

Effective Comparison of Protein Structures Based on Extended PSAML

확장된 PSAML을 통한 효과적인 단백질 구조 비교

Jinhong Kim^{1*}, Geontae Ahn¹, Suhyun Lee², Myungjoon Lee¹

¹ School of Computer Engineering & Information Technology, Ulsan University, Ulsan, Korea

² School of Computer & Information Technology, Changwon National University, Changwon, Korea

E-mail: avenue@ulsan.ac.kr

Abstract

단백질 구조를 비교하는 방법은 단백질 구조를 표현하는 기술에 따라 다양하게 존재한다. 일반적인 단백질 구조 정렬방법은 단백질 구조를 원자 또는 Residue를 기준으로 표현하고, 표현된 두 구조사이의 일치된 부분을 찾는 방법과 단백질 구조를 단백질 이차구조 요소로 표현하고 표현된 두 단백질 구조를 정렬하는 방법으로 크게 구분된다. 이러한 단백질 구조 비교 방법은 단백질 구조의 유사성을 측정하는 과정에서 많은 시간을 요구할 뿐만 아니라 PDB에 저장된 데이터가 증가함에 따라 보다 많은 단백질과 비교가 요구된다. 따라서 대용량의 단백질 구조 데이터베이스를 대상으로 효율적으로 단백질의 유사 부분구조를 찾을 수 있는 방법이 필요하다.

본 논문에서는 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여, 기존의 단백질 이차구조 기반의 구조 표현 방법인 PSAML을 확장하여 단백질 이차구조가 가지는 공간상의 정보를 내포한 Topology String을 생성하고 이를 이용하여 대용량의 단백질 구조 데이터베이스에서 유사성이 높은 단백질 구조를 필터링하는 방법에 대하여 기술한다. Topology String은 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하여, 단백질 이차구조를 이용하여 구조 비교를 수행하기 이전에 유사성이 높은 단백질 구조를 신속하게 찾아내는데 효과적으로 적용될 수 있다.

능 과약을 위하여 그 중요성이 커지고 있다.

서 론

단백질 구조 비교 알고리즘은 단백질 구조 데이터베이스인 PDB(Protein Data Bank)[1] 데이터의 증대에 따라 단백질 기

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

단백질 구조를 비교하는 방법은 단백질 구조를 표현하는 방법에 따라 다양한 방법이 존재한다. 일반적인 단백질 구조 정렬 방법은 단백질 구조를 원자(Ca) 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]

이러한 단백질 구조 비교 방법은 단백질 구조의 유사성을 측정하는 과정에서 많은 시간을 요구할 뿐만 아니라 PDB 데이터베이스의 데이터가 증가함에 따라 보다 많은 단백질과 비교가 요구되며, 따라서 대용량의 단백질 구조 데이터베이스를 대상으로 효율적으로 단백질의 유사 부분구조를 찾을 수 있는 방법이 필요하다.

본 논문에서는 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여, 기존의 단백질 이차구조 기반의 구조 표현 방법인 PSAML[3]을 확장하여 단백질 이차구조가 가지는 공간상의 정보를 내포한 TS(Topology String)을 생성하고 이를 이용하여 대용량의 단백질 구조 데이터베이스에서 유사성이 높은 단백질 구조를 필터링하는 방법에 대하여 기술한다. 기존의 PSAML 기반의 단백질 구조 비교 방법[4]은 단백질 이차구조 구성요소와 이들 사이의 관계를 이용하여 단백질 구조 사이의 유사성이 높은 구조를 찾아내는 방법으로써 단백질 구조 데이터베이스의 용량이 증가함에 따라 더 많은 수행 시간이 요구된다. 이러한 단점을 해결하기 위하여 확장된 PSAML은 TS에 대한 정보를 제공하여 이를 이용한 효과적인 단백질 구조 비교 방법을 제공할 수 있다. TS는 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하여, 단백질 이차구조를 이용하여 구조 비교를 수행하기 이전에 유사성이 높은 단백질 구조를 신속하게 찾아내는데 효과적으로 적용될 수 있다.

PSAML: 단백질 이차구조 기반의 단백질 구조 표현 방법

PSAML은 단백질 구조를 구성하는 이차구조와 그들 사이의 관계를 이용하여 단백

질 구조를 추상화하여 표현할 수 있는 방법을 제공한다.

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

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

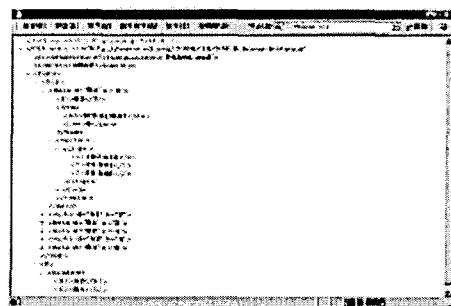
$$R = (\theta, \gamma, \nu, h, d), \text{ 단, } E_i, E_j \in S, i \neq j.$$

S는 단백질을 구성하는 이차구조의 집합을 나타낸다. T, C, A는 각각 이차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보를 나타낸다. R은 두 이차구조 사이에 정의되는 관계로써 다음과 같이 표현되고, 이차구조 사이의 관계는 <표 1>과 같다.

<표 1> 이차구조 사이의 관계

관계	의미	표현
θ	각도	$\theta(E_i, E_j) = \text{angle}(\theta_i)$
γ	거리	$\gamma(E_i, E_j) = \text{distance}(D)$
ν	길이차	$\nu(E_i, E_j) = \text{length}(l_i, l_j)$
h	수소결합	$h(E_i, E_j) = \{E, N\}, E_i \text{와 } E_j \text{는 } \beta\text{-strand}$
d	방향성	$d(E_i, E_j) = \{P, A\}, E_i \text{와 } E_j \text{는 } \beta\text{-strand}$

(그림 1)은 PSA에서 제공하는 구성요소와 관계정보를 바탕으로 생성된 PSAML의 한 예를 보여주고 있다. PSAML은 XML 스키마[5, 6]를 이용하여 PSA 정보를 표현하고 있다.



(그림 1) PSAML 문서의 예

확장된 PSAML에서 제공하는 Topology String

TS(Topology String)에 대한 정의

단백질을 이루는 각각의 이차구조를 하나의 문자로 매핑하여 생성되는 TS에 대한 정의는 다음과 같다.

● $TS(\text{protein_id}) = \{t_1, t_2, \dots, t_i\}$, 단, protein_id는 단백질 식별자, t_i 는 topology 문자, i 는 이차구조 개수이며 서열상의 순서

● $t_i = \{V \text{ or } D, A \text{ or } M, E \text{ or } F\}$, t_i 가 α -나선일 경우,

$= \{H \text{ or } N, G \text{ or } I, K \text{ or } L\}$, t_i 가 β -판상조각일 경우,

● t_i 의 값은 단백질 이차구조의 종류 및 위상학적 방향성에 따라 결정된다.<표 2>

● 생성된 TS는 x 축, y 축, 그리고 z 축을 기준으로 각 90° 변환하여 모두 24가지의 서로 다른 TS로 변환된다.<표 3>

PSAML 기반에서 TS(Topology String)의 생성과정

다음은 PSAML에서 제공하는 정보를 이용하여 TS를 추출하는 과정을 설명하고 있다.

① 아미노산 서열 순서대로 PSAML에서 이차구조 하나를 선택한다. 선택된 이차구조의 시작점을 원점으로 평행 이동시키고 이차구조의 끝점이 위치한 공간상의 위치와 이차구조의 종류에 따라 <표 2>에 제시된 문자로 변환한다. <표 2>에서 나타나는 TS를 이루는 문자들은 20가지의 아미노산 문자들 중에서 선택되었다. 이것은 보다 효과적인 TS 서열의 상동성 평가를 수행하기 위하여 NCBI Blast 프로그램[7]을 적용할 수 있도록 하기 위한 것이다.

② 기본적으로 생성된 Topology String은

<표 3>에서 제시된 변환표에 의하여 24가지의 서로 다른 TS으로 변환된다. 이때 변환은 각축으로 90° 의 방향으로 한다.

③ PDB ID 200L 단백질 구조는 12개의 단백질 이차구조로 이루어져 있다. 이 단백질의 TS은 ①과 ②에 기술된 방식에 따라 생성하면 "AGICDCAFCABC" (이외 23가지)과 같다.

<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

변환 방향											
+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

TS(Topology String) 데이터베이스 생성

단백질 이차구조 기반의 단백질 구조 비교를 수행하기 이전에 TS을 이용한 필터링 과정을 거친다. 이러한 필터링 과정에서 TS 서열의 상동성을 보다 효과적으로 빠르게 평가하기 위하여 NCBI Blast 프로그램[]을 활용한다. 이를 위하여 PSAML 데이터베이스로부터 생성된 TS 서열을 FASTA 형식으로 저장하고 이를 NCBI Blast에서 활용할 수 있는 형태의 데이터베이스로 변환한다.

현재 TS 데이터베이스는 36,768개의 서열을 저장하고 있으며, 이는 단백질 15,098개(체인으로 분리)에 대한 TS이다.

<표 4>은 1MBA를 이루고 있는 이차구조에 대한 3차원 정보 및 이를 바탕으로 생성된 TS를 보여주고 있다. 각 이차구조는 3차원 공간에 위치하는 시작점과 끝점에 대한 좌표 값을 가지며, 종류 및 아이디를 가진다. 각 이차구조의 아이디에 나타나는 인덱스는 아미노산에 나타나는 순서에 의하여 결정된다. 즉, 인덱스 숫자가 낮을수록 아미노산 서열상 앞에 위치한다.

1MBA의 이차구조 정보를 바탕으로 생성된 TS 서열은 FASTA 형식으로 저장한다. NCBI Blast 프로그램은 저장된 TS 파일을 읽어 서열 정렬을 수행한다.

<표 5>는 NCBI의 blastp 프로그램을 이용하여 1MBA의 TS와 상동성이 높은 단백질을 나타내고 있다. PDB ID가 5MBA, 4MBA, 3MBA, 2FAM, 그리고 2FAL는 1MBA의 TS와 완전 일치하였으며, 부분적으로 일치한 단백질도 추출되었다.

<표 5> 1MBA의 TS의 서열 정렬 결과

Protein Id:Chain	Score	Identities	Evalue
5MBA:null	21	100%	75
4MBA:null	21	100%	75
3MBA:null	21	100%	75
2FAM:null	21	100%	75
2FAL:null	21	100%	75
1QHA:A	19	75%	218
1OUT:B	17	75%	827
1JSW:A	17	75%	827
1BI7:A	17	75%	827

<표 6>은 NCBI blast 프로그램의 상동성 검사한 후, 추출된 단백질을 대상으로 1MBA와 이차구조 기반 단백질 구조 비교를 수행한 결과를 보여주고 있다. PDB ID가 5MBA, 4MBA, 3MBA, 2FAM, 그리고 2FAL는 1MBA와 아주 비슷한 구조를 이루고 있음을 알 수 있었다. 이는 1MBA 및 TS 서열 정렬에서 추출된 각 단백질의 각 이차구조가 3차원 공간에 벡터로 표현되었을 때 매우 유사함을 의미한다. 그리고 TS

상동성의 결과에서 일치성이 낮은 단백질들은 1MBA의 특정 부분 구조와 비슷한 구조를 내포하고 있음을 알 수 있었다. <표 6>에서 a2a1는 1MBA의 a2와 다른 단백질의 a1이 유사함을 나타낸다.

<표 6> 일치된 이차구조

ID	일치된 이차구조
1qha:A	a3a4 a5a10 a6a14 a8a11
1out:B	a2a15 a1a10 a3a8 a5a13 a6a14
1jsw:A	a6a6 a7a13 a4a16

결론

본 논문에서는 단백질 구조 비교를 보다 빠르고 효과적으로 수행하기 위하여, 기존의 단백질 이차구조 기반의 구조 표현 방법인 PSAML을 확장하여 단백질 이차구조가 가지는 공간상의 정보를 내포한 TS를 생성하고 이를 이용하여 대용량의 단백질 구조 데이터베이스에서 유사성이 높은 단백질 구조를 필터링하는 방법에 대하여 기술하였다.

TS(Topology String)은 단백질 이차구조를 하나의 문자로 기술하여 아미노산 순서와 위상학적인(공간적인) 정보를 바탕으로 단백질 구조를 표현하는 서열 정보이다. 이 정보는 서열 정렬 프로그램을 이용하여 유사성이 높은 단백질 구조를 추출하는데 유용하게 사용된다. TS를 이용한 단백질 구조 비교 방법은 단백질 이차구조를 이용하여 구조 비교를 수행하기 이전에 유사성이 높은 단백질 구조를 신속하게 찾아내는데 효과적으로 적용될 수 있다.

Acknowledgements

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

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*, 1999.
- [3] 김진홍, 안건태, 이수현, 이명준, "구조 비교를 위한 단백질 데이터의 XML 표현기법", *한국정보과학회 프로그래밍언어연구회*, 16권, 2호, 15-16, 2002.
- [4] 김진홍, 안건태, 조민수, 이수현, 이명준, "PSAML를 기반으로 한 단백질 구조 비교", *한국정보과학회 프로그래밍언어연구회*, 16권, 3호, 33-44, 2002.
- [5] D. C. Fallside, "XML Schema Part 0: Primer", *W3C*, May 2001.
- [6] W3C, "Document Object Model (DOM)," *WWW document* (<http://www.w3.org/DOM/>).
- [7] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997) "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." *Nucleic Acids Res.* 25:3389-3402.
- [8] Hiroaki KATO and Yoshimasa TAKAHASHI, "Automated Identification of Three-Dimensional Common Structural Features of Proteins?", *J. Chem. Software*, Vol. &, No. 4, p. 161-170, 2001.
- [9] Sampo Niskanen, Patric Ostergard, "Cliquier: routines for clique searching", 2002, <http://www.hut.fi/~pat/cliquier>