구조비교를 위한 단백질 데이터의 XML 표현기법

(An XML Representation of Protein Data for Structure Comparison)

김진홍*, 안건태*, 이수현**, 이명준*

*울산대학교 컴퓨터정보통신공학부
**창원대학교 컴퓨터공학과

*{avenue, java2u, mjlee}@mail.ulsan.ac.kr, **suhyun@sarim.changwon.ac.kr

요 약

현재의 단백질 구조비교 시스템들은 구조비교를 위하여 자신들의 알고리즘에 맞는 고유한 형식의 데이터를 이용하고 있으며, 이에 따라 호환성이나 상호작용성에 문제를 드러내게 되었다. 따라서, 이러한 문제를 해결하여 단백질 구조를 비교하는 시스템을 신속히 개발하기 위해서는 단백질 3차 구조를 표현하기 위한 데이터를 추출하여 XML과 같은 표준 형식으로 기술된 데이터를 제공하는 것이 바람직하다.

본 논문에서는 단백질 구조 및 유사성을 비교하기 위한 단백질 데이터의 XML 표현기법인 PSAML에 대하여 기술한다. PSAML은 단백질의 이차구조 구성요소와 그들 사이의 관계를 이용하여 단백질 구조를 기술하는 PSA라는 단백질 구조 모델을 이용하여 설계되었다.

1. 서론

최근 분자 생물학 기술의 발달과 인간유전체사업(Human Genome Project)의 연구를 통해 대량의 생물분자 정보 및 새로운 형태의 생물학 정보들이 산출되고 있다. 현재 단백질 서열, 구조, 패밀리에 관련된 몇몇의 데이터베이스가 공개되어 있지만 이들 대부분은 정보의 표현과 교환을 위하여 그들 고유의 데이터 형식을 정의하여 사용하고 있는 실정이다. 현재 단백질 구조 정보의 교환을 위하여 사용되는 대부분의 형식은 PDB[1]에서 보는 바와 같이 단순 텍스트 기반이고 정형화된 문법 명세가 부족하여 파싱(parsing) 에러 가능성을 내포하고 있다. 따라서, 이러한

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

문제를 해결하기 위하여 단백질 관련 정보를 표현하고 교환하기 위한 보다 효과적인 접근 방법이 요구되며, XML(eXtensible Markup Language)[2]은 이러한 문제를 해결하기 위한 이상적인 해결책을 제시해 준다. 그러나, CML이나 BIOML 등과 같은 현존하는 XML 기반의 언어들은 대부분 일반적인 목적을 위하여 고안되어서 구조비교나 예측과 같은 특수한 목적으로 사용하기를 원하는 사용자들의 요구를 충족시키지 못한다.

본 논문에서는 단백질 구조에 대한 표현 방법으로 이차구조 구성요소(secondary structure element)를 이용하는 PSA(Protein Structure Abstraction)에 대하여 소개한다. PSA로 정의되는 단백질 구조 데이터는 PSAML(PSA Markup Language)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 논문은 다음과 같이 구성되어 있다. 2장에서는 기존의 XML을 이용한 단백질 구조 표현과 단백질 구조를 비교하는 방법에 대한 연구들을 살펴보겠다. 3장에서는 이차구조와 그들 사이의 관계를 이용하여 단백질 구조를 표현하는 방법에 대하여 설명하고 4장에서는 기술된 단백질 추상화 표현을 기반으로 단백질 구조를 표현하는 XML 표현법(PSAML)에 대하여 설명하고자 한다. 마지막으로 5장에서는 결론 및 향후 연구 방향으로 끝을 맺고자 한다.

2. 관련연구

2.1 XML을 이용한 단백질 구조 표현

지난 수년 동안 유전자 발현[4]과 주석처리[5]와 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.

Protein Data Bank (PDB)

PDB[1]에서 제공하는 데이터 형식은 현재 가장 널리 알려진 것으로써 단백질 구조를 공개 데이터베이스에 등록하거나, 단백질 3차구조 뷰어인 RasMOL[7] 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 결과중의 하나로서 PDB에서는 단백질의 결정학(Crystallography) 정보를 포함하는 mmCIF 형식[8]의 데이터를 제공하고 있다. 그러나 mmCIF는 STAR라고 하는 구조화되지 않은 형식을 사용하고 있으며 mmCIF와 관련한 도구가 널리 제공되지 못하고 있는 실정이다.

Biopolymer Markup Language (BIOML)

BIOML[9]은 생물고분자 물질의 서열정보에 대한 주석처리를 위하여 고안된 언어이

다. BIOML은 단백질이나 유전자 같은 생물 고분자들로 구성된 알려진 생화합물에 대한 모든 실험정보에 대하여 전체 명세가 가능하도록 지원한다. BIOML의 궁극적인 목표는 생물분자에 대한 효과적인 주석처리를 위한 확장 가능한 프레임워크를 지원하는 것이고, 또한, 웹을 사용하는 과학자들 사이에 이러한 정보를 교환할 수 있는 공통의 방법을 제공하는 것이다. BIOML은 단백질 관련 데이터나 구조 정보만을 표현하기 위해 설계된 것은 아니며 BIOML 명세에서 제공하는 유연성은 다른 많은 유형의 정보를 표현하는 데도 효과적으로 이용될 수 있다. 따라서 문서 구조를 표현하기에 충분한 태그들을 제공하고 있지만, 고도의 유연성을 지원하기 위하여 지나치게 복잡한 데이터를 표현하고 있으며, 비 구조적인 문서를 만들어 내는 단점을 내재하고 있다.

Protein Markup Language (ProML)

ProML[10]은 단백질 서열, 구조, 패밀리(families)등에 관한 명세언어이다. 이 언어는 단백질 필수 정보를 표현하는데 있어서 이식성이 강하고 시스템 독립적이며, 기계적 파싱이 가능하고 가독성이 높은 특징을 가진다. ProML은 단백질들을 패밀리별로 그룹화하고 각각의 패밀리들이 내포한 공통적인 성질들 표현하는데 성공적으로 적용되어 왔다. 특히, ProML은 쓰레딩(threading)이나 그룹화에 사용되는 단백질의 속성들을 잘 표현해 준다. 이들 속성들에는 아미노산 서열, PROSITE 패턴[11], 이차구조 구성요소(나선, 판상조각, 루프), 삼차구조 데이터(3차원 좌표), 이황화 결합 정보 등이 포함된다.

Chemical Markup Language (CML)

CML[12]은 XML과 JAVA 기술을 이용하여 분자구조를 기술하는 언어이다. CML은 고분자 서열에서부터 무기화합물 및 양자화학의 연구에 이르기까지 광범위하게 이용되고 있다. CML 언어는 분자관련 문서들에 포함된 많은 이산 객체 정보를 완벽하게 처리하고 단백질 명세를 확장하기 위한 이상적인 기초를 제공해 준다. CML 파일에는 화학적 MIME 타입과 같은 특정 파일들이 포함될 수 있다. 따라서, 단백질에 대한 하나의 CML 파일은 하이퍼텍스트와 함께 PDB와 SWISS-PROT 파일 등을 포함할 수 있다. CML은 단백질 분자의 물리화학적인 구조를 표현할 수 있도록 지원한다는 이점이 있지만, 주석처리나 서열관련 데이터 및 SCOP[13]와 같은 구조적 분류 데이터를 표현하는데 단점을 지닌다.

2.2 단백질 구조비교

단백질 구조비교 분야에서는 단백질 모티프(motif)나 폴드 패밀리(fold family)[14]정보의 구별을 통한 비교 기법의 중요성이 점점 증대되고 있다. 구조비교 프로그램의 가장 기본적인 목표 중 하나는 알려진 단백질 쌍들에 대한 구조적인 유사도를 정량적으로 측정하는 것이다. 또한, 구조비교 프로그램은 단백질 구조의 본질이나 기능적인 메카니즘[15]에 대한 직관적인 의미를 제공한다.

단백질 구조를 비교하고 그들 사이의 유사도를 측정하기 위한 몇 가지 기법들이 있는데, 이 들은 단백질 구조[16]를 표현하는 방법에 따라, 단백질 구조간의 유사도를 계산하는 방법이 각각 다르다. 단백질 구조를 표

현하는 가장 일반적인 방법은 단백질 구조를 기본 유닛(원자, 잔기, 이차구조)으로 구분하고 유닛들을 분리하여 기술하고 그들 사이의 관계를 정의하는 것이다.

다음은 단백질 구조를 표현하고 그들 사이의 유사도를 측정하는 기존의 방법에 대하여 간략하게 기술한다.

DALI[17]는 단백질 구조를 내부 분자들 사이의 Cα-Cα 거리 매트릭스로 표현한다. DALI는 스코어링 함수(scoring function)에 의하여 계산된 각각의 거리 매트릭스로부터 유사 패턴에 대한 최적화 정렬을 한다.

LOCK[14] 알고리즘은 Cα 원자들 사이에 RMSD가 최소가 되는 점을 찾음으로서 두 구조의 최적의 겹침 포인터를 찾는 방법이다. LOCK은 정확하게 정렬된 잔기를 선택하기 위하여 재귀호출 기법에 의하여 일치하는 잔기의 쌍을 선택하게 되고 이들 사이의 RMSD를 최소화시킨다.

3dSEARCH[15] 알고리즘은 단백질의 구조를 이차구조 구성요소만으로 표현한 기법이다. 따라서, 계산속도는 빠르지만 이차 구조에 기반한 근사 정렬을 수행한다는 특징이 있다. 이 알고리즘은 컴퓨터 비전 분야에서 개발된 기하학적 해싱(geometric hashing) 기법을 기초로 하고 있다.

SARF2 알고리즘[18]은 단백질 구조를 단지 단백질 이차구조 구성요소로만 표현한다. 이 알고리즘은 단백질 이차구조를 전체 원자나 잔기의 연산대신 벡터로 표현한다. SARF2는 단백질 구조들 사이의 비교 가능한 쌍을 찾은 후, RMSD가 최소를 이루는 두 단백질 구조에 대한 이차구조 구성요소 집합들을 구한다.

위에 기술한 단백질 구조비교 알고리즘들은 단백질 구조비교를 위하여 PDB 데이터를 복잡한 처리과정을 통하여 재가공하여 만든 새로운 데이터를 이용한다. 따라서, 만약 표준 기술을 통하여 독립적으로 생성된 어떤 단백질 데이터를 이용하고자 하는 경우, 새로운 구조비교 시스템이나 기존 시스템들은 이러한 데이트를 자신들의 데이터 폼에 맞도록 변경하여야 하는 단점을 지닌다.

3. 단백질 구조의 추상화 표현

단백질 구조는 원자, 부분 구조(fragment), 또는 이차구조요소 등을 이용하여 다양한 방법으로 표현되고 있으며, 이러한 표현 방법에 따라 구조를 비교하는 방법이 달라진다. 기존의 다양한 단백질 표현 방법에 있어서, 단백질의 구조 분석 도구를 통하여 얻어진 데이터를 모두 이용하는 것보다 단백질 구조의 특징을 나타내는 대표적인 정보를 이용하는 것이 단백질의 폴딩과 구조를 이해하고 분석하는데 효과적인 도움을 줄 수 있다. 단백질 구조의 특징을 잘 표현하고, 추상화할 수 있는 방법은 단백질을 이루는 이차구조와 이들 이차구조 사이의 구조적인 특징을 반영하는 상대적인 관계로써 단백질 구조를 표현하는 것이다. 단백질 구조의 이차구조와 그들 사이에서 발견되는 상호적인 관계에 대한 정보는 생물학적 고분자화합물을 분석하여 3차원적 구조 정보를 제공하는 PDB 데이터를 바탕으로 만들어 낼 수 있다.

이 장에서는 단백질 구조를 구성하는 이차구조와 그들 사이의 관계를 이용하여 단백질 구조를 추상화하여 표현할 수 있는 PSA(Protein Structure Abstraction)에 대하여 기술한다.

PSA는 단백질의 구조를 이차구조와 그 들 사이의 관계를 이용하여 표현하며, 기술된 구조표현을 기반으로 두 단백질 사이의 유사한 부분 구조를 찾기 위한 공간적인 정보 및 생물학적인 정보를 가지고 있다.

한 단백질 구조를 표현하기 위해서, PSA는 구조를 결정하고 있는 이차구조에 대한 공간적인 정보를 표현한다. PSA에서 표현하는 단백질 3차원 구조의 표현은 공간상에 위치한 이차구조(나선, 판상조각)를 벡터(vector)로 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 이차구조쌍에 대한 각도, 거리, 길이, 그리고 수소 결합 및 방향성 등의 관계를 표현하고 있다.

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

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

(1) S, 단백질을 구성하고 있는 이차구조의 집합은 다음과 같이 정의된다.

$S = \{E_1, E_2, ..., E_k\}$, 단, k는 이차구조요소의 수

S는 하나의 단백질을 구성하고 있는 이차구조요소의 집합이다. S의 요소는 PDB 데이터에서 기술하고 있는 아미노산 서열 순서에 따라 이차구조의 인덱스가 결정된다. 이차구조요소는 3차원 공간에 위치하는 벡터로 대응된다.

(2) T, 이차구조의 종류에 대한 정보는 다음과 같이 정의된다.

$$T(E_i) = \begin{cases} \alpha, & E_i\text{가 } \alpha\text{-나선일 경우,} \\ \beta, & E_i\text{가 } \beta\text{-판상조각일 경우} \end{cases}$$

단, $E_i \in S$

T는 이차구조의 종류에 대한 정보를 가지고 있다. 이차구조의 종류에는 α-나선(α-helix) 또는 β-판상조각(β-strand)이 있다. 이러한 이차구조에 대한 정보는 PDB 데이터를 기반으로 추출할 수 있다.

(3) C, 삼차원 위치 정보는 다음과 같이 정의된다.

$C(E_i) = (o, e)$, 단, $E_i \in S$, o와 e는 E_i의 시작잔기와 끝잔기의 좌표값

C는 이차구조를 삼차원 공간에 위치한 벡터에 대한 정보를 나타내고 있다. 하나의 이차구조를 3차원 공간에 벡터로 표현할 때, 벡터에 대한 정보는 시작점과 끝점에 대한 좌표값이다. 이러한 이차구조를 벡터로 표현하기 위한 정보는 PDB 데이터를 분석함으로써 얻어질 수 있다.

(4) A, 아미노산 서열에 대한 정보는 다음과 같이 정의된다.

$A(E_i) = (AA, l)$, 단, $E_i \in S$, AA는 아미노산 서열, l은 양의 정수

A는 이차구조를 구성하고 있는 아미노산 서열(AA)과 아미노산 서열의 길이(l) 정보를 가지고 있다.

(5) R, 두 이차구조사이에 정의되는 관계는 다음과 같이 표현된다.

$R = (\Theta, \gamma, \nu, h, d)$, 단, $E_i, E_j \in S, i \neq j$

R은 두 단백질 구조를 비교할 때 사용될 수 있는 관계를 표현하고 있으며, 각각의 R은 다음과 같이 기술된 요소들로 정의된다:

(6) Θ, 두 이차구조인 E_i와 E_j 사이의 각도 관계는 다음과 같이 기술된다.

$\Theta(E_i, E_j) = angle(\Theta 1, \Theta 2, \Theta 3, \Theta 4)$

Θ는 두 이차구조사이에 다음과 같은 네 가지의 각도를 나타내고 있다. Θ1과 Θ2는 두 이차구조(Ei와 E_j)에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 각도로서, 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, Θ1과 Θ2는 각각은 E_i와 L사이의 각도와 E_j와 L사이의 각도를 말한다.

그리고, E_i의 끝점을 시작점으로 하고, E_j의 시작점을 끝점으로 하는 벡터를 V라고 할 때, Θ3은 E_i와 V가 이루는 각도이며, Θ4는 E_j와 V가 이루는 각도이다.

(7) γ, 두 이차구조인 E_i와 E_j의 거리 관계는 다음과 같이 정의된다.

$\gamma(E_i, E_j) = distance(Dmid, Dmaxi, Dmini, Dmaxj, Dminj)$

γ은 두 이차구조인 E_i와 E_j 사이의 상대적인 거리에 대한 관계로써 다섯 가지의 값을 가진다. $Dmid$는 3차원 공간에서 두 이차구조의 중점들간의 거리를 기술하고 있다. 반면에, 나머지 거리관계는 두 이차구조에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 거리관계이다. 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, $Dmaxi$, $Dmini$은 각각 E_i와 L사이의 최대거리 및 최소거리 값을 가지고, $Dmaxj$, $Dminj$은 각각 E_j와 L사이의 최대거리 및 최소거리 값을 가진다.

(8) ν, 두 이차구조인 E_i와 E_j의 각각의 길이는 다음과 같이 정의된다.

$\nu(E_i, E_j) = length(l_i, l_j)$

ν는 각 이차구조 E_i와 E_j의 길이를 나타낸다. 이차구조는 공간상의 벡터로 표현되기 때문에 이차구조의 공간상의 길이는 쉽게 계산될 수 있다. 이러한 각각의 이차구조에 대한 길이에 대한 정보는 단백질 구조를 비교하는 방법에 적용될 수 있다.

(9) h, 두 이차구조인 E_i와 E_j 사이에 수소 결합의 유무는 다음과 같이 정의된다

$$h(E_i, E_j) = \begin{cases} \text{'E', } E_i\text{와 } E_j \text{ 사이에 수소결합이 있는 경우,} \\ \text{'N', 그렇지 않은 경우} \end{cases}$$

단, E_i와 E_j는 β-판상조각

h는 두 이차구조인 E_i와 E_j 사이의 수소결합의 유무를 나타내고 있다. 이때 수소결합은 β-판상조각인 이차구조사이에 정의된다. h는 β-판상조각인 이차구조요소인 E_i와 E_j 사이에 수소 결합이 있는 경우에는 'E', 없는 경우에는 'N' 값을 가진다. 이러한 이차구조 사이의 수소 결합 관계에 대한 정보는 PDB 데이터에서 정의하는 β-병풍 구조를 나타내는 정보를 분석하여 얻어질 수 있다. 즉, β-병풍 구조를 이루고 있는 β-판상조각들 사이에는 수소결합이 있기 때문이다.

(10) d, 두 이차구조요소인 E_i와 E_j 사이에 나

타나는 방향성, 은 다음과 같이 정의된다.

$$d(E_i, E_j) = \begin{cases} \text{'P', } E_i\text{와 } E_j\text{의 방향성이 평행한 경우,} \\ \text{'A', } E_i\text{와 } E_j\text{의 방향성이 역방향인 경우,} \end{cases}$$

단, E_i와 E_j는 β-판상조각.

d는 두 이차구조인 E_i와 E_j 사이에 나타나는 방향성을 나타낸다. 이러한 방향성은 두 이차구조가 β-판상조각인 경우에 나타난다. d는 두 이차구조의 방향성이 같은 경우에는 'P', 다른 경우에는 'A'을 가진다.

PSA에서는 한 단백질 구조를 표현하는데 있어서 그 단백질을 구성하는 이차구조의 특징을 기술하고, 또한 그 사이의 다양한 관계를 이용하여 3차원 구조 특징을 가진 단백질 구조를 정의하고 있다. PSA의 단백질 구조 표현 기법은 기존의 이차구조 기반의 단백질 구조비교를 수행하는 SARF2[18]와 LOCK[14]의 구조 표현을 모두 내포하고 있다. 이러한 한 단백질의 구조를 표현하고 표현된 정의를 기반으로 일반적인 클러스터링 알고리즘(clustering algorithms)[19, 20]을 적용하여 다른 특징을 가진 단백질 구조비교 시스템을 개발할 수 있다.

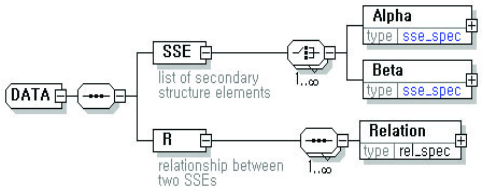
4. PSAML: PSA 마크업 언어

단백질 구조를 표현을 위한 PSA 표현을 XML로 표현하기 위하여 XML 스키마(XML schema)[6]를 이용하였다. XML 스키마는 XML 문서의 구조를 기술할 수 있는데, 스키마의 목적은 XML 문서의 블럭을 작성하는데 있어서 문법에 맞는 블럭을 정의하는데 있다.

4.1 단백질 구조를 표현하기 위한 스키마

PSAML 문서는 식별(identity) 부분과 데이터 부분으로 구성된다. 식별 부분은 단백질의 주석을 나타내고 있으며, 데이터 부분은 단백질을 구성하고 있는 구성요소에 대한 기술과 더불어 그들 사이의 관계를 나타내고 있다.

데이터 부분의 모델을 그림 1에 보였다. 이 그림은 XML SPY[21]에 의하여 생성되었는데, 그림에서, 사각형, 한쪽 둥근 사각형, 양쪽 둥근 사각형은 각각 XML의 요소(element), 데이터 타입(data type), 구성자(compositor)를 나타낸다. 구성자는 두 가지 종류가 있으며, 스위치 표시(⊕)와 직선형태 표시(⋯)는 각각 선택적 요소(choice)와 순차적 요소(sequence)를 나낸다. 구성자에 대한 범위를 지정하는 부호(1..∞)는 최소 한번 이상으로 그 하위 요소를 가질 수 있다는 것을 의미한다.



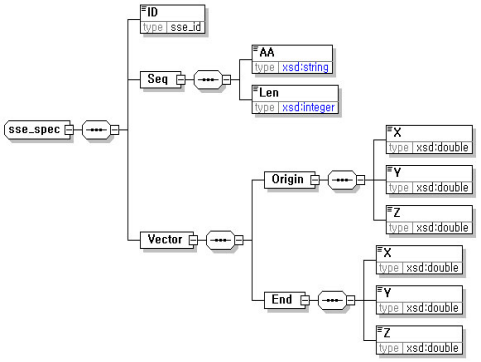
(그림 1) PSAML의 데이터 모델

데이터 부분은 <SSE>과 <R>의 두 요소(elements)를 가지고 있다. <SSE> 요소는 단백질을 구성하고 있는 모든 이차구조요소의

각각을 기술하며 이차구조를 형성하고 있는 아마노산의 서열에 대한 정보와 3차원적인 공간정보를 포함한다. <R> 요소는 단백질을 구성하고 있는 모든 구성요소의 각각의 쌍에 대하여 각도, 거리, 방향성과 같은 관계들은 표현한다.

SSE 세션 (SSE session)

SSE 세션에서는 α-나선을 정의하는 <Alpha> 요소와 β-판상조각를 정의하는 <Beta> 요소를 가진다. PSA의 타입 정보를 나타내는 *T*는 태그 이름 자체로 인코딩된다. 그림 2에서 SSE 세션에 대한 각각의 요소들의 구성 형태를 보였다.

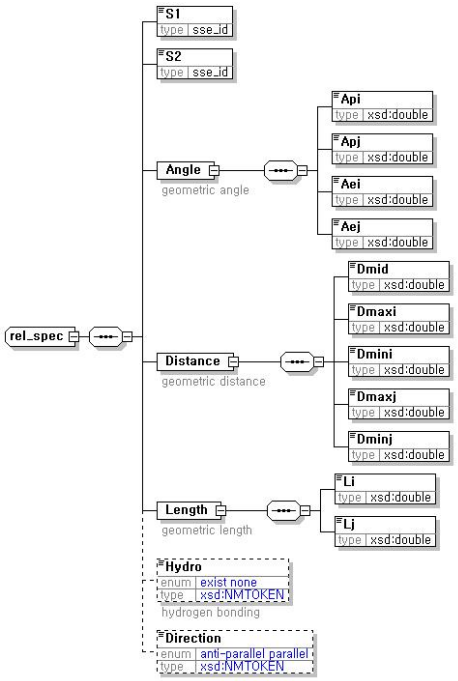


(그림 2) SSE의 구성요소

하나의 이차구조요소에 대한 식별자(identifier)는 <ID>에 기술된다. 단백질 일차구조인 아미노산 서열을 기술하는 PSA의 *A*는 <Seq> 태그에 대응되어 기술된다. <Seq> 태그의 하위에 존재하는 <AA> 태그는 연속적인 아미노산 서열에 대한 정보를 가지고 있으며, <Len> 태그는 이 서열의 길이, 즉 아미노산의 알파벳 수를 가지고 있다. PSA에서 벡터에 대한 3차원적 정보를 가지고 있는 *C*에 대한 정보는 <Vector> 태그에 기술된다. <Vector> 태그는 벡터의 시작점과 끝점에 대한 정보를 가지는 <Origin>과 <End> 태그로 구성된다.

관계 세션 (Relation session)

PSAML로 기술되는 문서는 하나의 단백질을 구성하고 있는 이차구조요소 집합에서 가능한 모든 쌍에 대한 관계에 대한 정보를 기술하고 있다. 이러한 두 이차구조요소 사이에 기술되는 관계 정보는 두 단백질을 비교하고 유사한 부분 구조를 파악하는데 사용될 수 있다. 두 이차구조 사이에 형성되는 관계는 그림 3에서 기술한 것처럼 *rel_spec* 타입을 가진 <Relation> 태그로 기술된다. 그림 3에서 <Hydro>와 <Direction>과 같이 점선으로 된 사각형은 선택적(optional) 요소를 의미한다.



(그림 3) 이차구조 요소사이의 관계 표현

하나의 관계는 <S1>와 <S2> 사이의 벡터에 대한 관계와 수소결합에 대한 관계 정보를 기술하고 있다. 여기서 수소결합에 대한 관계정보는 두 이차구조요소가 β-판상조각일 때 고려된다. <Relation>의 하위 태그들은 PSA의 *R*에 대한 모든 정보를 가지고 있다 (표 1 참조).

<표 1> PSA의 *R*과 <Relation> 대응

PSA 요소	XML 태그
Θ	<Angle>
ɣ	<Distance>
v	<Length>
h	<Hydro>
d	<Direction>

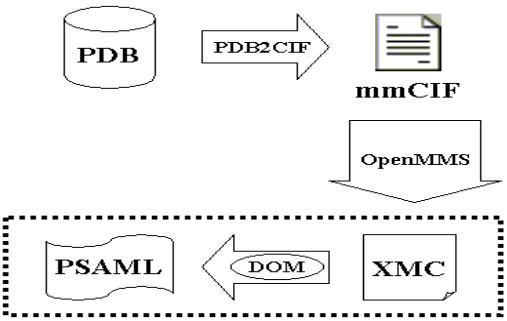
PSAML 문서에 대한 간단한 예를 그림 5에 보였다.

4.2 PSAML 문서로 변환

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.

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 프로그램들이 있다 [22].

OpenMMS 툴킷[23]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. 이러한 기능을 제공하는 OpenMMS 툴킷은 자바를 이용하여 구현되었으며 무료로 제공되고 있다.



(그림 4) PSAML 문서로 변환하는 과정

PSAML의 문서를 생성하는 과정에서 XMC 파일 형태는 OpenMMS을 이용하여 생성된

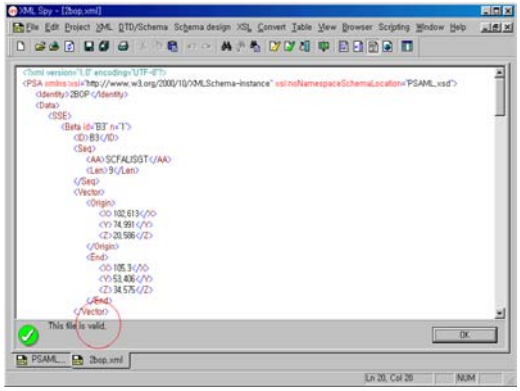
것이다. 그림 4는 PDB 데이터를 PSAML 형태의 문서로 변환하는 전반적인 단계를 보여주고 있다. DOM(Document Object Model)[24]은 응용 프로그램과 스크립트에서 문서의 내용, 구조, 스타일(style) 등을 동적으로 접근하거나 변경할 수 있는 플랫폼-독립적과 언어-중립적인 인터페이스를 제공한다. 그림 5는 생성된 PSAML 문서의 예이다.

(그림 5) PSAML 문서의 예

4.3 PSAML의 활용

PSA와 PSAML를 구현한 중요한 목적은 PDB에서 제공하는 X-선 회절과 NMR 기술을 이용하여 나온 저수준의 데이터에서 단백질 구조를 비교에 필요한 데이터를 기술하기 위해서이다. PSAML에서 기술되는 정보를 이용하여 두 단백질의 이차구조요소 사이에 다양한 상대적인 관계를 기술할 수 있으며, 또한 이러한 관계들과 현존하는 일반적인 클러스터링 알고리즘을 이용한다면, 단백질 구조를 비교하는 새롭고 다양한 시스템이 구현될 수 있다.

XML 문서 형태로 기술하는 방법의 주된 장점으로는 XML 문서를 처리할 수 있도록 제작된 기존의 프로그램을 이용할 수 있다는 점이다. 예를 들어, 작성된 문서에 대한 유효성 검사와 다양한 형태로 데이터를 표현하여 보여주는 기능은 XML SPY 등을 이용할 수 있다(그림 6참조). XML 문서에 표현된 데이터를 파싱하고 유효성을 평가하기 위한 프로그램들은 상업적인 또는 비상업적인 형태[25]로 널리 배포되고 있다. 또한 XML 문서에 기술된 요소 또는 문서 전체에 대하여 접근하여 처리할 수 있는 표준 API를 이용할 수 있다.



(그림 6) 문서의 유효성 검사

사용자는 다양한 방법으로 PSAML 문서를 작성하고 검사하고 분석할 수 있다. 예를 들어, PSAML의 규칙을 분석하여 데이터를 다루는 전용 프로그램이 개발되어 사용될 수도 있고, 마이크로소프트 익스플로러 5와 같이 XML 문서를 처리할 수 있는 브라우저에서도 PSAML 문서를 직접적으로 보여줄 수 있다. 앞으로 PSAML 문서를 효과적으로 보여줄 수 있도록 CSS(Cascading Style Sheet)를 개발할 예정이다. 개발될 CSS는 문서의 요소와 화면 표시와의 매핑을 제공하

며, 이를 통하여 PSAML 문서는 데이터를 저장하기 위한 원래의 기능과 함께 데이터를 보여주기 위한 기능을 제공할 것이다.

또한 PSAML 문서의 내용을 도식적으로 보여줄 수 있는 렌더링(rendering) 프로그램을 개발할 예정이다. 개발될 프로그램은 PSAML 문서에서 원자의 종류와 그 삼차원 위치에 대한 정보를 뽑아내어 분자의 이미지를 다양한 형태로 보여줄 수 있다.

4.4 다른 표현들과의 비교

기존의 단백질 데이터 형식과 PSAML을 비교하여 표 2에 나타내었다. BIOML과 CML은 단백질을 나타내기 위한 표기로 보기는 힘들지만, 단백질 관련 정보를 포함할 수 있기 때문에 비교대상에 넣었다. 각각 형식의 목적은 서로 다르기 때문에 본 비교표로서 상대적인 우월성을 판단하기는 힘들다. 본 논문에서 제안하는 PSAML은 구조 비교를 위한 2차구조들 사이의 관계를 기술한다는 점에서 다른 형식들에 장점이 있다. 표에서 이중원(◎)은 아주 좋음을, 원(○)은 좋음을, 삼각형(△)은 보통을, 엑스(×)는 나쁨을 나타낸다.

<표 2> 다른 형식들과의 비교

	PDB	BIOML	ProML	CML	PSAML
XML 기반	×	○	○	○	○
메타데이터	-	DTD	DTD	Schema	Schema
서열 데이터	○	×	○	△	△
주석	○	◎	○	△	×
구조 데이터	○	×	○	◎	○
구조간의 관계	×	×	×	×	○
각종 도구	많음	적음	없음	많음	없음

5. 결론

본 논문에서는 XML 표준을 통하여 단백질 구조를 표현하는 명세언어인 PSAML에 대하여 기술하였다. PSAML 언어는 PSA 단백질 데이터 모델을 기반으로 설계되었다. PSA는 현재 많은 단백질 구조비교 시스템에서 유용하게 사용되는 단백질 구조를, 이차구조와 그들 사이의 관계로 기술하는 언어이다. PSA에서 제공하는 단백질 구조에 관한 정보를 이용하여 단백질 구조를 비교하는 여러 형태의 시스템을 개발할 수 있다.

이 언어는 단백질의 여러 특성들 중에 이차구조 구성요소에 초점을 맞추고 있으므로 PSAML 문서는 다른 XML 기반 표현들보다 간결하다. PSAML 문서는 PDB 데이터로부터 자동적으로 생성되며, 변환기는 XML 문서 객체 모델을 이용하여 구현되었다. 추가로, 단백질 구조의 공간정보, 벡터로 표현된 이차구조 구성요소, 그리고 구성요소들 사이의 관계 등도 PDB 데이터로부터 추출하였다.

앞으로 단백질 구조비교와 유사도 측정을 위한 시스템을 개발할 예정이다. PSA에 기반하여 기술된 단백질 구조 정보를 이용할 경우, 두 단백질 사이의 이차구조 구성요소들 간의 다양한 호환 관계를 정의할 수 있고, 유사한 단백질 부분 구조를 다양한 클러스터링 알고리즘을 이용하여 쉽게 찾을 수 있다.

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] T. Bray, J. Paoli, C. M. Sperberg-McQueen, and E. Maler (ed.), "Extensible Markup Language (XML) 1.0 (Second Edition)," W3C, Oct. 2000.

[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] MGED group, "MicroArray and Gene Expression (MAGE)," WWW document (http://www.mged.org/Workgroups/MAGE/mage.html).

[5] BioXML, "Genome Annotation Markup Elements (GAME)," WWW document (http://www.bioxml.org/Projects/game/).

[6] D. C. Fallside (ed.), "XML Schema Part 0: Primer," W3C, May 2001, (http://www.w3.org/TR/xmlschema0).

[7] R. Sayle and E. Milner-White, "RASMOL: biomolecular graphics for all," Trends in Biochemical Science, Vol. 20, pp. 374-376, 1995.

[8] 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).

[9] Proteomics Inc., "BioML-Biological Markup Language," WWW document (http://www.bioml.com/bioml/).

[10] D. Hanisch, R. Zimmer, and T. Lengauer, "ProML - the Protein Markup Language for specification of protein sequences, structures and families," German Conference on Bioinformatics' 01, Oct. 2001.

[11] K. Hoffman, P. Bucher, L. Falquet, and A. Bairoch, "The PROSITE database, its status in 1999," Nucleic Acids Research, Vol. 27, pp. 215-219, 1999.

[12] 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.

[13] A. Murzin, S. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequences and structures," Journal of Molecular Biology, Vol. 247, pp. 536-540, 1995.

[14] A. P. Singh and D. L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations," Proc. Intelligent Systems for Molecular Biology 97, 1997.

[15] A. P. Singh and D. L. Brutlag, Protein Structure Alignment: A Comparison of Methods, 1999.

[16] I. Eidhammer, I. Jonassen, and W. R. Taylor, "Structure Comparison and Structure Patterns," Report no 174, University of Bergen, 1999.

[17] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," Journal of Molecular Biology, Vol. 233, pp. 123-138, 1993.

[18] N. N. Alexandrov and D. Fischer, "Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures," Proteins: Structure, Function, and Genetics, Vol. 25, No. 3, pp. 354-365, 1996.

[19] D. Fischer, C.J. Tsai, R. Nussinov, and H. Wolfson. "A 3D sequence-independent representation of the protein data bank," Protein Eng., Vol. 8, pp. 981-997, 1995.

[20] G. Vriend and C. Sander, "Detection of common three-dimensional substr-uctures in proteins," Proteins: Structure, Function, and Genetics, Vol. 11, pp. 52-58, 1991.

[21] Altova Inc., "XML SPY," WWW document (http://www.xmlspy.com/).

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

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

[24] W3C, "Document Object Model (DOM)," WWW document (http://www.w3.org/DOM/).

[25] Apache Project, "Apache XML Project," WWW document (http://xml.apache.org/index.html).

김진홍

1999년 2월 울산대학교 전자계산학과 졸업(학사)
2001년 2월 울산대학교 컴퓨터정보통신 공학부 졸업(석사)
2001년 3월 ~ 현재 울산대학교 컴퓨터정보통신 공학부 박사과정
관심분야 : 생물정보학, 제한프로그래밍, 협업지원 시스템, 이동에이전트 시스템 등.

안건태

1999년 2월 울산대학교 전자계산학과 졸업(학사)
2001년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(석사)
2001년 3월 ~ 현재 울산대학교 컴퓨터·정보통신공학부 공학박사과정
관심분야 : 생물정보학, 협업지원 시스템, 분산시스템, 이동에이전트 시스템 등.

이수현

1987년 2월 광운대학교 전자계산학과 졸업(학사)
1989년 2월 한국과학기술원 전산학과 졸업(석사)
1994년 8월 한국과학기술원 전산학과 졸업(박사)
1994년 9월 ~ 1996년 2월 한국전자통신연구원 선임연구원
1996년 3월 ~ 현재 창원대학교 컴퓨터공학과 부교수
관심분야 : 프로그래밍언어, 제한프로그래밍, 생명정보학 등.

이명준

1980년 2월 서울대학교 수학과 졸업(학사)
1982년 2월 한국과학기술원 전산학과 졸업(석사)
1991년 8월 한국과학기술원 전산학과 졸업(박사)
1982년 3월 ~ 현재 울산대학교 컴퓨터정보통신 공학부(교수)
1993년 8월 ~ 1994년 7월 미국 버지니아대학 교환교수
관심분야 : 프로그래밍언어, 분산 객체 프로그래밍 시스템, 병행 실시간 컴퓨팅, 인터넷 프로그래밍시스템, 생물정보학 등.