이차구조요소 기반의 부분구조 검색을 위한 단백질 구조 비교 시스템

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

Protein Structure Comparison System for Searching Substructures Based on Secondary Structure Elements

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 단백질 이차구조 및 그들 사이의 관계를 이용하여 단백질 구조를 표현하는 PSAML과 이를 이용하여 표현된 단백질 구조를 비교하는 시스템인 S4E(Search Substructures of Secondary Structure Elements)에 관하여 기술한다. S4E 시스템은 단백질 이차구조와 그들 사이의 관계(각도, 거리, 길이)를 이용하여 표현된 단백질 구조를 비교하여 유사성이 높은 부분을 찾는 기능을 제공한다. 또한 S4E 시스템은 이차구조 기반의 단백질 구조 데이터베이스(PSAML 데이터베이스) 및 웹 기반 사용자 인터페이스를 제공하여 사용자가 쉽고 효과적으로 단백질 구조 비교를 할 수 있다.

1. 서 론

단백질 구조 비교 방법은 단백질의 구조적인 특징에 따라 단백질 구조를 분류하고 공통의 부분 구조를 찾아내는데 활용되고 있으며, 새로운 단백질의 기능을 파악하기 위하여 유용하게 사용되고 있다. 이러한 단백질 구조 비교 방법은 단백질 구조를 표현하는 방법 및 유사한 구조를 파악하는 방법에 따라 다양하게 존재한다.[1]

대표적인 단백질 구조 알고리즘은 단백질 구조의 내부 분자들 사이의 거리 정보를 동적 프로그래밍 기법을 이용한 DALI[2], Cα 원자들 사이에 RMSD가 최소가 되는 부분을 찾는 LOCK[3], 단백질 이차 구조의 3차원 위치 정보를 유사 부분을 찾기 위하여 기하학적 해싱 기법 사용하는 3dSEARCH[4], 그리고 단백질 이차구조 사이의 거리 및 각도 관계를 이용한 SARF2[5] 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 효과적인 단백질 구조 비교를 위한 단백질 구조 표현 방법이 개발되어야 하고, 구조 비교 시 요구되는 많은 데이터를 효과적으로 처리할 수 있도록 새로운 방법이 개발되어야 한다.

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

PSAML(Protein Structure Abstraction Markup Language)[6,7]은 단백질의 2차구조와 2차구조 사이에서 발견되는 상호 관계를 이용하여 단백질 구조를 표현하는 방법을 제공하는 PSA(Protein Structure Abstraction)를 기준으로 단백질 구조를 표준화된 문서 표현 양식인 XML로 기술할 수 있는 언어이다.

본 논문에서는 단백질 이차구조 및 그들 사이의 관계를 이용하여 단백질 구조를 표현하는 PSAML를 이용하여 입력된 단백질 구조와 유사성이 높은 부분구조를 찾는 구조 비교 시스템인 S4E(Search Substructures of Secondary Structure Elements)에 관하여 기술한다. S4E 시스템은 단백질 이차구조와 그들 사이의 관계(각도, 거리, 길이)를 이용하여 표현된 단백질 구조를 비교하여 유사성이 높은 부분을 찾는 기능을 제공하며, 이차구조 기반의 단백질 구조 데이터베이스(PSAML 데이터베이스) 및 웹 기반 사용자 인터페이스를 제공하여 사용자가 쉽고 효과적으로 단백질 구조 비교를 할 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 PSAML 기반의 단백질 구조 표현 방법에 대하여 기술한다. 3장에서는 PSAML 기반의 단백질 구조 비교 시스템인 S4E과 그 결과에 대하여 살펴본다. 마지막으로 4장에서는 결론 및 향후 연구 방향으로 끝을 맺고자 한다.

2. 이차구조 기반의 단백질 구조 표현

2.1 PSA

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.

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

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.

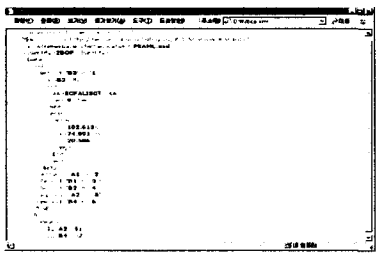
<표 1> 이차구조 사이의 관계

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

2.2 PSAML

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.



(그림 1) PSAML 문서의 예

3. 단백질 구조 비교 시스템

단백질 구조 비교 시스템인 S4E 시스템은 PSAML을 이용하여 표현된 단백질 구조를 비교하는 기능을 제공하며, PDB 데이터베이스에 포함된 단백질 구조를 PSAML 형식의 단백질 구조 표현을 저장하고 있는 PSAML 데이터베이스를 제공하여 입력된 단백질 구조와 유사성이 높은 부분 구조를 가진 단백질 구조를 찾는 기능을 제공한다. 또한, S4E 시스템은 단백질 구조 비교를 쉽게 이용할 수 있도록 웹 인터페이스를 제공한다.

3.1 단백질 구조 비교 시스템의 구조

S4E 시스템은 사용자의 서비스 요청 및 구조 비교 결과를 보여주는 인터페이스 모듈, 단백질 서열상 유사성이 높은 구조를 필터링하는 필터링 모듈, PSAML 데이터를 가지고 단백질 구조 비교를 수행하는 구조비교 모듈, 그리고 PSAML 형식으로 기술된 단백질 구조 데이터베이스인 PSAML 데이터베이스로 구성되어 있다.

① *인터페이스 모듈*
사용자로부터 비교하려는 단백질 구조 표현 정보(PSAML과 서열정보)와 구조 비교에 사용되는 인자를 입력하는 웹 인터페이스를 생성하는 기능을 제공한다. 또한 구조 비교 모듈의 실행 결과를 보여 주는 웹 페이지를 생성한다.

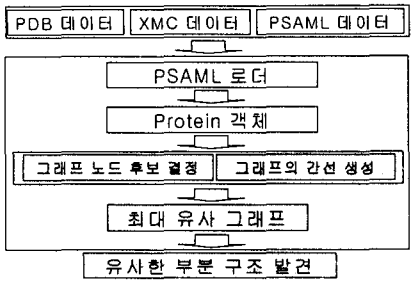
② *필터링 모듈*
사용자로부터 입력된 단백질 서열 정보를 이용하여 단백질 서열상 가장 유사성이 많은 SCOP에서 정의한 Family 정보를 추출한다. 추출된 Family에 속한 모든 단백질과 사용자가 입력한 단백질 구조와 비교를 수행한다.

③ *구조 비교 모듈*
PSAML로 기술된 단백질 구조를 비교하는 기능을 제공한다. 두 단백질사이의 유사성이 높은 부분 구조를 찾는 기능 및 PSAML 데이터베이스에 저장된 단백질 구조와 비교하는 기능을 제공한다. 단백질 구조를 비교하는 과정은 3.2절에서 보다 자세히 기술하고 있다.

④ *PSAML 데이터베이스*
PSAML 데이터베이스는 XML 데이터베이스인 Apache Xindice를 이용하여 효과적으로 SCOP에서 제공하는 분류 정보를 기반으로 PSAML 문서를 저장하고 검색할 수 있는 방법을 제공하고 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.

3.2 단백질 구조 비교 과정

S4E 시스템의 *구조 비교 모듈*에서 제공하는 단백질 구조 비교 과정은 (그림 2)와 같다.



(그림 3) 단백질간 구조 비교 방법

입력되는 PDB 형태의 단백질 구조 데이터는 구현된 변환 도구를 통하여 PSAML 문서로 변환된다. 단백질 구조 비교를 수행하는 방법은 변환된 PSAML을 읽어 Protein 객체를 생성한다. 생성된 Protein 객체는 PSAML에서 정의하는 2차구조의 특징(아이디, 타입, 3차원 좌표)과 2차구조 사이의 관계(각도, 거리, 길이 차이, 수소결합 유무, 방향성)에 대한 정보를 가진다. 생성된 두 Protein 객체로부터 단백질 간 유사성을 내포하는 유사성 그래프를 생성하여 모든 노드들 사이의 간선이 존재하는 부분 그래프를 찾는 알고리즘을 이용하여 최대로 유사한 부분 구조를 파악할 수 있다.

○ 유사성 그래프의 정의

PSAML 데이터를 기반으로 단백질 구조간의 유사성을 내포하는 유사성 그래프 G는 <표 2>와 같이 정의된다.

<표 2> 유사성 그래프 정의

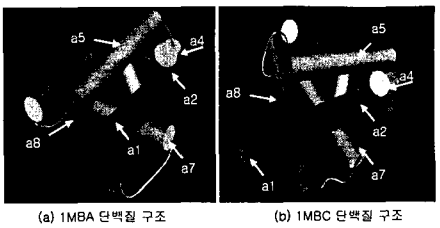
$G(A, B) = \{V, E\}$, A, B는 단백질 구조
$V = \{(ai, bj) \mid ai \in A, bj \in B\}$
$E = \{[(ai, bk),(aj, bl)] \mid$ 2차구조 유사성 비교$\}$

<표 2>에서, 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.

3.3 단백질 구조 비교 결과

S4E 시스템을 이용하여 PDB ID 1MBA와 단백질 구조 비교를 수행하였다. 1MBA는 8개의 α-나선으로 이루어진 단백질로써 Myoglobin 계열에 속하며, 산소를 저장하는 기능을 담당한다.

[그림 3]과 <표 3>는 S4E 시스템을 이용하여 PDB ID 1MBA와 1MBC 단백질 구조에서 찾은 부분 구조를 보여주고 있다.



(그림 4) 유사한 단백질 2차구조들

<표 3> 일치된 이차구조

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

4. 결론 및 향후과제

본 논문에서는 PSAML 기반의 단백질 구조 비교 시스템인 S4E에 대하여 기술하였다.

S4E 시스템은 단백질 이차구조 및 그들 사이의 관계를 이용하여 단백질 구조를 표현하는 PSAML를 이용하여 입력된 단백질 구조와 유사성이 높은 부분구조를 찾는 서비스를 제공한다. 제안된 구조 비교 시스템은 단백질 이차구조와 그들 사이의 관계(각도, 거리, 길이)를 이용하여 표현된 단백질 구조를 비교하여 유사성이 높은 부분을 찾는 기능을 제공하며, 이차구조 기반의 단백질 구조 데이터베이스(PSAML 데이터베이스) 및 웹 기반 사용자 인터페이스를 제공하여 사용자가 쉽고 효과적으로 단백질 구조 비교를 할 수 있다.

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

[참고문헌]

[1] I Eidhammer, I Jonassen, W R. "Structure Comparison and Structure Patterns", Reports in Informatics, 7, 1999.
[2] L. Holm and C. Sander, "Protein structure comparison by alignment of distance matrices," Journal of Molecular Biology, Vol. 233, pp. 123-138, 1993.
[3] 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.
[4] A. P. Singh and D. L. Brutlag, "Protein Structure Alignment: A Comparison of Methods", 1999.
[5] 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.
[6] 김진홍, 안건태, 변경익, 윤형석, 이수현, 이영준, "단백질 3차 구조의 추상적인 표현기법", 한국정보과학회, '2001 가을 학술발표논문집(B) 제 28권 2호, 595-597, 2001.
[7] 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.