PSAML을 이용한 단백질 구조 비교 시스템
(A Protein Structure Comparison System based on PSAML)

김 진 홍 † 안 건 태 † 변 상 희 † 이 수 현 †† 이 명 준 †††
(Jin-Hong Kim) (Geon-Tae Ahn) (Sang-Hee Byun) (Su-Hyun Lee) (Myung-Joon Lee)

요 약 단백질 구조에 대한 유사성과 특이성에 대한 이해는 단백질의 기능을 파악하는데 있어 중요한 역할을 하고 있기 때문에, 많은 단백질 구조를 비교하는 시스템이 개발되고 있다. 그러나 이러한 시스템들은 단백질 구조 비교를 위한 자신의 알고리즘에 맞게 PDB에서 제공하는 데이타를 가공해야 한다. 더욱이 PDB 데이타베이스에 저장된 데이타가 증가함에 따라 대용량의 단백질 구조 데이타베이스를 대상으로 주어진 단백질과 유사한 부분구조를 찾는 시스템은 보다 많은 계산량이 필요하여진다.

본 논문에서는 XML 데이타베이스인 eXist를 이용하여 PSAML 문서를 제공하는 PSAML 데이타베이스에 기반을 둔 WS4E(A Web-Based Searching Substructures of Secondary Structure Elements) 단백질 구조 비교 시스템을 소개한다. PSAML(Protein Structure Abstraction Markup Language)은 XML 기반의 단백질 구조 표현 기법으로서 단백질의 2차구조 구성요소와 그들 사이의 관계를 이용하여 단백질 구조를 정형화된 방법으로 기술한다. 구축된 PSAML 데이타베이스를 이용하여, WS4E는 PSAML로 표현된 단백질 구조에서 유사한 부분 구조를 찾는 웹서비스를 제공한다. 또한, PSAML 데이타베이스에서 비교 대상이 되는 단백질의 숫자를 감소시키기 위하여, 단백질 2차구조가 가지는 공간상의 정보를 이용하여 하나의 단백질 구조를 표현하는 기법인 topology string을 이용하였다.

키워드 : PSAML, topology string, 단백질 구조 비교, XML, WS4E

Abstract Since understanding of similarities and differences among protein structures is very important for the study of the relationship between structure and function, many protein structure comparison systems have been developed. But, unfortunately, these systems introduce their own protein data derived from the PDB(Protein Data Bank), which are needed in their algorithms for comparing protein structures. In addition, according to the rapid increase in the size of PDB, these systems require much more computation to search for common substructures in their databases.

In this paper, we introduce a protein structure comparison system named *WS4E*(A Web-Based Searching Substructures of Secondary Structure Elements) based on a *PSAML database* which stores *PSAML* documents using the *eXist* open XML DBMS. PSAML(Protein Structure Abstraction Markup Language) is an XML representation of protein data, describing a protein structure as the secondary structures of the protein and their relationships. Using the PSAML database, the WS4E provides web services searching for common substructures among proteins represented in PSAML. In addition, to reduce the number of candidate protein structures to be compared in the PSAML database, we used topology strings which contain the spatial information of secondary structures in a protein.

Key words : PSAML, topology string, Protein Structure Comparison, XML, WS4E

· 본 연구는 한국과학재단 목적기초연구(R01-2001-000-00535-0) 지원으로 수행됨
† 학생회원 : 울산대학교 컴퓨터·정보통신공학부
avenue@ulsan.ac.kr
java2u@ulsan.ac.kr
heeya@mail.ulsan.ac.kr
†† 종신회원 : 창원대학교 컴퓨터공학과 교수
suhyun@sarim.changwon.ac.kr
††† 종신회원 : 울산대학교 컴퓨터·정보통신공학부 교수
(Corresponding author임)
mjlee@ulsan.ac.kr
논문접수 : 2004년 7월 20일
심사완료 : 2004년 12월 29일

1. 서 론

최근 분자 생물학 기술의 발달과 인간유전체사업(Human genome project)의 연구를 통해 대량의 생물분자 정보 및 새로운 형태의 생물학 정보들이 산출되고 있다. 특히 PDB(Protein Data Bank)[1]는 가장 널리 알려진 단백질 구조 데이타베이스로서 단백질을 이루고 있는 아미노산 서열 및 3차원 구조에 대한 정보를 제공하고 있으며, 이의 데이타 증가 속도는 날이 갈수록 늘

어나고 있다. 현재 PDB에서 제공하는 생물학 정보를 바탕으로 미지의 단백질 기능을 파악하려는 연구가 활발히 진행되고 있다[2].

단백질 구조 비교 도구는 단백질의 구조적인 특징에 따라 단백질 구조를 분류하는 분야와 공통의 부분 구조를 찾는 분야에 활용되어 새로운 단백질의 기능을 파악하는데 유용하게 사용되고 있다. 대표적인 단백질 구조 비교 도구에는 분자들 정보를 바탕으로 동적 프로그래밍 기법을 이용하여 유사한 부분 구조를 찾는 DALI[3], Ca 원자들 사이의 RMSD의 값이 최소가 되는 부분을 찾는 LOCK[4], 단백질 2차구조의 3차원 위치 정보를 이용한 기하학적 해싱 기법을 사용하는 3dSearch[5], 그리고 단백질 2차구조 사이의 거리 및 각도 정보를 이용한 SARF2[6] 등이 있다.

대부분의 단백질 구조 비교 도구들은 PDB 데이타베이스에서 제공하는 데이타에서 자신들의 도구에 필요한 중간 데이타를 추출하는 과정이 필요하다. 이러한 과정에서 PDB 데이타의 표현 양식이 단순 텍스트 기반이고 정형화된 문법 명세가 부족하여 파싱(parsing) 에러를 내포할 가능성이 매우 높다. 따라서 이러한 문제를 해결하기 위하여 단백질 구조 정보를 표준화된 방법을 이용하여 표현하고 단백질 구조 비교에 대한 시스템을 효과적으로 개발하기 위하여 효율적으로 구조비교에 필요한 데이타를 가공하여 제공하는 표준화된 접근 방법이 요구된다.

PDB 데이타베이스에 저장된 데이타가 증가함에 따라 대용량의 단백질 구조 데이타베이스를 대상으로 주어진 단백질과 유사한 부분구조를 찾는 시스템은 많은 계산량을 요구하고 있다. 현재 SARF2는 두 개의 단백질 간의 구조 비교를 수행하는 웹서비스만을 제공하고 있으며, DALI는 사용자로부터 단백질 정보를 입력받아 로컬 컴퓨터에서 단백질 구조 비교를 수행하여 그 결과를 사용자에게 전달하고 있다.

본 논문에서는 단백질 구조를 비교하여 유사한 부분 구조를 찾기 위하여 XML 기반의 단백질 구조 표현 기법인 PSAML(Protein Structure Abstraction Markup Language)[7]과 웹기반 단백질 구조 비교 시스템인 WS4E(A Web-Based Searching Substructures of Secondary Structure Elements)에 대하여 기술한다. WS4E는 PSAML 데이타베이스를 기반으로 PSAML로 표현된 단백질 구조에서 유사한 부분 구조를 찾는 웹서비스를 제공하며, topology string을 이용하여 PSAML 데이타베이스에서 비교 대상이 되는 단백질의 수를 감소시키는 옵션을 함께 제공한다.

PSAML은 2차구조 구성요소(secondary structure element)의 특징과 그들 사이에 정의되는 관계(각도, 거리, 길이)를 이용하여 XML 형태로 단백질 구조를 표현한다. PSAML은 개발된 번환기를 통하여 자동으로 생성되며, 단백질의 여러 특성들 중에서 2차구조 구성요소에 초점을 맞추고 있으므로 다른 XML 기반 표현들보다 간결하다. 그리고 PDB 데이타를 분석하는 과정 없이 단백질 구조를 비교하는 여러 형태의 시스템을 개발하는데 활용될 수 있다. topology string은 단백질 2차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하는 방법으로서 대용량의 단백질 구조 데이타베이스에 저장된 단백질 구조와 비교하는 과정을 보다 빠르게 수행할 수 있도록 활용될 수 있다.

웹기반 단백질 구조 시스템인 WS4E는 PSAML로 표현된 두 개의 단백질 구조에서 유사성이 높은 부분을 찾는 기본 비교 기능을 제공하며, 구축된 대용량 단백질 구조 데이타베이스인 PSAML 데이타베이스에서 단백질 구조 비교를 신속하고 효과적으로 비교할 수 있도록 비교 대상 단백질의 수를 줄일 수 있는 방법과 사용자 중심의 편리한 웹 인터페이스를 제공한다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 XML을 이용하는 단백질 구조 표현과 단백질 구조를 비교하는 방법에 대하여 살펴보겠다. 3장과 4장에서는 2차구조와 그들 사이의 관계를 이용하여 단백질 구조를 표현하는 PSAML과 topology string에 대하여 기술하고, 5장에서는 PSAML 기반의 단백질 구조 비교 시스템인 WS4E에 대하여 설명하고자 한다. 6장에서는 WS4E의 실행결과에 대하여 분석하며, 7장에서는 결론과 향후 연구 방향을 제시한다.

2. 관련연구

2.1 XML을 이용한 단백질 구조 표현

지난 수년 동안 유전자 발현[9]과 주석처리[10]와 같은 생물정보학 분야의 데이타 표현을 위한 다양한 XML 기반의 데이타 형식이 개발되었다[11]. 특히, 단백질과 관련한 XML 표현법도 다수 개발되었으며, 이들은 단백질 구조를 단일 표준 데이타로 표현할 수 있도록 지원한다.

• Protein Data Bank (PDB)

PDB에서 제공하는 데이타 형식은 현재 가장 널리 알려진 것으로서 단백질 구조를 공개 데이타베이스에 등록하거나, 단백질 3차구조 뷰어인 RasMOL[12] 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 데이타의 무결성과 일관성을 높이기 위한 노력

들이 있어 왔고, 그 결과중의 하나로서 PDB에서는 XML 문서와 단백질의 결정학(Crystallography) 정보를 포함하는 mmCIF 형식[13]의 데이타를 제공한다. 그러나 mmCIF는 STAR라는 이름을 가진 구조화되지 않은 형식을 사용하고 있으며 mmCIF와 관련한 도구가 널리 제공되지 못하고 있는 실정이다.

• Biopolymer Markup Language (BIOML)

BIOML[14]은 생물고분자 물질의 서열정보에 대한 주석처리를 위하여 고안된 언어이다. BIOML은 단백질이나 유전자 같은 생물고분자들로 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 데이타나 구조정보만을 표현하기 위해 설계된 것은 아니며 BIOML 명세에서 제공하는 유연성은 다른 많은 유형의 정보를 표현하는 데도 효과적으로 이용될 수 있다. 따라서 문서구조를 표현하기에 충분한 태그들을 제공하고 있지만, 고도의 유연성을 지원하기 위하여 지나치게 복잡한 데이타를 표현하고 있으며, 비 구조적인 문서를 만들어 내는 단점을 내재하고 있다.

• Protein Markup Language (ProML)

ProML[15]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[16], 2차구조 구성요소(나선, 판상조각, 루프), 삼차구조 데이타(3차원 좌표), 이황화 결합 정보 등이 포함된다.

• Chemical Markup Language (CML)

CML[17]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 CML파일은 하이퍼텍스트와 함께 PDB와 SWISS-PROT 파일 등을 포함할 수 있다. CML은 단백질분자의 물리화학적인 구조를 표현할 수 있도록 지원한다는 이점이 있지만, 추석처리나 서열관련 데이타 및 SCOP[18]와 같은 구조적 분류 데이타를 표현하는데 단점을 지닌다.

2.2 단백질 구조 비교 방법

단백질 구조비교 분야에서는 단백질 모티프(motif)나 폴드 패밀리(fold family)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 메카니즘에 대한 직관적인 의미를 제공한다.

단백질 구조를 비교하고 그들 사이의 유사도를 측정하기 위한 몇 가지 기법들이 있다. 단백질 구조[19]를 표현하는 방법에 따라, 단백질 구조간의 유사도를 계산하는 방법이 각각 다르다. 단백질 구조를 표현하는 가장 일반적인 방법은 단백질구조를 기본 유닛(원자, 잔기, 2차구조)으로 구분하고 유닛들을 분리하여 기술하고, 그들 사이의 관계를 정의하는 것이다.

DALI는 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 알고리즘은 Cα 원자들 사이에 RMSD가 최소가 되는 점을 찾음으로써 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 알고리즘은 단백질의 구조를 2차구조 구성요소만으로 표현한 기법이다. 따라서, 계산속도는 빠르지만 2차구조에 기반한 근사 정렬을 수행한다는 특징이 있다. 이 알고리즘은 컴퓨터 비전 분야에서 개발된 기하학적 해싱(geometric hashing) 기법을 기초로 하고 있다.

SARF2 알고리즘은 단백질 구조를 단지 단백질 2차구조 구성요소로만 표현한다. 이 알고리즘은 단백질 2차구조를 전체 원자나 잔기의 연산대신 벡터로 표현한다. SARF2는 단백질 구조들 사이의 비교 가능한 쌍을 찾은 후, RMSD가 최소를 이루는 두 단백질 구조에 대한 2차구조 구성요소 집합들을 구한다.

위에 기술한 단백질 구조비교 알고리즘들은 단백질 구조비교를 위하여 PDB 데이타를 복잡한 처리과정을 통하여 재가공하여 만든 새로운 데이타를 이용한다. 따라서 만약 표준 기술을 통하여 독립적으로 생성된 어떤 단백질 데이타를 이용하고자 하는 경우, 새로운 구조비

교 시스템이나 기존 시스템들은 이러한 데이타를 자신들의 데이타 폼에 맞도록 변경하여야 하는 단점을 지닌다.

3. PSAML을 이용한 단백질 구조 표현

단백질 구조는 원자, 부분 구조(fragment), 또는 2차 구조요소 등의 특징을 이용하여 표현되고 있으며, 이러한 표현 방법에 따라 구조를 비교하는 방법이 달라진다. PSAML(Protein Structure Abstraction Markup Language)은 단백질의 구조를 형성하는 단백질 2차구조를 기본 요소로 하여 단백질 구조의 특성을 표현할 수 있는 방법과 2차구조들 사이의 생물학적, 3차원적 구조적인 관계를 기술하는 구성요소를 제공한다.

하나의 단백질 P에 대하여, PSAML에 의하여 표현되는 구조는 2차구조를 기술하기 위한 S, T, C, A라는 구성요소와 한 단백질 구조에 속하는 임의의 두 2차구조 쌍에 대한 각도, 거리, 길이, 그리고 수소 결합 및 방향성 등의 관계를 기술하는 R 구성요소를 이용하여 기술될 수 있다.

3.1 PSAML의 구성요소

- ***S*** *구성요소*는 한 단백질의 구조를 구성하는 2차구조의 집합을 기술하고 있다. S의 요소인 E_i의 인덱스는 PDB 데이타에서 기술하고 있는 아미노산 서열 순서에 의하여 결정된다. 그리고 각각의 S의 요소는 3차원 공간 데이타를 기반으로 공간상에 위치하는 벡터로 대응된다.
- ***T*** *구성요소*는 2차구조의 종류에 대한 정보를 나타내고 있다. 2차구조의 종류에 따라 단백질을 구성하고 있는 부분적인 구조의 모양이 달라진다. 이러한 2차구조의 종류를 기술함으로써 단백질 구조를 비교할 때 부분 구조를 찾는데 유용하게 이용될 수 있다. 이러한 2차구조에 대한 정보는 PDB 데이타를 기반으로 정보를 추출할 수 있다.
- ***C*** *구성요소*는 2차구조 요소의 3차원 공간상의 위치 정보를 나타내고 있다. 이 구성요소는 2차구조를 3차원 공간에 위치한 벡터에 대한 정보를 나타내고 있다. 하나의 2차구조를 3차원 공간에 벡터로 표현할 때, 벡터에 대한 정보는 시작점과 끝점에 대한 좌표 값이다.
- ***A*** *구성요소*는 2차구조를 구성하고 있는 아미노산 서열과 길이에 대한 정보를 나타내고 있다.
- ***R*** *구성요소*는 두 단백질 구조를 비교할 때 사용되는 관계를 표현하고 있다.
- ***θ*** *구성요소*는 두 2차구조인 E_i과 E_j 사이의 각도 관계를 나타내고 있다. θ는 두 2차구조사이에 다음과 같

표 1 PSAML의 단백질 3차구조 표현 방법

$PSAML(P) = (S, T, C, A, R)$ ---------- {1}
S : 단백질 P의 구조를 구성하는 2차구조의 집합
T : 2차구조의 종류에 대한 정보
C : 2차구조 요소의 3차원 공간상의 위치 정보
A : 2차구조를 구성하고 있는 아미노산 서열과 길이에 대한 정보
R : 임의의 두 2차구조 쌍에 대한 각도, 거리 길이, 그리고 수소 결합 및 방향성 등의 관계

표 2 PSAML의 구성요소

구성요소	정 의
S	$S = \{E_1, E_2, ..., E_k\}$, 단, k는 2차구조요소의 수.
T	$T(E_i) = \alpha$, E_i가 α-나선일 경우, $= \beta$, E_i가 β-판상조각일 경우, $E_i \in S$.
C	$C(E_i) = (o, e)$, 단, $E_i \in S$, o와 e는 E_i의 시작 잔기와 끝 잔기의 좌표 값.
A	$A(E_i) = (AA, l)$, 단, $E_i \in S$, AA는 아미노산 서열, l은 양의 정수.
R	$R = (\theta, v, v, h, d)$, 단, $E_i, E_j \in S$, $i \neq j$.
θ	$\theta(E_i, E_j) = angle(\theta_1, \theta_2, \theta_3, \theta_4)$.
d	$d(E_i, E_j) = distance(D_{mid}, 2D_{maxi}, 2D_{mini}, 2D_{maxj}, 2D_{minj})$.
v	$v(E_i, E_j) = length(l_i, l_j)$.
h	$h(E_i, E_j)$ = 'E', E_i와 E_j 사이에 수소결합이 있는 경우, = 'N', 그렇지 않은 경우, 단, E_i와 E_j는 β-판상조각.
Φ	(E_i, E_j) = 'P', E_i와 E_j의 방향성이 평행한 경우, = 'A', E_i와 E_j의 방향성이 역방향인 경우, 단, E_i와 E_j는 β-판상조각.

은 네 가지의 각도를 나타내고 있다. θ_1과 θ_2는 두 2차구조(E_i과 E_j)에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 각도로서, 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, θ_1과 θ_2는 각각은 E_i,와 L사이의 각도와 E_j와 L사이의 각도를 말한다. 그리고 E_i의 끝점을 시작점으로 하고, E_j의 시작점을 끝점으로 하는 벡터를 V라고 할 때, θ_3는 E_i와 V가 이루는 각도이며, θ_4는 E_j와 V가 이루는 각도이다.

- ***d*** *구성요소*는 두 2차구조인 E_i와 E_j의 거리 관계를 나타낸다. d는 두 2차구조인 E_i와 E_j사이의 상대적인 거리에 대한 관계로써 다섯 가지의 값을 가진다. D_{mid}는 3차원 공간에서 두 2차구조의 중점들 간의 거리를 기술하고 있다. 반면에, 나머지 거리관계는 두 2차구조에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 거리관계이다. 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, $2D_{maxi}$, $2D_{mini}$은 각각 E_i와 L사이의 최대거리 및 최소거리 값을 가지고, $2D_{maxj}$, $2D_{minj}$은 각각 E_j와 L사이의 최대거리 및 최소거리 값을 가진다.
- ***v*** *구성요소*는 두 2차구조인 E_i와 E_j의 각각의 길이를 나타낸다. v는 각 2차구조 E_i와 E_j의 길이를 나타낸다. 2차구조는 공간상의 벡터로 표현됨으로써 쉽게 2차구조의 공간상의 길이는 계산되어 질 수 있다. 이러한 각각의 2차구조에 대한 길이에 대한 정보를 이용하여 단백질 구조를 비교할 때 이 관계를 이용하여 다양한 비교 방법을 만들어 낼 수 있다.
- ***h*** *구성요소*는 두 2차구조인 E_i와 E_j 사이에 수소 결합의 유무를 나타낸다. 이때 수소결합은 β-판상조각인 2차구조 사이에 정의된다. h는 β-판상조각인 2차구조 요소인 E_i와 E_j 사이에 수소 결합이 있는 경우에는 'E', 없는 경우에는 'N' 값을 가진다. 이러한 2차구조 사이의 수소 결합 관계에 대한 정보는 PDB 데이타에서 정의하는 β-병풍 구조를 나타내는 정보를 분석하여 얻어질 수 있다. 즉, β-병풍 구조를 이루고 있는 β-판상조각들 사이에는 수소결합이 있기 때문이다.
- ***ϕ*** *구성요소*는 두 2차구조요소인 E_i와 E_j 사이에 나타나는 방향성을 나타낸다. ϕ 구성요소는 두 2차구조가 β-판상조각인 경우에 나타난다. 이 구성요소는 2차구조의 방향성에 따라 그 값이 결정된다. 두 2차구조의 방향성이 같은 경우에는 'P', 다른 경우에는 'A'을 가진다.

3.2 단백질 구조를 표현하기 위한 스키마

PSAML은 단백질 구조를 표현을 위한 3.1절에서 기술한 구성요소를 XML 스키마(schema)를 이용하여 XML 문서 형태로 변환한 데이타 형식을 제공한다. 제공되는 PSAML은 단백질 구조의 표현에 대한 데이타에 대한 유효성과 구조적인 방법으로 저장할 수 있는 기능을 제공한다. PSAML은 두 단백질의 2차구조 요소 사이에 다양한 상대적인 관계를 기술할 수 있는 방법을 제공함으로써, 현존하는 일반적인 클러스터링 알고리즘을 이용하여 새롭고 다양한 단백질 구조를 비교하는 시스템이 구현될 수 있다.

PSAML 문서는 *Identity 세션*과 *Data 세션*으로 구성된다. Identity 세션은 단백질의 주석을 나타내고 있으며, data 세션은 단백질을 구성하고 있는 구성요소에 대한 기술과 더불어 그들 사이의 관계를 나타내고 있다. 그리고 Data 세션은 <SSE>와 <R>의 요소(elements)를 가진다. 그림 1은 PSAML 문서 표현에서 Data 세션의 모델을 보여주고 있다.

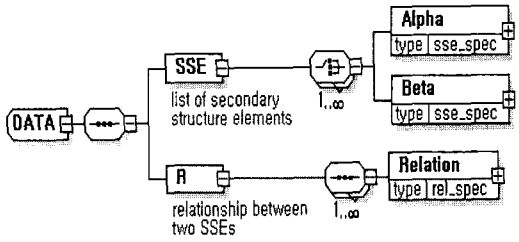


그림 1 PSAML의 Data 모델

- SSE 세션에서는 α-나선을 정의하는 <Alpha> 요소와 β-판상조각를 정의하는 <Beta> 요소를 가진다. 단백질 2차구조의 타입 정보를 나타내는 T는 태그 이름 자체로 사용되고 있다.

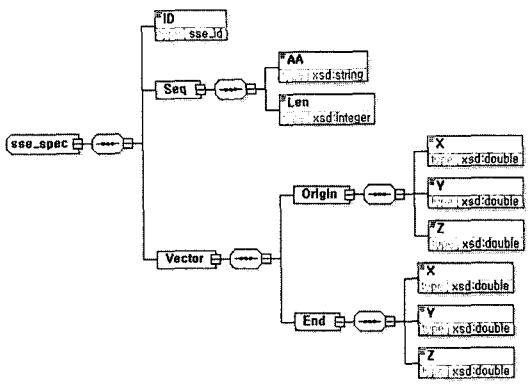


그림 2 SSE의 구성요소

그림 2에서 SSE 세션에 대한 각각의 요소들의 구성형태를 보여 주고 있다. 하나의 2차구조 요소에 대한 식별자(identifier)는 <ID>에 기술된다. 단백질 일차구조

인 아미노산 서열을 기술하는 A 구성요소는 <Seq> 태그에 대응되어 기술된다. <Seq> 태그의 하위에 존재하는 <AA> 태그는 연속적인 아미노산 서열에 대한 정보를 가지고 있으며, <Len> 태그는 이 서열의 길이, 즉 아미노산의 서열의 길이를 제공한다. 벡터에 대한 3차원적 정보를 가지고 있는 C 구성요소에 대한 정보는 <Vector> 태그에 기술된다. <Vector> 태그는 벡터의 시작점과 끝점에 대한 정보를 가지는 <Origin>과 <End> 태그로 구성된다.

- 관계 세션(Relation section)은 하나의 단백질을 구성하고 있는 2차구조 요소 집합에서 가능한 모든 쌍에 대한 관계에 대한 정보를 기술하고 있다. 이러한 두 2차구조 요소 사이에 기술되는 관계 정보는 두 단백질을 비교하고 유사한 부분 구조를 파악하는데 사용될 수 있다. 두 2차구조 사이에 형성되는 관계는 그림 3에서 기술되는 것처럼 rel_spec 타입을 가진 <Relation> 태그에 기술된다. 그림 3에서 <Hydro>와 <Direction>과 같이 점선으로 된 사각형은 선택적(optional) 요소를 의미한다.

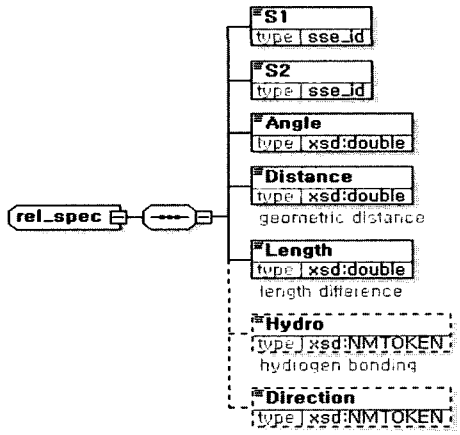


그림 3 2차구조 요소 사이의 관계 표현

표 3 PSAML의 R과 <Relation> 대응

PSAML 구성요소	XML 태그
θ	<Angle>
d	<Distance>
v	<Length>
h	<Hydro>
ϕ	<Direction>

3.3 PDB 데이타에서 PSAML 데이타로 변환하는 구조표현 변환기

PDB2PSAML은 PDB 데이타에서 PSAML 언어에 필요한 정보를 추출하여 XML 형태의 데이타로 변환하는 구조표현 변환기이다. PDB 데이타를 PSAML 형태의 문서로 변환하는 전반적인 단계는 그림 4와 같다.

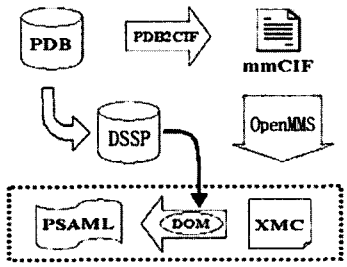


그림 4 PSAML 문서로의 변환 과정

mmCIF 데이타는 고분자 화합물에 대한 구조 데이타를 표현하는데 있어서 결정학 정보를 가지고 있다. mmCIF 데이타 형태에서 데이타 영역에 기술되는 각 데이타 아이템은 유일한 데이타 이름으로 대응되는데, mmCIF 데이타 이름은 mmCIF 사전에 나열되고 정의된다. 그리고, PDB 데이타 파일을 mmCIF 데이타 파일로 변환하는 여러 프로그램들이 있다.

OpenMMS 툴킷[20]은 mmCIF 형태의 파일에 기술된 단백질과 핵산으로 기술되는 고분자 화합물에 대한 데이타를 분석할 수 있는 프로그램들을 제공하고 있다. 이 툴킷은 또한 mmCIF 데이타 파일을 읽어들여 같은 형태의 관계형 데이타베이스 및 XML 형태의 파일로 변환하는 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)[21]은 응용 프로그램과 스크립트에서 문서의 내용, 구조, 스타일 등을 동적으로 접근하거나 변경할 수 있는 플랫폼-독립적인 기능과 언어-중립적인 인터페이스를 제공한다. PDB2PSAML의 구체적인 동작은 그림 5와 같다.

PDB2PSAML은 먼저 입력된 PDB 파일을 OpenMMS를 이용하여 XMC 파일을 생성한다. 그리고 DOM 파서를 이용하여 생성된 XMC 파일을 파싱한 후 DOM 트리를 생성한다. 그 다음 단계로 생성된 DOM 트리를 재귀적인 방법으로 각 노드를 탐색하면서 원하는 정보가 있는 노드의 텍스트 데이타를 배열과 변수에 저장한다. 2차구조의 정보와 2차구조 사이의 관계 정보는 DOM 트리와 DSSP 파일의 정보를 바탕으로 계산된다. DSSP(Dictionary of Protein Secondary Structure)[22]는 Kabsch와 Sander에 의해 제안된 방법으로

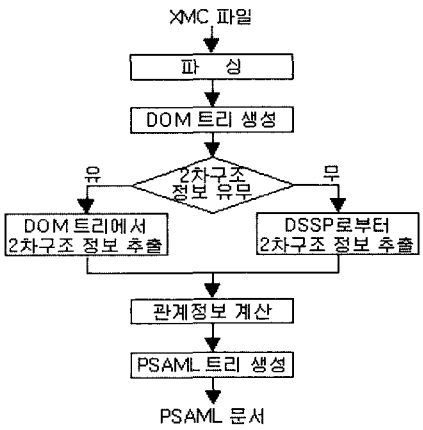


그림 5 PDB2PSAML 변환도구의 동작

PDB 데이타로부터 2차구조에 관련된 정보를 산출해 낼 수 있다. 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 4는 DSSP 파일과 PSAML의 관련 부분을 나타내고 있다.

표 4 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 데이타를 산출하고 새로 생성된 PSAML 트리에 각 노드 및 텍스트 데이타를 추가한다. 마지막으로 PSAML 트리를 재귀적으로 탐색하면서 각 노드와 데이타를 출력하게 된다. 변환도구는 JAVA로 구현되었으며, XMC 파일의 파싱은 Apache XML 프로젝트에서 제공하는 XML 파서인 Xerces[23]를 이용하였다.

3.4 PSAML 데이타베이스

PSAML 기반의 단백질 구조 데이타베이스는 공개 XML 데이타베이스인 eXist[24]를 이용하여 효과적으로 SCOP 데이타베이스에서 제공하는 분류 정보를 기반으로 PSAML 문서를 저장하고 검색할 수 있는 방법을 제공하고 있다. SCOP은 단백질이 지닌 구조적인 유사성과 분류학적인 관계를 기반으로 단백질에 대한 구조 분류 정보를 체계적으로 제공하는 데이타베이스이다. SCOP의 계층적인 구조 분류는 분류 기준에 따라 11개의 클래스(class)로 나뉘어 지며, 각각의 클래스는 2차구조의 구성과 토폴로지(topology)에 의해 폴드(fold)로 나뉘어 진다(그림 3). 폴드는 다시 수퍼패밀리로, 수퍼패밀리는 패밀리로, 그리고 패밀리는 도메인으로 나뉘어 진다. 이러한 구조 분류정보는 PDB ID를 통하여 추출될 수 있다. PSAML 기반 단백질 구조 데이타베이스는 제공된 SCOP 분류 정보를 eXist에서 제공하는 컬렉션(collection)으로 정의하고 여기에 분류된 PSAML 형식의 단백질 구조 정보를 저장한다. 하나의 컬렉션에 속한 단백질 구조 정보는 eXist에서 제공하는 질의 방법을 통하여 비교적 용이하게 접근할 수 있다.

4. Topology string을 이용한 단백질 구조 표현

topology string은 단백질 2차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하는 방법[25]으로서 PSAML에서 정의된 2차구조들이 3차원 공간상 위상학적인 정보를 바탕으로 생성된다. topology string은 대용량의 단백질 구조 데이타베이스에 저장된 단백질 구조와 비교하는 과정을 보다 빠르게 수행할 수 있도록 활용될 수 있다.

4.1 topology string의 정의

단백질을 이루는 각각의 2차구조를 하나의 문자로 변환하여 생성되는 topology string에 대한 정의는 다음과 같다.

- TS(protein_id) = $\{t_1, t_2, ..., t_i\}$, 단, protein_id는 단백질 식별자, t_i는 topology 문자, i는 2차구조 개수이며 서열상의 순서
- t_i = {V or D, A or M, E or F}, t_i가 α-나선일 경우,
  = {H or N, G or I, K or L}, t_i가 β-판상조각일 경우,
- t_i의 값은 단백질 2차구조의 종류 및 위상학적 방향성에 따라 결정된다(표 5).

표 5에서 표현된 topology string을 이루는 문자들은 20가지의 아미노산 문자들 중에서 선택되었다. 이것은 보다 효과적인 topology string 서열의 상동성 평가를 수행하기 위하여 NCBI Blast 프로그램[26]을 적용할 수 있도록 하기 위한 것이다. PSAML 기반으로 생성된 topology string은 X축, Y축, 그리고 Z축을 기준으로 각 90° 회전하여 모두 24가지의 서로 다른 topology string으로 변환된다. 3차원 공간상에서 존재하는 2차구조는 바라보는 관점에 따라 다른 topology 문자로 변환될 수 있다. 본 논문에서는 하나의 2차구조에 대한 topology 문자를 생성할 때 총 24가지의 방향을 고려하여 생성하였다. 표 6은 각 축의 90° 회전에 따라 하나의

표 5 2차구조 변환 규칙

위상학적 방향성		단백질 2차구조	
		α	β
+x	위쪽	V	H
-x	아래쪽	D	N
+y	오른쪽	A	G
-y	왼쪽	M	I
+z	앞쪽	E	K
-z	뒤쪽	F	L

표 6 회전에 따른 변환 규칙

변환 방향											
+x(90°)				+y(90°)				+z(90°)			
α		β		α		β		α		β	
V	V	H	H	V	F	H	L	V	M	H	I
D	D	N	N	D	E	N	K	D	A	N	G
A	F	G	L	A	A	G	G	A	V	G	H
M	E	I	K	M	M	I	I	M	D	I	N
E	A	K	G	E	V	K	H	E	E	K	K
F	M	L	I	F	D	L	N	F	F	L	L

topology 문자의 변환 규칙을 나타내고 있다.

4.2 topology string 생성 방법

그림 6은 PSAML에서 제공하는 각각의 2차구조의 3차원 공간인 시작점과 끝점에 대한 정보를 이용하여 topology string을 추출하는 과정을 보여주고 있다.

① 아미노산 서열 순서대로 PSAML에서 2차구조 하나를 선택한다(그림 6-(2)).

② 선택된 2차구조의 시작점을 원점으로 평행 이동한다(그림 6-(3)).

③ 평행 이동 이후, 2차구조의 끝점 값에서 X, Y, 그리고 Z의 절대값 중에서 가장 큰 값을 선택한다. 선택된 값에 따라 위상학적 방향성과 2차구조의 종류에 따라 표 5에 제시된 문자로 변환한다(그림 6-(4)).

④ PSAML에 기술된 각각의 2차구조를 선택하고 ①-③ 과정을 반복 수행하여 topology string을 생성한다.

⑤ ④에서 생성된 topology string을 표 6에서 제시된 변환표에 의하여 24가지의 서로 다른 topology string으로 변환된다. 이때 변환은 각축으로 90°의 방향으로 한다(그림 6-(5)).

4.3 topology string 데이타베이스

PSAML 데이타베이스로부터 topology string 데이타베이스를 생성하는 변환기를 제작하였다. 개발된 변환기는 PSAML 데이타에서 topology string 서열을 FASTA[27] 형식으로 저장하고 이를 NCBI Blast에서 활용할 수 있는 형태의 데이타베이스로 변환한다.

현재 topology string 데이타베이스는 단백질 15,098개에 대한 36,768개의 topology string을 저장하고 있다.

5. 웹기반 단백질 구조 비교 시스템

웹기반 단백질 구조 시스템인 WS4E(Web-based Searching Substructures of Secondary Structure Elements)는 PSAML로 표현된 두 개의 단백질 구조에서 유사성이 높은 부분을 찾는 기본 비교 기능을 제공하며, topology string 데이타베이스를 이용하여 질의 단백질과 유사성이 높은 단백질 구조를 가지는 구조 분류 정보를 추출할 수 있는 방법을 제공한다. 그리고 사용자는 웹 인터페이스를 통하여 단백질 구조 비교를 수행하고 그 결과를 신속하게 확인할 수 있다.

5.1 WS4E 시스템의 구조

WS4E는 사용자 인터페이스, topology string을 이용한 필터링 모듈, 두 단백질간의 구조 비교를 수행하는 구조 비교 모듈, 그리고 PSAML 및 topology string 데이타베이스로 구성되어 있다(그림 7).

WS4E는 사용자가 입력한 PSAML 형식의 단백질과 유사한 부분 구조를 빠르게 찾기 위하여, 입력 단백질의 topology string을 생성하고, NCBI Blast 프로그램을 이용하여 입력된 단백질의 topology string과 유사성이

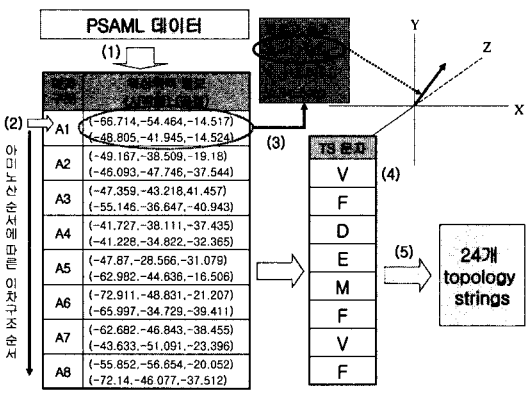


그림 6 topology string 생성과정

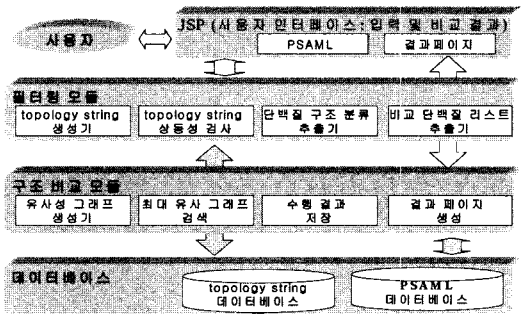


그림 7 WS4E 시스템 전체구조

높은 단백질 구조를 추출한 후, 추출된 단백질이 속하는 SCOP Fold를 찾는다. 찾은 SCOP Fold에 속한 단백질들을 대상으로 2차구조 단백질 구조 비교를 수행한다.

5.1.1 사용자 인터페이스 모듈

사용자 인터페이스를 통해서 사용자는 단백질 구조 비교를 수행하고 그 결과를 신속하게 확인할 수 있다. 그리고 단백질 구조 비교에 사용되는 인자(각도, 거리, 길이)를 변경할 수 있는 인터페이스 및 유사한 부분 구조를 포함하는 단백질 구조에 대한 정보를 PDB 및 SCOP 사이트에서 보다 정확한 정보를 확인할 수 있는 인터페이스를 제공한다.

5.1.2 필터링 모듈

필터링 모듈은 5.1.3절에서 수행하는 단백질 2차구조 기반의 단백질 구조 비교를 수행하기 이전에 입력된 단백질과 유사한 부분 구조를 가지는 단백질을 추출하는 기능을 제공한다. 이는 대용량의 단백질 구조 데이타를 포함하는 PSAML 데이타베이스에서 비교 대상 단백질의 수를 줄여 보다 효과적으로 빠르게 유사한 부분 구조를 포함하는 단백질 구조를 찾기 위한 것이다.

그림 8은 topology string과 단백질 구조 분류 정보를 이용하여 입력 단백질과 유사한 부분 구조를 가진 단백질 구조를 추출하는 단계를 보여주고 있다.

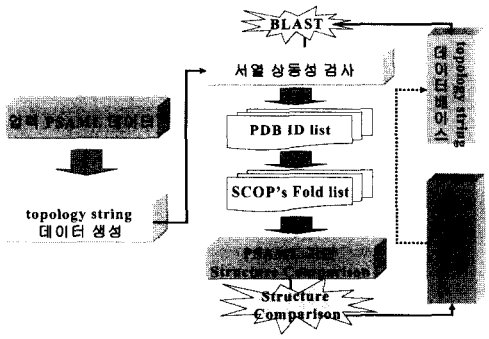


그림 8 topology string을 이용한 단백질 구조 비교 방법

① *입력 단백질(PSAML)의 topology string 생성*

4.2절에 기술된 과정을 통하여 입력된 PSAML 데이타에서 2차구조 요소 및 관계 정보를 이용하여 topology string을 생성한다.

② *topology string 서열의 유사성 측정*

생성된 입력 단백질의 topology string과 topology string 데이타베이스에 저장된 모든 단백질의 topology string과 서열 상동성 검사를 NCBI에서 제공하는 Blast 프로그램을 이용하여 수행한다. 이때 사용되는 치환 매트릭스(BLOSUM)는 임의의 한 문자가 임의의 문자로 치환될 때의 동일한 비율로 적용할 수 있도록 수정되었다.

③ *비교 대상 단백질 추출*

②의 결과에서 입력 topology string과 유사성이 높은 단백질 구조(PDB ID)를 추출한다. Blast 프로그램의 결과 파일을 분석하여 서열 상동성이 일정값(기본값은 75%) 이상인 단백질들을 추출한다. 추출된 단백질은 topology string 측면에서 입력된 단백질과 유사성이 높은 단백질 구조를 가진다.

④ *SCOP Fold 정보 추출*

③과정에서 추출된 단백질 리스트에서 각 단백질이 속하는 SCOP의 Fold 정보를 추출하여 입력 단백질과 2차구조 기반 단백질 구조 비교(5.1.3절)를 수행할 단백질들을 추출한다. SCOP 데이타베이스는 유사한 구조를 가지고 있는 단백질들을 같은 Fold에 분류하여 저장하고 있다. 이 과정에서는 ② 과정에서 NCBI의 Blast 프로그램을 이용하여 추출된 단백질 구조들 이외에 SCOP의 Fold에 속한 단백질 구조들을 추출함으로써, topology string을 이용하여 배제될 수 있는 입력 단백질과 유사한 구조를 가진 단백질들을 고려할 수 있다.

5.1.3 구조 비교 모듈

구조 비교 모듈에서는 5.1.2절에서 기술된 topology string 및 SCOP의 Fold 정보를 통하여 추출된 단백질 구조와 입력 단백질 구조에서 유사한 부분구조를 찾는 기능을 수행한다.

그림 9는 두 단백질 구조에서 유사한 부분 구조를 찾는 과정으로서, 개발된 유사성 그래프 생성 알고리즘[28]을 이용하여 두 단백질의 PSAML 정보를 바탕으로 유사한 부분구조를 내포하는 유사성 그래프를 생성한 후, 모든 노드 사이에 간선이 존재하는 부분 그래프[29]를 찾는 알고리즘[30]을 이용하여 최대 유사 부분 구조를 파악할 수 있다.

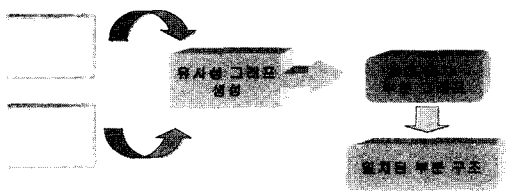


그림 9 PSAML 기반 단백질 구조 비교 과정

PSAML 데이타를 기반으로 단백질 구조간의 유사성을 내포하는 유사성 그래프 G는 표 7과 같이 정의된다.

표 7 유사성 그래프 정의

$G(A, B) = \{V, E\}$, A, B는 단백질 구조
$V = \{(a_i, b_j) \mid a_i \in A, b_j \in B\}$
$E = \{[(a_i, b_k),(a_j, b_l)] \mid$ 2차구조 유사성 비교$\}$

표 8 비교 인자 및 비교 식

관계	비교 식
타입	$T(a_i) = T(b_k),\ T(a_j) = T(b_l)$
각도	$\lvert\theta(a_i, a_j) - \theta(b_k, b_l)\rvert < \theta_d$
거리	$\lvert d(a_i, a_j) - d(b_k, b_l)\rvert < d_d$
길이	$\lvert v(a_i, a_j) - v(b_k, b_l)\rvert < v_d$

표 7에서, 유사성 그래프 G는 단백질 A와 단백질 B의 2차구조의 특징 및 관계를 이용하여 유사성이 있는 2차구조를 표현하고 있다. V와 E는 각각 그래프 G의 노드와 간선의 집합을 나타내고 있다. V에 속한 각 노드는 단백질 A의 한 2차구조와 단백질 B의 한 2차구조의 쌍으로 이루어져 있다. E에 속한 각 노드 사이의 간선은 노드에 포함된 단백질 2차구조 간의 관계가 유사하면 존재한다.

그림 10(a)는 유사성 그래프의 예를 보여주고 있다. 노드 (a_1, b_3)과 노드 (a_5, b_4) 사이의 간선은 단백질 A에 속하는 2차구조 a_1와 a_5에 존재하는 관계와 단백질 B에 속하는 b_3와 b_4에 존재하는 관계가 유사할 때 생성한다. 이 경우, 단백질 A의 a_1과 단백질 B의 b_3 및 단백질 A의 a_5와 단백질 B의 b_4가 유사하다는 것을 의미한다. 그림 10에서는 유사성 그래프 노드 표현을 a_ib_l과 같이 기술한다.

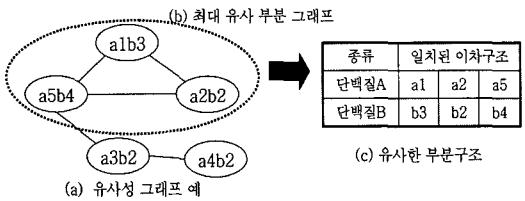


그림 10 유사성 그래프와 최대 유사성 그래프

단백질 A와 B에서 유사성 그래프를 생성하는 과정은 다음과 같다.

① 단백질 A와 B 각각에서 두 2차구조 요소를 선택한다. 이때, 2차구조 사이의 거리가 일정 거리(threshold distance) 이내에 있는 2차구조를 선택한다.

② 선택되는 2차구조의 쌍은 [a_i, a_j]와 [b_k, b_l]과 같은 형태이며, (a_i, b_k), (a_j, b_l), (a_j, b_k), (a_i, b_l)와 같은 유사성 그래프 노드가 생성된다.

③ ②에서 생성된 유사성 그래프 노드 사이의 간선이 존재할 수 있는지 결정한다. 이때 표 8에 기술된 각 관계의 인자 값(θ, d, v)보다 작고 같은 타입일 때 생성된다.

④ 단백질 A에서 형성되는 2차구조의 모든 쌍과 단백질 B에서 형성되는 2차구조의 모든 쌍에 대하여 ①-③ 과정을 반복하여 유사성 그래프 G를 형성한다.

그림 10(a)와 같은 유사성 그래프에서 그림 10(b)와 같은 모든 노드사이에 간선이 존재하는 부분 그래프들에서 노드가 제일 많은 최대 유사 부분 그래프를 찾기 위하여 C. Bron와 J. Kerbosch가 제안한 하나의 그래프에서 모든 clique(모든 노드 사이에 간선이 존재하는 그래프)를 찾는 알고리즘[30]을 이용하였다.

6. WS4E 시스템의 실행 결과 및 분석

WS4E 시스템은 Intel(R) XEON(TM) 프로세서 2.0GHz와 1G 메모리를 탑재한 리눅스 운영체제에서 운용되고 있으며, 현재 PSAML 데이타베이스는 15,098개의 PDB 데이타(체인으로 구분)로부터 생성된 PSAML 데이타를 930개의 SCOP Fold 정보를 기준으로 분류하여 제공하고 있다. topology string 데이타베이스는 NCBI Blast 프로그램이 이용 가능한 형태이며, 362,352개의 topology strings을 제공하고 있다.

6.1 PSAML 기반 단백질 구조 비교 결과

PSAML 기반으로 표현된 단백질 구조를 비교하는 방법을 이용하여 PDB ID *1MBA*(그림 11(a))와 *1MBC*(그림 11(b)) 단백질 간 구조를 비교하였다. *1MBA*와 *1MBC*는 각각 8개의 α-나선으로 이루어진 단백질이다. 두 단백질은 Myoglobin 계열에 속하며, 산소를 저장하는 기능을 담당한다.

(a) 1MBA 단백질 구조 (b) 1MBC 단백질 구조

그림 11 비교 대상 단백질 구조

• PSAML 문서 생성

*1MBA*와 *1MBC*의 PDB 정보를 바탕으로 PSAML 형식으로 변환하는 변환도구를 이용하여 PSAML 형식의 데이타를 생성한다. 그림 12는 생성된 *1MBA*와 *1MBC*의 PSAML 문서를 보여주고 있다.

• 유사성 그래프 및 최대 유사 부분 그래프

생성된 *1MBA*와 *1MBC*의 PSAML 문서를 읽어 Protein 객체를 생성한 후, 두 단백질 구조 사이에 유사한 2차구조 쌍에 대한 정보를 나타내는 유사성 그래프를 생성하고, 생성된 유사성 그래프에서 최대 유사 부분 그래프를 찾는다. *1MBA*와 *1MBC*의 유사 부분구조는

```
<?xml version="1.0" encoding="UTF-8" ?>
- <PSA>
  <Desc>1MBA</Desc>
  - <DATA>
    - <SSE>
      + <Alpha id="A1" n="1">
      + <Alpha id="A2" n="2">
      + <Alpha id="A3" n="3">
      + <Alpha id="A4" n="4">
      + <Alpha id="A5" n="5">
      + <Alpha id="A6" n="6">
      + <Alpha id="A7" n="7">
      + <Alpha id="A8" n="8">
      </SSE>
    - <R>
      - <Relation s1="A1" s2="A2">
          <S1>A1</S1>
          <S2>A2</S2>
          <Angle>97.66143748273083</Angle>
          <Distance>17.853649192067795</Distanc
          <Length>1.1079969833610832</Length>
        </Relation>
      - <Relation s1="A1" s2="A3">
          <S1>A1</S1>
```


(a) 1MBA


```
<?xml version="1.0" encoding="UTF-8" ?>
- <PSA>
  <Desc>1MBC</Desc>
  - <DATA>
    - <SSE>
      + <Alpha id="A1" n="1">
      + <Alpha id="A2" n="2">
      + <Alpha id="A3" n="3">
      + <Alpha id="A4" n="4">
      + <Alpha id="A5" n="5">
      + <Alpha id="A6" n="6">
      + <Alpha id="A7" n="7">
      + <Alpha id="A8" n="8">
      </SSE>
    - <R>
      - <Relation s1="A1" s2="A2">
          <S1>A1</S1>
          <S2>A2</S2>
          <Angle>106.48443308139</Angle>
          <Distance>18.173678564210306</Distance
          <Length>0.11796603473912981</Length>
        </Relation>
      - <Relation s1="A1" s2="A3">
          <S1>A1</S1>
          <S2>A3</S2>
```


(b) 1MBC

그림 12 두 단백질의 PSAML

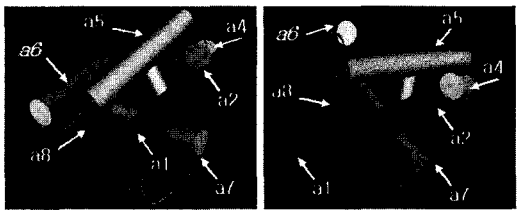
표 9 일치된 2차구조

ID	일치된 2차구조					
1MBA	a_1	a_2	a_4	a_5	a_7	a_8
1MBC	a_1	a_2	a_4	a_5	a_7	a_8

표 9와 그림 13에서 보여주고 있다. 그림 13은 VRML [31]을 이용하여 *1MBA*와 *1MBC*의 유사한 부분구조를 보여주고 있다.

6.2 topology string 기반 단백질 구조 비교 결과

topology string을 이용하여 *1MBA*와 관련된 단백질 구조 분류 정보를 추출하고, 이를 기반으로 PSAML 데



(a) 1MBA 단백질 구조 (b) 1MBC 단백질 구조

그림 13 두 단백질간 유사 부분 구조

이타베이스에 있는 단백질들과 6.1절에서 기술된 단백질 구조 비교를 수행하여 유사한 부분 구조를 찾아보았다.

• topology string 생성 및 서열 상동성 검사

표 10은 *1MBA*를 이루고 있는 2차구조에 대한 3차원 정보 및 이를 바탕으로 생성된 topology string을 보여주고 있다. 각 2차구조는 3차원 공간에 위치하는 시작점과 끝점에 대한 좌표 값을 가지며, 종류 및 아이디를 가진다. 각 2차구조의 아이디에 나타나는 인덱스는 아미노산에 나타나는 순서에 의하여 결정된다. 즉, 인덱스 숫자가 낮을수록 아미노산 서열상 앞에 위치한다.

표 11은 NCBI의 Blast 프로그램을 이용하여 *1MBA*의 topology string과 상동성이 높은 단백질을 나타내고 있다. PDB ID가 *5MBA*, *4MBA*, *3MBA*, *2FAM*, 그리고 *2FAL*는 *1MBA*의 topology string과 완전 일치하였으며, 부분적으로 일치한 단백질(*1QHA*, *1OUT*, *1JSW*)도 추출되었다.

• 단백질 분류 정보인 SCOP의 Fold 추출

표 12는 NCBI Blast 프로그램을 이용하여 추출된 단백질들에서 topology string 일치도가 75% 이상인 단백질을 대상으로 SCOP의 Fold를 추출한 결과를 보여

표 10 1MBA의 2차구조 정보

2차구조요소	(시작점좌표)(끝점좌표)	TS 문자
a_1	(66.714, 54.464, 14.517) (-48.805,-41.945,-14.524)	V
a_2	(49.167, 38.509, 19.18) (-46.093,-47.746,-37.544)	F
a_3	(47.359, 43.218,41.457) (-55.146,-36.647, 40.943)	D
a_4	(41.727, 38.111, 37.435) (-41.228, 34.822, 32.365)	E
a_5	(47.87, 28.566, 31.079) (-62.982,-44.636,-16.506)	M
a_6	(72.911, 48.831, 21.207) (-65.997,-34.729, 39.411)	F
a_7	(62.682, 46.843, 38.455) (-43.633,-51.091,-23.396)	V
a_8	(-55.852, 56.654, 20.052) (-72.14, 46.077, 37.512)	F

표 11 1MBA의 topology string의 서열 정렬 결과

Protein Id:Chain	Score	Identities
5MBA:null	21	100%
4MBA:null	21	100%
3MBA:null	21	100%
2FAM:null	21	100%
2FAL:null	21	100%
1QHA:A	19	75%
1OUT:B	17	75%
1JSW:A	17	75%

표 12 1MBA과 유사한 단백질 구조를 포함할 SCOP's Folds

Protein Id:Chain	SCOP' Fold
5MBA:null	Globin-like
4MBA:null	Globin-like
3MBA:null	Globin-like
2FAM:null	Globin-like
2FAL:null	Globin-like
1QHA:A	Ribonuclease H-like motif
1OUT:B	Globin-like
1JSW:A	L-aspartase-like

주고 있다. 추출된 SCOP의 Fold에 속한 단백질 구조를 추출함으로써 topology string의 상동성 검사를 통하여 배제된 입력 단백질과 구조적으로 유사한 단백질들을 포함하여 2차구조 기반 단백질 구조를 수행할 수 있도록 하였다. PDB ID가 *5MBA*, *4MBA*, *3MBA*, *2FAM*, 그리고 *2FAL*이 1MBA와 2차구조 기반 단백질 구조 비교를 수행하였을 때 아주 비슷한 구조를 이루고 있음을 VRML을 통하여 알 수 있었다. 이는 *1MBA*를 구성하는 각 2차구조와 topology string 서열 정렬을 통하여 추출된 각 단백질의 각 2차구조가 3차원 공간에 벡터로 표현되었을 때 매우 유사함을 의미한다.

• 추출된 단백질 구조와 구조 비교 수행

표 13은 topology string 서열 정렬에서 추출된 단백질 중에서 topology string의 유사도의 정도가 낮은 *1QHA*, *1OUT*, 그리고 *1JSW*를 대상으로 *1MBA*와 2차구조 기반 단백질 구조 비교를 수행한 결과를 부분적으로 보여주고 있다. 표 13에서 나열된 topology string 상동성의 결과에서 일치성이 낮은 단백질들은 *1MBA*의 특정 부분 구조와 비슷한 구조를 내포하고 있음을 알 수 있었다. 표 13에서 a_2a_1은 *1MBA*의 a_2와 다른 단백질의 a_1이 유사함을 나타낸다.

표 13 일치된 2차구조

ID	일치된 2차구조
1QHA:A	a3a4 a5a10 a6a14 a8a11
1OUT:B	a2a15 a1a10 a3a8 a5a13 a6a14
1JSW:A	a6a6 a7a13 a4a16

6.3 실행 속도 향상을 위한 방법

WS4E 시스템의 특정 SCOP Fold에 속하는 단백질들을 대상으로 유사한 부분 구조를 찾는 과정에서, 가장 많은 실행 시간을 요구하는 부분은 PSAML 데이터베이스에서 해당 단백질을 가져오는 과정과 PSAML 데이타를 분석하여 자바 객체를 생성하는 과정이다(표 14). 이 과정은 텍스트 기반의 PDB 파일에서 단백질 구조 비교에 필요한 정보를 추출하는 과정에 해당한다. 개발된 시스템은 표 14에 기술된 PSAML 데이타베이스에서 PSAML을 읽어 자바 객체를 생성하는 시간을 줄이기 위하여, 미리 PSAML 데이타에서 자바 객체를 생성하고 Java Object Serialization[32] 기술을 이용하여 디스크에 저장한다. 표 15에 기술된 바와 같이 저장된 데이타를 읽어 자바 객체를 생성하는 과정은 표 14에서 기술된 PSAML 데이타베이스에서 읽어 자바 객체를 생성하

표 14 SCOP Fold별 단백질 구조 비교 수행 시간(PSAML 데이타베이스사용)

SCOP Fold	#PSAML	#SSE (평균)	#R (평균)	PSAML DB 자바객체화	유사성 그래프	최대유사성 그래프	총시간
Globin-like	517	8	29	53528	273	1300	55190
Ribonuclease H-like motif	115	24	338	128125	333	1691	130195
L-aspartase-like	21	22	150	16301	54	341	16715

*시간단위: millisecond

표 15 SCOP Fold별 단백질 구조 비교 수행 시간(자바 객체화 사용)

SCOP Fold	#PSAML	#SSE (평균)	#R (평균)	자바 객체 읽기	유사성 그래프	최대유사성 그래프	총시간
Globin-like	517	8	29	2089	260	1401	3861
Ribonuclease H-like motif	115	24	338	2507	341	1924	4815
L-aspartase-like	21	22	150	400	55	384	859

*시간단위: millisecond

는 과정 보다 매우 짧은 시간이 소요되었다. 표 14와 표 15에 기술된 측정 시간은 실행 결과를 파일에 기록하는 시간을 포함하지 않은 시간이며, 각도 관계 정보만을 이용하여 유사한 부분 구조를 찾는데 소요된 시간이다.

6.4 타 시스템과 성능 및 특징 비교

대표적인 단백질 2차구조 기반 구조 비교 시스템에는 3dSearch, TOPS[33], 그리고 SARF2 등이 있다. 3dSearch는 현재 서비스가 제공되고 있지 않은 관계로 SARF2, TOPS 그리고 구조를 구성하는 원자 정보를 이용하여 구조 비교를 수행하는 DaliLite[34]와 실행 속도 및 특징을 비교하였다. 각 시스템에서 제공하는 구조 비교 도구의 실행 속도를 측정하기 위하여, 10개의 2차 구조를 가지는 단백질 구조 90개를 가지는 데이타를 준비하였으며, 각 도구를 이용하여 PDB ID 1MXB와 준비된 데이타에 속한 단백질 구조들과 10번 구조 비교를 수행하여 그 평균을 구하였다. Intel(R) XEON(TM) 프로세서 2.0GHz와 1G 메모리를 사용하는 개인용 컴퓨터 환경에서 수행한 결과, 평균적으로 TOPS는 6.45초, SARF2는 84.19초, 그리고 DaliLite는 217.95초로 측정되었으며, PSAML 기반의 구조 비교 방법은 TOPS 다음으로 빠른 속도인 28.39초로 측정되었다(표 16).

TOPS는 DSSP 파일을 분석하여 2차구조의 종류, 방향성 그리고 결합에 대한 정보를 내포하는 tops 파일을 생성하고, 이를 바탕으로 문자열 형태로 단백질 구조를 표현한다. 예를 들어 PDB ID 1MBA는 "NhHhhh-HhHC 6:8R"과 같이 기술된다. TOPS는 단백질 구조를 문자열 형태로 표현하고 제한 프로그래밍 기법을 이용하여 일정한 패턴을 찾는다. 이 방법은 다른 도구들에 비하여 빠르게 유사한 부분 구조를 찾지만, 빠른 실행속도에 비하여 실제 구조 비교에 있어 2차구조의 3차원 정보를 활용하지 않는 관계로 단백질 3차 구조가 다르다고 하더라도 유사한 부분 구조로 파악하는 경우가 많다. DaliLite는 PSAML 기반 단백질 구조 비교 방법, TOPS, 그리고 SARF2와는 달리 단백질을 구성하는 원자 정보를 바탕으로 유사한 부분 구조를 찾으므로 수행시간은 오래 걸리지만 보다 자세한 정보를 바탕으로 구조 비교를 수행한다.

그림 14는 SARF2를 이용하여 *1MBA*와 *1MBC*의 단백질 구조에서 유사한 부분 구조를 찾은 결과의 일부분을 보여주고 있다. 단백질 구조의 3차원 형태를 확인할 수 있는 VRML을 이용하여 개발된 PSAML 기반의 단백질 구조 비교 방법(그림 13)과 비교한 결과, SARF2의 결과인 그림 14(a)는 유사하지 않은 2차구조를 포함하고 있으며, 그림 14(b)는 유사한 2차구조를 찾지 못하고 있다. 또한 그리고 SARF2는 두 개의 단백질 간의 구조 비교를 수행하는 서비스만을 제공하는 관계로 단백질 데이타베이스를 대상으로 한 성능 비교가 불가능하였으며, 검색된 유사부분 구조를 확인할 수 있는 도구의 지원이 부족하였다.

단백질 1MBA와 1MBC의 비교 결과

```
141 Ca-atoms ( 96%), rmsd = 1.91,  26% identical residues

alignment of  9 SS elements: 9 alpha and  0 beta

  a01a02a03a04a05a06a07a08a09
  a01a01a02a03a04a05a07a07a09
```

(a) 9개의 2차 구조가 일치된 경우

```
42 Ca-atoms ( 28%), rmsd = 1.88,  11% identical residues

alignment of  3 SS elements: 3 alpha and  0 beta

  a07a08a09
  a07a07a09
```

(b) 3개의 2차구조가 일치된 경우

그림 14 SARF2의 실행결과 일부분

기존의 단백질 구조 비교 시스템과 WS4E를 비교하여 표 17에 나타내었다. WS4E는 2차구조의 특징 및 관계 정보를 이용하여 구조 비교를 위한 정보를 제공하는 PSAML 문서를 기반으로 하는 단백질 구조 비교 도구

표 16 단백질 구조 비교 도구의 실행 속도 측정 (단위: 초)

수행회수	TOPS	SARF2	DaliLite	WS4E
1	6.158	84.798	229.116	28.203
2	7.466	83.738	215.031	28.539
3	6.165	84.009	216.796	28.338
4	6.476	84.570	217.156	28.380
5	6.342	84.658	217.531	28.387
6	6.192	84.257	217.481	28.402
7	6.105	83.547	217.218	28.362
8	6.764	84.912	216.125	28.546
9	6.750	84.144	216.336	28.147
10	6.102	83.227	216.704	28.549
평균	6.49	76.06	218.09	25.93

표 17 기존 단백질 구조 비교 시스템과 비교

특 징	SARF2	TOPS	DaliLite	WS4E
구조 표현 단위	2차구조	2차구조	원자	2차구조
구조 비교 요소	각도, 거리	방향성, 결합	원자의 특징,거리	각도,거리,길이
1:1 구조 비교 서비스	제공	제공	제공	제공
1:N 구조 비교 서비스	제공안함	제공	제공	제공
구조 결과 제공 방법	웹 문서	웹 문서	E-Mail	웹 문서
구조 데이타베이스	제공안함	제공	제공	제공
입력파일 형태	텍스트	텍스트	텍스트	XML

를 제공하고 대용량 단백질 구조 데이타베이스에서 비교 대상 단백질의 수를 줄일 수 있는 방법을 제공한다는 특징이 있다.

7. 결 론

본 논문에서는 단백질 구조를 비교하여 유사한 부분 구조를 찾기 위하여 XML 기반의 단백질 구조 표현 기법인 PSAML에 대하여 기술하였다. 그리고 PSAML 데이타베이스를 기반으로 PSAML로 표현된 단백질 구조에서 유사한 부분 구조를 찾는 웹서비스를 제공하는 웹기반 단백질 구조 비교 시스템인 WS4E에 대하여 기술하였다.

PSAML(Protein Structure Abstraction Markup Language)은 2차구조 구성요소(secondary structure element)의 특징과 그들 사이에 정의되는 관계(각도, 거리, 길이)를 이용하여 XML 형태로 단백질 구조를 표현한다. PSAML은 XML 스키마를 이용하여 XML 기반 언어의 요소를 정의하고, PDB나 다른 단백질 관련 XML 데이타 형식을 이용하는 것보다 간결하면서 구조적으로 단백질 구조 정보를 표현할 수 있다. PSAML에서 제공하는 단백질 구조에 관한 정보를 이용하여 단백질 구조를 비교하는 여러 형태의 시스템을 개발할 수 있다. PSAML은 단백질의 여러 특성들 중에서 2차구조 구성요소에 초점을 맞추고 있으므로 다른 XML 기반 표현들보다 간결하다. PSAML은 PDB 데이타로부터 개발된 변환기를 통하여 자동적으로 생성된다. topology string은 단백질 2차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하는 방법으로서 PSAML에서 정의된 2차구조들이 3차원 공간상 위상학적인 정보를 바탕으로 생성된다. topology string은 PSAML 데이타베이스에서 주어진 단백질 구조와 비교될 대상 단백질의 수를 줄일 수 있는 방법에 활용될 수 있다.

개발된 WS4E(Web-based Searching Substructures of Secondary Structure Elements)는 PSAML로 표현된 두 개의 단백질 구조에서 유사성이 높은 부분을 찾는 기본 비교 기능을 제공하며, 구축된 대용량 단백질 구조 데이타베이스인 PSAML 데이타베이스에서 단백질 구조 비교를 신속하고 효과적으로 비교할 수 있도록 주어진 단백질의 구조와 유사성이 높은 단백질 구조를 필터링하는 방법을 제공한다.

향후 보다 효과적으로 구조 비교 서비스를 제공하기 위하여 병렬 컴퓨팅 기술을 활용하여 신뢰성이 보장되는 단백질 구조 비교 시스템을 개발할 예정이다.

참 고 문 헌

[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] Helen B, T. Bhat, Philip B., Zukang F., Gary G., Helge W., and John W., "The Protein Data Bank and the challenge of structural genomics," Nature Structural Biology, Vol.7, pp.957-959, 2000.

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

[4] A. P. Singh and D. L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations," Intelligent Systems for Molecular Biology 97, Vol.5, pp.284-293, 1997.

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

[6] 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.

[7] 김진홍, 안건태, 이수현, 이명준, "구조비교를 위한 단백질 데이타의 XML 표현기법", 한국정보과학회 프로그래밍언어연구회, 제16권, 제2호, pp.15-16, 2002.

[9] MGED group, MicroArray and Gene Expression (MAGE), WWW document (http://www.mged.org/Workgroups/MAGE/mage.html), 2004.

[10] BioXML, Genome Annotation Markup Elements (GAME), WWW document (http://www.bioxml.org/Projects/game/), 2003.

[11] V. Guerrini and D. Jackson, "Bioinformatics and Extended Markup Language (XML)," Online Journal of Bioinformatics, Vol.1, No.1, pp.12-21, 2000.
[12] R. Sayle and E. Milner-White, "RASMOL: biomolecular graphics for all," Trends in Biochemical Science, Vol.20, pp.374-376, 1995.
[13] 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.
[14] Proteomics Inc., BioML:Biological Markup Language, WWW document (http://www.bioml.com/bioml/), 2004.
[15] D. Hanisch, R. Zimmer, and T. Lengauer, "ProML: the Protein Markup Language for specification of protein sequences, structures and families," In Silico Biol, Vol.2, No.3, pp.313-324, 2002.
[16] 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.
[17] 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.
[18] 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.
[19] I. Eidhammer, I. Jonassen, and W. R. Taylor, Structure Comparison and Structure Patterns, Report no 174, University of Bergen, 1999.
[20] D. S. Greer, J. D. Westbrook, and P. E. Bourne, OpenMMS: An Ontology Driven Architecture for Macromolecular Structure, Objects in Bio and Cheminformatics, 2001.
[21] W3C, Document Object Model (DOM), WWW document (http:// www.w3.org/DOM/), 2004.
[22] 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.
[23] The Apache Software Foundation, Xerces: XML parsers in Java, Apache XML Project, WWW document (http://xml.apache.org/), 2004.
[24] Akmal B. Chaudri, Awais Rashid, Roberto Zicari, XML Data Management: Native XML and XML-Enabled Database Systems, Addison Wesley Professional, 2003.
[25] Martin AC, "The ups and downs of protein topology; rapid comparison of protein structure," Protein Eng. Vol.13, No.12, pp.829-837, 2002.
[26] Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res., No.25, pp.3389-3402, 1997.
[27] David W. Mount, Bioinformatics Sequence and Genome Analysis, Gold Spring Harbor Laboratory Press, pp.31-32, 2001.
[28] Su-Hyun Lee, Jin-Hong Kim, Geon-Tae Ahn, and Myung-Joon Lee, "Efficient Generation of Compatibility Graphs for Two Sets With an Ordered Attribute," Information Sciences, (submitted), 2004.
[29] Hiroaki KATO and Yoshimasa TAKAHASHI, "Automated Identification of Three-Dimensional Common Structural Features of Proteins," J. Chem. Software, Vol.7, No.4, pp.161-170, 2001.
[30] Sampo Niskanen, Patric Ostergard, Cliquer: routines for clique searching, WWW document (http://www.hut.fi/~pat/cliquer), 2002.
[31] VRML Plugin, VRML Plugin and Browser Detector, WWW document (http://cic.nist.gov/vrml/vbdetect.html), 2002.
[32] Sun Microsystems, Java Object Serialization Specification, WWW document (http://java.sun.com/j2se/1.4/docs/guide/serialization/spec/serialTOC.doc.html), 2003.
[33] D. Gilbert, D. Westhead, J. Viksna, and J. Thornton, A computer system to perform structure comparison using TOPS representations of protein structure, Comput. Chem., Vol.26, pp.23-30, 2001.
[34] Holm, L., Park, J. DaliLite workbench for protein structure comparison, Bioinformatics, Vol.16, pp.566-567, 2000.

김 진 홍

1999년 2월 울산대학교 전자계산학과 졸업(학사). 2001년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(석사). 2005년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(박사). 2005년 3월~현재 울산대학교 컴퓨터·정보통신공학부 객원교수. 관심분야는 생물정보학, 제한프로그래밍, 협업지원 시스템, 이동에이전트 시스템 등

안 건 태

1999년 2월 울산대학교 전자계산학과 졸업(학사). 2001년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(석사). 2005년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(박사). 관심분야는 생물정보학, 협업지원 시스템, 분산시스템, 이동에이전트 시스템 등

변 상 희

2003년 2월 울산대학교 컴퓨터·정보통신공학과 졸업(학사). 2005년 2월 울산대학교 대학원 컴퓨터·정보통신공학부 졸업(석사). 2005년 2월~현재 온넷기술 주식회사 연구원. 관심분야는 생물정보학, 협업지원 시스템, 웹 프로그래밍, 유·무선 통신프로토콜 등

이 수 현

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. 관심분야는 프로그래밍언어, 제한프로그래밍, 생명정보학 등

이 명 준

1980년 2월 서울대학교 수학과 졸업(학사). 1982년 2월 한국과학기술원 전산학과 졸업(석사). 1991년 8월 한국과학기술원 전산학과 졸업(박사). 1982년 3월~현재 울산대학교 컴퓨터·정보통신공학부 교수. 1993년 8월~1994년 7월 미국 버지니아대학 교환교수. 2005년 1월~현재 미국 캘리포니아 주립대학 교환교수. 관심분야는 프로그래밍언어, 분산 객체 프로그래밍 시스템, 병행 실시간 컴퓨팅, 인터넷 프로그래밍 시스템, 생물정보학 등