정보 공유를 위한 BSML 기반의 생물학 데이터 변환기

김영억[0] 정광수 정영진 차효성 류근호
충북대학교 데이터베이스연구실
{kimuks[0], ksjung, yjjeong, kkido, khryu}@dblab.chungbuk.ac.kr

The Biological Data Converter based on BSML for Sharing Information

Young Uk Kim[0], Kwang Su Jung, Young Jin Jung, Hyo Soung Cha, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

현재 생물학 연구실에서 시퀀싱 실험을 통해 생성되거나 또는 공개용 생물 데이터베이스로부터 획득된 유전체 및 단백질 정보는 각각 이질적인 데이터형식을 사용하고 있다. 이 때문에, 생물정보를 분석하여 상호간의 정보를 효율적으로 사용하기 위해서는 공통된 형식의 데이터 표준화작업이 필수적이다. 그리고 이러한 이질적 데이터 형식에 대한 표준화 연구의 미비로 인하여 플랫 파일간의 정보공유에 어려움을 겪고 있다. 따라서, 이 논문에서는 다양한 유전체 및 단백질 정보를 관리·공유하기 위해 이질적인 포맷간의 맵핑 과정을 통하여 BSML(Bioinformatic Sequence Markup Language) 형태로 변환하고, 이를 객체관계형 데이터베이스(Object Relational DataBase)에 저장하는 시스템을 개발하였다. 그리고, 개발된 시스템은 생물정보 데이터의 표준화를 위해 개발된 XML(Extend Markup Language) 기반의 BSML을 이용함으로써 효율적으로 생물학 데이터들 간의 정보를 공유할 수 있으며, 개인 생물학 데이터베이스 구축이나 다양한 생물학적 데이터를 통합관리하는 시스템에서 유용하게 쓰일 수 있다.

1. 서 론

인간게놈프로젝트의 초안 발표 이후 유전자 및 단백질의 기능을 밝히기 위한 유전체 및 단백질체학 연구가 활발해졌다. 이와 같이 생명정보 분야의 발전과 더불어 생물학 데이터는 폭발적으로 증가했으며, 과거로부터 연구되어 온 방대한 양의 생물학 데이터들이 다양한 공개용 생물 데이터베이스에 축적되었다.

그러나, 공개용 생물학 데이터베이스의 데이터들은 데이터를 작성한 사람이나 기관 또는 연구 목적 등의 차이로 인해 구조적으로나 내용적, 그리고 의미적으로 볼 때 이질적인 면이 많이 존재한다[1]. 그리고, 이렇게 다양한 공개용 생물 데이터베이스로부터 얻어진 생물학 데이터를 변환·분석하는 작업은 실험생물학자들에게 어려운 과제로 대두되고 있다[2].

따라서, 이 논문에서는 좀 더 효율적으로 생물학 데이터들 간의 데이터 교환 및 공유를 위해 BSML[3] 기반인 생물학 변환기를 개발하였다. 그리고, XML[4]기반인 BSML 스키마가 객체 모델을 지향하기 때문에, 생성된 정보들을 ORDB에 저장하였다. 다양한 생물학 데이터들의 상호교환을 용이하게 하기 위한 생물학 관련 XML표준들[5]의 하나인 BSML을 사용하여 다양한 생물 데이터베이스로부터 추출된 이질적인 생물학 데이터들의 상호교환 및 변환·분석을 용이하게 할 수 있다. 또한 시퀀싱된 서열 데이터를 다른 연구자와 교환하기 위해, 공통 포맷으로 포맷변환을 자유롭게 할 수 있다.

이 연구는 2003년도 KISTEP의 특정연구개발과제의 연구비 지원으로 수행되었음

2. 관련연구

2.1. 공개용 생물학 데이터베이스

현재 대용량의 DNA(Deoxyribo Nucleic Acid)와 단백질에 대한 서열, 구조, 실험 및 참조 정보에 대한 데이터베이스를 구축하여 웹 상에서 제공하고 있다. 그러나, 대표적인 생물학 데이터베이스인 GenBank[6], EMBL(European Molecular Biology Laboratory), PDB(Protein Data Bank), SCOP(Structural Classification of Proteins) 등에서는 각각 이질적인 포맷으로 데이터를 표현하고 있으며 대표적인 서열 유사성 검색 시스템인 BLASTA에서는 FASTA라는 데이터 포맷을 사용하고 있다. 이러한 이질적인 플랫 파일 형태에서는 필드의 의미와 데이터 타입에 대한 일치되지 않은 점이 많이 있다. 그리고, 데이터베이스마다 제공하는 형식이 다르므로 전문가가 아니라면 매뉴얼을 참조해야만 내용을 이해할 수 있다.

2.2. BSML 개요

BSML은 의미있는 생물학적 관계를 얻기 위한 방법으로 바이오인포매틱스 연구정보에 관한 개방된 XML 언어이다. DTD(Document Type Definition)에 의해 표현되는 XML의 응용으로써 사용자가 조정할 수 있는 표나 그림들로써 서열정보를 보여준다. 특히 서열의 표현 방법(Visualization) 뿐만 아니라 서열에 대한 생체 물리학적인 특성까지 표현 가능하다. 또한 다양한 생물학 파일포맷의 상호변환과 플랫폼 독립성을 통한 정보의 상호 운영성을 제공해 준다.

이러한 장점을 보유한 BSML은 게놈 연구정보를 표현·교환하거나 sequence data, feature tables, literature references를 포함하는 생물분자를 표현하는 분야에 응용되며, 생물학적 서열을 가장 많이 포함하고 있는 NCBI의 GenBank에서는 서열 데이터를 XML형식으로 나타내는 표준으로 이용하고 있다. 따라서, 이 논문에서도 이질적인 생물학 데이터의 정보 공유를 위하여 BSML을 채택하였다.

3. 시스템 구조

BSML 기반의 생물학 데이터 변환기의 전체적인 구성은 그림 1과 같다. 공개용 생물학 데이터베이스로부터 생물학자에 의한 자료수집을 통해 획득한 이질적인 공개용 생명정보 데이터를 파싱하여 BSML 공통모델로 표현하고, ORDB인 생물학 통합 데이터베이스에 저장한다. 데이터 편집기에 의해 수정된 데이터는 최종 포맷 변환기를 통해 BSML 뿐만 아니라, FASTA나 GenBank형식의 파일로 저장할 수 있다.

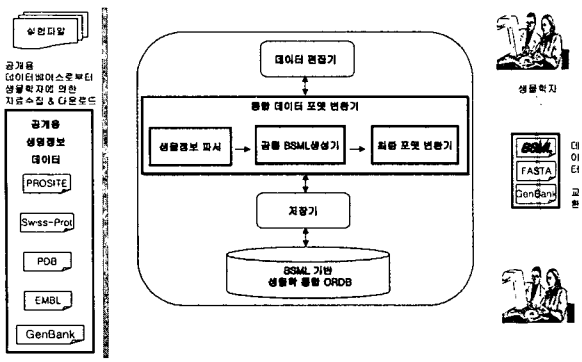

그림 1. 시스템 구조

3.1. 통합데이터 포맷 변환기

통합데이터 포맷 변환기는 다양한 생물정보 데이터베이스로부터 수집된 이질적인 파일 포맷을 변환하는 역할을 하며, 생물정보파서, 공통BSML생성기, 최종포맷변환기로 구성된다. 생물정보 파서는 다양한 생물정보 플랫 파일의 필드 및 데이터 추출을 위해 각 플랫파일에 맞는 파서기능을 담당한다. 공통BSML생성기는 생물정보파서에 의해 추출된 필드 및 데이터를 그림 2와 같이 다른 포맷간의 매핑 관계에 따라서 BSML 형태로 생성하는 역할을 한다. 최종 포맷변환기는 BSML 뿐만 아니라, 생물학자가 원하는 PDB, FASTA, GenBank, Swiss-Prot, PIR 등과 같은 포맷 생성을 위한 변환기이다.

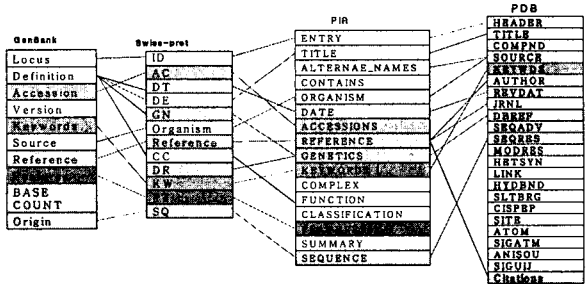

그림 2. 이질적인 포맷간의 매핑정보

3.2. 데이터 편집기

데이터편집기는 주석이나 서열 데이터같은 생물정보에 대하여 생물학자가 편집을 할 수 기능을 제공한다. XML 문서를 다루기 위한 기법으로는 DOM(Document Object Model)을 이용하였다. DOM은 문서의 구조에 대한 풍부한 표현력과 XML 문서를 생성 및 조작할 수 있는 장점이 있어서 BSML을 조작하기에 적합하다. 먼저 공통BSML생성기에 의해 생성되거나 ORDB에 저장된 BSML객체를 읽어 들인다. 그리고, 그림 3과 같이 XML 파서에 의해 트리 구조를 만든다. DOM API를 이용하여 엘리먼트, 텍스트, 에트리뷰트 내용을 추출한 후 BSML을 조작(추가, 삭제, 갱신)한다.

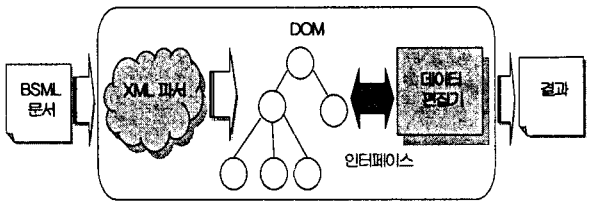

그림 3. DOM을 이용한 BSML 문서변환

3.3. 저장기

저장기는 대용량의 생물학적 데이터를 체계적이고 효율적으로 저장하는 기능을 담당한다. 공통BSML생성기에 의해 생성된 BSML객체를 그림 4와 같이 ORDB 스키마 형태로 저장한다. ORDB는 RDBMS(Relational DataBase)의 데이터 모델을 그대로 활용하여 어렵고 까다로운 OODBMS (Object Oriented DataBase)의 데이터 모델링 문제를 해결했고, 기존의 RDBMS를 기반으로 하는 많은 생물학 데이터베이스 시스템과 호환이 가능하다. 그리고, 상속이나 다형성과 같은 OODBMS의 개념들을 이용하므로 데이터베이스에 유연성을 부여할 수 있다.

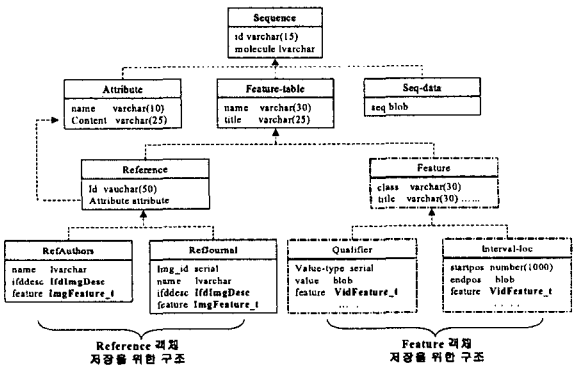

그림 4. BSML의 ORDB 스키마

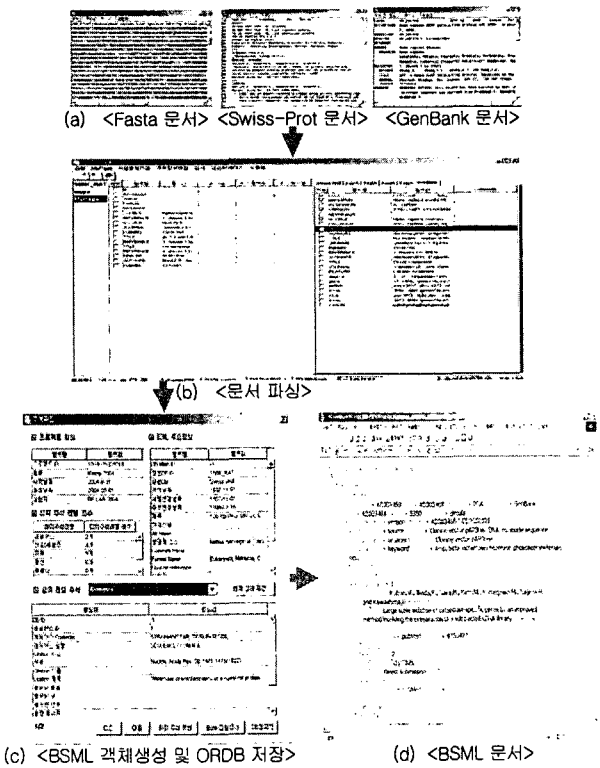

그림 5. Swiss-Prot 포맷을 BSML 파일로 변환한 결과

4. 구현

구현된 시스템의 구현환경은 Pentium PC 850MHz 시스템에서 ORDB를 지원해주는 Oracle 9.2.0.4.0버전[7]을 이용하였다. 그리고, 플랫폼 독립적으로 시스템을 실현하기 위해 JAVA 언어를 사용하였고, BSML Type 저장을 위해 Oracle 사의 Java Document Model(DOM) API를 참조했다.

구현한 생물학 데이터 변환기의 가용성을 보이기 위해, Swiss-Prot ID가 100K_RAT인 플랫파일을 BSML로 변환하였다. 그림 5는 100K_RAT Swiss-Prot 플랫파일을 파싱하여 BSML 객체로 생성하고, ORDB에 저장한 후, 이를 BSML 문서파일로 변환한 결과이다.

5. 결 론

현재 여러 분야에서 XML을 이용한 표준화 작업이 이루어지고 있으며 웹 상에서 서로 다른 실험과 소스로부터 생성된 이질적인 데이터를 교환하고 있는 생명 정보학에서도 XML을 이용한 데이터의 표준화가 필요하다.

이 논문에서는 이질적인 생명정보 플랫 파일사이에 정보 공유를 위해 BSML 기반으로 ORDB에 저장하는 변환기를 개발하였다. 개발된 시스템을 기반으로 여러 이질적인 플랫파일들을 BSML 이나 다른 플랫파일 형태로 변환할 수 있을 뿐만 아니라, 서로 간의 정보 공유가 가능하다.

향후 연구로는 생물학자 상호간에 정보를 공유하기 위해 P2P 서비스를 지원하는 에이전트 기술을 추가하고자 한다.

참 고 문 헌

[1] A. Silvescu, "Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed Biological Data Sources", Proceedings of the IJCAI-2001 Workshop on Knowledge Discovery.

[2] Andreas D. Baxevanis, "Bioinformatics", Wiley, 2001.

[3] http://www.bsml.org/

[4] David C. Fallside, "XML Schema Part 0", Primer, 2001.

[5] http://www.visualgenomics.ca/gordonp/xml

[6] http://www.ncbi.nih.gov/Genbank/

[7] http://otn.oacle.com/software/index.html