17-3-R03

Topology String을 이용한 단백질 구조 비교 방법

(Protein Structure Comparison Using the Topology String)

김 진 홍*, 안 건 태*, 이 수 현**, 이 명 준*

*울산대학교 컴퓨터정보통신공학부, **창원대학교 컴퓨터공학과

*{avenue, java2u, mjlee}@mail.ulsan.ac.kr, **suhyun@sarim.changwon.ac.kr

요 약

단백질 구조 비교 방법은 단백질 구조의 유사성을 측정하는 과정에서 많은 시간을 요구할 뿐만 아니라 PDB 데이터베이스에 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 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.

1. 서론

인터넷은 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_a) 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 Topology String[4]을 생성하고 이를 이용하여 대용량의 단백질 구조 데이터베이스에서 유사성이 높은 단백질 구조를 필터링하는 방법에 대하여 기술한다. 기존의 PSAML 기반의 단백질 구조 비교 방법[5]은 단

백질 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 해결하기위하여 확장된 PSAML은 Topology String에 대한 정보를 제공하여 이를 이용한 효과적인 단백질 구조 비교 방법을 제공할 수 있다. 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.

2. PSAML : 단백질 이차구조 기반의 단백질 구조 표현 방법

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

PSAML은 단백질 구조를 구성하는 이차구조와 그들 사이의 관계를 이용하여 단백질 구조를 추상화하여 표현하는 PSA를 이용하여 하나의 단백질 표현을 XML[6,7] 형태로 제공한다.

한 단백질 구조를 표현하기 위해서, PSA는 구조를 결정하고 있는 이차구조를 3차원 공간상의 벡터(Vector)로 표현하여 공간적인 정보 및 임의의 두 이차구조 쌍에 대한 각도, 거리, 길이, 그리고 수소 결합 및 방향성 등의 관계에 대한 정보를 이용한다.

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

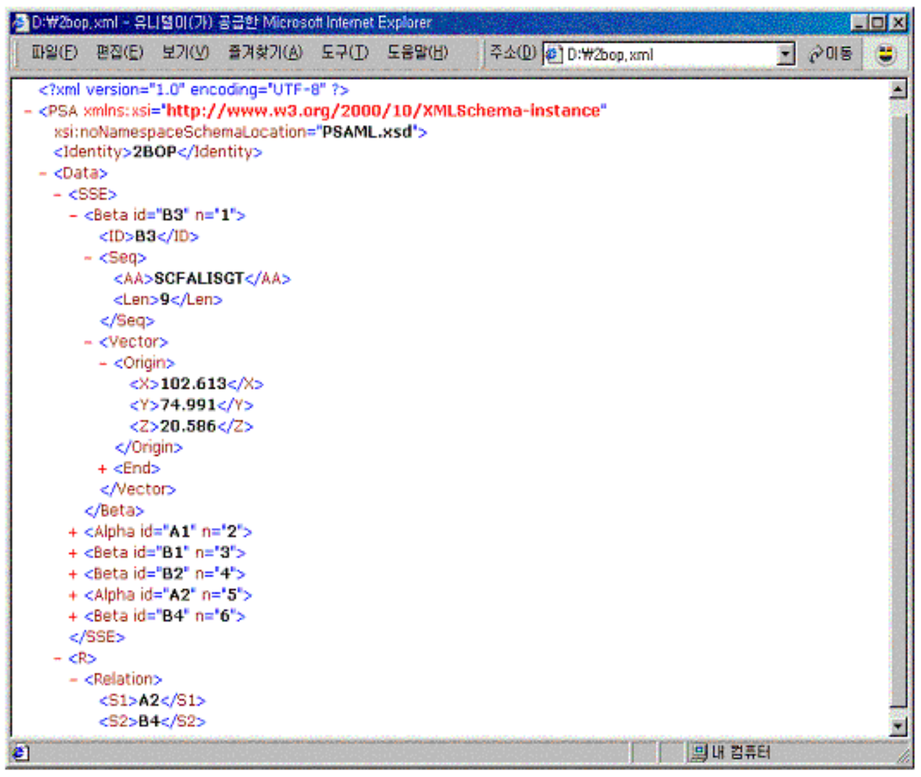
$$PSA(P) = (S, T, C, A, R)$$
$$R = (\theta, \gamma, \nu, h, d),\ 단,\ Ei, Ej \in S,\ i \neq j.$$

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

<표 1> 이차구조 사이의 관계

관계	의 미	표 현
θ	각 도	$\theta(Ei, Ej) = angle(\theta I)$
γ	거 리	$\gamma(Ei, Ej) = distance(D)$
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

(그림 1)은 PSA에서 제공하는 구성요소와 관계정보를 바탕으로 생성된 PSAML의 한 예를 보여주고 있다. PSAML은 XML 스키마[6,7]를 이용하여 PSA 정보를 표현하고 있다.



(그림 1) PSAML 문서의 예

PSAML 문서는 식별(Identity) 부분과 데이터(Data) 부분으로 나누어 단백질 구조에 대한 정보를 표현한다. 식별 부분은 단백질의 주석을 나타내며, 데이터 부분은 단백질을 구성하고 있는 이차구조 요소에 대한 기술과 더불어 그들 사이의 관계를 나타낸다.

데이터 부분은 <SSE>과 <R>의 두 요소(elements)를 가진다. <SSE>요소는 단백질을 구성하고 있는 모든 이차구조 요소에 대한 정보를 포함한다. <R>요소는 단백질을 구성하

고 있는 모든 구성요소의 각각의 쌍에 대하여 각도, 거리, 방향성과 같은 관계들을 표현한다.

3. Topology String 정의 및 생성

3.1 Topology String에 대한 정의

단백질을 이루는 각각의 이차구조를 하나의 문자로 변환하여 생성되는 Topology String에 대한 정의는 다음과 같다.

○ TS(protein_id)={t1, t2, ..., ti}, 단, protein_id는 단백질 식별자, ti는 topology 문자, i는 이차구조 개수이며 서열상의 순서
○ ti={V or D, A or M, E or F}, ti가 α-나선일 경우,
={H or N, G or I, K or L}, ti가 β-판상조각일 경우,
○ ti의 값은 단백질 이차구조의 종류 및 위상학적 방향성에 따라 결정된다.(표 2)

(표 2)에서 나타나는 Topology String을 이루는 문자들은 20가지의 아미노산 문자들 중에서 선택되었다. 이것은 보다 효과적인 Topology String 서열의 상동성 평가를 수행하기 위하여 NCBI Blast 프로그램[8]을 적용할 수 있도록 하기 위한 것이다.

PSAML 기반으로 생성된 Topology String은 X축, Y축, 그리고 Z축을 기준으로 각 90° 회전하여 모두 24가지의 서로 다른 Topology String으로 변환된다. 3차원 공간상에서 존재하는 이차구조는 바라보는 관점에 따라 다른 Topology 문자로 변환될 수 있다. 본 논문에서는 하나의 이차구조에 대한 Topology 문자를 생성할 때 총 24가지의 방향을 고려하여 생성하였다. (표 3)은 각 축의 90° 회전에 따라 하나의 Topology 문자의 변환 규칙을 나타내고 있다.

3.2 PSAML 기반에서 Topology String의 생성과정

다음은 PSAML에서 제공하는 정보를 이용하여 Topology String을 추출하는 과정이다.

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.
② 기본적으로 생성된 Topology String은 (표 3)에서 제시된 변환표에 의하여 24가지의 서로 다른 Topology String으로 변환된다. 이때 변환은 각축으로 90°의 방향으로 한다.
③ PDB ID 200L 단백질 구조는 12개의 단백질 이차구조로 이루어져 있다. 이 단백질의 Topology String은 ①과 ②에 기술된 방식에 따라 생성하면 “AGIMDMAFMAVM”(이외 23가지)과 같다.

<표 2> 이차구조 변환 규칙

위상학적 방향		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

3.3 Topology String 데이터베이스 생성

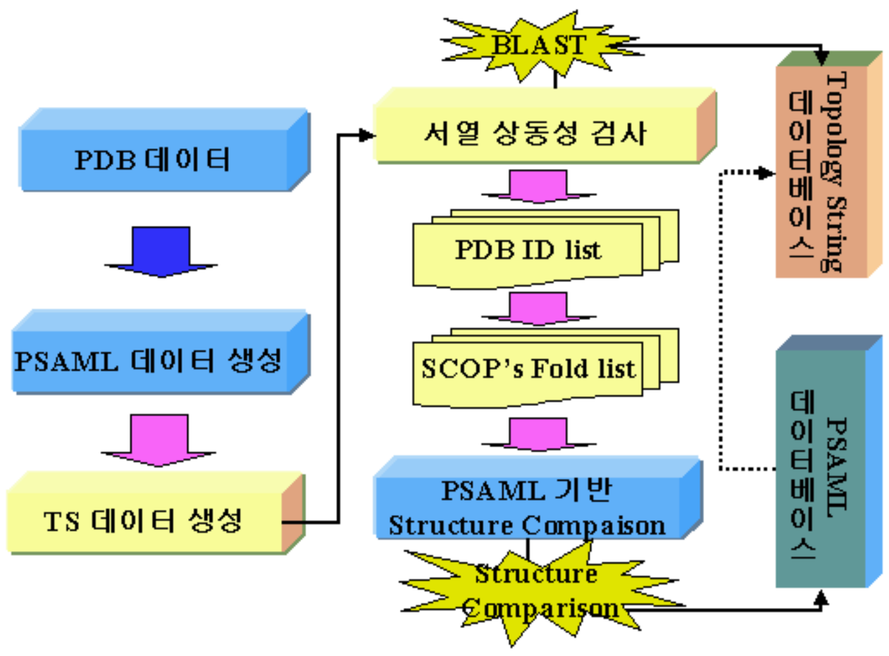
단백질 이차구조 기반의 단백질 구조 비교를

수행하기 이전에 Topology String을 이용한 필터링 과정을 거친다. 이러한 필터링 과정에서 Topology String 서열의 상동성을 보다 효과적으로 빠르게 평가하기 위하여 NCBI Blast 프로그램[8]을 활용한다. 이를 위하여 PSAML 데이터베이스로부터 Topology String 데이터베이스를 생성하는 변환기를 제작하였다. 개발된 변환기는 PSAML 데이터에서 Topology String 서열을 FASTA[9] 형식으로 저장하고 이를 NCBI Blast에서 활용할 수 있는 형태의 데이터베이스로 변환한다.

현재 Topology String 데이터베이스는 단백질 15,098에 대한 36,768개의 Topology String을 저장하고 있다.

4. Topology String을 이용한 단백질 구조 비교 방법

Topology String을 이용한 단백질 구조 비교 방법은 NCBI Blast 프로그램을 이용하여 입력된 단백질의 Topology String과 유사성이 높은 단백질 구조를 추출한 후, 추출된 단백질이 속하는 SCOP Fold[10]를 찾는다. 찾은 SCOP Fold에 속한 단백질들을 대상으로 이차구조 단백질 구조 비교를 수행한다. (그림 2)는 Topology String을 이용하여 단백질 구조 비교를 수행하는 단계를 보여주고 있다.



(그림 2) Topology String을 이용한 단백질 구조 비교 방법

○ 입력 단백질의 PSAML 및 Topology String 생성

입력된 PDB 데이터에서 PSAML 데이터를 생성한다. 생성된 PSAML 데이터에서 이차구조 요소 및 관계 정보를 이용하여 Topology String을 생성한다. PSAML 데이터에는 단백질을 구성하는 이차구조에 대한 종류에 대한 정보가 아미노산 서열 순으로 나타나 있다. 선택된 이차구조는 3차원 공간상의 각도 정보를 바탕으로 하나의 문자로 변환된다. 변환된 Topology String 서열은 아미노산 순서에 따라 생성된다.

○ Topology String 서열의 유사성 측정

생성된 단백질의 Topology String과 유사성이 높은 단백질을 Topology String 데이터베이스에서 NCBI에서 제공하는 Blast 프로그램을 이용하여 추출한다. NCBI의 Blast 프로그램은 아미노산 서열을 정렬하는 프로그램으로써 빠르게 결과를 보여준다.

○ 비교 대상 단백질 추출

NCBI Blast 프로그램을 통한 입력된 Topology String과 유사성이 높은 단백질 구조(PDB ID)를 추출한다. 추출된 단백질은 Topology String 측면에서 입력된 단백질과 유사성이 높은 단백질 구조를 가진다.

○ SCOP Fold 정보 추출

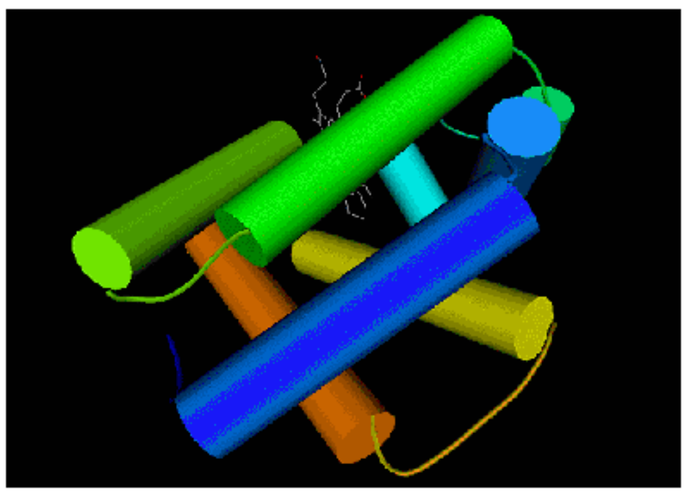
NCBI의 Blast 프로그램을 통하여 추출된 단백질 리스트에서 각 단백질이 속하는 SCOP의 Fold 정보를 추출하여 입력 단백질과 이차구조 기반 단백질 구조 비교의 실행 대상을 추출한다. SCOP 데이터베이스는 유사한 구조를 가지고 있는 단백질들을 같은 Fold에 분류하여 저장하고 있다. NCBI의 Blast 프로그램을 이용하여 추출된 단백질 구조들 이외에 SCOP의 Fold에 속한 단백질 구조들을 추출하는 방법은 Topology String을 이용하여 배제될 수 있는 입력 단백질과 유사한 구조를 가진 단백질들을 고려할 수 있는 방법을 제공한다.

○ 이차구조 기반 단백질 구조 비교

Topology String 및 SCOP의 Fold 정보를 통하여 추출된 단백질 구조와 입력된 단백질 구조를 이용하여 단백질 이차구조 기반의 단백질 구조 비교 수행한다. 단백질 이차구조 기반의 단백질 간 구조 비교 방법[5]은 두 단백질의 PSAML 정보를 바탕으로 유사한 부분구조를 내포하는 유사성 그래프를 생성한 후, 모든 노드 사이에 간선이 존재하는 부분 그래프[11]를 찾는 기존의 알고리즘[12]을 이용하여 최대 유사 부분 구조를 파악할 수 있다.

5. 수행 결과

Topology String 기반의 단백질 구조 비교 방법을 이용하여 1MBA와 유사한 부분 구조를 찾아보았다. 1MBA(그림 3)는 8개의 α-나선으로 이루어진 단백질로써 Myoglobin 계열에 속하며, 산소를 저장하는 기능을 담당한다.



>1MBA:_ MYOGLOBIN (MET) ($P*H 7.0) - CHAIN _
XSLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKSVADIKASPK
LRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHVGFGVGSAQFENVRSMFPGFVASV
AAPPAGADAAWTKLFGLIIDALKAAGA

(그림 3) 1MBA의 구조와 서열

(표 4)은 1MBA를 이루고 있는 이차구조에 대한 3차원 정보 및 이를 바탕으로 생성된 Topology String을 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.

(표 4) 1MBA의 이차구조 정보

이차구조 요소	(시작점좌표) (끝점좌표)	TS 문자
A1	(-66.714,-54.464,-14.517) (-48.805,-41.945,-14.524)	V
A2	(-49.167,-38.509,-19.18) (-46.093,-47.746,-37.544)	F
A3	(-47.359,-43.218,41.457) (-55.146,-36.647,-40.943)	D
A4	(-41.727,-38.111,-37.435) (-41.228,-34.822,-32.365)	E
A5	(-47.87,-28.566,-31.079) (-62.982,-44.636,-16.506)	M
A6	(-72.911,-48.831,-21.207) (-65.997,-34.729,-39.411)	F
A7	(-62.682,-46.843,-38.455) (-43.633,-51.091,-23.396)	V
A8	(-55.852,-56.654,-20.052) (-72.14,-46.077,-37.512)	F

(표 5)는 NCBI의 Blast 프로그램을 이용하여 1MBA의 Topology String과 상동성이 높은 단백질을 나타내고 있다. PDB ID가 5MBA, 4MBA, 3MBA, 2FAM, 그리고 2FAL는 1MBA의 Topology String과 완전 일치하였으며, 부분적으로 일치한 단백질도 추출되었다.

(표 5) 1MBA의 Topology String의 서열 정렬 결과

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

(표 6)은 NCBI Blast 프로그램을 이용하여 추출된 단백질들에서 Topology String 상으로 75% 이상인 단백질을 대상으로 SCOP의 Fold를 추출

한 결과를 보여주고 있다. 추출된 SCOP의 Fold에 속한 단백질 구조를 추출함으로써 Topology String의 상동성 검사를 통하여 배제된 입력 단백질과 구조적으로 유사한 단백질들을 포함하여 이차구조 기반 단백질 구조를 수행할 수 있도록 하였다.

(표 6) 1MBA과 유사한 단백질 구조를 포함할 SCOP's Folds

Protein Id:Chain	SCOP's Fold
5MBA:null	Globin-like
4MBA:null	Globin-like
3MBA:null	Globin-like
2FAM:null	Globin-like
2FAL:null	Globin-like
1QHA:A	Ribonuclease H-like motif
1OUT:B	Globin-like
1JSW:A	L-aspartase-like
1BI7:A	beta-hairpin-alpha-hairpin repeat

PDB ID가 5MBA, 4MBA, 3MBA, 2FAM, 그리고 2FAL는 1MBA와 이차구조 기반 단백질 구조 비교를 수행하였을 때 아주 비슷한 구조를 이루고 있음을 알 수 있었다. 이는 1MBA 및 Topology String 서열 정렬에서 추출된 각 단백질의 각 이차구조가 3차원 공간에 벡터로 표현되었을 때 매우 유사함을 의미한다.

(표 7)은 Topology String 서열 정렬에서 추출된 단백질 중에서 Topology String의 유사도의 정도가 낮은 단백질들을 대상으로 1MBA와 이차구조 기반 단백질 구조 비교를 수행한 결과를 부분적으로 보여주고 있다.

(표 7) 일치된 이차구조

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

(표 7)에서 나열된 Topology String 상동성의 결과에서 일치성이 낮은 단백질들은 1MBA의 특정 부분 구조와 비슷한 구조를 내포하고 있음을 알 수 있었다. (표 7)에서 a2a1는 1MBA의 a2와 다른 단백질의 a1이 유사함을 나타낸다.

6. 결론 및 향후 연구 방향

본 논문에서는 단백질 구조를 보다 빠르고 효과적으로 비교하기 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 기술하였다.

Topology String은 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. Topology String을 이용한 단백질 구조 비교 방법은 단백질 이차구조를 이용하여 구조 비교를 수행하기 이전에 유사성이 높은 단백질 구조를 신속하게 찾아내는데 효과적으로 적용될 수 있다.

감사의 글

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, p235~242, 2000.

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

[3] 김진홍, 안건태, 이수현, 이명준, "구조비교

를 위한 단백질 데이터의 XML 표현기법", *한국정보과학회 프로그래밍언어연구회*, 16권, 2호, p15~16, 2002.

[4] The ups and downs of protein topology; rapid comparison of protein structure, *Protein Eng.* p829~837, 2000

[5] 김진홍, 안건태, 조민수, 이수현, 이명준, "PSAML를 기반으로 한 단백질 구조 비교", *한국정보과학회 프로그래밍언어연구회*, 16권, 3호, p33~44, 2002.

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

[7] W3C, "Document Object Model (D OM)", WWW document(http:// www.w3.org/DOM/).

[8] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". *Nucleic Acids Res.* p3389~3402, 1997

[9] David W. Mount, "Bioinformatics Sequence and Genome Analysis", *Gold Spring Harbor Laboratory Press*, p31~32, 2001

[10] Alexey M., Steven B., Tim H. and Cyrus Ch., "SCOP : A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", *J. Mol. Biol.*, p536-540, 1995.

[11] Hiroaki KATO and Yoshimasa TAKAHASHI, "Automated Identification of Three-Dimensional Common Structural Features of Proteins", *J. Chem. Software*, Vol. &, No. 4, p161~170, 2001.

[12] Sampo Niskanen, Patric Ostergard, "Cliquer: routines for clique searching", http://www.hut.fi/~pat/cliquer, 2002

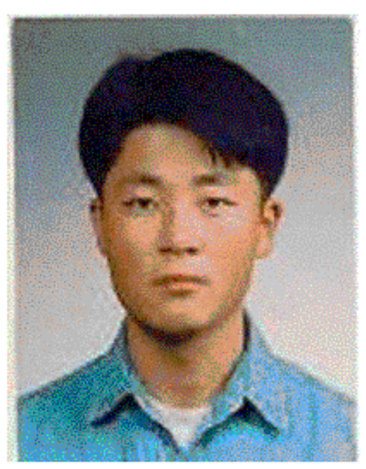
김 진 홍

1999년 2월 울산대학교 전자계산학과 졸업(학사)

2001년 2월 울산대학교 컴퓨터정보통신 공학부 졸업(석사)

2001년 3월~현재 울산대학교 컴퓨터정보통신 공학부 박사과정

관심분야 : 생물정보학, 제한프로그래밍, 협업지원 시스템, 이동에이전트 시스템 등



안 건 태

1999년 2월 울산대학교 전자계산학과 졸업(학사)

2001년 2월 울산대학교 컴퓨터·정보통신공학부 졸업(석사)

2001년 3월~현재 울산대학교 컴퓨터·정보통신공학부 공학박사과정

관심분야 : 생물정보학, 협업지원 시스템, 분산 시스템, 이동에이전트 시스템 등

이 수 현

1987년 2월 광운대학교 전자계산학과 졸업(학사)

1989년 2월 한국과학기술원 전산학과 졸업(석사)

1994년 8월 한국과학기술원 전산학과 졸업(박사)

1994년 9월~1996년 2월 한국전자통신연구원 선임연구원

1996년 3월~현재 창원대학교 컴퓨터공학과 부교수

관심분야 : 프로그래밍언어, 제한프로그래밍, 생명정보학 등

이 명 준

1980년 2월 서울대학교 수학과 졸업(학사)

1982년 2월 한국과학기술원 전산학과 졸업(석사)

1991년 8월 한국과학기술원 전산학과 졸업(박사)

1982년 3월～현재 울산대학교 컴퓨터정보통신공학부(교수)

1993년 8월～1994년 7월 미국 버지니아대학 교환교수

관심분야 : 프로그래밍언어, 분산 객체 프로그래밍 시스템, 병행 실시간 컴퓨팅, 인터넷 프로그래밍시스템, 생물정보학 등