제19회 한국정보처리학회 춘계학술발표대회 논문집 제10권 제1호 (2003.5)

통합 모티프 자원 교환을 위한 XML 생성기 구현

이범주, 김영균, 최은선, 류근호
충북대학교 데이터베이스연구실
{bjlee, ygkim, eschoi, khryu}@dblab.chungbuk.ac.kr

Implementation of XML Generator for Exchanging Integrated Motif Resources

Bum Ju Lee, Young Gyun Kim, Eun Sun Choi, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

최근 생물정보학 분야에서는 이질적인 데이터 형식으로 제공되는 모티프 자원들을 하나의 자원으로 통합하고자 하는 많은 연구가 진행되고 있다. 이러한 통합을 위해 웹 기반 cross-reference를 이용한 논리적 통합이 주로 사용되어져 왔고, 이러한 이질적인 데이터의 상호교환을 위한 표준 형식으로 XML을 정의하여 이용하기 시작하였다. 그러나 이러한 웹 기반 cross-reference를 이용한 논리적 통합은 복잡한 질의 처리 문제, 중복된 데이터베이스 핸들링 문제 등을 지니고 있다. 따라서 이 논문에서는 이러한 문제점들을 개선하기 위해 물리적으로 모티프 자원들을 하나의 통합 데이터베이스로 구축하였고, 통합된 모티프 자원의 표준적 상호 교환을 위해 XML 생성기를 구현하였다.

1. 서론

모티프는 분자의 기능을 예측할 수 있는 특정한 서열의 패턴이나 구조적인 특징을 가진다[1, 2, 3, 4]. 지난 10년간 개발된 모티프 데이터베이스들로는 PROSITE 데이터베이스[8, 13], PRINTS 데이터베이스[5, 7, 11], Pfam 데이터베이스[6] 등 매우 다양한 모티프 데이터베이스들이 독자적으로 개발, 발전되어져 왔다. 또한 최근에 이르러 이러한 이질적인 데이터 구조로 생성된 여러 모티프 데이터베이스들의 통합을 위해 웹 기반 Cross-reference를 이용한 논리적 통합이 주로 사용되었다[3, 15, 22]. 또한 생물정보학 데이터들의 상호 교환을 위해 기존의 플랫파일과 같은 형태를 벗어나 XML[19] 형식으로 배포되어가고 있다. 이러한 예로 XML 정의를 통하여 생물정보 데이터 교환을 위한 BioML과 같은 마크업 언어가 출현하였다[18].

그러나 웹 기반 cross-reference를 이용한 논리적 통합 및 검색 시스템은 엔트리 상호간 데이터 구조를 변경하지 않고 관련된 엔트리간에 유연한 통합을 지원할 수 있는 장점에 비해, 복잡한 질의 처리 문제, Cross-reference된 과도한 엔트리들의 수, 네트워 과부화 등과 같은 문제점들을 지니고 있다[2, 23].

이 논문에서는 위에 기술한 문제들에 대한 해결 방안으로 단백질 모티프들의 Annotation 정보, 3차 구조 정보 및 분류 정보 등을 물리적으로 하나의 통합된 DBMS를 사용하여 저장함으로써 효율적 관리를 위한 기반을 마련한다. 또한 이렇게 통합된 모티프 자원의 상호 교환을 위하여 XML 생성기를 구현하였다. 따라서 웹 기반 cross-reference 통합에서 나타나는 복잡한 질의 처리 문제와 중복된 데이터베이스들의 핸들링 문제들에 대한 해결책을 제시하며, 통합된 모티프 자원을 XML 형식으로 재 생성하여 상호 교환이 가능하도록 하였다.

2. 관련연구

2.1 InterPro 데이터베이스

단백질 패밀리, 도메인, functional site들에 대한

※ 이 논문은 2002년도 한국학술진흥재단의 지원에 의하여 연구되었음(KRF-2002-072-AM1013)

물리적 통합 문서 자원을 목적으로 생성된 InterPro 데이터베이스는 PRINTS, PROSITE, Pfam, ProDom과 같은 시그네쳐 데이터베이스들에 대한 검색 진단 데이터와 문서들을 하나의 집중된 자원으로 통합하였다.

통합 메소드로 parent/child와 contains/found_in을 사용한 이 데이터베이스의 각 엔트리는 functional description, annotation, literature reference를 포함하고 있다.

이 데이터베이스 버전 2.0(2001년도)에서 총 6,804개의 regular expression, profile, fingerprint, HMMs를 포함한 약 3,000개의 엔트리를 포함하고 있으며, 웹상에서 엔트리 데이터와 매치 데이터를 XML 형식으로 배포하고 있다[1, 2, 4].

2.2 생물정보 데이터의 특성과 XML

다양한 소스 및 생물학 실험에서 생성되는 데이터들의 타입은 이종 형태의 매우 다양한 포맷으로 제작되며 규칙적인 업데이트와 빈번한 수정이 요구된다. 이러한 데이터들은 데이터 자체로 종료되지 않고, 새로운 데이터를 생성하기 위한 데이터로 사용가능하며 데이터 자체 분석만으로도 새로운 데이터를 생성한다. 따라서 위와 같은 데이터를 관리하기 위한 데이터베이스는 이종 형태의 데이터들을 상호 교환 및 검색할 수 있어야하며, 빠른 스키마 변경을 지원할 수 있는 유연성을 가지고 있어야하고, 데이터 자체만을 저장하는 것이 아니라 데이터에 대한 메타 정보들까지 저장하여야한다[14, 20].

이렇게 다양한 생물정보들의 상호 교환을 위해 생성된 BioML(BIOpolymer Markup Language)은 생물 정보학에서 단백질과 nucleotide sequence 등에 대한 실험 정보와 같은 주석처리 제공을 목적으로 XML[19]을 기반으로 Washington University와 Institute for Marine Bioscience에서 정의하여 운영되고 있다[18].

3. 모티프 자원 통합

이 논문에서는 이질적 데이터 포맷의 모티프 자원을 하나로 통합하기 위해 PRINTS, Pfam, Prosite 데이터베이스에서 제공하는 각각의 플랫파일을 분석하고, 이를 분해 및 합병하였다. 이러한 과정의 수행은 그림 1과 같다.

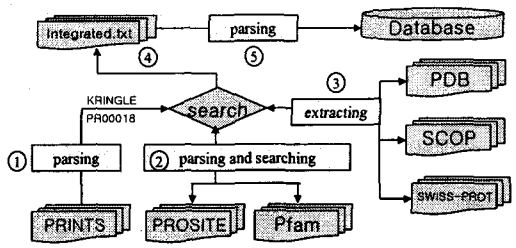

그림 1. 모티프 자원 통합 순서

첫째, PRINTS 플랫파일에서 각 ID, Accession number, PROSITE reference, Pfam reference 라인 항목을 파싱한다. 둘째, 파싱한 PROSITE reference와 Pfam reference 항목을 각각 PROSITE, Pfam 플랫파일에서 검색한다. 이때 검색 후 동일한 엔트리가 나타나면 해당 엔트리에서 정보를 추출한다. 셋째, Pfam플랫파일에 존재하는 PDB[9]와 SCOP에 해당하는 reference 항목을 파싱한 후 그 항목을 이용하여 PDB 플랫파일에서 3차 구조 정보를 추출해 내고 SCOP은 ID만을 추출해 낸다. 넷째, 이렇게 추출된 모든 정보들은 새로운 하나의 플랫파일에 새롭게 저장된다. 마지막으로, 이렇게 생성된 새로운 플랫파일은 다시 파싱과정을 거쳐 데이터베이스에 저장된다. 이러한 데이터베이스의 스키마는 그림 2와 같다.

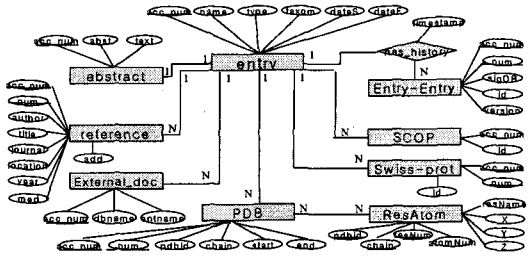

그림 2. 통합 데이터베이스 구축을 위한 E-R 다이어그램

4. 자원교환을 위한 XML

3장에서 각각 독립적으로 존재하는 모티프 자원들을 하나의 물리적 자원으로 통합하여 관계형 데이터베이스에 저장하였다. 그러나 이러한 자원들은 또 다른 사용자를 위하여 상호 교환되어져야 한다. 따라서 현재 상호 문서 교환의 표준이 되고있는 XML의 형태로 이미 통합된 자원들을 재 생성한다. 이러한 XML은 간단한 데이터교환, 특정 분야에 대한 편리한 마크업 언어의 설계, 스스로 설명이 가능한 데이터와 같은 장점으로 인해 널리 쓰이고 있다[16,

17]. 이 논문에서 자원 상호 교환을 위해 구현한 XML의 DTD는 InterPro 데이터베이스에서 사용된 DTD를 표준으로 생성하였다.

5. 구현 및 평가

우리는 각 모티프 데이터베이스에서 제공하는 플랫 파일들을 하나로 통합된 새로운 플랫파일로 생성하기 위해 Window 2000 환경에서 C언어를 사용하였으며, 통합된 플랫파일을 오라클 데이터베이스에 삽입하기 위해 ProC언어를 사용하였다. 또한 시스템 기종으로는 Sun사의 Enterprise 250을 사용하였으며, 운영체제로는 Sun Solaris 7(5.7), DBMS로는 Oracle 8i를 이용하였다. 또한 XML 생성기를 위해 Windows 환경에서 Java 1.3을 이용하여 구현하였다. XML 생성기와 데이터베이스와의 연결을 위해서 JDBC를 사용하였고 Graphic User Interface를 위해서는 Java Swing을 사용하였다.

통합에 이용한 맴버 데이터베이스들 즉, Prosite, Pfam, PRINTS의 엔트리들은 다음과 같다.

① PRINTS에서 제공하는 1,410개의 fingerprint들

② Prosite에서 제공하는 1,510개에 해당하는 rule, regular expression, profile들

③ Pfam-A.seed에서 제공하는 3,849개의 엔트리들

이러한 엔트리들을 이용하여 3장에서 기술한 통합 과정을 거쳐 5,670개의 새로운 엔트리로 재구성하였다.

이러한 엔트리들은 그림 3과 같이 검색이 가능하다. 그림 3은 Dihydrofolate reductas signature에 관한 내용들을 웹 상에서 검색한 결과 인터페이스이다.

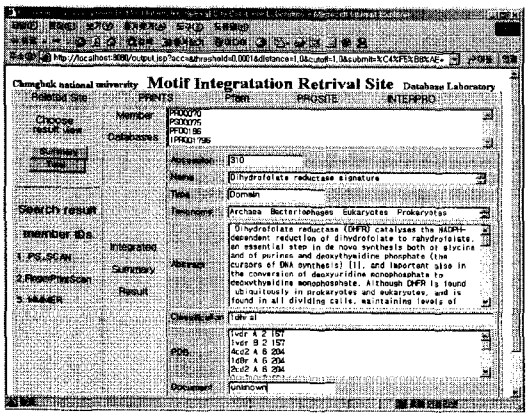

그림 3. 통합된 모티프 자원 검색 결과 인터페이스

이러한 통합된 모티프 자원 교환을 위해 그림 4와 같은 XML 생성기를 구현하였다.

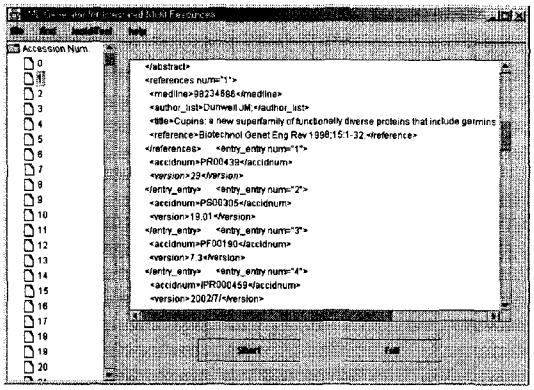

그림 4. XML 생성기 인터페이스

그림 4에서 보듯이 XML 생성기의 특징은 크게 두 가지로 나뉜다. 첫째, 하나의 엔티티만을 선택하여 XML 형식으로 생성 및 출력하는 기능과 둘째, 필요시에 전체 엔트리들을 하나의 XML 파일로 모두 생성하는 기능이다.

6. 결론

이 논문에서는 각각의 다양한 데이터 형식으로 성장해온 모티프 데이터베이스들에서 나타나는 이질적인 검색 문제와 웹 기반 cross-reference 통합에 따른 복잡한 질의처리, 중복된 데이터베이스 엔트리 핸들링 문제들을 해결하기 위해 모티프 자원들에 대한 물리적 통합 연구에 대하여 다루었다. 따라서 모티프의 Annotation 정보와 3차 구조 정보 및 분류 정보를 통합하여 하나의 DBMS에 저장하였고, 통합된 모티프 자원들을 상호 교환하기 위해 아래와 같은 과정을 수행하였다.

■ 각 모티프 데이터베이스들에서 제공하는 플랫 파일을 분석

■ 각 엔트리에 대한 모티프 3차 구조 정보와 분류 정보 통합

■ 개체-관계 모델링을 통해 Oracle 8i를 사용하여 통합 데이터베이스 구축

■ 통합된 모티프 자원 상호 교환을 위한 XML 생성기 구현

따라서, 웹 기반 통합에 따른 복잡한 질의 처리 문제, 중복된 데이터베이스들의 핸들링 문제, 기존의 데이터베이스 검색시 사용자가 겪는 이질적 검색환

경 및 반복 접근 문제를 해결하였고 기존의 웹 기반 통합 검색에서 지원하지 못했던 단백질의 3차 구조 정보, 분류 정보, 샘플 정보의 지원을 가능케 하였다. 또한 통합된 모티프 자원들을 XML 생성기를 통해 재 생성하여 XML 형식의 상호 교환이 가능하도록 하였다.

참고문헌

[1] R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, E. Birney, M. Biswas, P. Bucher, L. Cerutti, F. Corpet, M. D. R. Croning, R. Durbin, L. Falquet, W. Fleischmann, J. Gouzy, H. Hermjakob, N. Hulo, L. Jonassen, D. Kahn, A. Kanapin, Y. Karavidopoulou, R. Lopez, B. Marx, N. J. Mulder, T. M. Oinn, M. Pagni, F. Servant, C. J. A. Sigrist and E. M. Zdobnov, "The InterPro database, an integrated documentation resource for protein families, domains and functional sites", Nuleic Acids Research, Vol.29, No.1, pp.37-40, 2001.

[2] M. R. Wilkins, K. L. Williams, R. D. Appel, D. F. Hochstrasser, "Proteome Research: New Frontiers in Functional Genomics", Springer-Verlag Berlin Heidelberg, pp.109-175, 1997.

[3] Minoru Kanehisa, "Post-Genome Informatics", Oxford university press, pp.35-47, 2000.

[4] David W. Mount, "Bioinformatics : Sequence and Genome Analysis", Cold Spring Harbor Laboratory Press, pp.45-48, 2001.

[5] T. K. Attwood, M. E. Beck, D. R. Flower, P. Scordis, N. Selley, "The PRINTS protein fingerprint database in its fifth year", Nucleic Acids Research, Vol.26, No.1, pp.304-308, 1998.

[6] Alex Bateman, Evan Birney, Lorenzo Cerruti, Richard Durbin, Laurence Etwiller, Sean R. Eddy, Sam Griffiths-Jones, Kevin L. Howe, Mhairi Marshall, Erik L. L. Sonnhammer, "The Pfam Protein Families Database", Nucleic Acids Research, Vol.30, No.1, pp.276-280, 2002.

[7] T. K. Attwood, H. Aviison, M. E. Beck, M. Bewley, A. J. Bleasby, F. Brewster, P. Cooper, K. Degtyarenko, A. J. Geddes, D. R. Flower, M. P. Kelly, S. Lott, K. M. Measures, D. J. Parry-Smith, D. N. Perkins, P. Scordis, D. Scott, C. Worledge, "The PRINTS Database of Protein Fingerprints: A Novel Information Resource for Computational Molecular Biology", J. Chem. Inf. Comput. Sci. 37, pp.417-424, 1997.

[8] Laurent Falquet, Marco Pagni, Philipp Bucher, Nicolas Hulo, Christian J. A. Sigrist, Kay Hofmann, Amos Bairoch, "The PROSITE database, its status in 2002", Nucleic Acids Research, Vol.30, pp.235-238, 2002.

[9] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne, "The Proten Data Bank", Nucleic Acids Research, Vol.18, pp.235-242, 2000.

[10] Ramez Elmasri, Shamkant B. Navathe, "Fundamentals of Database Systems", Addison-Wesley, Reading, Massachusetts, 2000.

[11] T. K. Attwood, M. J. Blythe, D. R. Flower, A. Gaulton, J. E. Mabey, N.Maudling, L. McGregor, A. L. Mitchell, G. Moulton, K. Paine, P. Scordis, "PRINTS and PRINTS-S shed light on protein ancestry", Nucleic Acids Research, Vol.30, No.1, pp.239-241, 2002.

[12] Doug Brutlag, "Protein Structure & Motifs", Biochemistry 201, Molecular Biology, 2000.

[13] Cynthia Gibas, Per Jambeck, "Developing Bioinformatics Computer Skills", O'REILLY, pp.290-295, 2001.

[14] Attwood, "The Babel of Bioinformatics", Science 290, pp.471-473, 2000.

[15] Barbara Eckman, Julia Rice, Bill Swope, "Heterogeneous Data and Algorithm Integration in Bioinformatics", ISMB, 10th International Conference Tutorial, 2002.

[16] Steven Holzner, "Inside XML", New Reders Publish, pp.33-40, 2000.

[17] Bill Brogden Chris Minnick, "JAVA Developer's Guide to E-Commerce with XML and JSP", SYBEX Inc, pp.3-10, 2001.

[18] http://www.bioml.com/BIOML/, 2003.

[19] http://www.w3.org/pub/WWW/TR/WD-xml. html, 2003.

[20] 류근호, "유전체 데이터베이스와 EST 데이터베이스", 연구개발 정보센터, 지식정보인프라 10월호, pp.48-61, 2000.

[21] 박경현, 김록원, 양은주, 최은선, 류근호, "반구조적 데이터의 효율적인 최소 경계 스키마 추출 기법", 한국정보과학회 학술발표논문집, 제27권, 제2호, 2000.

[22] 이범주, 최은선, 류근호, "모티프 자원 통합 데이터베이스 구축 및 메타엔진 설계", 한국정보처리학회 추계학술발표대회논문집, 제9권 제2호, pp.1877-1880, 2002.

[23] 이범주, 최은선, 류근호, "모티프 자원 통합을 위한 데이터베이스 구축", 한국정보과학회 추계학술대회 발표논문집, 제9권, 제2호, pp.160-162, 2002.

[24] 이범주, 최은선, 류근호, "모티프 자원 통합을 이용한 단백질 모티프 예측 시스템 구현", 한국정보처리학회 논문지, 제10 D권, 제3호, 2003.