논•문 17-1-3

SCOP 분류 정보를 활용한 단백질 구조 비교 (Protein Structure Comparison Using the Protein Classification Information of SCOP)

김 진 홍*, 안 건 태*, 변 상 희*, 이 수 현**, 이 명 준*
*울산대학교 컴퓨터정보통신공학부, **창원대학교 컴퓨터공학과
*{avenue, java2u, heeya, mjlee}@mail.ulsan.ac.kr, **suhyun@sarim.changwon.ac.kr

요 약

PSAML 기반의 단백질 구조 비교 방법은 단백질의 2차구조 구성요소와 그들 사이의 관계를 이용하여 단백질 구조를 기술하고, 두 단백질 구조 사이의 유사성 그래프를 생성한 후, 최대 유사 부분 그래프를 찾는 방법을 이용하여 유사한 부분 구조를 찾는다. 그러나, PSAML 기반의 단백질 구조 데이터베이스의 데이터 량이 증가함에 따라 단백질 구조 비교에 요구되는 계산 량을 줄이는 효과적인 방법이 요구된다.

본 논문에서는 PSAML 기반의 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여 단백질 폴드(Fold) 또는 패밀리(Family)를 대표하는 아미노산 서열 정보 및 단백질 구조 분류 정보를 활용하는 방법에 대하여 기술한다. 개선된 PSAML 기반의 단백질 구조 비교 방법은 단백질 아미노산 서열을 이용하여 서열상 유사성이 높은 단백질 구조(Domain)를 필터링하고, 그 결과를 이용하여 추출된 단백질 구조 분류에 속한 모든 단백질 구조와 비교할 수 있는 방법을 제공한다.

1. 서론

단백질 기능을 파악하기 위하여 단백질의 구조적인 특징에 따라 단백질 구조를 분류하고 공통의 부분 구조를 찾아내는데 활용되고 있는 단백질 구조 비교 방법은 단백질 구조를 표현하는 방법에 따라 다양하다[1].

PSAML(Protein Structure Abstraction Markup Language)[2,3]은 단백질의 2차구조와 2차구조 사이에서 발견되는 상호 관계를 이용하여 단백질 구조를 표현하는 방법을 제공하는 PSA(Protein Structure Abstraction)[2,3]를 기준으로 단백질 구조를 표준화된 문서 표현 양식인 XML[4]로 기술할 수 있는 언어이다. PSAML 기반의 단백질 구조 비교 방법[5]은 2차구조의 및 이들 사이의 관계를 비교하여 효과적으로 두 단백질 사이의 유사

한 부분 구조를 찾을 수 있도록 하여 준다.

현재 새롭게 밝혀지는 단백질 3차구조 정보의 증가량이 날이 갈수록 높아짐에 따라 단백질 구조 비교 방법은 보다 효과적으로 빠르게 결과를 산출할 수 있어야 한다. 이는 단백질 구조 비교 시 요구되는 많은 데이터를 효과적으로 처리할 수 있는 방법과 고성능의 계산 능력을 가진 고급 장비가 요구된다. 또한, 효과적인 단백질 구조 비교를 위한 단백질 구조 표현 방법이 요구된다.

본 논문에서는 PSAML 기반의 단백질 구조 데이터베이스를 이용한 단백질 구조 비교에 요구되는 성능향상을 위한 방법에 대하여 기술한다. 이를 위하여 단백질 폴드(Fold) 또는 패밀리(Family)를 대표하는 아미노산 서열 정보(Representative Sequences)를 추출하는 방법 및 단백질 구조 분류 정보(SCOP[6], Pfam[7])를 활용하여 유사한 단백질 구조를 필터링(filtering)하는 방법에 대하여 알아본다.

아미노산 서열의 상동성과는 관계없이 3차원적으로 비슷한 구조를 가지는 단백질들이 존재하고 있다. 제안된 단백질 구조 비교 방법은 아미노산 서열상으로는 관련성이 적지만 3차 구조적으로 관련성이 높은 단백질들을 추출하기 위하여 폴더 정보를 추출한다. 추출된 폴더 정보는 아미노산 서열상의 유사성과 관계없이 3차구조가 유사한 단백질에 관한 정보를 제공한다. 이를 바탕으로 제안된 방법은 새로운 단백질 구조와 관련성이 매우 낮은 단백질들을 미리 비교 대상에서 제외시켜 보다 빠른 단백질 구조 비교를 수행할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 PSAML 기반의 단백질 구조 비교 방법에 대하여 기술한다. 3장에서는 새로운 단백질과 유사한 단백질 구조 분류 정보를 추출하는 방법에 대하여 기술한다. 4장에서는 추출된 단백질 구조 정보를 이용한 PSAML 기반 단백질 구조 비교 방법과 그 결과에 대하여 살펴본다. 마지막으로 5장에서는 결론 및 향후 연구 방향으로 끝을 맺고자 한다.

2. PSAML 기반 단백질 구조 비교

개발된 PSA 및 PSAML 기반의 단백질 구조 비교 방법은 2차구조의 특징 및 이들 사이의 관계를 비교하여 두 단백질 사이의 유사한 부분 구조를 찾을 수 있다.

2.1. PSA와 PSAML

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

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

$$PSA(P) = (S, T, C, A, R)$$
$$R = (\Theta, \gtrless, \vee, h, d),\ 단,\ Ei, Ej \in S,\ i \neq j.$$

S는 단백질을 구성하는 2차구조의 집합을 나타낸다. T, C, A는 각각 2차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보를 나타낸다. R은 두 2차구조사이에 정의되는 관계로써 다음과 같이 표현되고, 이차구조 사이의 관계는 <표 1>과 같다.

<표 1> 이차구조 사이의 관계

관계	의미	표 현
Θ	각도	$\Theta(Ei, Ej) = angle(\Theta l)$
ɣ	거리	$\gamma(Ei, Ej) = distance(D)$
v	길이차	$v(Ei, Ej) = length(li, lj)$
h	수소결합	*h*(*Ei*, *Ej*) = {E, N}, *Ei*와 *Ej*는 β-strand
d	방향성	*d*(*Ei*, *Ej*) = {P, A}, *Ei*와 *Ej*는 β-strand

PSAML은 단백질 구조의 표현을 위한 PSA 표현을 XML로 표현하기 위하여 XML 스키마(XML schema)[4]를 이용하여 XML로 기술할 수 있는 언어이다[그림 1].

(그림 1) PSAML 문서의 예

2.2. 기존 방법의 문제점

PSAML 기반 단백질 구조 비교 방법은 입력되는 두 단백질의 구조적 유사 부분을 찾아내는 방법이다.

현재의 단백질 구조 비교 방법은 두 단백질 사이의 구조를 비교하는 방법으로써 대용량의 단백질 구조 데이터베이스를 대상으로 하지 않고 있다. 대용량의 단백질 구조 데이터베이스를 대상으로 단백질 구조를 비교할 경우, 구조적 유사성이 전혀 없는 단백질 구조들과도 비교가 수행되어야 함으로 많은 실행시간이 요구된다. 따라서 이러한 문제점으로 해결하기 위하여 기존의 단백질 구조 비교 서비스[8,9] 중에는 요청된 서비스의 결과를 실시간이 아닌 전자메일로 제공하는 경우가 많다.

3. 단백질 서열로부터 연관된 구조 분류 정보 분석

단백질 도메인은 특정한 기능을 담당하는 단백질의 부분 구조이다. 단백질 도메인 기반의 단백질 구조 비교 방법은 입력된 단백질 서열과 관련된 단백질 분류(Fold 또는 Family)에 대한 정보를 추출한 후, 이 분류에 속한 모든 단백질 구조와 입력 단백질 구조를 비교하는 방법이다.

입력된 단백질 서열과 서열상으로 유사성이 높은 단백질 구조 분류 정보는 Pfam[7] 및 SCOP[6] 데이터베이스를 활용하여 추출될 수 있다.

3.1. 도메인 및 구조 분류 정보

Pfam은 많은 단백질 구조가 가지는 공통의 도메인(Domains) 및 패밀리(Families)에 대한 정보를 제공하고 있으며, SCOP 데이터베이스는 이미 구조가 밝혀진 단백질을 유사한 구조를 가진 그룹으로 분류한 정보를 제공하고 있다.

3.1.1. Pfam 데이터베이스

Pfam은 현재 파악되어 있는 단백질 구조에서 공통으로 많이 나타나는 단백질 도메인과 패밀리에 대한 정보를 제공하고 있다.

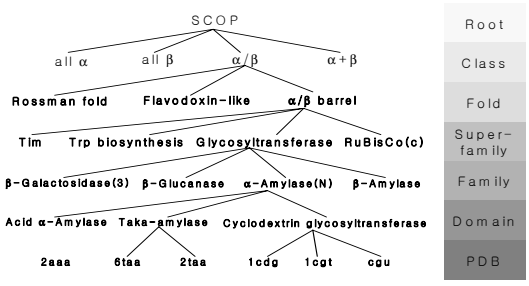
Pfam에서 제공하는 도메인 및 패밀리 정보는 많은 단백질 서열에서 공통으로 나타나

는 보존된 영역(conserved region)을 다중 서열 정렬을 통하여 분석하여 얻어진다. 이러한 단백질 서열과 연관된 단백질 부분 구조에 대한 표현 방법은 HMM(Hidden Markov models)[10]을 이용하고 있다.

3.1.2. SCOP 데이터베이스

SCOP은 단백질이 지닌 구조적인 유사성과 분류학적인 관계를 기반으로 단백질들을 체계적으로 분류해 놓은 데이터베이스이다.

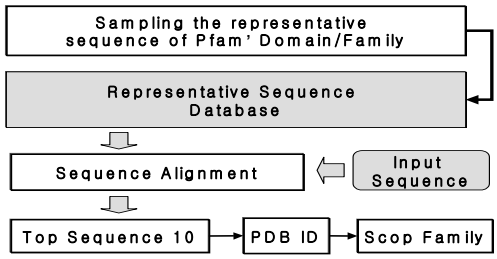
SCOP의 계층적인 구조 분류는 분류 기준에 따라 11개의 클래스(class)로 나뉘어 지며, 각각의 클래스는 2차구조의 구성과 토폴로지(topology)에 의해 폴드(fold)로 나뉘어 진다(그림 3). 폴드는 다시 수퍼패밀리로, 수퍼패밀리는 패밀리로, 그리고 패밀리는 도메인으로 나뉘어 진다. PDB[11]의 모든 단백질은 다른 단백질들과 비교되어 구조적 유사성을 가지는 그룹으로 분류된다.



(그림 2) SCOP 구조 분류 정보

3.2. 단백질 서열로부터 구조 분류 정보 추출 과정

[그림 3]은 입력된 단백질 서열로부터 SCOP 구조 분류 정보를 추출하는 과정이다.



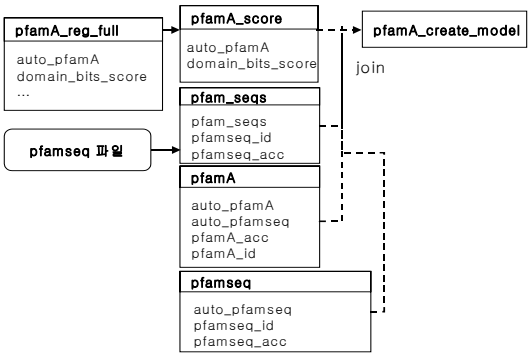
(그림 3) 단백질 서열로부터 구조 정보 추출 과정

Pfam에서 제공하는 각 도메인 및 패밀리를 대표할 수 있는 대표 서열 데이터베이스는 각 도메인 및 패밀리를 생성할 때 사용된 서열 중에서 가장 유사성이 많은 서열을 추출하여 생성된다. SCOP 구조 분류 정보인 패밀리 정보는 Pfam 도메인 정보에서 추출된 PDB ID를 통하여 얻어진다.

3.2.1. 공통 도메인 및 패밀리의 대표 서열 추출 과정

Pfam 데이터베이스는 2003년 2월 현재 5,193개의 도메인 및 패밀리 정보를 관계형 데이터베이스 및 일반 텍스트 기반 형태로 제공하고 있다. Pfam의 도메인 및 패밀리 정보는 HMM 형태로 표현되고 있으며, 이를 생성하기 위한 단백질 서열 정보 또한 제공하고 있다.

[그림 4]는 Pfam의 각 도메인을 대표하는 대표 서열 데이터베이스를 생성하는 과정에 사용된 테이블을 보여주고 있다.



(그림 4) Pfam 도메인 및 패밀리 대표 서열 추출에 사용된 테이블

Pfam에서 제공하는 도메인의 대표 서열을 생성하기 위하여 제공된 pfamA_reg_full 테이블의 domain_bit_score 필드 값을 이용한다. 이 필드는 각 도메인의 HMM과 생성 시에 사용된 서열 사이의 유사성 정도를 나타내는 값을 나타낸다. 각 도메인을 대표하는 서열은 각 도메인을 생성하는 과정에서 사용된 서열 중에서 domain_bit_score 값이 가장 낮은 값을 가진 서열이다.

다음은 Pfam에서 제공된 정보를 이용하여 각 도메인을 대표하는 대표 서열 데이터베이스를 생성하는 과정이다.

① pfamA_score 테이블 생성

Pfam에서 제공하는 pfamA 테이블은 각 도메인에 대한 정보를 나타내며, pfamA_reg_full 테이블은 각 도메인과 각 도메인을 생성할 때 사용된 서열사이의 관계 정보를 제공하고 있다. 새롭게 생성되는 pfamA_socre 테이블은 pfamA에 나타난 도메인을 지정하는 auto_pfamA와 각 도메인과 생성 시에 사용된 서열과의 유사성 정도를 나타내는 domain_evalue_score값을 저장한다. domain_evalue_score 값이 낮을수록 도메인과 유사성이 많다는 것을 나타낸다.

② pfam_seqs 테이블 생성

Pfam에서 제공하는 pfamseq 텍스트 파일은 각 도메인을 생성할 때 사용된 모든 서열 정보를 제공한다. pfam_seqs 테이블은 pfamseq 파일에서 제공하는 모든 서열을 각 서열을 구별할 수 있는 필드인 pfamseq_acc와 pfamseq_id를 이용하여 제공한다.

③ pfamA_create_model 테이블 생성

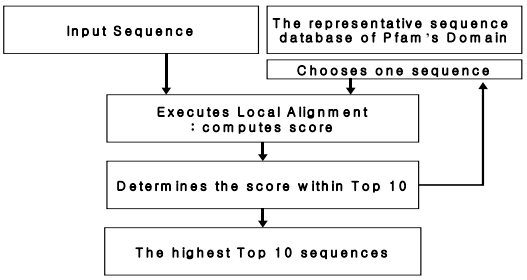
새롭게 생성된 pfamA_create_model 테이블은 Pfam에서 제공하는 각 도메인을 대표하는 서열을 가진다. ①에서 생성한 pfam_score 테이블의 각 도메인의 대표 서열을 결정하는 domain_evalue_score 필드 값을 이용하여 pfamA_reg_full에서 대표 서열을 나타내는 auto_pfamseq를 가져온다. auto_pfamseq를 이용하여 ②에서 생성한 pfam_seqs 테이블에서 한 도메인(auto_pfamA)의 대표 서열을 구할 수 있다.

3.2.2. 서열 정렬 방법을 통한 유사성이 높은 단백질 서열 정보 추출 과정

단백질 서열 정렬 방법은 입력된 두 단백질 서열 사이의 유사성 정도를 측정하는데 사용되고 있다. 본 논문에서는 입력된 단백질 서열과 유사성이 높은 단백질 구조 정보(도메인 및 패밀리)를 대표하는 서열을 찾기 위하여 기존의 부분 정렬(local alignment) 방법[12]을 이용하였다.

[그림 5]는 입력 서열과 유사성이 가장 높

은 10개의 Pfam 도메인을 대표하는 서열을 추출하는 과정을 보여 주고 있다. 10개의 도메인은 입력 서열과 관련된 SCOP 분류 정보를 추출하는데 현재로써는 충분한 개수로 추정된다. 추출되는 도메인 수를 작게 할 경우, 입력된 단백질과 구조적으로 유사한 단백질을 제외시키는 경우가 발생할 수 있다.



(그림 5) 서열 정렬 과정

입력 서열과 유사성이 높은 Pfam의 도메인을 대표하는 10개의 서열을 추출하여 SCOP에서 구조 분류 정보를 추출할 수 있는 pfamA_acc 값을 얻는다. pfamA_acc 필드 값은 Pfam에서 제공하는 도메인 및 패밀리와 관련된 PDB 데이터를 추출하는데 사용된다.

3.2.3. SCOP 단백질 구조 분류 정보 추출 과정

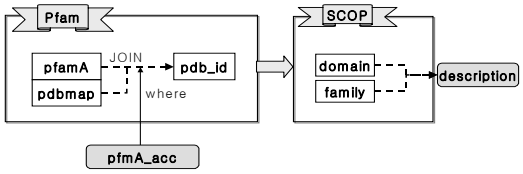
입력 서열과 연관된 SCOP의 단백질 분류 정보는 Pfam에서 제공하는 각 도메인 및 패밀리에 대한 정보를 바탕으로 얻어진다.

<표 2>는 Pfam에서 제공하는 도메인 정보 중에서 다른 생물학적 데이터베이스의 연결을 위한 정보를 나타내고 있다.

<표 2> Pfam의 도메인 및 패밀리 정보

종 류	설 명
Smart	도메인에 대한 HMM 정보
PDB ID	도메인이 발견되는 PDB 데이터
Interpro	도메인에 대한 기능 설명

<표 2>에서 기술하는 Pfam의 도메인 및 패밀리의 정보 중에서 PDB ID는 입력된 단백질 서열과 구조적으로 비슷한 SCOP 구조 분류 정보를 추출하는데 사용된다.

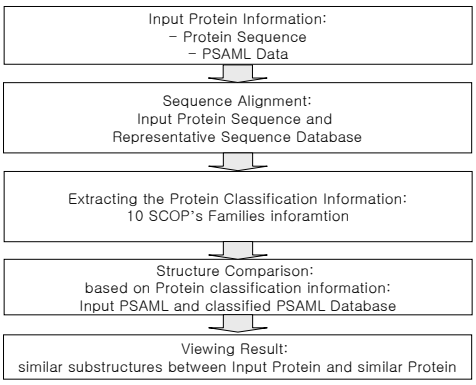


(그림 6) SCOP 분류 정보 추출

[그림 6]은 Pfam과 SCOP 데이터베이스를 이용하여 입력된 서열과 연관된 단백질 구조 분류 정보를 추출하는 과정을 보여주고 있다. 서열 정렬을 통하여 나온 10개의 유사성이 높은 서열의 pfamA_acc 값을 이용하여 pfamA와 pdbmap 테이블을 이용하여 PDB ID를 얻어낸다. pfamA 테이블은 도메인과 연관된 서열에 대한 정보를 기술하며, pdbmap 테이블은 Pfam에서 도메인 및 패밀리와 PDB 데이터베이스에서 제공하는 단백질 사이의 연결 관계를 나타낸다. SCOP 데이터베이스는 PDB 데이터베이스에서 제공하는 모든 단백질에 대한 구조 분류 정보를 제공한다. 이러한 구조 분류 정보는 PDB ID를 통하여 추출될 수 있다.

4. SCOP 기반의 단백질 구조 비교 방법

SCOP에서 제공하는 단백질 구조 분류 정보를 이용한 PSAML 기반 구조 비교 방법은 입력된 단백질 서열과 연관된 SCOP 구조 분류 정보를 추출하고 이 분류에 속한 모든 단백질과 구조 비교를 수행한다.



(그림 7) 구조 분류 기반 단백질 구조 비교

[그림 7]은 SCOP에서 제공하는 단백질 구조 분류 정보를 이용하여 입력된 단백질 구조와 유사한 부분 구조를 가진 단백질을 찾는 전체과정을 보여주고 있다.

4.1. PSAML 기반의 단백질 구조 데이터베이스

PSAML은 XML 기술을 이용하여 단백질 구조 정보를 나타내는 언어로써, 이를 이용하여 기술된 단백질 구조 정보는 XML 파일로써 저장된다.

PSAML 기반의 단백질 구조 데이터베이스는 XML 데이터베이스인 Apache Xindice[13]를 이용하여 효과적으로 SCOP에서 제공하는 분류 정보를 기반으로 PSAML 문서를 저장하고 검색할 수 있는 방법을 제공하고 있다.

2003년 2월 현재 SCOP에서는 1,940개의 패밀리 구조 분류 정보를 가지고 있다. PSAML 기반 단백질 구조 데이터베이스는 제공된 SCOP 분류 정보를 Xindice에서 제공하는 컬렉션(collection)으로 정의하고 여기에 분류된 PSAML 형식의 단백질 구조 정보를 저장한다. 하나의 컬렉션에 속한 단백질 구조 정보는 Xindice에서 제공하는 질의 방법을 통하여 비교적 용이하게 접근할 수 있다.

4.2. SCOP 단백질 구조 분류 기반 단백질 구조 비교 과정

SCOP 단백질 구조 분류 기반 단백질 구조 비교 방법은 자바[14] 언어를 이용하여 구현되었으며, [그림 8]과 같은 유사성 그래프[15]에서 최대 유사 부분 구조는 Maximal Clique를 찾는 방법[16]을 이용하여 구할 수 있다.

다음은 SCOP 단백질 구조 분류 기반 단백질 구조 비교 방법이다.

① 입력된 단백질 서열과 Pfam의 도메인 대표 서열 데이터베이스와 서열 정렬을 수행하여 가장 유사한 서열 10개를 추출한다.

② 추출된 각각의 서열이 대표하는 Pfam의 도메인 정보를 알아낸다.

③ ②에서 추출된 각 도메인에 속한 PDB ID를 이용하여 Pfam 도메인에 대한 각각의 SCOP 구조 정보(family)를 알아낸다.

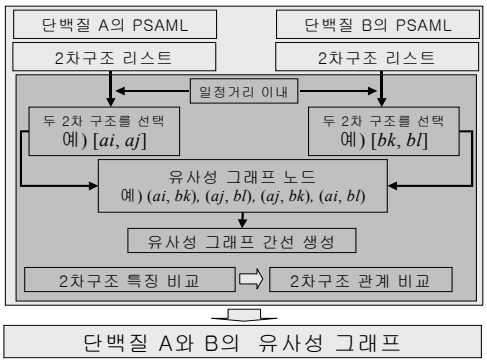
④ 추출된 SCOP 구조 분류를 사용자에게 보여주고, 사용자로부터 단백질 구조 비교를

수행할 구조 분류를 입력받는다.

⑤ 선택된 SCOP 구조 분류에 속한 모든 단백질 구조(PSAML)와 입력된 단백질 구조(PSAML)사이에 유사한 부분 구조를 찾아낸다.

⑥ 찾은 최대 유사 부분구조를 보여 준다.

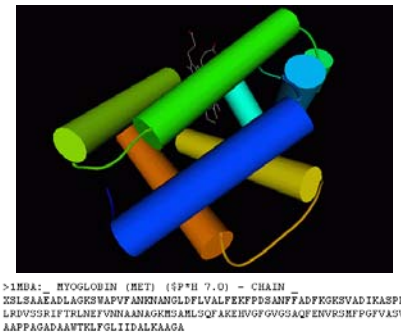
⑤, ⑥과정에서 수행되는 두 단백질 구조 사이의 구조 비교 방법은 [5]에 보다 자세히 기술되어 있다.



(그림 8) 유사성 그래프 생성 방법

4.3. 단백질 구조 비교 결과

SCOP 구조 분류 정보 기반 단백질 구조를 비교하는 방법을 이용하여 PDB ID 1MBA([그림 9])을 이용하여 수행하였다. 1MBA는 8개의 α-나선으로 이루어진 단백질로써 Myoglobin 계열에 속하며, 산소를 저장하는 기능을 담당한다.



(그림 9) 1MBA의 구조와 서열

<표 3>은 입력된 1MBA의 서열을 Pfam의 도메인을 대표하는 서열 데이터베이스를 대상으로 서열 정렬을 수행하여 유사성이 높은 Pfam 도메인 및 PDB ID를 보여주고 있다.

<표 3> 입력서열과 유사한 Pfam 도메인

유사성 순위	도메인	PDB ID	
		종류	총갯수
1	Globin	101m, 1a6m, ..., 1mbc ,..	311
2	Condensation	없음	0
3	pp-binding	1acp, 1af8, 1dv5, ...	5
4	ALA_synthase	1bs0, 1dj9, 1dje	3
5	NB-ARC	없음	0
6	mRNA_cap_C	1ckm, 1ckn, 1cko	3
7	ACR_tran	없음	0
8	CheR	1af7, 1bc5	2
9	CBM_15	1gny	1
10	Surface_Ag_2	없음	0

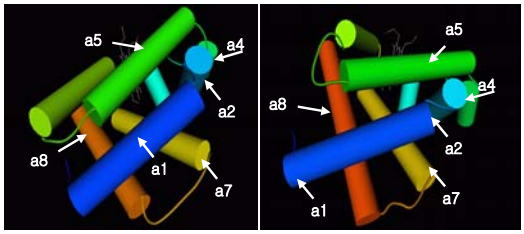
Pfam에서 정의된 도메인 정보에 PDB ID 정보가 없는 경우, SCOP 구조 분류 정보를 추출할 수 없으므로 다른 도메인 정보를 참조로 SCOP 구조 분류 정보를 추출한다.

<표 4>는 입력된 1MBA의 단백질 서열과 관련성이 있는 SCOP 분류 정보를 PDB ID를 이용하여 추출한 결과이다.

<표 4> PDB ID와 SCOP 구조 분류 정보

PDB ID	SCOP 구조 분류 정보
1a6m	Globins
1dv5	apo-D-alanyl carrier protein
1af8	Acyl-carrier protein (ACP)
1dj9	omega-Amino acid:pyruvate amino transferase-like
1ckm	RNA guanylyltransferase, N-terminal domain
1af7	Chemotaxis receptor methyltransferase CheR, C-terminal domain
1gny	CBM15
1f80	Holo-(acyl carrier protein) synthase ACPS
1bs2	Arginyl-tRNA synthetase, N-terminal 'additional' domain
1cqx	Ferredoxin reductase FAD-binding domain-like

[그림 10]과 <표 5>는 PSAML 기반 단백질 구조 비교 방법[5]을 이용하여 SCOP 분류에 속한 단백질 구조와 비교한 결과를 보여주고 있다.



(a) 1MBA 단백질 구조 (b) 1MBC 단백질 구조

(그림 10) 유사한 단백질 2차구조들

<표 5> 일치된 이차구조

SCOP 분류	ID	일치된 이차구조					
Globins	1MBA	a1	a2	a4	a5	a7	a8
	1MBC	a1	a2	a4	a5	a7	a8

● 기존 단백질 구조 비교 방법과 비교

추출된 SCOP 분류 정보가 올바른 단백질 비교 대상을 포함하고 있는지 파악하기 위하여 Topscan[17]과 비교하였다. Topscan은 입력된 단백질 구조와 유사한 단백질을 찾기 위하여 단백질 데이터베이스에 저장된 모든 단백질 구조와 비교한다.

Topscan을 실행한 결과, 1MBA와 비슷한 구조(60% 이상의 상동성)를 가진 단백질들은 <표 4>에서 제시한 1MBA의 서열을 이용하여 추출된 SCOP 분류 정보인 *'Globins'*에 속하였다. 제안된 단백질 비교 방법에서 추출한 단백질 구조 분류 정보는 Topscan에서 찾은 입력 단백질과 유사성 정도가 높은 단백질들을 모두 포함하고 있음을 알 수 있다.

5. 결론 및 향후 연구 방향

본 논문에서는 PSAML 기반의 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여 단백질 폴드(Fold) 또는 패밀리(Family)를 대표하는 아미노산 서열 정보(Representative Sequences) 및 단백질 구조 분류 정보(SCOP, Pfam)를 활용하는 방법에 대하여 기술하였다.

SCOP은 현재까지 밝혀진 단백질 구조 분류 정보를 제공하고 있으며, Pfam은 많은 단백질에서 자주 나타나는 도메인에 대한 정보를 제공하고 있다. 개선된 PSAML 기반의 단백질 구조 비교 방법은 단백질 아미노산 서열을 이용하여 서열상 유사성이 높은 단백질 구조(Domain)를 필터링하고, 그 결과를 이용하여 추출된 단백질 구조 분류에 속한 모든 단백질 구조와 비교하도록 하여 준다.

향후 PSAML 기반으로 기술된 단백질 구

조 비교 서비스를 단백질 구조 정보를 필터링 하는 방법을 적용하여 웹 정보 시스템을 통하여 서비스할 수 있도록 할 계획이다. 그리고, 제한 프로그래밍 기법을 이용하여 보다 효과적으로 비교할 수 있도록 할 예정이다.

감사의 글

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

참고문헌

[1] I Eidhammer, I Jonassen, W R. "Structure Comparison and Structure Patterns", *Reports in Informatics*, 7, 1999.

[2] Su-Hyun Lee, Jin-Hong Kim, Geon-Tae Ahn, and Myung-Joon Lee, "An XML Representation of Protein Data for Efficient Structure Comparison", *Second ICIS*, No. 1, pp. 313, 2002.

[3] 김진홍, 안건태, 이수현, 이명준, "구조비교를 위한 단백질 데이터의 XML 표현 기법", *한국정보과학회 프로그래밍언어연구회*, 16권, 2호, 15-16, 2002.

[4] D. C. Fallside, "XML Schema Part 0: Primer", *W3C*, May 2001.

[5] 김진홍, 안건태, 조민수, 이수현, 이명준, "PSAML를 기반으로 한 단백질 구조 비교", *한국정보과학회 프로그래밍언어연구회*, 16권, 3호, 33-44, 2002.

[6] 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.

[7] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL, "The Pfam Protein Families Database", *Nucleic Acids Research*, 30(1), 276-280, 2002

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

[9] 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.

[10] R. Hughey and A. Krogh, "Hidden Markov models for sequence analysis: Extension and analysis of the basic method", *CABIOS*,12(2), 95-107, 1996

[11] H. M. Berman, J. D. Westbrook, Z. F- eng, 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.

[12] Durbin et al, "Biological Sequence Analysis", *CUP*, 1998, http://www.dina.kvl.dk /~sestoft/bsa.html

[13] "The version 1.0 of Xindice", *The Apache Software Foundation*, 2003, http://xml.apache.org/xindice/

[14] John Lewis, William Loftus, "Java Software Solutions Fundations of Program Design", *ADDISON-WESL- EY*, 1998

[15] Hiroaki KATO and Yoshimasa TAK-

AHASHI, "Automated Identification of Three-Dimensional Common Structural Features of Proteins", *J. Chem. Software*, Vol. &, No. 4, p. 161-170, 2001.

[16] Sampo Niskanen, Patric Ostergard, "Cliquer:routines for clique searching", 2002, http://www.hut.fi/~pat/cliquer.html

[17] Martin, A. C. R. "The Ups and Downs of Protein Topology; Rapid Comparison of Protein Structure", *Protein Engineering*, 13, 829-837, 2000

김진홍

1999년 2월 울산대학교 전자계산학과 졸업(학사)
2001년 2월 울산대학교 컴퓨터정보통신 공학부 졸업(석사)
2001년 3월 ~ 현재 울산대학교 컴퓨터정보통신공학부 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.

변상희

2003년 2월 울산대학교 전자계산학과 졸업(학사)
2003년 3월 ~ 현재 울산대학교 컴퓨터정보통신공학부 석사과정
관심분야 : 생물정보학, 분산시스템

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 컴퓨터정보통신공학부(교수)
1993년 8월~1994년 7월 미국 버지니아대학 교환교수
관심분야 : 프로그래밍언어, 분산 객체 프로그래밍 시스템, 병행 실시간 컴퓨팅, 인터넷 프로그래밍시스템, 생물정보학 등.