단백질 3차 구조의 추상적인 표현기법

김진홍[o*] · 안건태[*] · 변경익[*] · 윤형석[*] · 이수현[**] · 이명준[*]
*울산대학교 컴퓨터 · 정보통신공학부
**창원대학교 컴퓨터공학과

A Technique for Abstract Representation of Protein Tertiary Structure

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

요 약

오늘날 인간 유전체 프로젝트(Human Genome Project)의 완성은 인간의 모든 유전자 서열정보를 제공하게 되었으며, 이러한 데이터를 바탕으로 생명현상과 관련된 산업 및 연구가 각광받게 되었다. 특히 생명체의 특정 기능을 파악하기 위한 단백질 3차 구조에 대한 연구가 활발히 진행중이다. 본 논문에서는 단백질 3차 구조를 추상적으로 기술 할 수 있는 표현기법을 기술한다. 제안된 표현기법은 단백질 2차 구조요소(α-나선구조와 β-병풍구조)를 이용하여 인접한 구성요소간의 접힘(folding)에 대한 관계를 기술하여 추상적인 단백질 3차 구조를 표현한다. 제안된 표현기법으로 기술된 추상적 단백질 3차 구조 표현은 단백질 구조에 대한 보다 빠른 이해와 다른 단백질 구조와 비교될 수 있는 장점을 지닌다.

1. 서 론

생물정보학(Bioinformatics)[1]은 인터넷과 그것의 파급효과만큼이나 빠르게 우리의 관심을 가지는 새로운 분야가 되어 가고 있다. 최근 많은 분자 생물학 기술의 발달과 *인간 유전체 프로젝트(Human Genome Project)*의 완성으로 많은 유전자 서열정보 및 새로운 형태의 생물학 정보들이 산출되었다. 따라서, 생물정보의 기초자료인 염기 및 단백질 서열과, 서열과 관련된 다차원 구조 데이터들에 대한 연구는 생명현상과 직접적으로 관련이 있으므로 그 중요성이 크다. 특히 단백질 서열과 구조에 관한 연구는 생명체내 핵심적인 역할을 수행하는 단백질의 기능을 파악하는데 중요한 부분이다. 이러한 단백질의 기능을 파악하기 위한 연구는 단백질의 3차원구조의 파악 및 해석이 필요하다[2]. 단백질은 *아미노산*이라 불리는 분자로 구성되어 있으며 20개의 아미노산의 조합으로 구성된 중합체이다. 이들 아미노산들이 구성환경이나 아미노산의 특성에 따라 접힘 현상이 발생하여 다양한 3차원 구조를 형성하게 된다[3]. 이론적으로 100개의 아미노산을 가진 단백질의 3차원 구조는 두 아미노산의 *결합각*에 따라 약 10개의 상태가 존재할 수 있어, 10^{100}가지의 3차원 구조가 가능하다. 이러한 모든 구조를 파악하기 위한 시간은 한가지 구조를 파악하기 위한 시간을 1 *femtosecond*(1^{-15}초)로 생각할 경우, 무려 10^{85}초(우주의 나이 약 10^{18}초)의 시간이 요구된다. 일반적으로 단백질은 약 150 ~ 250개의 아미노산을 가진다.

단백질 3차원 구조의 유사성과 차이점을 이해하는 것은 염기서열, 구조 그리고 기능 사이의 관계에 대한 연구를 위해서는 매우 중요한 부분이라고 할 수 있다.

많은 단백질에 대한 구조 분석 과정에서 이들 구조들 사이에 규칙성이 존재한다는 것을 밝혀냈으며, 대부분의 단백질은 *이차구조 구성요소(Secondary Structure Element)*인 *나선형 구조*나 *판상형 구조*를 가지며 단지, 이들 이차구성 요소의 성질, 개수, 순서적인 결합관계, 그리고 공간적인 정렬상태에 따라 다양한 접힘 구조를 가지는 단백질이 형성되는 것이다.[4][5]

본 논문에서는 단백질의 이차구성요소 및 이들 사이의 연관관계를 정의함으로써 단백질의 구조를 효과적으로 기술하는 추상적 단백질 3차 구조 표현에 대하여 설명한다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 2차 구조의 구성요소와 이 구성요소사이에 3차원 구조 표현을 하기 위한 관계에 대하여 살펴보고, 이들을 이용하여 추상적 단백질 3차 구조 표현을 표현기법에 대하여 기술한다. 끝으로 3장은 결론 및 향후 연구방향에 대해여 기술한다.

2. 추상적 단백질 3차 구조 표현 기법

단백질 구조는 구조를 형성하는 요소에 따라 몇 가지 구조로 나눌 수 있다. 단백질을 구성하는 아미노산 서열 자체를 *1*

†본 연구는 한국과학재단 목적기초연구사업 지원으로 수행되었음.

*차구조*라 하며, 폴리펩티드 사슬이 *α-나선(α-helix)*과 *β-병풍(β-pleated sheet)*을 형성하는 구조를 *2차구조*라 한다. 그리고 삼차원 공간에서 단일 폴리펩티드 사슬의 *접힘(folding)* 구조를 나타내는 3차원적 형태를 ***3차구조***라 하고, 한 개 이상의 폴리펩티드 사슬이 복합체를 이루는 형태를 ***4차구조***라 한다.

단백질의 정확한 기능을 알아내기 위해서 단백질 3차구조를 정확하게 파악하는 것이 중요하다. 단백질 3차구조는 단백질을 구성하는 아미노산의 3차원 데이터를 바탕으로 파악될 수 있다. 현재 몇 연구단체에서 그 기능이 파악된 단백질에 대한 3차원 정보를 제공하고 있으며 대표적인 단백질 구조의 3차원 정보를 제공하는 데이터베이스로 Protein Data Bank(http://www.rcsb.org/pdb/) 데이터베이스[6]가 있다.

단백질 3차 구조를 보다 빠르고 쉽게 파악하기 위해서는 단백질 3차 구조를 표현하기 위한 3차원 정보를 모두 표현하기 보다 단백질 2차 구성요소인 α-나선구조와 β-병풍구조를 이용하여 이들 사이의 접힘(folding) 정보를 이용하여 단백질 3차 구조를 표현하는 것이 효율적이다.[7]

본 장에서는 단백질에 대한 3차원 정보를 바탕으로 단백질 2차 구성요소인 α-나선구조와 β-병풍구조를 파악하고 파악된 구성요소 사이의 관계를 정의하여 추상적인 단백질 3차 구조를 표현하는 방법을 기술한다.

2.1 단백질 2차 구조를 표현하는 구성요소

단백질 2차 구조를 표현하고 있는 구성요소에는 α-나선 구조와 β-병풍 구조가 있다. 이들 구조는 단백질의 일반적인 접힘 형태를 나타내고 있다. α-나선은 한 가닥의 폴리펩티드 사슬이 단단하게 감긴 원통모양의 구조이다. α-나선은 폴리펩티드 주 사슬의 N-H기와 C=O기 사이의 수소결합에 의해서 안정된다. 이때의 수소결합은 한 펩티드결합의 C=O기와 네 번째 있는 다른 펩티드결합의 N-H기 사이에서 형성된다. β-병풍은 수소결합의 국소적이고 협동적인 형성으로 인하여 생긴 구조로써, 얇고 접혀진 종이조각을 나란히 놓은 형태를 가지고 있다. β-병풍 구조는 평행과 역평행 형태로 존재할 수 있다.

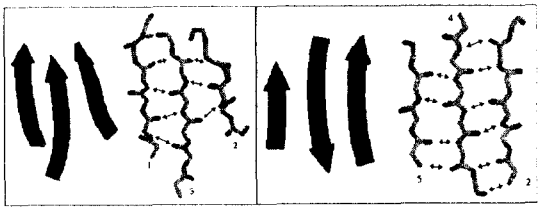
[그림 1] α-나선 구조와 β-병풍 구조

2.2 단백질 3차 구조 표현을 위한 2차 구성요소간의 관계

추상적 단백질 3차 구조는 단백질 2차 구조 구성요소간의 접힘(folding)에 관한 정보를 이용하여 나타낼 수 있다. 추상적 단백질 3차 구조 표현을 위하여 단백질 2차 구성요소간의 위상학적인 정보는 다음과 같은 구성요소사이의 관계를 기술하여 표현될 수 있다.

2.2.1 β-병풍 구조의 방향성(수소 결합) 관계

β-병풍 구조는 인접한 폴리펩티드 사슬들이 수소결합의 방향에 따라 평행(parallel)과 역평행(antiparallel) 형태로 존재한다. 평행 β-병풍 구조는 인접사슬과의 방향이 같은 방향(N→C 또는 C→N)이며, 역평행 β-병풍 구조는 서로 반대 방향으로 나아간다. 즉, 인접한 폴리펩티드 사슬의 수소 결합의 방향인 아미노기(N)에서 카르복실기(C)의 방향성에 따라 평행 또는 역평행 β-병풍 구조가 생긴다.



[그림 2] 평행 β-병풍 구조 [그림 3] 역평행 β-병풍 구조

2.2.2 구성요소의 아미노산 개수

단백질 2차 구조의 구성요소는 단백질 1차 구조를 구성하는 몇 개의 아미노산으로 구성이 된다. 단백질 2차 구조를 구성하는 아미노산의 개수는 α-나선 구조와 β-병풍 구조의 상대적인 크기를 알아 낼 수 있는 정보를 제공한다.

2.2.3 구성요소의 상대적 위치 관계

단백질 3차 구조를 표현하기 위해서 단백질 2차 구조 구성요소 사이의 공간적 위치 표현이 중요하다. 공간적 위치에 대한 정보는 서로 인접한 단백질 2차 구조 구성요소간의 공간적 위치를 표현하여 서로 어떠한 모양으로 접힘(folding)이 되었는지 표현 할 수 있다.

2.3 추상적 단백질 3차구조 표현

단백질 3차 구조의 추상적인 구조를 위한 표현 기법은 단백질을 이루는 아미노산의 위상 정보를 바탕으로 단백질 2차 구조 구성요소와 두 구성요소간의 상대적인 위치 정보 및 관계를 표현한다.

추상적 단백질 3차 구조를 표현을 위한 기본 구성요소들은 다음과 같다.

S : 단백질 구성요소의 집합

E : 단백질 구성요소(알파, 베타)

H : 베타의 방향성(평행, 역평행)

L : 구성요소의 아미노산 개수

T : 구성요소의 상대적인 위치

추상적 단백질 3차원 구조인 PATS(Protein Abstract Tertiary Structure)는 (S, H, L, T)으로 다음과 같은 구성요소간의 관계를 이용하여 표현된다.

① $S = (E_1, E_2, \ldots, E_k)$, $S_\alpha = \{\alpha_1, \alpha_2, \ldots, \alpha_j\}$, $S_\beta = \{\beta_1, \beta_2, \ldots, \beta_n\}$, $1 \le i \le k$, $1 \le j \le k$, $1 \le n \le k$, $E_i \in (S_\alpha \cup S_\beta)$

단백질 2차 구성요소는 모든 α-나선 구조를 나타내는 S_α 와 모든 β-병풍 구조를 나타내는 S_β 에 속한다. S는 단백질 1차 구조인 아미노산의 순서에 의한 단백질 2차 구성요소의 연결 순서를 나타낸다.

② $H = \{(Ei, \delta, Ej) \mid Ei, Ej, \in S_\beta, \delta = \{P, A\}\}$

H 관계는 인접한 단백질 2차 구성요소인 두 β-병풍 구조의 방향성을 나타낸다. 즉, 두 병풍 구조 사이의 수소결합의 방향성을 나타내는 δ는 평행(parallel)한 관계를 가질 때 P, 역평행(antiparallel)한 관계를 가질 때 A를 가진다.

③ $L = \{(Ei, x) \mid Ei \in (S_\alpha \cup S_\beta), x \in 양의\ 정수\}$

L은 단백질 2차 구조 구성요소를 이루는 아미노산의 개수를 나타낸다.

④ T = {(Ei, ω, θ, Ej) | Ei, Ej ∈ $S_\alpha \cup S_\beta$, $i < j$, ω ∈ {0, n, w, e, s, ne, nw, se, sw}, θ ∈ {+, 0, -)}

T 관계는 두 단백질 2차 구조 구성요소의 상대적인 위상정보를 나타낸다. 상대적인 위상 정보란 한 구성요소를 기준으로 공간적으로 상대 구성요소의 위치를 말한다. T 관계는 S에 속한 모든 구성요소 *Ei*를 기준으로 *ω(같은 위치:0, 동:e, 서:w, 남:s, 북:n, 북동:ne, 북서:nw, 남동:se, 남서:sw)* 방향으로 *θ(앞쪽:+, 같은 평면:0, 뒷쪽:-)* 깊이에 *Ei*를 제외한 모든 구성요소 *Ej*에 대한 상대적 위치를 모두 표현한다. 단, *Ei*는 *Ej*보다 단백질 1차 구조인 아미노산 서열상 앞에 나타나는 단백질 2차 구성요소이다.

다음 예에서 추상적 단백질 3차원 구조 표현 기법을 이용하여 PDB 아이디(ID)가 2bop인 단백질을 기술하였다.

2bop PATS = (S, H, C, L, X), where

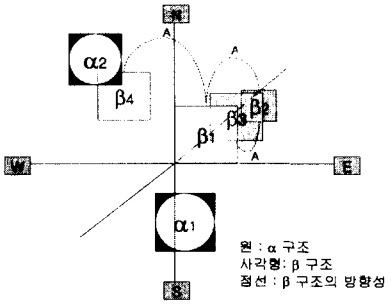
$S = \{\beta_1, \alpha_1, \beta_2, \beta_3, \alpha_2, \beta_4\}$

$S_\alpha = \{\alpha_1, \alpha_2\}$, $S_\beta = \{\beta_1, \beta_2, \beta_3, \beta_4\}$

$H = \{(\beta_1, A, \beta_2), (\beta_2, A, \beta_3), (\beta_1, A, \beta_4)\}$

$L = \{(\alpha_1, 11), (\alpha_2, 10), (\beta_1, 9), (\beta_2, 4), (\beta_3, 8), (\beta_4, 8)\}$

$T = \{ (\beta_1, s, -, \alpha_1), (\beta_1, 0, -, \beta_2), (\beta_1, 0, -, \beta_3), (\beta_1, nw, -, \alpha_2), (\beta_1, nw, +, \beta_4), (\alpha_1, ne, 0, \beta_2), (\alpha_1, ne, +, \beta_3), (\alpha_1, nw, +, \alpha_2), (\alpha_1, n, +, \beta_4) (\beta_2, 0, -, \beta_3), (\beta_2, nw, +, \alpha_2), (\beta_2, w, +, \beta_4) (\beta_3, nw, 0, \alpha_2), (\beta_3, nw, +, \beta_4), (\alpha_2, se, +, \beta_4)\}$



[그림 4] 단백질 2bop의 추상적 3차 구조

3. 결론

본 논문에서는 단백질 2차 구조 구성요소인 α-나선 구조와 β-병풍 구조를 기반으로 구성요소 사이의 관계를 정의하여 추상적 단백질의 3차 구조를 기술하는 표현기법을 제시하였다. 제시된 표현기법은 단백질의 비교 및 단백질간의 상호작용을 설명하는데 유용할 것으로 기대된다.

앞으로 추상적 단백질 3차 구조 표현 기법을 이용하여 단백질 구조 데이터베이스 구축과 추상적 단백질 3차 구조로 표현된 단백질 구조를 비교하는 알고리즘을 개발할 예정이다.

[참고문헌]

[1] David Gilbert, Rolf Backofen, Roland H. C. Yap, "Introduction to the Special Issue on Bioinformatics", Constraints, pp. 139-139, Volume 6,Issue 2/3, June 2001

[2] 박상대, 필수 세포생물학, 교보문고, 2001

[3] 생화학교재편찬위원회, 생화학, 청문각, 1999

[4] T.P Flores, D.M. Moss, and Thornton J.M, An algorithm for automatically generating protein topology cartoons. Protein Engineering, 7(1):31-37, 1994

[5] W. Kabsch and C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-banded and geometrical features. Biopolymers, 22:2577-2637, 1983

[6] 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

[7] D. R. Westhead, T. W. F. Slidel, T. P. J. Flores and J. M. Thornton, Protein structural topology: automated analysis, diagrammatic representation and database searching, Protein Sci, 897-904, 1999