단백질 이차 구조에 기반을 둔 단백질 구조 정렬 방법

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

A Method for Protein Structure Alignment based on Protein Secondary Structure

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

요 약

단백질 구조를 정렬하는 방법은 단백질의 모티프 또는 폴드를 찾는데 사용되고 있으며, 기능적 또는 구조적으로 연관된 단백질을 분류하는데 유용하게 사용되고 있다. 본 논문에서는 단백질 이차 구조(α-나선 구조와 β-병풍구조)를 기반으로 하는 단백질 구조 정렬 방법에 대하여 기술한다. 제안된 단백질 이차 구조 요소 기반의 정렬방법은 단백질 구조를 단백질 이차 구조 요소와 그들 사이의 관계(수소결합, 상대적 위치)를 이용하여 표현하고, 표현된 두 개의 구조를 단백질 이차 구조 요소와 그들 사이의 관계만을 이용하여 비교하는 방법으로 기존의 방법보다 빨리 정렬할 수 있다.

1. 서 론

단백질 구조 데이터베이스인 PDB(Protein Data Bank)[1]는 실험적으로 결정된 생물학적 데이터를 가지고 있으며, 그 데이터의 증가량은 날이 갈수록 증대되고 있다. 급속하게 증가하는 단백질 구조 데이터를 분석하여 단백질의 모티프(motif)와 유사성이 있는 폴드(fold)를 기준으로 분류하는 과정에서 단백질 정렬 기법이 사용되고 있다.[2]

단백질 구조는 원자(atom), Residues와 단백질 이차 구조 요소(Secondary Structure Elements)등과 같은 최소 구성단위로 표현될 수 있다.[2] 하나의 단백질 구조에서 각 최소 구성단위는 다른 최소 구성단위와 상대적인 위치, 또는 아미노산 서열상의 위치 등[3]의 속성을 가질 수 있다. 더욱이 최소 구성단위는 자신을 구성하는 아미노산의 화학적 특징을 추가하여 보다 상세히 기술될 수 있다.[4]

단백질 구조에 대한 정렬 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.[5]

본 논문에서는 단백질 이차 구조 요소를 기반으로 표현[6,7]된 두 단백질 구조를 정렬하는 방법에 대하여 기술한다. 제안된 단백질 구조 정렬 방법은 단백질 이차 구조 요소와 그들 사이의 관계를 이용하여 단백질 구조를 정렬하는 방법이다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 이차 구조 요소와 이 구조 요소 사이의 관계를 이용한 단백질 구조 표현에 대하여 살펴보고, 3장에서는 표현된 단백질 구조사이의 정렬 방법에 대하여 알아본다. 끝으로 4장은 결론 및 향후 연구방향에 대해여 기술한다.

2. 단백질 구조 표현 방법

단백질 구조는 단백질의 골격을 이루는 C_α, Residues, 또는 이차 구조 요소 등으로 표현될 수 있다. 본 논문에

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

서는 단백질 구조를 단백질 이차 구조 요소로 표현하고 이들 사이의 관계를 이용하여 기술한다. 단백질 구조를 단백질 이차 구조 요소로 표현하기 위하여, 그 기능이 파악된 단백질에 대한 3차원 정보를 제공하는 PDB 데이터베이스를 이용한다.

2.1 이차 구조 요소를 이용한 단백질 구조 표현

단백질 3차원 정보는 단백질 구조를 형성하는 단백질 이차 구조 요소인 α-나선구조와 β-병풍구조를 구별하는데 이용될 뿐만 아니라, 이들 이차 구조 요소 사이의 관계를 분석하는데 사용된다. (그림 1)은 단백질 구조를 표현하기 위한 기본 구성요소들을 보여주고 있다.

S : 단백질 이차 구조 요소의 집합
E : 단백질 이차 구조 요소(알파, 베타)
H : 수소 결합(베타의 방향성)
L : 단백질 이차 구조 요소의 아미노산 개수
T : 단백질 이차 구조 요소의 상대적인 위치

(그림 1) 단백질 구조 표현을 위한 기본 구성요소

H 구성요소는 베타 구조에서 수소 결합의 방향성(순방향, 역방향)을 나타낸다. T 구성요소는 하나의 단백질 이차 구조 요소를 기준으로 다른 단백질 이차 구조 요소의 상대적인 위치를 나타낸다. 상대적인 위치 관계는 3차원 공간의 위치관계를 나타내고 있다.

2.2 단백질 구조 표현 예

다음은 PDB 식별자가 2bop인 단백질 구조를 단백질 이차 구조 요소와 이들 사이의 관계를 이용하여 표현한 예이다.

2bop PATS = (S, H, C, L, X), where
$S = \{\beta_1, \alpha_1, \beta_2, \beta_3, \alpha_2, \beta_4\}$
$S_\alpha = \{\alpha_1, \alpha_2\}$, $S_\beta = \{\beta_1, \beta_2, \beta_3, \beta_4\}$
$H = \{(\beta_1, A, \beta_2), (\beta_2, A, \beta_3), (\beta_1, A, \beta_4)\}$
$L = \{(\alpha_1, 11), (\alpha_2, 10), (\beta_1, 9), (\beta_2, 4), (\beta_3, 8), (\beta_4, 8)\}$
$T = \{ (\beta_1, 0, -, -, \alpha_1), (\beta_1, 0, 0, -, \beta_2), (\beta_1, 0, 0, -, \beta_3),$
$(\beta_1, +, -, -, \alpha_2), (\beta_1, +, -, +, \beta_4), (\alpha_1, 0, +, -\beta_2),$
$(\alpha_1, 0, +, -, \beta_3), (\alpha_1, +, -, +, \alpha_2), (\alpha_1, +, 0, +, \beta_4)$
$(\beta_2, 0, 0, -, \beta_3), (\beta_2, +, +, \alpha_2), (\beta_2, +, -, +, \beta_4)$
$(\beta_3, +, -, 0, \alpha_2), (\beta_3, +, -, +, \beta_4), (\alpha_2, -, +, +, \beta_4)\}$

(그림 2) 단백질 구조 표현의 예

(그림 2)에서 상대적인 위치 관계를 표현하는 $(\beta_1, 0, -, -, \alpha_1)$의 의미는 α_1이 β_1의 X축 방향으로는 같은 위치(0), Y축 방향으로 아래 쪽(-), 그리고 Z축 방향으로 뒤쪽(-)에 위치함을 의미한다.

3. 단백질 구조 정렬 방법

단백질 이차 구조 요소와 관계로 표현된 대상 단백질과 질의 단백질의 정렬은 두 단백질 이차 구조 요소간의 유사성을 측정하는 함수를 이용하는 동적 프로그래밍(Dynamic Programming)[8]을 사용한다. 동적 프로그래밍은 Smith-Waterman 알고리즘[9]을 사용하여 대상 단백질 구조와 질의 단백질 구조에서 부분적으로 가장 길게 정렬되는 부분 범위를 찾아낸다. 정렬된 단백질 구조의 상대적인 위치 관계는 임의의 이차 구조 요소를 중심으로 변환 될 수 있다. 질의 단백질의 상대적 위치 관계를 변환하여 두 단백질 구조 사이에서 가장 일치하는 부분을 찾아낸다.

3.1 두 단백질 이차 구조 요소사이의 유사성 측정

두 단백질 이차 구조 요소의 유사성(Similarity) 측정을 위한 함수는 단백질 이차 구조 요소간의 속성 값을 비교하여 결정된다.

① S_1: 두 개의 이차 구조 요소간의 형태를 비교한다.
- S_1=5 : 같은 형태
- S_1=0 : 같지 않은 형태

② S_2: 두 개의 이차 구조 요소간의 수소결합 관계가 일치하는지 알아본다.
- S_2=5 : 모두 수소 결합이 있는 경우, 모두 소수 결합이 없는 경우
- S_2=0 : 한쪽 이차 구조 요소에만 수소결합이 있는 경우

③ S_3: 두 개의 이차 구조 요소간의 상대적 위치 관계가 일치하는지 알아본다.
- S_3=5 : 상대적 위치가 일치하는 경우
- S_3=0 : 상대적 위치가 일치하지 않은 경우

④ S_4 : 두 개의 이차 구조 요소간의 아미노산 길이 차이를 알아본다.
- S_4=2 : 길이 차이가 5 미만인 경우
- S_4=0 : 길이 차이가 5 이상인 경우

두 개의 이차 구조 요소간의 유사성(Similarity)
$$= S_1 + S_2 + S_3 + S_4$$

3.2 단백질 이차 구조 요소의 상대적 위치 관계 변환

(그림 2)에서 보여진 단백질 구조의 상대적 위치 정보는 아미노산 서열상의 첫 번째 단백질 이차 구조 요소를

기준으로 기술하였다. 상대적 위치 정보는 기준 단백질 이차 구조 요소를 변경함에 따라 변경된다.

다음은 기존의 상대적인 위치 관계를 임의의 단백질 이차 구조 요소 E_n을 기준으로 상대적인 위치 정보를 변경하는 방법이다.

· 기존의 상대적 위치 관계

◇ (E_p, D_{p1}, D_{p2}, D_{p3}, E_n)과 같이 표현

◇ 단, {E_n, E_p} ∈ 단백질 이차 구조 요소들, p≠n, D_1, D_2, D_3는 각각 X축, Y축, Z축 방향의 상대적 관계

· E_n을 기준으로 하는 새로운 상대적 위치 관계 표현

◇ (E_n, D_{n1}, D_{n2}, D_{n3}, E_p)과 같이 표현

◇ 단, D_{n1}, D_{n2}, D_{n3}의 값(기존 상대적 위치 관계와 비교)

→ E_n과 E_p가 서로 변환 된 경우: 기존 상대적 위치 관계와 반대의 값을 가진다. (+, -, 0는 각각 -, +, 0로 변환

→ E_n과 E_p가 같은 경우: 기존 상대적 위치 관계와 같은 값을 가진다.(+, -, 0는 각각 +, -, 0로 유지)

3.3 단백질 구조 정렬 과정

단백질 이차 구조 요소와 관계로 표현된 두 단백질 구조를 정렬하는 방법은 다음과 같다.

1. 두 개의 단백질 구조를 정렬하기 위한 초기 위치를 찾아낸다.
   - 질의 단백질 구조와 대상 단백질 구조의 이차 구조 요소 사이의 유사성 값을 3.2절에서 기술된 유사성 측정 함수를 이용하여 구한다.
   - 유사성 값을 이용하여 질의 단백질과 대상 단백질을 동적 프로그래밍(Dynamic Programming)에 적용시킨다.
   - 동적 프로그래밍을 적용시킨 결과로 두 단백질 구조간의 최적으로 정렬된 부분 구조를 찾아내고, 이 부분 구조에 포함된 질의 단백질 구조의 이차 구조 인자들을 알아낸다.
2. 1의 과정에서 찾은 이차 구조 인자를 이용하여 질의 단백질의 상대적 위치 관계를 변경 후, 유사성 값을 계산한다.
   - 최적으로 정렬된 부분 구조에 포함된 질의 단백질 구조의 이차 구조 인자를 차례로 선택한다.
   - 선택된 이차 구조 인자를 중심으로 질의 단백질의 상대적 위치 관계를 변환한다.
   - 질의 및 대상 단백질의 유사성 값을 계산한다.
3. 2에서 나온 유사성 값 중 최고 높은 유사성을 나타내는 구조를 찾아낸다.

4. 결 론

본 논문에서는 단백질 이차 구조로 표현된 두 단백질 구조를 정렬하는 방법에 대하여 기술하였다. 단백질 구조는 3차원 구조 데이터에서 단백질 이차 구조 요소(α-나선 구조와 β-병풍 구조)를 알아내고, 이들 사이의 관계를 정의함으로써 표현된다. 그리고 단백질 이차 구조 기반의 단백질 정렬 방법은 단백질 이차 구조 요소와 이들 사이의 관계를 이용하여 단백질 구조 정렬을 수행한다. 따라서 기존의 방법보다 빠르게 두 단백질 구조를 정렬할 수 있다.

추후연구 과제로는 단백질 구조를 구성하는 원자들의 정보를 본 논문에서 제시한 단백질 구조 정렬 방법에 효과적으로 적용하여 보다 정확한 단백질 구조를 비교하는 알고리즘을 개발할 예정이다.

[참고문헌]

[1] Berman, H.M., Westbrook, J., Feng., Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., "The Protein Data Bank". Nucleic Acids Research 28, 235-242, 2000.

[2] Holm, L. and Sander, C. "Protein structure comparision by alignment of distance matrices". J.Mol. Biol., 233, pp. 123-138, 1993.

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

[4] Singh, A.P. and Brutlag, D.L. 1997. "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations". Proc. Intelligent Systems for Molecular Biology 1997.

[5] Ingvar Eidhammer, Inge Jonassen, William R. Taylor, "Structure Comparison and Structure Patterns", Reports in Informatics, July, 1999.

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

[7] 이근우, 이수현, 이명준, "제한 논리 프로그래밍 언어에서 DCG를 이용한 생물학적 서열의 구조 검색", 한국정보과학회, '2001 가을학술발표논문집(B) 제 28권 2호, 352-354, 2001.

[8] S. Needleman and C. Wunsch. "A general method applicable to the search for similarities in the amino acid sequence of two proteins". J. Mol. Biol., 48:443-454, 1970.

[9] Smith, T.F. and Waterman, M.S. "Identification of common molecular subsequences", J. Mol. Biol, 147:195-197, 1981.