BSML 기반의 유전자 데이터베이스와 변환기의 구축

윤애란 이수정 이희전 용환승
이화여자대학교 컴퓨터학과 데이터베이스 연구실
{diable6, crystal2, kkakkungyi, hsyong}@ewha.ac.kr

Development of Bioinformatic Database and Converting Tools based on BSML

Ae-Ran Youn Su-Jung Lee Hee-Jeon Lee Hwan-Seung Yong
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 바이오인포매틱스 분야의 발전에 따라 방대한 양의 유전체 데이터에 대한 연구가 진행되고 있으며, 이러한 데이터를 효율적으로 다루기 위해 다양한 형태의 파일과 데이터베이스들이 사용되고 있다. 하지만 표준화의 미비로 인하여 데이터의 관리와 변환에 어려움이 많다. 본 논문에서는 이러한 문제점을 해결하기 위하여 바이오인포매틱스 데이터를 다루기 위한 표준으로 다양한 XML 포맷들 중에서 BSML(Bioinformatic Sequence Markup Language)을 채택하고, Genbank 파일을 변환하여 관계형 데이터베이스에 저장하는 모듈을 개발한다. 또한 관계형 데이터베이스 형태의 유전체 데이터를 BSML 형태로, Genbank 파일 형태를 BSML 형태로 그리고 AGAVE(Architecture for Genomic Annotation)파일 형태를 BSML 형태로 변환하는 변환기를 개발하고자 한다.

1. 서 론

최근 빠르게 발전하고 있는 학문인 바이오인포매틱스는 생물학 데이터의 관리와 분석에 컴퓨터학 분야의 첨단 기술을 이용하여 이를 자동화, 전산화하는 응용 분야이다[1]. 바이오인포매틱스 분야의 발전과 더불어 과거 축적되어 온 방대한 양의 생물학 데이터에 대한 데이터베이스가 구축 되었고 다양한 생물학 데이터들의 상호교환을 용이하게 하기 위한 생물학 관련 XML 표준들[2] 또한 마련되고 있다. 특히 XML DTD는 표현력이 매우 우수하다는 장점을 지니고 있기 때문에 이를 기반으로 한 XML 표준화 작업이 활발하게 진행되고 있다[3].

본 논문에서는 현재 가장 널리 사용되고 있는 관계형 데이터베이스를 기반으로 유전체 데이터베이스를 구축하고, 유전체 데이터를 위한 XML 포맷인 BSML을 기반으로 다양한 형태의 생물학 데이터 포맷들간의 변환 소프트웨어를 개발한다. 이러한 시스템을 통해 좀 더 효율적으로 생물학 데이터들 간의 정보를 공유함으로써 유전 정보를 조직적으로 관리하지 못해 발생하는 시간과 비용의 낭비를 줄일 수 있을 것으로 기대된다.

본문의 구성은 다음과 같다. 제 1 장 서론에 이어 제 2 장에서는 XML 표준 중 하나인 BSML에 관한 개요와 기존의 변환 소프트웨어들에 대해 소개하였다. 제 3 장에서는 Genbank 파일을 기반으로 구축한 관계형 유전체 데이터베이스의 스키마와 구현 내용에 대해 기술하며 다양한 생물학 데이터 포맷간의 변환을 수행하는 3가지 변환 소프트웨어의 개발에 대하여 설명하였다. 본 논문에서 구현한 각각의 변환 소프트웨어의 시제품들은 현재 http://dblab.ewha.ac.kr을 통해 사용자 컴퓨터에 다운로드 받아 직접 사용해볼 수 있다. 4장에서는 결론 및 향후 과제에 대해 논의하였다.

2. 관련연구

최근의 바이오 서열데이터의 표준으로 각광 받고 있는 BSML에 대하여 간단히 소개하고, 본 연구와 관련하여 최근 동향을 살펴본다.

2.1 BSML 개요

BSML[4]은 DNA 구조와 같은 정보를 인코딩 하고 표현 하기 위한 언어로써 서열, 유전자, 전기 영동 젤, 다중 정렬 등과 같은 생물학적으로 의미 있는 객체들에 대한 내용을 그래픽하게 표현할 수 있으며, 서열 정보를 인코딩 할 수 있다는 점에서 여러 XML 형식들 중 다른 포맷에 비하여 완성도가 높고 구체적이다. 또한 BSML은 서열의 직접적인 표현 방법(data tables, visualization 등)뿐만 아니라 다른 형식들에서 이용할 수 없는 서열에 대한 많은 특징들 즉, 추상적인 annotation 그리고 서열 자체와 그에 대한 생체 물리학적인 특성까지 표현 가능하다. BSML DTD는 1997년도부터 사용되기 시작해서, 현재 버전 3.1이 나와있고 많은 응용프로그램과 데이터베이스들은 유전체 데이터의 교환과 시각화를 위해 사용되고 있다.

위와 같은 여러 가지 장점을 보유한 BSML은 Fujitsu, IBM, Labbook, Inc, EBI 등의 업계에서 바이오 서열 데이터를 표현하기 위한 새로운 표준으로 채택되었으며, 이를 기반으로 본 연구에서도 유전체 데이터 공유를 위한 XML표준으로 BSML을 채택하였다.

2.2 최근 동향

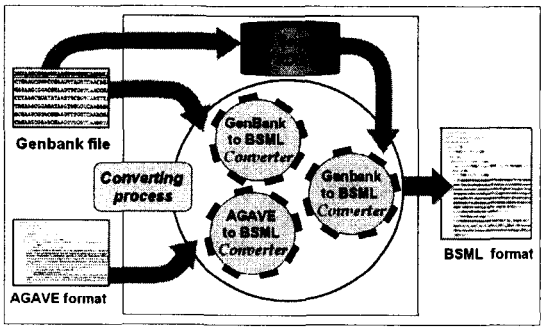
최근까지 개발된 여러 가지 표준간의 변환기 및 도구들은 Java나 Perl 모듈을 사용하여 개발되어 왔다. 이러한 도구들은 아직 패키지화 되어 있지 않을 뿐 아니라 주로 간단한 변환 작업만을 수행할 수 있다.

기존의 변환 소프트웨어들 중 Labbook, Inc에서 개발한 Genomic Browser가 가장 주목할 만한 프로그램이다[5].

본 연구는 2002년 농업생명과학원 연구비의 지원으로 수행된 결과임.

Genomic Browser는 Genbank나 플랫 파일 형태를 BSML 형태로 변환해 주며 프로그램 내부에 변환기와 그래픽 뷰어인 Genomic Viewer를 포함하고 있어 뛰어난 시각화 기능을 수행한다. 그 외에 AGAVE 포맷과 관련된 변환 프로그램으로 Game2agave.xsl와 Agave.pm이 있다. Game2agave.xsl은 XSLTscript를 사용하여 GAME 포맷을 AGAVE 포맷으로 변환해 주며 agave.pm은 Peral 모듈을 통하여 여러 가지 형태의 포맷(Fasta, EMBL, GenBank, Swiss, PIR, ace 등)들을 AGAVE로 변환해주는 기능을 제공한다. 그 외에 PROXIML에서 제공되는 자바 기반의 PDB2CML과 포트란 기반의 CIF2XML, Perl script를 통하여 Genbank 파일을 PISE XML로 변환해 주는 도구인 gb2xml, AGAVE와 BSML등의 다양한 포맷간의 변환을 지원하는 XEMBLE project[6]등의 변환기들이 존재한다.

3. BSML 기반, 유전체 데이터베이스와 변환기



[그림 1]전체 시스템 구성도

이 장에서는 관계형 데이터베이스 기반의 유전체 데이터베이스와 변환 소프트웨어에 대해 살펴보겠다. 시스템의 전체 구성도는 그림 1과 같다. Genbank 파일 혹은 AGAVE 파일은 각각의 파일을 지원하는 변환기를 통해 해당 출력 형태 (관계형 데이터베이스 혹은BSML)로 변환된다. 시스템 개발 환경은 Microsoft사의 Microsoft SQL-Server 2000이며, 개발 언어로는 Java를 사용하였고 사용자 인터페이스는 JBuilder 7.0으로 구현하였다.

3.1 Genbank 파일

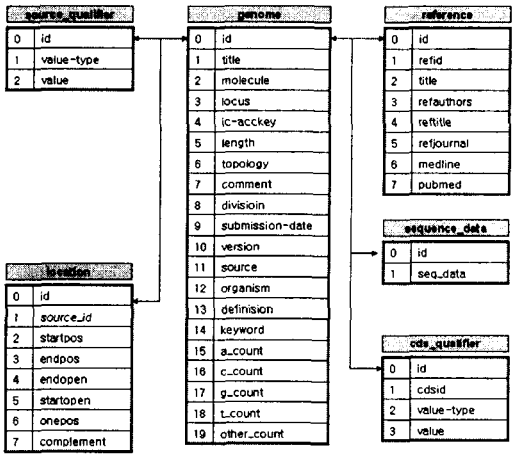
GenBank는 유럽의 분자생물학자 협회인 EMBL, 일본의 DDBJ와 함께 '국제 뉴클레오티드 서열 데이터베이스 합작'을 형성하여 자료를 교환 및 공유하고 있기 때문에, 이 세 곳의 자료는 포맷만 다를 뿐, 그 내용은 거의 같다고 할 수 있다. 그러나 Genbank가 이들 세계 3 대 유전자 서열 데이터베이스(Genbank, EMBL, DDBJ)중 가장 방대한 양의 데이터를 가지고 있으며 주요한 데이터베이스라 할 수 있기 때문에 본 연구에서는 Genbank를 유전체 데이터베이스 구축을 위한 원천 데이터로 사용하였다. Genbank데이터는 미국 NCBI의 Genbank ftp 사이트[7]에서 다운로드 받을 수 있다.

3.2 유전자 서열 데이터를 위한 관계형 데이터베이스 시스템

본 논문에서는 관계형 데이터베이스 기반으로 Genbank 파일을 파싱 하여 저장을 수행하는 시스템을 개발하였다.

3.2.1 유전체 관계형 데이터베이스의 스키마

유전체 데이터베이스를 위한 데이터베이스 스키마는 그림 2와 같이 6개의 테이블로 구성되어 있으며, 각 테이블은 GenBank 데이터의 accession 번호를 외부키(foreign key)로 하여 서로 연결된다[8]. 테이블의 저장된 데이터는 아래 표1과 같다.



[그림 2] 유전체 관계형 데이터베이스의 스키마

genome	해당되는 생물체의 서열이 언제 밝혀졌는지, 이 서열이 어떤 기능을 하고, sequence의 핵산의 구성이 어떻게 되는지에 대한 일반적 정보를 저장
reference	관련된 논문에 관한 내용, 논문의 저자, 제목, 저널의 이름과 같은 정보저장
location	Sequence에서 gene이 어느 위치에 존재하고 있는가에 대한 정보 저장
source_qualifier	서열의 길이, source organism의 이름, Taxon ID 번호 와 같은 각 레코드에 있는 Feature의 설명
cds_qualifier	전체 서열에서 단백질의 아미노산으로 바뀌는 서열부분의 대한 정보를 저장
sequence_data	Sequence정보를 저장.

[표 1] 테이블의 상세 내용

3.2.2 파싱 모듈

본 논문에서는 유전체 데이터베이스의 구축을 위해 가장 먼저 GenBank 플랫 파일을 파싱 하는 작업을 수행하였다. 이러한 파싱 작업을 위해 BioJava 라이브러리를 이용하였는데, BioJava[9]는 생물학 데이터 처리를 위한 Java 툴을 제공하기 위한 공개 표준 프로젝트이다. Genbank 플랫 파일의 데이터 포맷은 키워드와 내용으로 구성되어 있다. 주어진 키워드에 대해 관련 내용이 기술되는 형식을 취하고 있다. 키워드는 반복되어 나타나는 키워드와 반복이 없는 키워드로 나누어 진다. 반복 없는 키워드는 어떠한 Genbank 파일에서도 오직 한 번씩만 나타나는 것으로서, 이는 BioJava 라이브러리 내에서 제공되는 Genbank 플랫 파일 지원 클래스를 이용하여 파싱 할 수 있다. 그러나 반복되는 키워드의 경우는 각 Genbank 파일마다 각기 다른 형태를 가지며, 반복되는 횟수에도 차이가 있다. 반복되는 키워드 부분은 Genbank 데이터의 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 파싱 할 수 있도록 하였다.

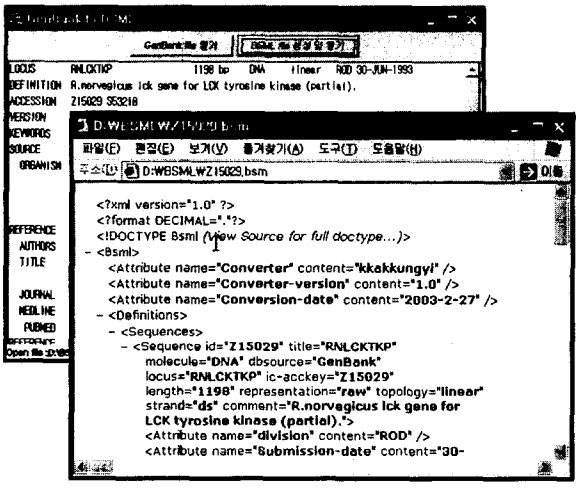
3.2.3 관계형 데이터베이스 기반의 저장 모듈

관계형 데이터베이스 기반의 저장 모듈에서는 NCBI Genbank ftp 사이트에서 다운로드 받은 Genbank 파일[Genbank file(*.gbk)]이 존재하는 폴더의 경로를 파싱 모듈의 입력 값으로 전해 받은 후, 선택된 폴더의 모든 Genbank 파일을 차례로 파싱 한 후 그림 2와 같은 여섯 개의 테이블로 구성된 데이터베이스에 저장한다.

질의 인터페이스를 통해 써버에 저장된 유전체 데이터에 대한 질의나 접근 명령이 주어지면, 시스템은 JDBC를 통해 SQL 서버에 접속 하고 활성화된 질의 입력 창에 입력된 질의에 해당하는 서열 정보, 즉 표 1에서 보여진 genome 테이블의 내용 중 Genbank 파일의 일반적인 정보(locus, 서열의 길이, molecule type, division, geometry)를 검색하여 보여준다[10].

3.3 변환 소프트웨어

이번 절에서는 다양한 생물학 데이터 포맷끼리의 변환을 수행하는 변환 소프트웨어에 대해 살펴보겠다.



[그림 3] Genbank 데이터형식을 BSML 형식으로 변환하는 변환 소프트웨어의 GUI

3.3.1 관계형 데이터베이스 기반의 Genbank에서 BSML로의 변환기

본 논문에서는 3.2절에서 개발된 관계형 유전체 데이터베이스를 기반으로 관계형 데이터 형태로 변환 및 저장된 Genbank 파일의 정보를 다시 유전체 XML 표준인 BSML 형태로 변환하는 변환기를 개발하였다. 얻고자 하는 시퀀스의 locus를 입력 값으로 받아 관계형 데이터베이스의 데이터를 검색한 후 추출된 데이터에 해당하는 BSML 파일을 익스플로러 브라우저를 통해 보여주며 생성된 BSML파일은 디폴트 디렉토리에 저장된다.

3.3.2 Genbank에서 BSML로의 변환기

BSML 포맷으로의 변환을 위해 우선 Labbook, Inc에서 개발한 Genomic viewer의 결과물인 BSML DTD를 분석하였다. 이 분석 결과를 기반으로 Genbank 플랫 파일의 파싱 결과의 각 항목들에 대해 데이터베이스를 사용하지 않고 변환을 수행하였다. BSML로의 변환 수행 시, 파싱 결과로 얻은 항목 중 특히 반복이 심한 feature 부분에 대해서는 각각에 id를 부여하여 방대한 feature 정보를 쉽게 관리, 분석할 수 있게 하였다.

3.3.3 AGAVE에서 BSML로의 변환기

AGAVE[11]포맷을 BSML 포맷으로 변환하기 위하여 AGAVE와 BSML 각각에 대한 표준 DTD가 요구된다. 이러한 점을 해결하기 위하여 본 논문에서는 EMBL-EBI에서 개발한 XEMBL Viewer의 AGAVE 파일과 BSML 파일의 DTD를 비교 분석하고, 분석한 결과를 기반으로 하여 변환 작업을 수행하였다. 우선 변환을 수행하기 전 AGAVE의 DTD와 BSML의 DTD를 비교하여 엘리먼트와 어트리뷰트 및 각각의 항목들을 분석하고 일치하는 항목으로 매핑 한다. 그 다음 BSML 파일로 변환할 AGAVE 파일을 XML파서(xerces)를 이용해 DOM 파싱 하여 파일 전체의 내용을 읽어 들인다. 파싱 결과 얻은 각각의 항목들에 대하여, DTD 분석 결과를 기반으로 AGAVE의 포맷과 일치하는 각각의 BSML의 엘리먼트, 어트리뷰트 및 텍스트 등으로 변환하는 작업 수행한다. 이러한 변환 과정은 데이터베이스를 사용하지 않고 직접적으로 수행되며, 이 과정을 통하여 새로운 BSML 형식의 XML 파일이 생성된다. 위와 같은 파싱 및 변환 작업을 통해서 기존의 AGAVE 포맷을 BSML로 변환하여 XML 포맷끼리도 변환 가능하도록 하였다.

4. 결론 및 향후 과제

본 논문에서는 유전체 데이터의 효율적인 관리 및 저장을 위하여 관계형 데이터베이스 기반의 유전체 데이터베이스 시스템과 변환기들을 개발하였다. 특히 일정하지 않은 형태의 Genbank 파일 포맷에 대한 관계형 데이터베이스 스키마를 정의하였고, 파싱 기능이 완벽하지 못한 BioJava 라이브러리를 수정하여 기존의 변환기에서는 제공하지 못하는 대량의 반복되는 키워드에 대한 파싱 작업을 수행하였다. 개발된 시스템을 기반으로 Genbank 데이터들을 관계형 데이터베이스 형태 및 BSML로 변환, 저장할 수 있을 뿐만 아니라, 다양한 생물학 데이터 포맷 사이의 효율적인 정보 교환 및 공유가 가능하다.

향후 과제로는 앞에서 언급한 Genbank 파싱 과정에서의 메모리 과부하 문제에 대한 연구가 필요하다. 파싱 과정에서의 메모리 과부하 문제는 대용량의 생물학 데이터 관리 시스템의 성능과 근본적으로 밀접한 관련이 있기 때문에 이러한 과부하 문제에 대한 좀 더 깊이있는 연구가 필요하다. 또한 개발된 XML 기반 변환기들을 기반으로 한 유전체 데이터베이스간의 연동 시스템을 구축함으로써 좀 더 통합적이고 효율적으로 유전체 데이터를 관리 및 공유할 수 있을 것이다.

5. 참고문헌

[1]Cynthia Gibas, Bioinformatics Computer Skills, O`REILLY, 2002.
[2]http://www.visualgenomics.ca/gordonp/xml/
[3] Tisdall, James, Beginning Perl for Bioinformatics, O`REILLY, 2001.
[4]http://www.bsml.org/
[5]http://www.labbook.com/
[6]http://www.ebi.ac.uk/xembl/
[7]ftp://ncbi.nlm.nih.gov/GenBank/genome/
[8]Guochun Xie, Reynold DeMarco, Richard Blevins Yuhong Wang, Storing biological sequence databases in relational form, BIOINFORMATICS. Vol.16, No.3, pp.288-289, 2000.
[9]http://www.biojava.org/
[10]John A. Crow, " Design and Implementation of a Simple Relational Database for GenBank-Derived Data" , Technical Report, Academic Health Center, University of Minnesota, 2001.
[11]http://www.agavexml.org/