제18회 한국정보처리학회 추계학술발표대회 논문집 제9권 제2호 (2002.11)

PDB 데이터에서 PSAML로의 변환도구 개발

조민수*, 이수현*, 이명준**
*창원대학교 컴퓨터공학과
**울산대학교 컴퓨터정보통신공학부
e-mail: oops@pl.changwon.ac.kr

Development of a Translator from PDB Data to PSAML

Min-Su Cho*, Su-Hyun Lee*, Myung-Joon Lee**
*Dept. of Computer Engineering, Changwon National University
**School of Computer Engineering & Information Technology, University of Ulsan

요 약

현재의 단백질 구조비교 시스템들 사이의 호환성이나 상호작용성의 문제를 해결하고 단백질 구조를 비교하는 시스템을 신속히 개발하기 위해서 단백질 3차구조를 표현하기 위한 데이터를 추출하여 XML과 같은 표준 형식으로 기술된 데이터를 제공하는 것이 바람직하다. 이에 따라 단백질의 2차구조 구성요소와 그들 사이의 관계를 이용하여 단백질 구조를 기술하는 PSA가 제안되었으며, PSA를 기반으로 하여 단백질 데이터의 XML 표현기법인 PSAML이 제안되었다. 본 논문에서는 PSAML 데이터의 생성을 위하여 PDB에서 제공되는 데이터를 PSAML 형식으로 변환시키는 도구를 설계하고 구현하였다. 변환도구는 XML DOM과 Java를 이용하여 구현되었으며, 생성된 데이터는 단백질 구조 및 유사성을 비교하기 위한 단백질 구조비교 시스템에서 사용될 수 있다.

1. 서 론

PDB(Protein Data Bank)[1]는 실험으로 밝혀진 생물체의 고분자 구조들에 대한 정보를 모아 놓은 데이터베이스로서 2002년 9월 10일 현재 18691개의 단백질에 대한 서열과 구조 정보를 제공하고 있다. PDB에 포함된 정보들은 X-선 회절이나 NMR에 의하여 정밀하게 얻어지므로 단백질의 구조 연구에 있어서 매우 활발히 사용되고 있다. 그러나 PDB 데이터의 형식은 단순한 텍스트의 형태로서 정형화된 문법 명세가 부족하여 파싱(parsing) 에러 가능성을 내포하고 있으므로 PDB 데이터를 직접 활용하기 위해서는 상당한 노력이 필요하다.

따라서 단백질 관련 정보를 표현하고 교환하기 위한 보다 효과적인 접근방법이 요구되며, XML은 이러한 문제를 해결하기 위한 이상적인 해결책을 제시해 준다. 그러나, CML이나 BIOML 등과 같이 현존하는 XML 기반의 언어들은 대부분 일반적인 목적을 위하여 고안되어서 구조비교나 예측과 같은 특수한 목적으로 사용하기를 원하는 사용자들의 요구를 충족시키지 못한다.

특히 단백질의 접힘(folding)과 구조를 이해하고 분석하는 데 있어서 단백질 데이터의 전체를 이용하는 것보다 단백질 구조의 특징을 나타내는 대표적인 정보를 이용하는 것이 효과적이다. 단백질의 2차구조는 단백질 구조의 핵심적인 부분이기 때문에 많은 연구자들이 이용하고 있다.

단백질 구조에 대한 표현 방법으로 2차구조 구성요소를 이용하는 PSA(Protein Structure Abstraction)[2]가 제안되었다. PSA로 정의되는 단백질 구조 데이터는 PSAML(PSA Markup Language)[2]로 표현되어 XML 형태로 저장된다. PSAML은 XML 스키마를 이용하여 XML 기반 언어의 요소를 정의하고, PDB나 다른 단백질 관련 XML 데이터 형식을 이용하는 것보다 간결하면서 구조적으로 단백질 구조 정보를 표현할 수 있다.

본 논문에서는 PDB에서 제공하는 데이터 형식을 PSAML 형식으로 변환하는 변환도구를 기술한다. 본 논문에서 구현한 변환도구를 이용하여 PDB로부터 원하는 단백질의 정보를 쉽게 PSAML로 변환할 수 있으므로, PSA에서 제공하는 단백질 구조에 관한 정보를 이용하여 단백질 구조를 비교하는 여러 형태의 시스템을 개발할 수 있다. 변환도구는 XML 문서 객체 모델(DOM)을 이용하여 구현되었다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 단백질 구조를 표현하는 형식인 PDB와 PSA에 대해 설명하고

† 본 연구는 한국과학재단 목적기초연구(R01-2001-000-00535-0) 지원으로 수행되었음.

PSA의 XML 표현법인 PSAML에 대해 설명하고자 한다. 3장에서는 mmCIF 형식을 PSAML 형식으로 변환하는 도구의 설계 및 구현과정, 그리고 그 결과를 살펴본다. 마지막으로 4장에서는 결론 및 향후 연구 방향으로 끝을 맺는다.

2. 관련연구

지난 수년 동안 유전자 발현과 주석처리와 같은 특정 생물정보학 분야의 데이터 표현을 위한 다양한 XML 기반의 데이터 형식이 개발되었다[3]. 특히, 단백질과 관련한 XML 표현법도 다수 개발되었으며, 이들은 단백질 구조를 단일 표준 데이터로 표현할 수 있도록 지원한다.

2.1 PDB

PDB에서 제공하는 데이터 형식은 현재 가장 널리 알려진 것으로써 단백질 구조를 공개 데이터베이스에 등록하거나, 단백질 3차구조 뷰어인 RasMOL, Jmol 등과 같은 단백질 구조에 연관된 다양한 도구들 사이의 정보 교환을 위해서 많이 이용되고 있다. 그러나, PDB 파일에 저장된 자료들은 텍스트 방식으로 저장되어 자료의 모호성이나 불일치성이 발생할 가능성이 있다. 이에 따라 PDB 데이터의 무결성과 일관성을 높이기 위한 노력들이 있어 왔고, 그 결과 중의 하나로서 PDB에서는 단백질의 결정학 정보를 포함하는 mmCIF 형식[4]의 데이터를 제공하고 있다. 그러나 mmCIF는 STAR라고 하는 구조화되지 않은 형식을 사용하고 있으며 mmCIF와 관련한 도구가 널리 제공되지 못하고 있는 실정이다.

DSSP(Dictionary of Protein Secondary Structure)[5]는 Kabsch와 Sander에 의해 제안된 방법으로 PDB 데이터로부터 2차구조에 관련된 정보를 산출해 낼 수 있다. DSSP 데이터베이스에는 PDB에 있는 모든 단백질 항목에 대한 2차구조가 정의되어 있다. 또한 PDB 데이터 파일로부터 DSSP 파일을 생성하는 프로그램을 제공하고 있는데, PDB 데이터로부터 제공되는 단백질 원자의 좌표들로부터 기하학적 특징과 2차구조를 정의한다. DSSP에서는 2차구조의 표현을 위하여 다음과 같은 코드를 사용한다.

코드	의미
H	alpha helix
G	3-helix (3/10 helix)
I	5 helix (pi helix)
E	extended strand
B	residue in isolated beta-bridge
T	hydrogen bonded turn
S	bend

2.2 PSA

PSA는 단백질 구조를 구성하는 2차구조와 그들 사이의 관계를 이용하여 단백질 구조를 추상화하여 표현할 수 있는 방법을 제공한다.

한 단백질 구조를 표현하기 위해서, PSA는 구조를 결정하고 있는 2차구조에 대한 공간적인 정보를 표현한다. PSA에서 표현하는 단백질 3차원 구조의 표현은 공간상에 위치한 2차구조(나선; helix, 판상조각; strand)를 벡터로 표현한다. 즉, 한 벡터는 3차원 공간상의 시작점과 끝점에 대한 정보 및 길이에 대한 정보로 표현된다. 그리고 다른 단백질과 비교하여 유사한 부분 구조를 찾기 위하여, 한 단백질 구조에 속하는 임의의 두 2차구조 쌍에 대한 각도, 거리, 길이, 그리고 수소 결합 및 방향성 등의 관계를 표현하고 있다.

하나의 단백질 P에 대하여, 추상화된 표현은 다음과 같이 기술될 수 있다.

$$PSA(P) = (S, T, C, A, R)$$

S는 단백질을 구성하는 2차구조의 집합을 나타낸다. T, C, A는 각각 2차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보를 나타낸다. R은 두 2차구조사이에 정의되는 관계로서 다음과 같이 표현된다.

$$R = (\Theta, \aleph, \nu, h, d), \text{단, } Ei, Ej \in S,\ i \neq j.$$

여기에서 각 구성요소의 의미와 표현은 <표 1>과 같다.

<표 1> 2차구조 사이의 관계

관계	의미	표 현
Θ	각도	$\Theta(Ei, Ej) = angle(\Theta)$
$\aleph$	거리	$\aleph(Ei, Ej) = distance(D)$
ν	길이차	$\nu(Ei, Ej) = length(li, lj)$
h	수소결합	$h(Ei, Ej)$ = {E, N}, Ei와 Ej는 판상조각
d	방향성	$d(Ei, Ej)$ = {P, A}, Ei와 Ej는 판상조각

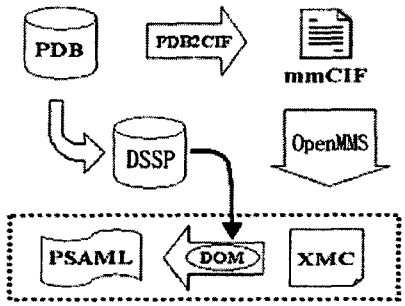
PSAML은 단백질 구조를 표현을 위한 PSA 표현을 XML로 표현하기 위하여 XML 스키마(XML schema)를 이용하여 XML로 기술할 수 있는 언어이다. PSAML 문서는 식별(Identity) 부분과 데이터(Data) 부분으로 구성된다. 식별 부분은 단백질의 주석을 나타내고 있으며, 데이터 부분은 단백질을 구성하고 있는 구성요소에 대한 기술과 더불어 그들 사이의 관계를 나타내고 있다.

데이터 부분은 <SSE>과 <R>의 두 요소(elements)를 가지고 있다. <SSE> 요소는 단백질을 구성하고 있는 모든 2차구조 요소의 각각을 기술하며 2차구조를 형성하고 있는 아마노산의 서열에 대한 정보와 3차원적인 공간정보를 포함한다. <R> 요소는 단백질을 구성하고 있는 모든 구성요소의 각각의 쌍에 대하여 각도, 거리, 방향성과 같은 관계들을 표현한다.

3. 변환도구의 설계 및 구현

PDB는 X-선 회절과 NMR 기술로부터 얻어진 생물체의 고분자 구조에 대한 데이터를 제공하고 있으며, 이러한 데이터는 XML 형태의 문서를 작성하는데 기본적인 데이터를

제공한다. 그러나, 현재의 PDB 데이터 형태는 줄과 필드에 대한 제한이나 과다한 REMARK 필드를 가지고 있는 등 기계적으로 파싱하는데 어려움을 야기하는 많은 제한을 가지고 있다. 본 논문에서는 PDB 데이터에서 시스템적인 방법으로 XML 문서를 만들기 위하여 mmCIF와 관련한 도구를 사용하여 변환 도구를 작성하였다. PDB 데이터를 PSAML 형태의 문서로 변환하는 전반적인 단계는 (그림 1)과 같다. 그림에서 점선으로 표시된 부분이 본 논문에서 구현한 부분이다.



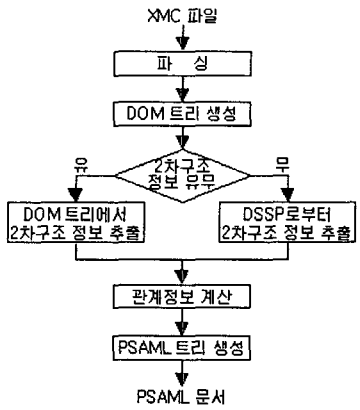
(그림 1) PSAML 문서로의 변환 과정

mmCIF 데이터는 고분자 화합물에 대한 구조 데이터를 표현하는데 있어서 결정학 정보를 가지고 있다. mmCIF 데이터 형태에서 데이터 영역에 기술되는 각 데이터 아이템은 유일한 데이터 이름으로 대응되는데, mmCIF 데이터 이름은 mmCIF 사전에 나열되고 정의된다. 그리고, PDB 데이터 파일을 mmCIF 데이터 파일로 변환하는 프로그램[6]들이 있다.

OpenMMS 툴킷[7]은 mmCIF 형태의 파일에 기술된 단백질과 핵산으로 기술되는 고분자 화합물에 대한 데이터를 분석할 수 있는 프로그램들을 제공하고 있다. 이 툴킷은 또한 mmCIF 데이터 파일을 읽어들여 같은 형태의 관계형 데이터베이스 및 XML 형태의 파일로 변환하는 기능을 제공하고 있다. 또한 코바(CORBA) 서버에 연결된 응용프로그램에게 이진 형태의 MMS 데이터를 직접적으로 전달할 수 있는 기능을 제공하고 있다. mmCIF 형태의 파일을 다른 형태로 변환은 mmCIF 사전에 기술된 용어를 기준으로 작성된 중앙 집중적인 메타모델을 이용한다. PSAML의 문서를 생성하는 과정에서 XMC 파일 형태는 OpenMMS을 이용하여 생성된 것이다.

DOM(Document Object Model)은 응용 프로그램과 스크립트에서 문서의 내용, 구조, 스타일 등을 동적으로 접근하거나 변경할 수 있는 플랫폼-독립적인 기능과 언어-중립적인 인터페이스를 제공한다.

변환도구의 구체적인 동작은 (그림 2)와 같다.



(그림 2) 변환도구의 동작

먼저 DOM 파서를 이용하여 XMC 파일을 파싱한 후 DOM 트리를 생성한다. 그 다음 단계로 생성된 DOM 트리를 재귀적인 방법으로 각 노드를 탐색하면서 원하는 정보가 있는 노드의 텍스트 데이터를 배열과 변수에 저장한다. 2차구조의 정보와 2차구조 사이의 관계 정보는 DOM 트리와 DSSP 파일의 정보를 바탕으로 계산된다. DOM 트리는 mmCIF 정보를 그대로 가지고 있으므로 <표 2>에서와 같은 규칙에 의하여 정보를 추출한다.

<표 2> mmCIF와 PSAML의 관련 태그

mmCIF	PSAML
_entry.id	Identity
_entity_poly_seq.mon_id	AA
_atom_site.label_seq_id _atom_site.label_atom_id _atom_site.Cartn_x _atom_site.Cartn_y _atom_site.Cartn_z	X, Y, Z Angle Distance Length
_struct_conf.beg_label_seq_id _struct_conf.end_label_seq_id	Len (Alpha)
_struct_conf.id _struct_sheet_range.id	S1, S2, ID
_struct_sheet.nember_strands _struct_sheet_hbond.range_1_beg_label_seq_id _struct_sheet_order.range_id_1 _struct_sheet_order.sense	Hydro Direction
_struct_sheet_range.beg_label_seq_id _struct_sheet_range.end_label_seq_id	Len (Beta)

만약 XMC 파일에 PSAML를 생성하는데 필요한 정보, 즉 단백질의 2차구조에 관련된 태그가 존재하지 않는 경우도 있는데 그럴 경우에는 DSSP 파일에서 2차구조 정보를 추출하게 된다. <표 3>은 DSSP 파일과 PSAML의 관련 부분을 나타내고 있다.

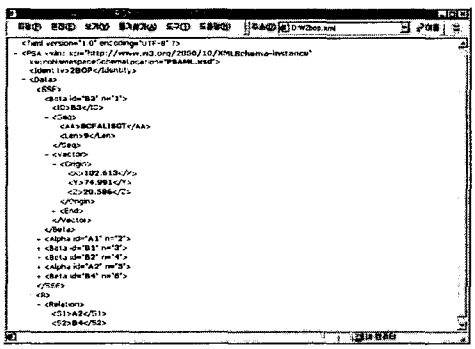
<표 3> DSSP와 PSAML의 관련 부분

DSSP	PSAML
RESIDUE	Len
AA	AA
STRUCTURE	S1, S2, ID
BP1, BP2	Hydro Direction
X-CA Y-CA Z-CA	X, Y, Z Angle Distance Length

얻어진 정보를 이용하여 필요한 데이터를 산출하고 새로 생성된 PSA 트리에 각 노드 및 텍스트 데이터를 추가한다. 마지막으로 PSA 트리를 재귀적으로 탐색하면서 각 노드와 데이터를 출력하게 된다.

변환도구는 JAVA로 구현되었으며, XMC 파일을 파싱하는데 있어서 Apache XML 프로젝트에서 제공하는 XML 파서 Xerces를 이용하였다.

생성된 PSAML 문서는 XML 형태의 문서이므로 Internet Explorer 5.5 등과 같은 도구에서 트리 형태로 쉽게 확인할 수 있다. (그림 3)은 생성된 PSAML 문서의 예이다.



(그림 3) PSAML 문서의 예

4. 결론

본 논문에서는 PDB에서 제공되는 mmCIF 형식의 파일을 PSAML 형식으로 변환시키는 변환도구를 설계하고 구현하였다. 변환도구를 이용하여 PDB로부터 원하는 단백질의 정보를 쉽게 PSAML로 변환할 수 있으므로, PSA에서 제공하는 단백질 구조에 관한 정보를 이용하여 단백질 구조를 비교하는 여러 형태의 시스템을 개발할 수 있다.

앞으로 생성된 PSAML 파일을 이용하여 단백질 구조비교와 유사도 측정을 위한 시스템을 개발할 예정이다. 또한 PSAML 형태로 표현된 단백질 구조를 논리적 표현으로 변환하는 방법과 제한 프로그래밍 기법을 이용하는 방법을 개발할 예정이다.

참고문헌

[1] H. M. Berman, J. D. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The Protein Data Bank," Nucleic Acid Research, Vol. 28, No. 1, pp. 235-242, 2000.

[2] Su-Hyun Lee, Jin-Hong Kim, Geon-Tae Ahn, and Myung-Joon Lee, "An XML Representation of Protein Data for Efficient Structure Comparison," Prec. of International Conference on Computer and Information Science, pp. 313-319, 2002.

[3] V. Guerrini and D. Jackson, "Bioinformatics and Extended Markup Language (XML)," Online Journal of Bioinformatics, Vol. 1, No. 1, pp. 12-21, 2000.

[4] P. Bourne, H. Berman, B. McMahon, K. Watenpaugh, J. Westbrook, and P. Fitzgerald, "The Macromolecular Crystallographic Information File (mmCIF)," Methods in Enzymology, Vol. 277, pp. 571-590, 1997, (http://www.sdsc.edu/pb/cif/papers/methenz.html).

[5] W. Kabsch and C. Sander, "Dictionary of Protein Secondary Stucture: Pattern Recognition of Hydrogen-Bonded and Geometrical Features," Biopolymers, Vol. 22, pp. 2577-2637, 1983.

[6] H. Bernstein, F. Bernstein, and P. Bourne, "pdb2cif: Translating PDB Entries into mmCIF Format," J. Appl. Cryst., Vol. 31, pp. 282-295, 1998.

[7] D. S. Greer, J. D. Westbrook, and P. E. Bourne, "OpenMMS: An Ontology Driven Architecture for Macromolecular Structure," Objects in Bio and Chem-informatics, 2001.