SCOPML과 SCOPBrowser

윤 형 석[0] · 황 의 윤 · 안 건 태 · 김 진 홍 · 이 명 준
울산대학교 컴퓨터정보통신공학부

SCOPML and SCOPBrowser

Hyeong-Seok Yoon[0] · Eui-Yoon Hwang · Geon-Tae Ahn · Jin-Hong Kim · Myung-Joon Lee
School of Computer Engineering & Information Technology, University of Ulsan

요 약

포스트지놈 시대에 있어서 가장 주된 연구는 단백질의 구조적 유사성이나 분류학적인 연관성을 밝히는 것이다. SCOP 단백질 구조 분류는 이러한 목적을 위하여 3차원 구조가 알려진 단백질에 대한 구조적, 분류학적 관계에 대해 상세한 정보를 제공한다. 그러나 SCOP의 데이터는 단순 텍스트 기반의 자료만 제공되고 있어서, 이를 이용한 다른 분석 도구를 개발하거나 유용한 정보 추출을 할 경우 그 작업이 매우 힘들며 오류 발생의 확률이 높다.

본 논문에서는 단백질 구조 관련 연구자들이 SCOP 데이터를 보다 효과적으로 이용할 수 있도록 구조화된 문서의 표준인 XML을 이용하여 개발된 SCOPML에 대하여 기술한다. 그리고 SCOPML을 이용하여 SCOP 데이터에 대한 효율적인 검색을 지원하는 SCOPBrowser의 개발에 대해 기술한다.

1. 서론

단백질은 대부분 분류학적으로 공통의 기원을 가지거나 구조적으로 유사한 경우가 많다. 단백질 사이의 구조와 분류학적인 관계 정보는 인간 유전체 사업의 결과로 생성된 대량의 서열정보를 번역하고 단백질의 기능을 밝혀내는 데 중요한 역할을 담당한다. SCOP(Structural Classification of Proteins)[1] 단백질 구조 분류 데이터베이스는 이러한 구조적인 정보를 이해하고 접근할 수 있도록 하기 위하여 구축된 대표적인 데이터베이스이다. 현재 SCOP은 단백질 분류 데이터들을 단순 텍스트 형태로 제공하고 있으며 HTML(HyperText Markup Language) 기반으로 웹을 통하여 서비스하고 있다. 이러한 분류정보에 대한 연구는 생물학에 있어서 중요한 의미를 가지지만 자료의 제공 형식이 텍스트 파일이고 검색 또한 HTML 기반이어서 데이터를 다양하게 활용하기 어려운 형편이다. 이러한 문제를 해결하기 위하여 단백질 구조 분류 정보를 보다 효과적으로 표현하고 교환하기 위한 접근방법이 요구된다.

XML(eXtensible Markup Language)은 이러한 문제를 해결하기 위한 이상적인 해결책을 제시해준다. 이미 생물정보학의 여러 분야에서 XML기술을 이용한 연구가 활발해지고 있으며[2], 대표적인 연구로는 유전자 정보에 대한 XML 표준인 BSML (Bioinformatic Sequence Markup Language)[3], 분자구조 기술을 위한 언어로서 단백질 명세를 확장하기 위한 이상적인 기초를 제공하는 CML(Chemical Markup Language)[4], 단백질 서열, 구조, 패밀리(families)등에 관한 명세 언어인 ProML(Protein Markup Language)[5], 그리고 온톨로지(ontology) 기반의 객체 메타모델인 OpenMMS(Open Macromolecular Structure)[6] 등이 있다.

본 논문에서는 단순 텍스트 기반의 SCOP 정보를 구조화된 문서의 표준인 XML을 사용하여 개발된 SCOPML에 대하여 기술한다. 그리고 SCOPML의 응용시스템으로 SCOP 데이터베이스를 보다 효율적이고 용이하게 검색할 수 있는 SCOPBrower에 대하여 기술한다. SCOPML은 SCOP의 자료구조를 XML DTD(Document Type Definition)로 정의하여 SCOP데이터 구조가 가진 정보를 구조화된 문서로 나타낸다. SCOPBrower는 SCOPML에 의하여 생성된 XML 데이터를 이용하여 SCOP 단백질 분류 구조에 대한 정보 분석 및 효율적인 검색 기능을 제공한다.

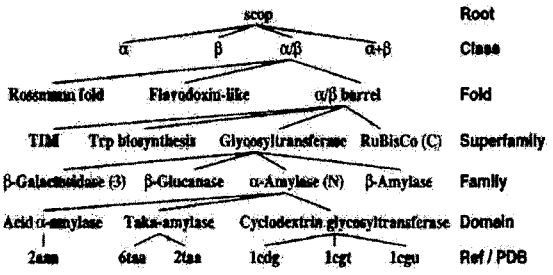
본 논문의 구성은 다음과 같다. 2장에서는 SCOP 데이터베이스에 대해서 간략히 살펴보고, 3장에서는 SCOPML의 설계 및 구현에 대하여 설명한다. 4장에서는 SCOPBrowser의 구현에 대하여 기술하고, 끝으로 5장에서는 결론과 향후과제에 대하여 기술한다.

2. SCOP 데이터베이스

SCOP은 단백질이 지닌 구조적인 유사성과 분류학적인 관계를 기반으로 단백질들을 체계적으로 분류해 놓은 데이터베이스이다. SCOP에는 이미 구조가 밝혀진 단백질들에 대한 자료가 저장되어 있어 미지의 단백질에 대한 구조를 밝히거나 기능을 예측하는 연구에 많이 활용되고 있다.

[그림1]은 SCOP의 계층구조를 나타낸 것이다. PDB(Protein Data Bank)[7]의 모든 단백질은 다른 단백질들과 비교되어 구조적 유사성(structural similarities)을 가지는 그룹으로 분류된다.

† 본 연구는 2002년도 울산대학교 연구비로 연구되었음



[그림1] SCOP의 계층구조

3. SCOPML

SCOPML은 구조적인 웹 문서 표준인 XML 기술을 이용하여 SCOP 데이터를 기술하기 위한 마크업 언어이다. SCOP 데이터베이스가 제공하는 단백질의 계층적인 분류를 효과적으로 기술하고 있으며, 이를 기반으로 단백질 구조 분류와 관련된 응용프로그램은 보다 용이하게 개발될 수 있다.

3.1 SCOPML의 DTD 설계

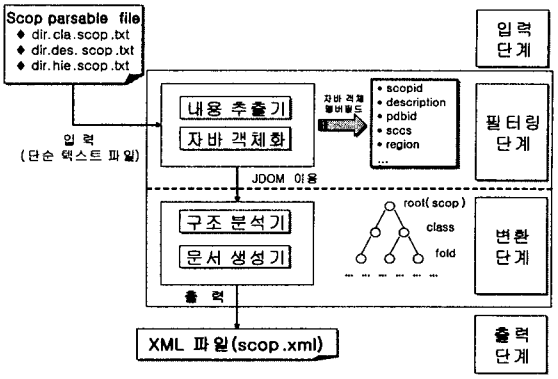
SCOPML DTD[그림2]는 SCOP 데이터베이스에서 나타내고 있는 계층구조를 Element로 표현하고, 각각의 하위 Element를 식별하기 위해 sunid 혹은 sccs 라는 속성 값을 가진다. sunid는 SCOP에서 제공하는 계층구조분류에 대한 식별자이며, sccs는 계층구조에 대한 간략화 된 표현이다.

```
<!ELEMENT scop (class+)>
<!ELEMENT class (#PCDATA | fold)*>
<!ATTLIST class sunid CDATA #REQUIRED
                sccs CDATA #REQUIRED>
<!ELEMENT fold (#PCDATA | superfamily)*>
<!ATTLIST fold sunid CDATA #REQUIRED
               sccs CDATA #REQUIRED>
<!ELEMENT superfamily (#PCDATA | family)*>
<!ATTLIST superfamily sunid CDATA #REQUIRED
                      sccs CDATA #REQUIRED>
<!ELEMENT family (#PCDATA | protein)*>
<!ATTLIST family sunid CDATA #REQUIRED
                 sccs CDATA #REQUIRED>
<!ELEMENT protein (#PCDATA | species)*>
<!ATTLIST protein sunid CDATA #REQUIRED>
<!ELEMENT species (#PCDATA | domain)*>
<!ATTLIST species sunid CDATA #REQUIRED>
<!ELEMENT domain (sid, pdb, region)>
<!ATTLIST domain sunid CDATA #REQUIRED>
<!ELEMENT sid (#PCDATA)>
<!ELEMENT pdb (#PCDATA)>
<!ELEMENT region (#PCDATA)>
```

[그림2] SCOPML의 DTD

3.2 SCOPML 변환

SCOP 데이터의 SCOPML 변환과정은 입력파일 분석 단계, 데이터 필터링 단계, XML 변환 단계, 그리고 XML 파일 생성 단계로 구성된다[그림3]. 입력파일 분석 단계에서는 SCOP의 텍스트 파일을 입력받아 내용 정보를 추출하고, 추출된 정보는 필터링 단계에서 자바 객체로 저장된다. XML 변환 단계에서는 필터링 단계에서 생성한 자바 객체를 JDOM(JAVA DOM)[8]을 이용하여 XML로 변환한다.



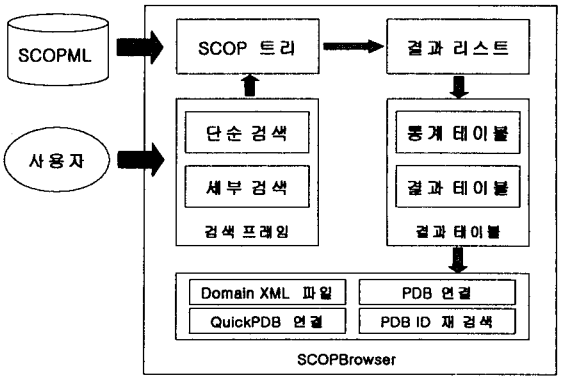
[그림3] SCOPML 변환 과정

4. SCOPBrowser

SCOPBrowser는 SCOPML 문서를 기반으로 SCOP 정보를 효율적으로 검색하고 브라우징 할 수 있는 도구이다. SCOPBrowser는 자바 언어와 JDOM 인터페이스를 이용하여 SCOPML 문서를 분석하고, SCOP 단백질 구조 분류를 트리 구조로 한 눈에 볼 수 있는 기능을 지원한다. 또한, 사용자가 원하는 데이터를 효율적으로 추출하여 보여주는 검색기능을 제공한다.

4.1 SCOPBrowser의 구조

SCOPBrowser는 내부적으로 SCOPML을 이용하여, SCOP 트리를 생성한다. SCOPML이 가지는 모든 데이터를 JDOM을 사용하여 분석한 다음, 트리로 변환하고 트리로부터 모든 기능을 수행한다.

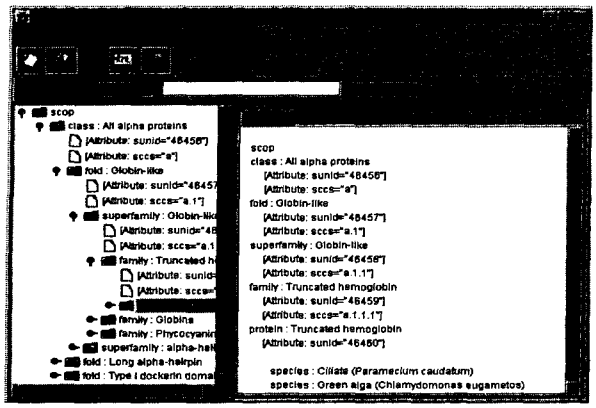


[그림4] SCOPBrowser의 구조

[그림4]는 SCOPBrowser의 전체적인 구조와 데이터의 흐름을 표현한 것이다. SCOPBrowser의 기능은 SCOP 트리를 중심으로 수행되어진다. 즉, 사용자로부터의 검색 요청은 SCOP 트리를 분석하여 결과 리스트를 얻고, 결과 리스트로부터 Table을 생성하게 된다. 또한, 사용자는 검색한 결과가 나타나는 Table로부터 다른 부가적인 기능을 이용할 수 있다.

4.2 SCOPBrowser의 구현 및 기능

SCOPBrowser는 SCOPML의 효과적인 검색 기능에 중점을 두어 개발되었다. SCOPBrowser는 JDOM을 사용하여 SCOPML로부터 자바객체인 JTree를 생성하고, 이 트리로부터 원하는 데이터를 검색할 수 있는 인터페이스를 가진다[그림5].



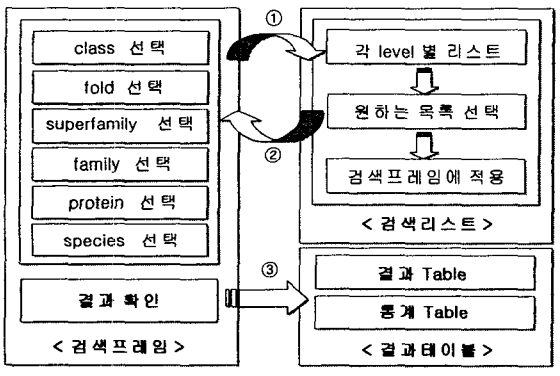
[그림5] SCOPBrowser의 기본 인터페이스

4.2.1 트리 뷰어

SCOPML은 SCOP의 단백질 구조 분류를 계층적으로 표현하고 있다. 따라서 SCOPBrowser는 SCOPML을 트리 형태로 변환하고, 생성된 트리로부터 SCOP에서 제공하는 모든 정보를 효과적으로 조회할 수 있다.

4.2.2 검색 기능

검색 기능은 단순검색기능과 세부검색기능으로 나누어진다. 단순검색기능은 sunid, sccs, scopid, pdbid를 이용하여 검색하는 기능이다. 그리고, 세부검색기능은 세부적으로 여러 가지 조건을 통하여 검색할 수 있도록 하여 사용자가 원하는 데이터와 더불어 결과에 대한 통계적 수치도 제공한다.



[그림6] 세부검색기능의 수행 절차

[그림6]은 검색프레임으로부터 검색 결과가 나오는 결과테이블까지의 과정을 설명하고 있다. 사용자는 각 level에서 검색을 원하는 데이터를 선택하고(①, ②)이 있고, 선택이 끝난 후 검색 결과를 확인(③) 한다.

4.2.3 결과 테이블에서의 기능

SCOPBrowser는 세부적이고 다양한 기능을 결과 테이블에서 팝업 메뉴를 통해 접근할 수 있다. 사용자는 팝업 메뉴를 이용하여 SCOPBrowser에 적용, 도메인정보를 XML파일로 보기, PDB 연결, QuickPDB 연결, PDB ID로 재 검색 등의 기능을 수행할 수 있다.

4.3 SCOPBrowser 검색 결과 분석

SCOP 데이터베이스에서 단백질 구조를 분류하는 기본 단위는 도메인(domain)이다. 현재 SCOP 사이트에서는 단순히 특정 분류에 대한 도메인 조회나 해당 단백질 도메인에 대한 정보 확인만이 가능하다. 하지만, 개발된 SCOPBrowser는 SCOP에서 제공하는 정보뿐만 아니라, 하나의 단백질을 이루는 도메인의 종류, 분류학적인 연관관계, 구조적으로 유사한 도메인들의 존재 유무, 그리고 여러 단백질에 공통으로 나타나는 도메인의 존재 등 다양한 통계자료를 산출할 수 있다.

5. 결론 및 향후 연구

지금까지 연구 결과로 축적된 생물정보를 효과적으로 관리하고 데이터의 교환을 용이하게 지원하기 위한 방법들이 최근 활발하게 연구되고 있다. 그 중 대표적인 것이 XML 기술을 기반으로 한 생물정보의 구조적 문서화 작업이다. 단순 텍스트 형식으로 저장된 자료들을 웹상의 구조적 문서 표준인 XML을 이용하여 재구성함으로써 데이터의 재활용성과 가용성을 높일 수 있다.

본 논문에서는 SCOP 데이터베이스에서 제공하는 데이터의 XML 표현 기법인 SCOPML과 효율적으로 SCOPML을 탐색할 수 있는 SCOPBrowser의 개발에 대하여 기술하였다.

향후 연구로는 SCOPML의 재사용성과 확장성을 고려하여 XML DTD를 XML 스키마(Schema) 구조로 재구성하고 SCOPBrower의 검색기능을 보다 세분화하여 다양한 통계자료를 산출할 수 있는 도구로 확장할 예정이다.

6. 참고문헌

[1] Alexey M., Steven B., Tim H. and Cyrus Ch., "SCOP: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", J. Mol. Biol.,247, p536-540, 1995.

[2] Guerrini VH. and Jackson D., "Bioinformatics and Extended Markup Language", Online Journal of Bioinformatics 1:12-21, 2000.

[3] http://www.bsml.org, "XML Data Standard for Genomics: The Bioinformatic Sequence Markup Language (BSML)DTD".

[4] P. Murray-Rust and H. Rzepa, "Chemical Markup Language and XML Part I. Basic principles", J. Chem. Inf. Comp. Sci, Vol.39, No.6,pp.928-942, 1999.

[5] Dniel H,Ralf Z. and Thomas L.,"ProML-The Protein Markup Language for specification of protein sequences, strucctures and families", German Conference on Bioinformatics 2001, Oct, 2001.

[6] http://openmms.sdsc.edu, "Corba, Relation Database and XML Software for Macromolecular Structure.

[7] John W., Zukang F. and Helen M., "The Protein Data Bank:unifying the archive", Nucleic Acids Research, Vol.30, No.1, pp245-248, 2002

[8] Loannides, D., "XML schema langueages: beyond DTD", Library Hi Tech, v.18, No.1, p9-14, 2000.