSQL 의 벌크 로딩 명령어(copy 명령어)를 사용하여 빠르게 데이터베이스에 로딩되도록 한다. 그리고, BLAST 의 인덱싱 명령어를 사용하여 염기 서열 데이터에 대해 인덱싱을 수행한다.

■ 검색 웹 페이지

검색 웹 페이지는 사용자가 특허 정보를 검색할 수 있도록 두 가지 측면의 인터페이스를 제공한다.

1) 특허 속성에 대한 검색에 대한 인터페이스
2) 염기 서열에 대한 유사성 검색에 대한 인터페이스

그림 2 는 본 시스템에서 제공하는 검색 웹 페이지를 나타낸다. 그림에서 보듯이 화면의 윗 부분은 특허 속성에 대한 검색 인터페이스를 나타내고, 화면의 아랫 부분은 염기 서열에 대한 유사성 검색에 대한 인터페이스를 나타낸다.

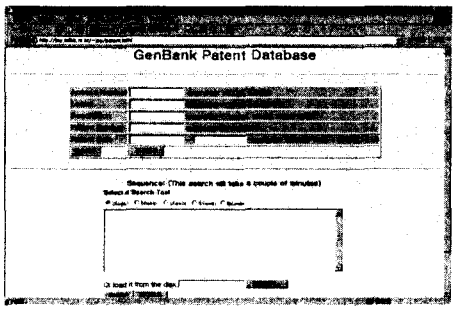

그림 2: GenBank 특허 데이터베이스 검색 웹 페이지.

현재 구현에서는 두 가지 기능을 동시에 이용하는 것은 허용하고 있지 않으나, 좀 더 정밀한 검색을 위해 제공되어야 될 기능으로 고려하여 구현을 예정중이다. 이를 위해서는 먼저 염기 서열에 대한 유사성 검색의 결과(Locus 번호 포함)를 PostgreSQL 데이타베이스의 테이블에 저장하고, 일반 속성 검색의 결과와 조인하는 기능을 구현하면 된다.

■ 웹 검색 모듈

본 모듈은 웹 페이지로부터 사용자 입력을 받아 이를 분석하여 PostgreSQL DBMS 와 BLAST 검색엔진을 검색하는 기능을 제공한다. 웹 서버와는 CGI 를 통해 통신하도록 하였으며, 구현한 언어로는 Perl[CT98]을 사용하였다. 그리고, CGI.pm [GGB00] Perl 모듈을 활용하여 간결하게 구현될 수 있

도록 하였다.

데이터베이스의 연결 인터페이스로는 DBI[DB00]를 사용하여 앞으로 DBMS 를 변경하더라도 검색 모듈은 변경하지 않도록 하였다. DBI 는 Perl 언어에서 제공하는 표준 데이터베이스 인터페이스로서, 현재 대부분의 DBMS 에서 DBI 인터페이스를 제공하고 있다.

BLAST 검색은 프로그래밍 인터페이스가 제공되지 않기 때문에 system 명령어를 사용하여 구현하였다. 이때 사용자의 입력이 파일로 저장해야만 BLAST 검색이 되므로, 사용자 질의가 들어올 때 마다 별도의 시퀀스를 저장하는 파일을 생성하고, 생성된 파일에 대해 BLAST 를 수행하였다. 이때 주의할 점은 생성해야 하는 파일의 이름이 매번 달라야 된다는 것이다. 이를 위해 POSIX 에서 제공하여 임시 파일 생성 루틴을 사용하였다.

■ **BLAST 검색 엔진[AGM+90]**

BLAST 는 DNA 시퀀스들에 대한 인덱싱 기능을 제공하고, 질의 시퀀스와 유사한 시퀀스를 인덱스를 이용하여 찾는 프로그램으로 NCBI 에서 개발된 대표적인 소프트웨어이다. 현재 BLAST, BLAST 2.0, Position Specific Iterated BLAST, BLAST 2 sequences against each other 등의 개선된 프로그램과 서비스가 제공중이다. 본 논문에서는 BLAST 프로그램(검색엔진)을 이용하여 GenBank 특허 시퀀스들에 대해 유사성 검색을 제공한다.

BLAST 는 제공하는 세부 기능은 다음과 같다.

- BLASTp: 단백질 서열간의 유사성 비교
- BLASTn: 염기 서열간의 유사성 비교
- BLASTx: 입력 염기 서열을 여섯 개의 프레임(frame)으로 변환 후 단백질 서열 데이터베이스로부터 유사성 비교
- Tblastn: 염기 서열 데이터베이스를 여섯 개의 프레임으로 변환 후 입력 단백질 서열과 유사성 비교
- Tblastx: 입력 염기 서열과 염기 서열 데이터베이스 모두를 여섯 개의 프레임으로 변환 후 유사성 비교

본 시스템에서는 웹 페이지에서 사용자가 세부 기능을 선택하게 하고, 검색 모듈에서 사용자가 선택한 옵션에 기반하여 BLAST 프로그램을 선택적으로 수행하도록 하였다.

BLAST 에서 제공하지 않는 Top-n 질의 등의 기능은 BLAST 검색 결과를 이용하여 상위 레벨에서 구현할 수 있다. 이러한 기능들은 본 저자들이 개발한 개선된 BLAST 검색엔진[JOY02]에 이미 구현되어 있으며 이러한 기능들을 본 시스템에 포팅하는 일과 어떠한 향상된 검색 기능이 필요한지는 향후 연구로 남겨놓는다.

■ **PostgreSQL DBMS**

PostgreSQL DBMS 는 UC Berkeley 에서 개발한 객체관계형 DBMS[SB99]이다. PostgreSQL DBMS 는 소스가 공개되어 있고 자유스럽게 사용할 수 있어 현재 많은 비영리 단체에서 사용되고 있다. 그러나, 키워드 검색, 염기 서열 등의 검색을 지원하지 않아 이러한 기능들을 원하는 환경에서는 사용자가 추가적으로 구현해야 되는 부담이 있다. 본 시스템에서는 키워드 검색을 위해서는 SQL[EN00] 질의 구성시에 '%' 연산자를 사용하였고, 염기 서열 검색을 위해서는 BLAST 프로그램을 사용하였다.

■ **결과 분석 모듈**

본 모듈은 BLAST 프로그램의 결과를 화면에 보기 좋게 변환하고 유용한 링크 정보를 추가하는 기능을 수행한다. 이를 위해 BLAST 검색 결과를 파싱하고, 검색 결과로 나온 염기 서열에 대해 이와 관련된 특허 사이트에 대한 링크를 추가 하였다.

5. 연구 이슈

본 시스템과 같이 DBMS 와 염기서열 검색 시스템을 함께 이용하여 구현할 때 고려해야 될 사항은 다음과 같다.

■ **DBMS 내에서 염기 서열 검색 기능 구현**

여러 검색 시스템(DBMS, BLAST, 키워드 검색 시스템)을 사용하여 구현하는 것이 아니라, 하나의 DBMS 내에서 이러한 기능이 한꺼번에 처리되도록 하는 것이 필요하다.

■ **데이터베이스 검색 결과를 GenBank 포맷으로 빠르게 변환**

데이터베이스 질의 결과를 GenBank 포맷으로 빠르게 변환하는 기능이 필요하다. 즉, 질의 결과를 다른 기관에 보내기 위해서는 표준 포맷인 GenBank 포맷으로 바꾸어야 되는데 이러한 기능이 빠르게 제공되는 것이 필요하다. 본 시스템에서는 하나의 추가적인 컬럼을 두어 원 GenBank 데이터(original GenBank data)도 함께 저장하는 방식을 취하고 있다. 이는 임시적인 해결책이므로 근본적인 해결책이 필요하다.

6. 결론 및 향후 연구

본 논문에서는 한국생명공학연구원에서 개발해오고 있는 웹 기반 GenBank 특허 데이터 검색 시스템의 설계와 구현에 대해서 소개하였다. 본 시스템은 1) 데이터변환 모듈, 2) 일반 속성 검색 모듈, 3) DNA 시퀀스 검색 모듈, 4) 결과 페이지 분석 모듈, 5) 웹 페이지들로 구성되며, 일반 속성 질의나 염기 서열 유사성 질의를 모두 빠르게 지원한다. 향후 연구로는 1) 향상된 검색 기능의 제공, 2) 웹상에서 특허 데이터를 등록 및 변경, 3) 새로운 DNA 검색 엔진과 결합 등이 있다.

참고 문헌

[AGM+90] Altschul, S. F., et. al., "Basic Local Alignment Search Tool," *Journal of Mol. Biol., Vol 215, pp. 403-410.*

[CT98] Christiansen, T. and Torkington, N., *Perl Cookbook*, O'Reilly, Aug. 1998.

[GGB00] Guelich, S., Gundavaram, S., and Birznieks, G., *CGI Programming with Perl*, O'Reilly, July 2000.

[DB00] Descartes, A., Bunce, T., *Programming the Perl DBI*, O'Reilly, Feb. 2000.

[EN00] Elmasri, R.A., and Navathe, S.B., *Fundamentals of Database Systems*, Addision-Wesley, 2000.

[GenBank01] http://www.ncbi.nlm.nih.gov/GenBank

[GenBank02] http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

[JOY01] http://joy.kribb.re.kr/~joy/patent.html

[JOY02] http://joy.kribb.re.kr/joy_blast/blast.html

[NCBI01] http://www.ncbi.nlm.nih.gov

[OWK01] Ostell, J. M., Wheelan, S.J., and Kans, J.A., "The NCBI Data Model in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins," John Wiley & Sons Publishing, pp. 19-44, 2001.

[POSTGRES] http://www.postgresql.org

[SB99] Stonebraker, M. and Brown, P., *Object-Relational DBMSs*, Morgan Kaufmann, 1999.