단백질 이차 구조 기반의 단백질간 구조 비교

김진홍[0*], 안건태*, 이수현**, 이명준*
*울산대학교 컴퓨터·정보통신공학부, **창원대학교 컴퓨터공학과
*{avenue[0], java2u, mjlee}@ulsan.ac.kr,
**suhyun@sarim.changwon.ac.kr

Pairwise Protein Structure comparison based on Protein Secondary Structure

Jin-Hong Kim[0*] Geon-Tae Ahn* Su-Hyun Lee** Myung-Joon Lee*
*School of Computer Engineering & Information Technology, University of Ulsan
**Dept. of Computer Science, Changwon National University

요 약

단백질의 3차원 공간상의 구조는 단백질 기능을 파악하는데 중요한 정보를 제공하고 있다. 단백질간 구조 비교 방법은 기능적 또는 구조적으로 연관된 단백질 분류 및 단백질 모티프(motif)를 찾는데 유용하게 사용되고 있다. 본 논문에서는 단백질 이차 구조(α-나선구조와 β-병풍구조)와 그들 사이의 관계(각도, 거리, 길이, 수소결합)를 기반으로 표현된 두 단백질 구조에서 유사한 부분 구조를 찾는 방법에 대하여 기술한다. 제안된 단백질간 구조 사이의 유사한 부분구조를 찾는 방법은 두 단백질 구조를 이차 구조와 그들 사이의 관계를 이용하여 그래프를 형성한 후, 최대 유사 서브 그래프를 찾는 방법을 이용하여 유사한 부분구조를 찾을 수 있다.

1. 서 론

단백질 구조 비교 알고리즘은 단백질 구조 데이터베이스인 PDB(Protein Data Bank)[1] 데이터의 증대에 따라 단백질 기능 파악을 위하여 그 중요성이 커지고 있다.

단백질 구조를 비교하는 방법은 단백질 구조를 표현하는 방법에 따라 다양한 방법이 존재한다. 일반적인 단백질 구조 정렬 방법은 단백질 구조를 원자($C\alpha$) 또는 Residues를 기준으로 표현하고, 표현된 두 구조사이의 일치된 부분을 찾는 방법과 단백질 구조를 단백질 이차 구조 요소로 표현하고 표현된 두 단백질 구조를 정렬을 하는 방법으로 크게 구분된다.[2]

현재 대표적인 단백질 구조 알고리즘은 단백질 구조의 내부 분자들 사이의 거리 정보를 동적 프로그래밍 기법을 이용한 DALI[3], Cα 원자들 사이에 RMSD가 최소가 되는 부분을 찾는 LOCK[4], 단백질 이차 구조의 3차원 위치정보를 유사 부분을 찾기 위하여 기하학적 해싱(geometric hashing)기법 사용하는 3dSEARCH[5], 그리고 단백질 이차구조 사이의 거리 및 각도 관계를 이용한 SARF2[6] 등이 있다.

본 논문에서는 단백질 이차 구조 요소를 기반으로 표현[7,8]된 단백질간 구조를 비교 방법에 대하여 기술한다. 제안된 방법은 두 단백질 구조를 이차 구조와 그들 사이의 관계를 이용하여 그래프를 생성한 후, 최대 유사 서브 그래프[9]를 찾는 방법을 이용하여 유사한 부분구조를 찾을 수 있다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 구조 표현 방법인 PSA[8] 및 PSAML[8]에 대하여 살펴보고, 3장에서는 PSAML 기반의 두 단백질 구조를 비교하는 방법에 대하여 설명한다. 끝으로 4장에서는 결론 및 향후 연구방향에 대해여 기술한다.

2. 단백질 구조 표현

단백질 구조 및 유사성을 비교하기 위한 단백질 구조

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

에 대한 표현이 필요하다. PSA 및 PSAML은 단백질 구조 비교를 위한 표현으로 PDB 데이터베이스에서 제공하는 데이터를 기반으로 생성된다.

2.1 PSA(Protein Structure Abstraction)

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

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

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

S는 단백질을 구성하는 이차구조의 집합을 나타낸다. T, C, A는 각각 이차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보를 나타낸다. R은 두 이차구조사이에 정의되는 관계는 다음과 같이 표현된다.<표 1>

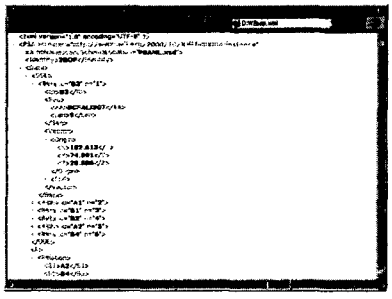
$$R = (\Theta, \gamma, v, h, d),\ 단,\ Ei, Ej \in S,\ i \neq j.$$

<표 1> 이차구조 사이의 관계

관계	의미	표 현
Θ	각도	$\Theta(Ei, Ej) = angle(\Theta I)$
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

2.2 PSAML(Protein Structure Abstraction Markup Language)

PSAML은 단백질 구조를 표현을 위한 PSA 표현을 XML로 표현하기 위하여 XML 스키마(XML schema)[10]를 이용하여 XML로 기술 할 수 있는 언어이다. (그림 1)은 PSA 기반의 PSAML 문서의 예를 보여주고 있다.



(그림 1) PSAML 문서의 예

3. 단백질 구조 비교 방법 및 프로그램

단백질 이차 구조 기반의 단백질간 구조 비교 방법은 주어진 PSAML 정보를 바탕으로 유사한 부분구조를 내포하는 그래프를 생성한 후, 기존의 Clique를 찾는 알고리즘[8]을 이용하여 최대로 유사한 부분 구조를 파악할 수 있다.

3.1 유사성을 내포한 그래프 생성 방법

PSAML 데이터를 기반으로 단백질 구조간의 유사성을 내포하는 그래프 G는 다음과 같이 정의된다.

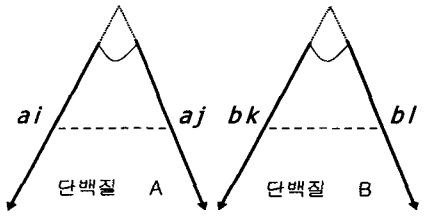
<표 2> 유사성을 내포한 그래프 표현

$G(A, B) = \{V, E\}$, A, B is 단백질 구조

$V = \{(ai, bj) \mid ai \in A, bj \in B\}$

$E = \{[(ai, bk), (aj, bl)] \mid T(ai)=T(bk),\ T(aj) = T(bl),$
$|\Theta(ai, aj)-\Theta(bk, bl)|<\Theta d,\ |\gamma(ai, aj)-\gamma(bk, bl)|<\gamma d$
$|v(ai, aj)-v(bk, bl)|<vd,\ h(ai, aj) = h(bk, bl)$
$d(ai, aj) = d(bk, bl)\ \}$

<표 2>에서, V와 E는 그래프 G의 노드 및 간선의 집합을 나타내고 있다. V에 속한 각 노드는 단백질 A의 한 이차 구조와 단백질 B의 한 이차 구조의 쌍으로 이루어져 있다. E에 속한 각 노드사이의 간선은 노드에 포함된 단백질 이차 구조간의 관계가 유사할 경우에 존재한다. 즉, 노드 (ai, bk)와 노드 (aj, bl) 사이의 간선은 이차구조 ai와 aj에 존재하는 관계와 bk와 bl에 존재하는 관계가 유사할 경우 생성된다. 이 경우, 단백질 A의 ai와 단백질 B의 bk 및 단백질 A의 aj와 단백질 B의 bl가 유사하다는 것을 의미한다. (그림 2)는 <표 1>에 기술된 노드에 포함된 이차 구조를 나타내고 있다.

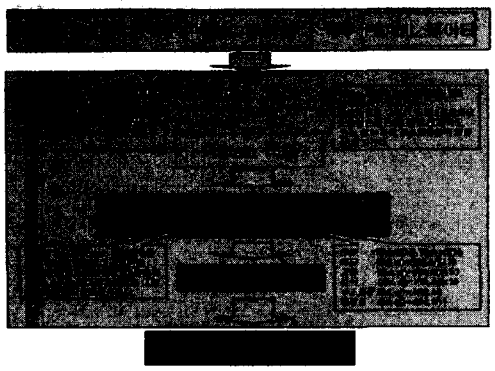


(그림 2) 그래프 G의 노드와 간선 관계

형성된 그래프 G에서 최대 유사한 부분 구조는 모든 노드사이에 간선이 존재하는 부분 그래프인 Clique를 찾음으로써 발견된다.

3.2 단백질간 구조 비교 프로그램

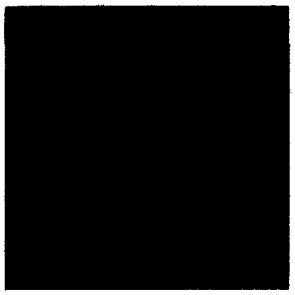
(그림 3)은 PSAML 데이터를 읽어 그래프를 생성하고, Clique를 찾아내는 프로그램의 과정을 보여주고 있다.



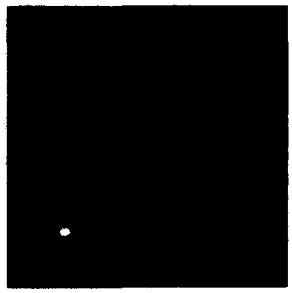
(그림 3) 프로그램 수행 과정

3.3 결과

단백질 구조 비교 프로그램을 이용하여 PDB ID 2BOP(그림 4)와 200L(그림 5) 단백질간 구조를 비교하였다. 2BOP는 알파 2개와 베타 4개, 그리고 200L은 알파 10개와 베타 2개로 구성된 단백질이다.



(그림 4) 2BOP 구조



(그림 5) 200L 구조

● 생성된 그래프의 간선의 집합

{a1a1-a2a3, a2a3-a1a1, a1a3-a2a1, a2a1-a1a3, a1a3-a2a5, ..., a2a5-a1a3, b3b1-b4b2, b4b2-b3b1}

● Clique = [a2a1,b3b2], [a2a9,b2b2], [b2b2,b3b1]

<표 3> 일치된 이차구조

ID	일치된 이차구조				
2BOP	a2	b3	a2	b2	b3
200L	a1	b2	a9	b2	b1

<표 3>은 2BOP와 200L 구조사이의 유사한 이차구조의 쌍을 보여주고 있다. 발견된 유사한 부분구조는 5가지 경우 중 하나이며, 이러한 유사한 부분 구조에서 보다 정확한 결과는 두 구조사이의 관계를 비교할 때 사용되는 인자 값의 결정에 좌우된다.

4. 결 론

본 논문에서는 단백질 이차 구조와 그들 사이의 관계를 이용하여 단백질 구조를 기술하는 PSAML 형태의 두 단백질 구조 데이터를 이용하여 두 단백질 구조 비교 방법에 대하여 기술하였다. 두 단백질 구조에서 유사한 부분 구조는 단백질 이차 구조의 정보(형태, 길이)와 그들 사이의 관계(각도, 거리, 길이)를 바탕으로 그래프 형태로 표현하여 최대 유사한 하위 그래프를 찾는 알고리즘을 이용하여 찾아진다.

추후연구 과제로는 PSAML 형태로 표현된 단백질 구조를 논리적 표현으로 변환하는 방법과 제한 프로그래밍 기법을 이용하여 보다 빠른 다중 단백질 구조 비교 방법을 개발할 예정이다.

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

[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," Proc. Intelligent Systems for Molecular Biology 97, 1997.

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

[6] 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, 변경익, 윤형석, 이수현, 이명준, "단백질 3차 구조의 추상적인 표현기법", 한국정보과학회, '2001 가을 학술발표논문집(B) 제 28권 2호, 595-597, 2001.

[8] 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

[9] C. Bron,J. Kerbosch. "Algorithm 457: Finding All Cliques of an Uundirected Graph". CACM, 16(9):575-577, 1973.

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