XPath 질의 처리를 적용한 단백질 데이터 통합 관리시스템 구축

차효성[0] 정광수 정영진 류근호
충북대학교 데이터베이스 연구실
{kkido[0],ksjeong, yjjeong, khryu}@dblab.chungbuk.ac.kr

Building a Integrated Protein Data Management System Using the XPath Query Process

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

요 약

최근 바이오 인포매틱스 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 관리 및 변환에 어려움이 많다.

따라서 이 논문에서는 시퀀싱을 통해 생성된 유전체 및 단백질 서열 데이터의 통합 저장 관리를 위해 서열 정보의 편집, 저장 및 검색과 서열 파일 포맷 변환을 수행하는 서열 정보관리 시스템의 구현을 목적으로 한다. 이러한 요구사항을 만족시키기 위해 바이오 인포메틱스 데이터를 다루기 위한 표준으로 BSML(Bioinformatic Sequence Markup Language)을 채택하고 이질적 플랫파일들은 DTD를 기반으로 BSML 스키마로 통합 및 저장한다. 그리고 객체 관계 데이터베이스 특성을 적용하여 XML 문서를 보다 쉽게 저장 관리하고 범위 또는 구조적 질의에 효율적인 XPath 질의 처리를 위한 시스템을 개발하였다.

1. 서론

현재 빠르게 발전하고 있는 학문인 바이오 인포메틱스는 생물학 데이터의 관리와 분석에 컴퓨터학 분야의 첨단 기술을 이용하여 이를 자동화, 전산화 하고 통계 하며 분석하는 응용 분야이다. 바이오 인포메틱스 분야의 발전과 더불어 과거 축적되었던 방대한 양의 생물학적 데이터에 대한 관리와 운영성에 대한 문제가 시급해지고, 그에 따른 다양한 생물학 데이터들의 상호 교환을 용이하게 하기 위한 생물학 관련 표준[1]들이 마련고 있다.

웹을 통한 시퀀싱 서비스가 제공되는 시점에서 국내 대부분 생물학 연구실에서는 시퀀싱된 서열 파일 정보 관리를 위한 소프트웨어가 존재하지 않아 파일 형태로 디스크에 저장하는 불편함이 있어 일관성 있게 서열 데이터가 관리 되지 않는다. 따라서 바이오 데이터의 무결성과 일치성 있는 관리가 어렵다. .

이 논문에서는 최근 가장 활발하게 사용되고 있는 XML 전용데이터 베이스를 기반으로 유전체 데이터베이스를 구축하고, 유전체 데이터를 위한 XML포맷인 BSML을 기반으로 이질적 플랫파일들 간에 변환 또는 편집, 저장 및 검색 소프트웨어를 개발한다. 이러한 시스템을 통해 좀 더 효율적으로 생물학 데이터들 간의 정보를 공유하고 실험을 통해 발생하는 버전 서열들을 관리 편집하며, BSML 형태로 저장하고 검색한다.

이러한 시스템을 통해 좀더 효율적으로 생물학 데이터들 간의 정보를 공유할 수 있고 그것들로 인한 시간과 비용의 낭비를 줄일 수 있기 위해서 XML 전용 데이터베이스를 사용하였다. 기존의 관계형 데이터베이스는 트리 기반의 XML 문서를 플렛한 테이블 또는 통합된 스키마 형태의 테이블에 저장하므로 모델 불일치 문제가 발생한다. 또한 문서를 검색할 때 고비용의 조인 연산이 필요하다. 하지만 XML 구조 속성을 기반으로 설계된 객체 데이터베이스는 트리기반의 XML문서를 저장할 때 모델 측면에서 매우 자연스럽다. 또한 XML 문서를 관계형 데이터베이스에 파싱하여 저장하는 데 드는 비용과 노력을 줄일 수 있다. 따라서 관계형 질의가 아닌 XML 문서 내에서 원하는 정보를 찾을 수 있는 XPath 질의가 필요하고, 보다 순서 및 구조화된 문서정보를 처리할 때에도 매우 유리하다.

2. 관련연구

생명정보학 데이터들이 다양한 플렛 파일로 존재하고 이러한 정보를 통합, 관리하기 위해 많은 연구가 진행되어왔다. 데이터 통합 기법에는 크게 세 가지로 구분 될 수 있다. 첫째, 현재 가장 많이 사용하고 있는 링크기반 통합기법은 플랫 파일 데이터베이스들을 www링크와 인덱스를 이용하여 연결하지만 실제적 통합이 아니기 때문에 링크 에러 발생가능 성이 높고 ad-hoc질의를 지원하지 않는다. 둘째, 미디에이터기반 기법은 데이터 소스에 대한 통합 뷰를 생성하고 통합 질의 메커니즘으로 질의 수행하지만 인터넷을 통한 응답 시간이 길고 데이터 재조직이 어렵다는 단점이 있다. 셋째, 데이터 웨어하우스 기반 통합기법은 통합스키마를 생성하고 모든 소스 데이터를 생성된 스키마에 따라 데이터 웨어하우스에 로딩하여 관리하는 기법이다. 대기시간이 필요 없고 네트워크 및 인터넷의 연결에 대한 의존도가 많지 않기 때문에 시스템에 대한 신뢰도가 우수하다.

이 연구는2003년도 KISTEP의 특정연구개발과제의 지원으로 수행되었음.

바이오 정보 시스템 간 생물학 데이터의 상호 교환을 편하게 하기 위하여 제안된 BSML[2]은 AGAVE, BIOML, DAS, GAME들과 다르게 DNA 구조와 같은 정보를 인코딩 하고 표현하는 방법이 여러 XML 형식들 중 다른 포맷에 비하여 완성도가 높고 구체적이다. 1997년에 시작된 BSML DTD는 2002년에 3.1 버전이 나왔으며 많은 응용프로그램과 데이터베이스들은 유전체 데이터의 교환과 시각화를 위해 사용되고 있다. 위와 같은 여러 가지 장점을 보유한 BSML은 미국 국립생물 정보센터인 NCBI[3], EMBL[4], PDB[5] 등에서 바이오 서열 데이터를 표현하기 위한 새로운 표준으로 채택되었으며 이를 기반으로 본 연구에서도 BSML을 데이터 웨어하우스 통합 기법을 기반으로 BSML DTD를 이용하여 통합 스키마로 구성하였다.

3. 시스템 설계

통합시스템 구조는 그림 1과 같이 단백질 데이터 관련 소스 데이터를 BSML DTD를 바탕으로 통합 스키마를 등록한 Oracle 9.2.0.4.0 XML DB와 여러 종류의 플랫 파일을 통합된 스키마로 변환 하여 객체 클래스 또는 파일 형태로 저장 할 수 있는 통합 데이터 포맷 변환기, 주석 및 버전 서열을 편집하거나 저장할 수 있는 MyPage 편집기, XPath검색 질의를 받아 구문 분석과 질의 처리를 담당하는 생물정보 검색 질의 처리기, 마지막으로 저장 및 관리를 수행하는 저장 관리기로 구성된다.

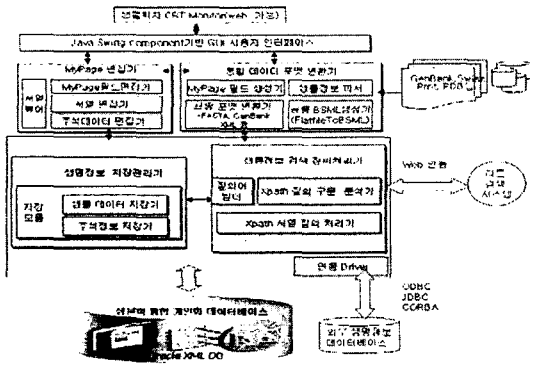

그림 1. XML DB를 이용한 통합시스템 구조

4. 단백질 데이터 통합 관리 시스템

4.1 통합 데이터 포맷 변환기

생명 정보학의 대표적인 데이터베이스에서 제공되는 GenBank, PDB, Swiss-prot, FASTA 등과 같은 플랫파일들을 파싱하여 그림 3과 같이 BSML 스키마로 생성할 수 있다. 포맷 변환기는 데이터를 추출하는 파싱 모듈과 객체로 저장한 후 BSML DTD에 맞게 통합된 스키마 구조에 따라 XML문서로 기술하는 FlatFileToBSML 모듈, 생성된 BSML구조에 따라 Genbank 또는 FASTA와 같은 플랫 파일의 형태로 복원이 가능한 포맷 변환기, MyPage 필드 생성기로 구성된다.

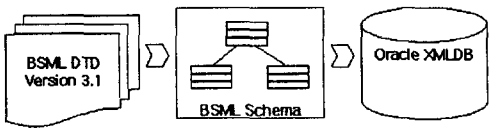

그림 2. 객체기반 변환 및 저장관계

그림 2는 BSML DTD를 객체 관계 데이터 베이스에 스키마를 등록하기 위한 변환 단계를 나타낸다.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
elementFormDefault="qualified">
<xs:element name="Aligned-chart-widget">
 <xs:complexType>
   <xs:sequence>
   <xs:element ref="Chart"/>
   <xs:element ref="Quantifier" minOccurs="0"/>
   <xs:element ref="Object" minOccurs="0"/>
   <xs:element ref="Resource" minOccurs="0" maxOccurs="unbounded"/>
   <xs:choice minOccurs="0" maxOccurs="unbounded">
   <xs:element ref="Attribute-list"/>
   <xs:element ref="Cross-reference"/>
   <xs:element ref="Link"/>
   <xs:element ref="Extended-link"/>
   <xs:element ref="Group-link"/>
  </xs:choice>
 </xs:sequence>
....
```

그림 3. DTD 기반으로 변환된 BSML 스키마

그림 4는 변환된 스키마를 Oracle XML DB에 등록한 후 테이블을 생성하고 생성된 테이블에 저장하는 SQL 질의 과정이다.

```
BEGIN
  DBMS_XMLSCHEMA.registerSchema(
    'http://www.w3.org/2001/XMLSchema/bsml3_1.xsd',
    getclobdocument('bsml3_1.xsd'), TRUE, TRUE, FALSE, FALSE
  );
END;

CREATE TABLE BSMLTABLE(
NAME VARCHAR2(30) PRIMARY KEY,
DOC XMLTYPE
) XMLTYPE COLUMN DOC
  XMLSCHEMA
  "http://www.w3.org/2001/XMLSchema/bsml3_1.xsd"
  ELEMENT "Aligned-chart-widget"

insert into BSMLTABLE values('bsml00_vdata.xml',
XMLType(getClobDocument('bsml00_vdata.xml')))
```

그림 4. XML DB에 스키마 등록, 테이블생성, 문서저장 과정

4.2 MyPage 편집기

서열 뷰어 모듈은 XPath질의를 통하여 얻어진 검색 정보를 기초로 선택한 후, 선택된 서열의 A, C, G, T 함량 및 서열 및 주석 정보를 나타낸다. MyPage 필드 편집기는 통합 데이터 포맷 변환기를 거쳐 파싱 또는 질의를 통하여 MyPage 필드 모듈이 활성 될 수 있다. 서열 편집기[6]는 염기의 구성 비율을 계산하는 Base Composite

연산, 특정 부분의 시작과 끝부분 위치를 지정하여 새로운 엔트리를 생성하는 Set Range연산, 상보 서열을 생성하는 Complement Sequence연산, 회전 연산인 Rolate연산, 마지막으로 DNA과 RNA사이에 전사되는 과정 즉, 티민(T)과 우라실(U)이 상호 전환되는 Rotate연산 모두 다섯 개의 연산들로 구성되었다. 주석데이터 편집기는 편집중인 데이터에 필요한 주석 정보를 추가한 후 저장한다.

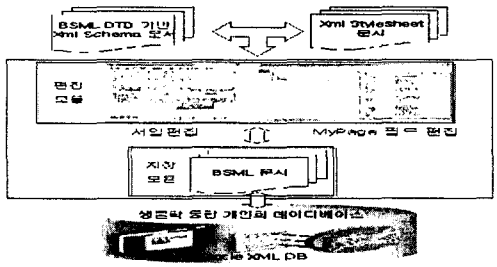

그림 5. MyPage편집기와 저장관리기의 모듈 관계

그림 5는 MyPage편집기와 저장관리기와의 모듈간의 관계로 BSML 문서 생성 및 저장과정을 보여준다.

4.3 생물정보 검색 질의 처리기 및 저장관리기

생물 정보 검색 질의 처리기는 검색 질의어 빌더, XPath 질의 구문 분석기, XPath 서열 질의 처리기, 검색 결과 브라우징 으로 구성된다. 검색 질의어 빌더를 통해 사용자로부터 원하는 질의어를 입력받아 XML DB에 XPath 질의로 질의를 요청하고 그 결과를 검색결과로 뷰잉한다. 또한 저장 관리기는 생물데이터 저장기, 주석정보 저장기로 구성한다. 그림 6은 저장된 BSML 문서에 대한 질의를 나타낸다. 스키마 참조 질의에 의해 XPath 질의는 복잡한 형태를 가지고 있어서 사용자에게 편리한 인터페이스로 검색할 수 있도록 하였다.

```
select extract(e.doc,'chs:GI//feature',
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"'),
extract(e.doc,'chahyosung:Aligned-chart-widget//humen',
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')
from BSMLTABLE e
where
 existsNode(e.doc,'chs:Aligned-chart-widget//feature[title="source"]',
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')=1
and
 existsNode(e.doc,'chs:Aligned-chart-widget//feature[title="mise_feature"]',
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')=1
```

그림 6. 저장된 BSML 문서에 대한 XPath 질의

5. 구현

Oracle 9.2.0.1.0 버전에서 XML DB내에 많은 양의 에러가 발생하여 9.2.0.4.0버전[7]으로 패치 하여 데이터베이스를 구축하였고, XMLType 저장을 위해 Oracle 사의 Java Document Model(DOM) API를 참조했다.

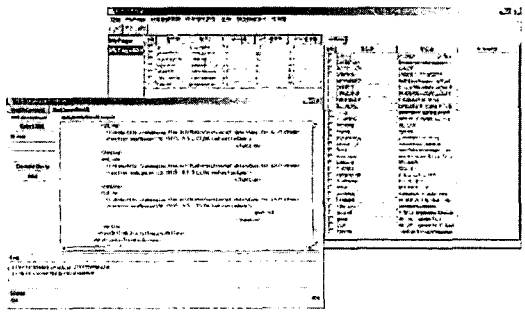

그림 7. 시스템 전체 화면 및 XPath를 이용한 검색화면

그림 7은 시스템 전체 화면과 XPath를 이용한 검색화면이다. 뒤쪽 화면 왼쪽에는 MyPage 편집기, 오른 쪽에는 플렛된 파일들이 파싱되어 BSML 스키마 구조에 맞게 뷰잉된다. 따라서 필요한 객체들을 왼쪽 MyPage 편집기로 이동하여 관리 하고 XPath질의로 검색할 수 있다.

6. 결론

생물학 정보 파일에 대한 수많은 표준화 노력에도 불구하고 바이오 인포메틱스 분야는 대량의 바이오 데이터가 일관성 있게 관리 저장되고 있지 못하는 실정이다. 또한 계층적 구조를 갖는 XML 데이터를 2차원 테이블의 집합인 관계 형 정보로 표현하는 RDBMS 설계에는 본질적인 한계가 있기 때문에 이 연구에서는 단백질 데이터의 효율적 관리를 위해 객체 관계 데이터베이스를 이용하였고 BSML DTD를 기반으로한 BSML 스키마로서의 저장, XPath 질의 처리를 통한 검색 시스템을 구현하였다. 또한 MyPage 편집기를 통해 파싱 과정을 거친 플랫한 여러가지 파일 포맷들은 편집 수정하여 BSML로 새롭게 생성 및 저장 관리 할 수 있고, XPath 질의 처리를 이용하여 범위 또는 구조 질의를 통해 보다 효율적인 검색이 가능하게 되었다. 이로서 단백질 구조 또는 바이오 Pathway, 바이오 온톨로지등의 구조 관련 분야 시스템 개발을 촉진 할 수 있다.

참고문헌

[1] http://www.visualgenomics.ca/gordonp/xml/
[2]' http://www.bsml.org/
[3] http://ncbi.nlm.nih.gov/
[4] G. Stoesser, "The EMBL nucleotide sequence database", Nucl. Acids. Res. 2001.
[5] http://www.rcsb.org/pdb/
[6] P. Sung-Hee, "Buliding Genome and Protein seuence information Mnagement System.",KOSTI, 2002.
[7] http://otn.oacle.com/software/index.html
[8] J. Ostell. "The NCBI data model. Chapter 2 in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins", B.F.F. New York: 2001.
[9] Robin Cover, "XML linking and addressing language", Oasis, 2001.