# 단백질 구조 비교를 위한 β-sheet의 분할

조민수*, 김진홍**, 이명준**, 이수현*
*창원대학교 컴퓨터·정보통신공학부
**울산대학교 컴퓨터정보통신공학부
e-mail: oops@pl.changwon.ac.kr

# Division of β-sheet for Protein Structure Comparison

Min-Su Cho*, Jin-Hong Kim**, Myung-Joon Lee**, Su-Hyun Lee*
*School of Computer & Information Technology, Changwon National University
**School of Computer Engineering & Information Technology, University of Ulsan

요 약

단백질의 이차구조 중에서 β-sheet의 구조를 비교하는데 있어서 비정확성의 문제가 있다. 이는 β-sheet가 그 구조적 특성 때문에 휘어지기 때문이다. 그래서 본 논문에서는 PSA에서 정의된 β-sheet를 좀 더 적절하게 추상화하기 위해서 β-sheet를 분할하는 방법을 제안하였으며, 이 방법을, DOM을 이용하여 JAVA로 구현하였다. 본 논문에서 제안한 방법은, 단백질의 이차구조의 정보를 포함하는 PSAML 데이터로부터 단백질의 구조 및 유사성을 비교하기 위한 단백질 구조비교 시스템에서 사용할 수 있다.

## 1. 서론

단백질 구조 데이터베이스인 PDB의 데이터가 급격히 증가하고 있으며, 단백질의 기능을 파악하기 위해서 다양한 방법이 시도되고 있다. 이미 알려진 단백질을 비롯하여 새로 발견되는 단백질의 기능을 알기 위해서, 단백질들의 구조를 비교하여 그 기능을 밝히는 연구가 많이 이루어지고 있다.

단백질 구조 비교는 단백질의 물리적·구조적인 특징에 따라 단백질 구조를 분류하고, 단백질의 공통 부분 구조를 찾아내는데 활용되고 있으며, 그 방법은 단백질 구조를 표현하는 방법에 따라 다양하다. 단백질의 접힘(folding)과 구조를 이해하고 분석하는 데에는 단백질 데이터의 전체를 이용하는 것보다 단백질 구조의 특징을 나타내는 대표적인 정보를 이용하는 것이 효과적이다. 일반적인 단백질 구조 표현 방법은 단백질 구조를 원자(C- $\alpha$), 잔기, 단백질 2차 구조 등의 특징을 이용하여 표현하는 것이며, 단백질의 2차구조는 단백질 구조의 핵심적인 부분이기 때문에 많이 이용되고 있다.

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

단백질 구조에 대한 표현 방법으로 2차구조의 구성요소를 이용하는 PSA(Protein Structure Abstraction)[1]가 제안되었다. PSA로 정의되는 단백질 구조 데이터는 PSAML(PSA Markup Language)[1]로 표현되어 XML 형태로 저장된다. PSAML은 PDB (Protein Data Bank)[2] 데이터베이스에서 제공하는 BETA mmCIF 데이터와 DSSP 데이터를 기반으로 생성되며, PDB나 다른 단백질 관련 XML 데이터 형식을 이용하는 것보다 간결하면서 구조적으로 단백질 구조 정보를 표현할 수 있다.

단백질의 2차구조 중에서 β-sheet는 그 구조적 특성 때문에 비틀려서 휘어지게 되는데, 이것은 PSA에서 정의한 방법으로 β-sheet의 구조를 비교할 때 비정확성을 초래한다. 이를 보완하기 위해서 본 논문에서는 β-sheet를 휘어진 정도에 따라 여러 개로 분할하여 추상화하는 방법을 제안하였다. 본 논문에서 제안하는 방법은 PDB로부터 PSAML로의 변환도구[3]의 구현에 적용될 수 있다.

본 논문은 다음과 같이 구성되어 있다. 2장에서는 단백질 구조를 표현하는 형식인 PSA에 대해서, 3장

에서는 단백질의 2차구조 중 $\beta$-sheet의 구조적 특성에 대해 설명하고자 한다. 4장에서는 $\beta$-sheet를 분할하는 방법에 대해서 설명하고, 그 결과를 살펴본다. 마지막으로 5장에서는 결론 및 향후 연구 방향으로 끝을 맺는다.

## 2. PSA

PSA는 단백질 구조를 구성하는 2차구조와 그들 사이의 관계를 이용하여 단백질 구조를 추상화하여 표현할 수 있는 방법을 제공한다.

한 단백질 구조를 표현하기 위해서, PSA는 구조를 결정하고 있는 2차구조에 대한 공간적인 정보를 표현한다. PSA에서 표현하는 단백질 3차원 구조의 표현은 공간상에 위치한 2차구조(나선; helix, 판상조각; sheet)를 벡터로 표현한다. 즉, 한 벡터는 3차원 공간상의 시작점과 끝점에 대한 정보 및 길이에 대한 정보로 표현된다. 그리고 다른 단백질과 비교하여 유사한 부분 구조를 찾기 위하여, 한 단백질 구조에 속하는 임의의 두 2차구조 쌍에 대한 각도, 거리, 길이, 그리고 수소 결합 및 방향성 등의 관계를 표현하고 있다.

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

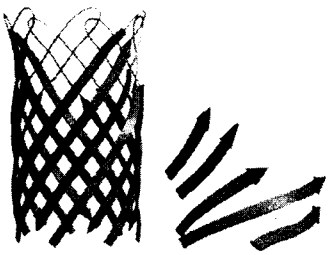
$$PSA(P) = (S, T, C, A, R)$$

S는 단백질을 구성하는 2차구조의 집합을 나타낸다. T, C, A는 각각 2차구조의 종류, 3차원상의 시작점과 끝점의 좌표값, 아미노산 서열 정보를 나타낸다. R은 두 2차구조사이에 정의되는 관계로서 각도 및 거리 관계 등을 나타낸다.

## 3. $\beta$-sheet의 구조적 특성

단백질의 2차구조는 폴리펩티드의 일부분의 지역적인 콘포메이션(원자의 공간배치)을 나타내는 것으로, 가장 잘 알려진 것은 $\alpha$-helix와 $\beta$-sheet이다.

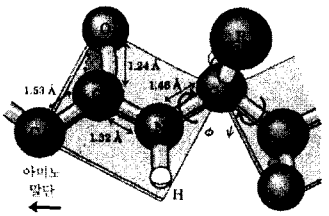
$\beta$-sheet에서는 폴리펩티드 사슬의 골격이 지그재그(zigzag)형으로 뻗어있는데, 이 폴리펩티드 사슬은 서로 평행으로 배치된 연속되는 주름과 비슷한 구조를 형성한다. $\beta$-sheet는 각 부분들이 오른손 감김 방식으로 약간 비틀려 있을 때 가장 안정되어 있으며, 이것은 $\beta$-sheet들 간의 배열과 그들 사이를 연결해 주는 폴리펩티드 연결의 방향에 영향을 미치게 된다. 또한 $\beta$-sheet의 비틀림은 여러 부분이 서로 합해져 형성되는 구조의 독특한 꼬임의 원인이 되는데, 이런 결과로 생긴 구조를 보이는 두 가지로 $\beta$-barrel과 비틀린 $\beta$-sheet(<그림 1>)이다[4].



<그림 1> $\beta$-barrel(좌)과
비틀린 $\beta$-sheet(우)

<그림 1>에서 보듯이 $\beta$-barrel이나 비틀린 $\beta$-sheet의 경우, 하나의 벡터만으로 표현하게 되면 그 구조정보를 나타내기에 부족하다는 것을 알 수 있다. 이를 보완하기 위해서 본 논문에서는 $\beta$-sheet를 분할하는 방법을 다음 장에서 제안하고 있다.

단백질은 폴리펩티드 사슬로 이루어져 있는데, <그림 2>에서 보는 바와 같이 $\alpha$-탄소(C-$\alpha$)는 3개의 결합에 의해서 떨어져 있으며[4], 두 개의 $\alpha$-탄소(C-$\alpha$) 사이의 거리는 대개 3.7~3.9 Å으로, 이것은 다음 장에서 $\beta$-sheet를 분할하기 위한 기준 값을 설정하는데 이용하게 된다.




<그림 2> 폴리펩티드
사슬상의 $\alpha$-탄소(C-$\alpha$)

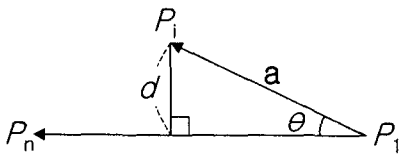
## 4. $\beta$-sheet의 분할

### 4.1 $\beta$-sheet 분할의 방법

본 논문에서는 $\beta$-sheet를 분할하는 것이 목적이므로 $\beta$-sheet의 서열정보가 3개 이상인 것들에 대해서만 다음의 방법을 적용하게 된다.

$\beta$-sheet를 분할하기 위해서 우선 하나의 $\beta$

-sheet를 구성하고 있는 n개의 아미노산에서 $\alpha$-탄소(C-$\alpha$)의 좌표값을 저장한다. 이 좌표값들은 PDB 파일에 포함되어 있으며, 3차원 공간상에 $n$개의 점으로 표시될 수 있는데, 이 때 각 점을 $P_1$, $P_2$, …, $P_n$이라고 하자. 그러면 이로부터 점 $P_1$을 시작점, 점 $P_n$을 끝점으로 하는 벡터 $\overrightarrow{P_1P_n}$을 구할 수 있다. 그런 다음, 구해진 벡터 $\overrightarrow{P_1P_n}$과, 점 $P_1$과 점 $P_n$을 뺀 나머지 점들($P_2$, $P_3$, …, $P_{n-1}$) 사이의 거리를 구해야 하는데, 이 거리 $d$는 직각삼각형에서의 삼각비의 정의에 따라 쉽게 구할 수 있다.




<그림 3> 벡터 $\overrightarrow{P_1P_n}$과 점 $P_i$ 사이의 관계

<그림 3>은 벡터 $\overrightarrow{P_1P_n}$과 임의의 점 $P_i$, 그리고 그 사이의 가장 가까운 거리 $d$를 보여주고 있으며, 이는 아래의 공식으로 구할 수 있다.

$$\sin\theta = \frac{d}{\|a\|},\quad d = \|a\|\sin\theta$$

여기서 a는 시작점이 점 $P_1$, 끝점이 점 $P_i$인 벡터이고, $\|a\|$는 벡터 a의 크기를 나타내며, 각 $\theta$는 두 벡터 $\overrightarrow{P_1P_n}$과 $\overrightarrow{P_1P_i}$가 이루는 각도를 말하는데 아래에 나오는 방법으로 구할 수 있다.

두 벡터가 이루는 각 $\theta$는 벡터의 유클리드 내적(*Euclidean inner product*)을 이용하여 구할 수 있다. u, v를 2차원 또는 3차원 공간상의 벡터, $\theta$를 이들이 이루는 각이라 할 때 u, v가 영이 아닌 벡터이면 다음과 같이 표기될 수 있으며, 이로부터 각 $\theta$를 구할 수 있다[5].

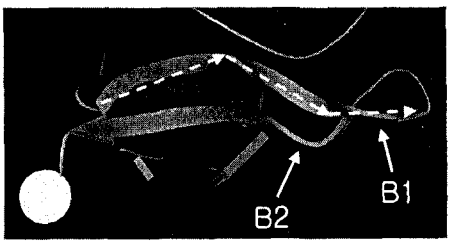
$$\cos\theta = \frac{u \cdot v}{\|u\|\|v\|} = \theta_1,\quad \mathrm{acos}(\theta_1) = \theta$$

앞에서 기술한 방법으로 벡터 $\overrightarrow{P_1P_n}$과 점 $P_i$ 간의 거리를 ($n-2$)개 얻을 수 있다. 여기서 구해진 거리가 기준 값인 3.8 Å 이상인 것이 없다면 분할하지 않고, 기준 값 이상인 것이 있을 경우에는 그 중에서 가장 값이 큰 점 $P_j$를 기준으로 해서 벡터 $\overrightarrow{P_1P_j}$와 $\overrightarrow{P_jP_n}$, 둘로 분할하게 된다. 그리고 분할된 $\overrightarrow{P_1P_j}$와 $\overrightarrow{P_jP_n}$는 위의 방법을 recursive하게 적용하여 분할하게 된다.

### 4.2 $\beta$-sheet 분할의 결과

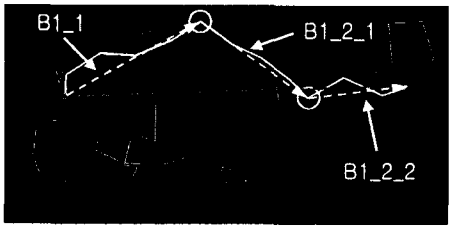
앞 절에서 기술한 방법을 JAVA 프로그램으로 구현하여 그 결과를 살펴보았다. 이 프로그램은 PDB에서 제공되는 정보를 XMC형태로 변환한 결과물을 입력으로 DOM 트리를 생성하고 본 논문에서 제안한 방법을 적용하여 최종 결과물인 PSAML 문서를 생성하게 된다.

2BOP라는 단백질은 두 개의 $\alpha$-helix와 네 개의 $\beta$-sheet로 이루어져 있는데, 그 구조는 <그림 4>과 같다. 여기서 $\alpha$-helix는 원통 모양, $\beta$-sheet는 굵은 화살표 모양으로 나타내고 있는데, $\beta$-sheet B1과 B2가 휘어진 모양을 확인할 수 있다. 굵은 화살표 위에 그어진 세 개의 점선 화살표는 하나의 $\beta$-sheet B1을 세 개의 벡터로 분할하여 추상화한 결과를 쉽게 알아볼 수 있도록 그려본 것이다.



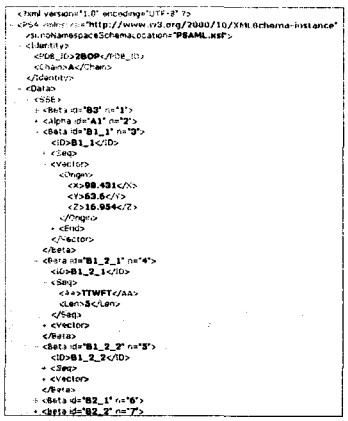

<그림 4> 2BOP 단백질 구조

단백질 2BOP의 $\beta$-sheet B1을 분할한 구체적인 결과는 <그림 5>에서 확인할 수 있으며, 이 그림은 http://www.rcsb.org에서 QuickPDB Applet으로 본 화면을 저장하여 수정한 것이다. 여기서 뒤에 희미하게 보이는 선들은 단백질 2BOP를 구성하고 있는 아미노산들의 $\alpha$-탄소(C-$\alpha$)들을 연결한 것인데, 붉은 색은 $\alpha$-helix를, 파란 색은 $\beta$-sheet를 나타내고 있다. 그리고 하늘색 굵은 선이 $\beta$-sheet B1인데, 이것이 먼저 B1_1과 B1_2로 분할된 뒤, 다시 B1_2가 B1_2_1과 B1_2_2로 분할된 결과가 세 개의 흰 색 점선 화살표이다. 가운데 붉은 색 화살표가 $\beta$-sheet B1을 하나의 벡터만으로 표현했을 경우를 보여주는데 분할을 한 것과의 현저한 차이를 알 수 있다.




<그림 5> (2BOP) β-sheet B1의 분할

분할된 각각의 벡터에 관한 정보를, JAVA 프로그램의 결과물로 생성된, 2BOP의 PSAML 문서(<그림 6>)에서 보여준다. 이 문서에서, β-sheet B1 외에도 B2의 분할된 정보 및 나머지 2차구조의 정보를 확인할 수 있다.



<그림 6> 2BOP의 PSAML

본 논문에서는 앞서 말한 기준 값을 3.8 Å으로 적용하여 결과를 확인하였다. <표 1>은 β-sheet를 포함하고 있는 1만7천여 개의 PDB 파일에 대해서 본 논문에서 기술한 방법을 적용한 결과를 보여주고 있다.

<표 1> β-sheet의 분할을 적용한 결과

| | 총 개수 | 분할된 개수 | % |
|---|---|---|---|
| 단백질(PDB_ID) | 17,339 | 10,480 | 60.44 |
| β-sheet | 374,969 | 57,631 | 15.37 |

위의 표에서 보듯이 실험한 PDB 파일은 17,339개로 37만여 개의 β-sheet가 존재하는데, 실험한 단백질의 약 60%, β-sheet의 약 15%가 분할된 것을 확인할 수 있다.

## 5. 결론

본 논문에서는 PSA에서 정의한 단백질의 2차구조 중에서 β-sheet를 좀 더 적절하게 추상화하기 위해서 β-sheet를 분할하는 방법에 대해서 기술하고 그 결과를 살펴보았다. 하나의 벡터로만 표현하기 어려운 β-sheet를 더 적절하게 나타내기 위해서 여러 개의 벡터로 분할하여 추상화된 정보를 제공함으로써 단백질의 구조 비교가 더 정확해 질 것이며, 이 정보를 이용하여 단백질 구조를 비교하는 여러 형태의 시스템을 개발할 수 있다.

본 논문에서는 기준 값을 3.8 Å으로 적용하였으나, 추가 실험을 통해서 좀 더 적절한 기준 값을 선정할 필요가 있다.

앞으로는 본 논문에서 제안된 방법을 적용시킨 프로그램으로 생성된 PSAML 파일을 이용하여 단백질 구조비교와 유사도 측정을 위한 시스템을 개발할 예정이다. 또한 PSAML 형태로 표현된 단백질 구조를 논리적 표현으로 변환하는 방법과 제한 프로그래밍 기법을 이용하는 방법을 개발할 예정이다.

## 참고문헌

[1] Su-Hyun Lee, Jin-Hong Kim, Geon-Tae Ahn, and Myung-Joon Lee, "An XML Representation of Protein Data for Efficient Structure Comparison," Proc. of International Conference on Computer and Information Science, pp. 313-319, 2002.

[2] 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.

[3] 조민수, 이수현, 이명준, "PDB 데이터에서 PSAML로의 변환도구 개발," 추계학술발표대회 논문집(하), 한국정보처리학회, 제 9권, 제 2호, 2002.

[4] D. L. Nelson, M. M. Cox, "Lehninger Principles of Biochemistry," Worth Publishers, 2000.

[5] 이장우, "알기쉬운 선형대수," (주)범한서적, 1998.