PSAML과 Topology String 데이터베이스를 이용한 웹 기반 단백질 구조 비교 시스템

김진홍[O*] 안건태* 변상희* 이수현** 이명준*
*울산대학교 컴퓨터정보통신공학부, **창원대학교 컴퓨터공학과
*{avenue[O], java2u, heeya, mjlee}@ulsan.ac.kr
**suhyun@sarim.changwon.ac.kr

A Web-Based Protein Comparison System Using PSAML and Topology String Databases

Jinhong Kim[O*] Geontae Ahn* Sanghee Byun* Suhyun Lee** Myungjoon Lee*
*School of Computer Engineering Information Technology, University of Ulsan
**Dept. of Computer Science, Changwon National University

요 약

단백질의 기능은 단백질의 구조에 따라 결정되며, 새로운 단백질의 기능을 파악하기 위하여 이미 밝혀진 단백질의 기능과 구조를 비교하는 방법이 사용되고 있다. 단백질 구조를 비교하는 방법은 단백질 구조를 표현하는 방법에 따라 다양하게 개발되고 있으며, 보다 효과적으로 관련된 연구자들이 자신의 연구에 활용하기 위해서는 빠르고 쉽게 활용할 수 있는 인터페이스를 제공하는 도구가 필요하다.

본 논문에서는 PDB 데이터베이스에서 제공하는 단백질 정보를 이용하여 PSAML 및 Topology String 데이터베이스를 구축하고 이를 바탕으로 웹 기반에서 단백질 구조 비교를 보다 빠르고 효과적으로 수행하는 시스템에 대하여 기술한다. PSAML 데이터베이스는 단백질 구조를 단백질 이차구조 및 그들 사이의 관계를 포함하는 PSAML 데이터를 제공하며, Topology String 데이터베이스는 단백질 구조를 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적(공간적) 정보를 포함하는 문자열로 단백질 구조 정보를 제공한다. 이를 이용하여 구축된 웹 기반 단백질 구조 비교 시스템은 Topology String 정렬 방법을 통하여 보다 빠르게 유사성이 높은 부분 구조를 찾는 방법을 제공한다.

1. 서 론

단백질 구조 비교 방법은 단백질의 구조적인 특징에 따라 단백질 구조를 분류하고 공통의 부분 구조를 찾아내는데 활용되고 있으며, 대표적인 단백질 구조 비교 알고리즘에는 단백질 구조의 내부 분자들 사이의 거리 정보를 동적 프로그래밍 기법을 이용한 DALI[1], Cα 원자들 사이에 RMSD가 최소가 되는 부분을 찾는 LOCK[2], 단백질 이차구조의 공간상의 정보를 바탕으로 하는 기하학적 해싱 기법 사용하는 3dSEARCH[3], 그리고 단백질 이차구조 사이의 거리 및 각도 관계를 이용한 SARF2[4] 등이 있다.

현재 새롭게 밝혀지는 단백질 3차구조 정보의 증가량이 날이 갈수록 높아짐에 따라 단백질 구조 비교 방법은 보다 효과적으로 빠르게 결과를 산출할 수 있어야 한다. 이를 위하여 새로운 단백질 구조 표현 방법이 개발되어야 하고, 구조 비교 시 요구되는 많은 데이터를 효과적으로 처리할 수 있도록 새로운 방법이 개발되어야 한다.

PSAML(Protein Structure Abstraction Markup Language)[5]은 단백질의 이차구조와 이차구조 사이에서 발견되는 상호 관계를 이용하여 단백질 구조를 표준화된 문서 표현 양식인 XML로 기술할 수 있는 언어이며, Topology String은 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적(공간적) 정보를 포함하는 문자열로 표현하는 단백질 구조 표현 방법이다.

본 논문에서는 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여, PSAML 및 Topology String 데이터베이스를 구축하는 방법과 이를 바탕으로 단백질 구조를 비교하여 유사한 부분구조를 찾는 웹 기반 단백질 구조 비교 시스템에 대하여 기술한다. Topology String 데이터베이스는 대용량의 단백질 구조 데이터베이스를 바탕으로 하는 단백질 구조 비교 방법에서 비교 대상의 단백질의 수를 줄이는데 유용하게 사용되었다. 개발된 웹 기반 단백질 구조 비교 시스템은 웹 기반의 편리한 인터페이스를 제공하며, 단백질 이차구조와 그들 사이의 관계(각도, 거리, 길이)를 이용하여 보다 빠르게 유사한 부분 구조를 찾을 수 있는 기능을 제공한다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 PSAML 및 Topology String 데이터베이스의 정의 및 그 생성 방법에 대하여 기술한다. 3장에서는 구축된 데이터베이스를 바탕으로 단백질 구조를 비교하는 웹 기반 시스템에 대하여 살펴본다. 마지막으로 4장에서는 결론 및 향후 연구 방향으로 끝을 맺고자 한다.

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

2. 단백질 구조 데이터베이스

2.1 PSAML 데이터베이스

PSAML 데이터베이스는 PDB 데이터베이스에서 제공하는 데이터를 개발된 변환기에 의하여 PSAML 형태의 데이터를 제공한다.

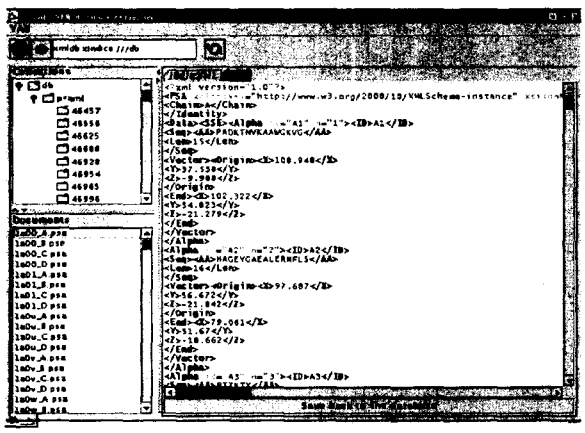
하나의 단백질 구조를 PSAML로 표현하기 위하여, 단백질을 구성하는 2차구조의 집합, 각 이차구조의 정보(이차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보), 그리고 <표 1>에서 제시된 이차구조 사이에 정의되는 관계 정보가 이용된다.

<표 1> 이차구조 사이의 관계

관계	의미	표 현
θ	각도	$\theta(Ei, Ej) = angle(\theta 1)$
γ	거리	$\gamma(Ei, Ej) = distance(D)$
v	길이차	$v(Ei, Ej) = length(li, lj)$
h	수소결합	$h(Ei, Ej)$ = {E, N}, Ei와 Ej는 β-strand
d	방향성	$d(Ei, Ej)$ = {P, A}, Ei와 Ej는 β-strand

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(보다 자세한 PSAML의 XML 스키마는 [5] 참조).

PSAML 데이터베이스는 XML 데이터베이스인 Xindice[6]를 이용하고 있으며, 효과적인 단백질 구조 비교를 지원하기 위하여 PSAML 데이터를 SCOP 분류 정보의 폴드(Fold) 정보를 바탕으로 분류하여 저장하고 있다. SCOP의 같은 폴드에 속한 단백질들은 매우 유사한 이차구조의 위상학적인 정보를 가지고 있다. (그림 1)은 구축된 PSAML 데이터베이스를 보여주고 있다.



[그림 1] PSAML 데이터베이스

2.2 Topology String 데이터베이스

Topology String은 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적(공간적) 정보를 포함하는 문자열로 표현하는 단백질 구조 표현 방법이다.

Topology String의 정의는 다음과 같다.

① TS(protein_id) = {t1, t2, ..., ti}, 단, protein_id는 단백질 식별자, ti는 topology 문자, i는 이차구조 개수이며 서열상의 순서

② ti = {V or D, A or M, E or F}, ti가 나선일 경우,
= {H or N, G or I, K or L}, ti가 판상조각일 경우,

③ ti의 값은 단백질 이차구조의 종류 및 위상학적 방향성에 따라 결정된다.<표 2>

④ 생성된 Topology String은 x축, y축, 그리고 z축을 기준으로 각 90° 변환하여 모두 24가지의 서로 다른 Topology String으로 변환된다.<표 3>

PSAML 데이터베이스에서 Topology String 데이터베이스를 생성하는 과정은 다음과 같다.

① 선택된 PSAML에서 아미노산 서열 순서에 의하여 이차구조를 하나씩 선택한다(선택된 이차구조는 3차원상의 벡터로 표현되어 있다). 선택된 이차구조의 시작점을 원점으로 평행 이동시키고 이차구조의 끝점이 위치한 공간상의 위치와 이차구조의 종류에 따라 <표 2>에 제시된 문자로 변환하여 Topology String을 생성한다.

② ①에서 생성된 Topology String을 <표 3>에서 제시된 변환표를 참조하여 23가지의 서로 다른 Topology String을 생성한다.(각축으로 90°의 방향으로 변환)

③ ①과 ②에서 생성된 총 24개의 Topology String을 FASTA 형태로 저장한다.

④ 저장된 데이터를 NCBI Blast 프로그램을 이용하여 서열 정렬을 할 수 있도록 Topology String 데이터베이스를 구축한다.

<표 2> 이차구조 변환 규칙

위상학적 방향성		단백질 이차구조	
		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 3>회전에 따른 변환 규칙

+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

3. 웹 기반 단백질 구조 비교 시스템

웹 기반 단백질 구조 비교 시스템은 Topology String 데이터베이스를 이용하여 입력된 단백질과 유사한 단백질 분류 정보를 추출함으로써 비교 대상이 되는 단백질의 수를 줄여 보다 빠르게 비교 결과를 제공한다.

3.1 전체 시스템 구조 및 비교 수행 과정

웹 기반 단백질 구조 시스템은 다음과 같은 구성요소를 가진다.

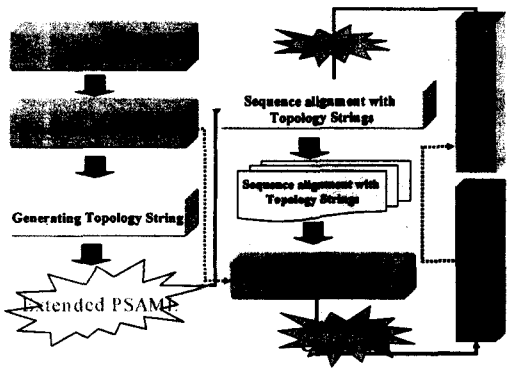
① *인터페이스 모듈*: 데이터(PSAML과 서열정보)를 입력하고 비교 결과를 사용자에게 보여주는 기능을 제공한다.

② *필터링 모듈*: Topology String을 이용하여 PSAML 데이터베이스에서 비교 대상 단백질의 수를 줄여 보다 빠른 비교를 수행할 수 있는 기능을 제공한다. 입력된 단백질의 PSAML에서 Topology string을 생성하고 NCBI의 Blast 프로그램[7]을 이용하여 Topology string 데이터베이스를 대상으로 서열 정렬을 수행한다. 서열상 상동성이 높은 단백질이 속한 SCOP 폴더 정보를 추출한다.

③ *구조 비교 모듈*: ②의 과정에서 추출된 단백질들과 입력 단백질 사이에 유사한 부분 구조를 찾는 기능을 제공한다.(보다 자세한 과정은 [8]을 참조)

④ *PSAML 및 Topology String 데이터베이스*: PDB에서 제공되는 단백질 구조 데이터를 개발된 PSAML 변환기 및 Topology String 변환기를 통하여 각 데이터베이스를 생성된다. 구축된 데이터베이스는 필터링 모듈에서 비교 대상 단백질의 수를 줄이는데 사용된다.

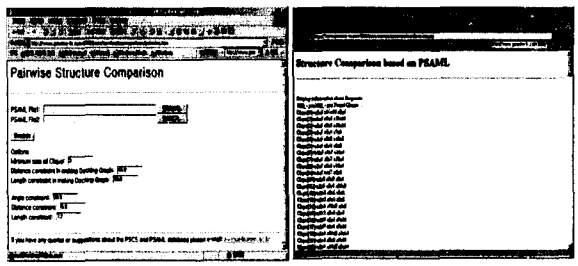
(그림 2)는 웹 기반 단백질 구조 비교 시스템에서 유사한 부분 구조를 찾는 과정을 보여주고 있다.



[그림 2] Topology String을 이용한 구조 비교 과정

3.2 Topology String 기반 단백질 구조 비교 결과

Topology string 기반의 단백질 구조 비교 방법을 평가하기 위하여 1MBA와 유사한 부분 구조를 찾는 실험을 수행하였다. 1MBA의 PSAML에서 생성된 Topology string은 "VFDEMFVF"이다. 그리고 Blast를 이용하여 서열의 상동성을 검사한 결과, 5MBA, 2FAM, 2FAL 등이 추출되었다. (그림 3)은 입력된 PSAML 데이터에서 부분적으로 유사한 구조를 찾은 결과를 보여주고 있다.



[그림 3] 웹 기반 단백질 구조 비교 시스템 실행 결과

4. 결론 및 향후과제

본 논문에서는 PDB 데이터베이스에서 제공하는 단백질 정보를 이용하여 PSAML 및 Topology String 데이터베이스를 구축하고 이를 바탕으로 웹 브라우저에서 단백질 구조를 보다 빠르고 효과적으로 비교할 수 있는 시스템에 대하여 기술하였다.

개발된 웹 기반 단백질 구조 비교 시스템은 PSAML 데이터에서 Topology String을 생성하고 이들 사이의 서열 정렬 방법을 통하여 PSAML 데이터베이스에서 비교 대상 단백질의 수를 줄일 수 있는 필터링 방법을 제공한다. 그리고 제안된 시스템은 사용자에게 편리한 웹 인터페이스를 제공하여 쉽게 단백질 이차구조 및 그들 사이의 관계(각도, 거리, 길이)를 이용하여 보다 빠르게 유사한 부분 구조를 찾을 수 있는 기능을 제공한다.

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

[참고문헌]

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

[2] 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.

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

[4] 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.

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

[6] M. Liotta, "Apache's Xindice Organizes XML Data Without Schema," 2003(http://www.devx.com/xml/article)

[7] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res. 25:3389-3402, 2001.

[8] 김진홍, 안건태, 이수현, 이명준, "단백질 이차 구조 기반의 단백질간 구조 비교", 한국정보과학회, 2002 가을 학술발표논문집(B) 제 29권 2호, pp. 613-615, 2002.