단백질 구조 비교를 위한 이차구조의 상관관계 계산

조민수[O]* 안건태** 이명준** 이수현*
*창원대학교 컴퓨터·정보통신공학부 **울산대학교 컴퓨터·정보통신공학부
*oops@pl.changwon.ac.kr

Calculation of Relation between Secondary Structures for Protein Structure Comparison

Min-Su Cho[O] Geon-Tae Ahn** 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

단백질 구조의 표현 방법을 정형화하고 호환성 및 상호작용성을 향상하기 위하여 단백질의 이차구조 구성요소와 그들 사이의 관계를 이용하여 단백질 구조를 기술하는 PSA가 제안되었다. 본 논문에서는 PSA에서 정의된 단백질의 이차구조 사이에 정의된 요소 중에서 네 가지의 각도관계와 다섯 가지의 거리관계를 계산하는 방법에 대하여 기술하였으며, 이를 자바로 구현하여 그 결과를 확인하였다. 본 논문에서 제안한 방법은 단백질의 이차구조 사이의 상관관계를 포함하는 PSAML 데이터로부터 단백질의 구조 및 유사성을 비교하기 위한 단백질 구조비교 시스템에서 사용할 수 있다.

1. 서 론

단백질의 접힘(folding)과 구조를 이해하고 분석하는데 있어서 단백질 데이터의 전체를 이용하는 것보다 단백질 구조의 특징을 나타내는 대표적인 정보를 이용하는 것이 효과적이다. 단백질의 2차구조는 단백질 구조의 핵심적인 부분이기 때문에 많은 연구자들이 이용하고 있다.

단백질 구조에 대한 표현 방법으로 2차구조 구성요소를 이용하는 PSA(Protein Structure Abstraction)[1]가 제안되었다. PSA로 정의되는 단백질 구조 데이터는 PSAML(PSA Markup Language)[1]로 표현되어 XML 형태로 저장된다. PSA 및 PSAML은 단백질 구조 및 유사성을 비교하기 위한 2차구조 기반 단백질 구조 표현으로 PDB[2] 데이터베이스에서 제공하는 데이터를 기반으로 생성된다. PSAML은 XML 스키마를 이용하여 XML 기반 언어의 요소를 정의하고, PDB나 다른 단백질 관련 XML 데이터 형식을 이용하는 것보다 간결하면서 구조적으로 단백질 구조 정보를 표현할 수 있다.

PSAML 데이터는 단백질 구조를 구성하는 2차구조와 그들 사이의 상관관계를 이용하여 단백질 구조를 추상화하여 표현할 수 있는 방법을 제공한다. 2차구조들 사이의 상관관계에는 각도, 거리, 길이 등이 포함되며, 이들 상관관계는 2차구조의 3차원적인 좌표 데이터로부터 계산되기 때문에 관점에 따라 다양한 계산방법이 존재한다. 본 논문에서는 PDB에서 제공하는 데이터 형식을 PSAML 형식으로 변환하는데 있어 2차구조들 사이의 상관관계를 계산하는 방법을 기술한다. 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 대해 설명하고자 한다. 3장에서는 2차구조 사이의 상관관계 계산의 방법에 대해서 설명하고, 그 결과를 살펴본다. 마지막으로 4장에서는 결론 및 향후 연구 방향으로 끝을 맺는다.

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

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; strand)를 벡터로 표현한다. 즉, 한 벡터는 3차원 공간상의 시작점과 끝점에 대한 정보 및 길이에 대한 정보로 표현된다. 그리고 다른 단백질과 비교하여 유사한 부분구조를 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 다음과 같이 표현된다.

$$R = (\theta, \gamma, \nu, h, d)$$

여기에서 각 구성요소의 의미는 다음과 같다. θ는 두 2차구조인 E_i와 E_j 사이의 각도 관계를 나타내고 있으며, γ는 두 2차구조인 E_i와 E_j의 거리 관계를 나타낸다. γ는 두 2차구조인 E_i와 E_j 사이의 상대적인 거리에 대한 관계로써 3차원 공간에서 두 2차구조의 중점들간의 거리를 기술하고 있다. ν는

두 2차구조인 E_i와 E_j의 각각의 길이를 나타내며, h는 두 2차구조인 E_i와 E_j 사이의 수소결합의 유무를 나타내고 있다. 그리고, d는 두 2차구조 요소인 E_i와 E_j 사이에 나타나는 방향성을 나타낸다.

두 이차구조인 E_i와 E_j사이의 각도관계는 다음과 같이 기술된다.

$$\theta(E_i,E_j) = angle(\theta_1,\theta_2,\theta_3,\theta_4)$$

θ는 두 이차구조 사이에 다음과 같은 네 가지의 각도를 나타내고 있다. θ_1과 θ_2는 두 이차구조 (E_i와 E_j)에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 각도로서, 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, θ_1과 θ_2, 각각은 E_i와 L사이의 각도와 E_j와 L사이의 각도를 말한다. 그리고, E_i의 끝점을 시작점으로 하고, E_j의 시작점을 끝점으로 하는 벡터를 V라고 할 때, θ_3은 E_i와 V가 이루는 각도이며, θ_4는 E_j와 V가 이루는 각도이다.

두 이차구조인 E_i와 E_j사이의 거리관계는 다음과 같이 기술된다.

$$\nu(E_i,E_j) = distance(D_{mid}, D_{\max i}, D_{\min i}, D_{\max j}, D_{\min j})$$

ν는 두 이차구조인 E_i와 E_j사이의 상대적인 거리에 대한 관계로써 다섯 가지의 값을 가진다. D_{mid}는 3차원 공간에서 두 이차구조의 중점들간의 거리를 기술하고 있다. 반면에, 나머지 거리관계는 두 이차구조에 평행한 평면에 투영한 두 벡터 사이에서 정의되는 거리관계이다. 투영된 두 벡터에 평행한 중심선을 L이라고 할 때, $D_{\max i}$, $D_{\min i}$은 각각 E_i와 L사이의 최대거리 및 최소거리 값을 가지고, $D_{\max j}$, $D_{\min j}$은 각각 E_j와 L사이의 최대거리 및 최소거리 값을 가진다.

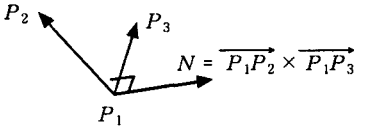
3. 이차구조 사이의 상관관계 계산

3.1 두 이차구조에 평행한 평면

각도와 거리를 계산하기 위해서는 먼저 두 이차구조 (E_i와 E_j)에 평행한 평면를 결정해야 한다. 공간상의 두 벡터에 평행한 평면은 수없이 많이 존재한다. 여기서는 수많은 평면 중에서 벡터 E_i를 포함하는 평면을 기준으로 벡터 E_j와 평행한 평면을 선택하여 그 평면 P로 벡터 E_j를 투영시킨다. 평면 P에 벡터 E_j를 투영시키기 위해서는 다음과 같은 과정을 거치게 된다.

첫 번째로 두 벡터 E_i와 E_j를 모두 원점으로 평행이동 시킨다. 이때 벡터의 시작점을 원점과 일치시킨다. 원점을 P_1이라 하고 이동한 두 벡터 E_i, E_j의 끝점을 각각 P_2, P_3라고 하면, 두 벡터 $\overrightarrow{P_1P_2}$와 $\overrightarrow{P_1P_3}$는 두 벡터 E_i와 E_j를 원점으로 평행 이동한 후의 벡터이다. 그러면 두 벡터 $\overrightarrow{P_1P_2}$와 $\overrightarrow{P_1P_3}$로 공간상의 평면이 만들어 지는데 이 평면 P_0은 평면 P와 평행하다.

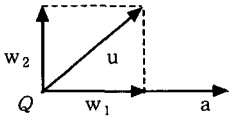
다음 단계로 평면에 수직인 법선 벡터를 구해야 하는데 법선 벡터 N은 <그림 1>과 같이 평면 위에 있는 두 벡터의 외적(cross product)으로 쉽게 구할 수 있다.



<그림 1> 법선 벡터 N을 구하는 방법

이제 구해진 법선 벡터 N을 이용하여 벡터 E_j를 평면 P로 투영시키게 되는데 이때 다음에 나오는 벡터의 분할이 이용된다.

만일에 두 벡터 u와 a가 시점이 어떤 점 Q에서 일치하도록 위치하였다면 u를 <그림 2>와 같이 분할할 수 있다. 그림에서 벡터 w_1를 u에서 a로의 직교사영(orthogonal projection of u on a)이라 하고 $proj_a u$로 표기한다. 또한 벡터 w_2를 a에 직교하는 u의 벡터성분(vector component of u orthogonal to a)이라 한다.



<그림 2> 벡터의 분할

여기서 벡터 E_j의 투영에 필요한 벡터는 w_2인데 다음과 같은 공식을 이용하여 구할 수 있다[4].

$$w_2 = u - proj_a u = u - \frac{u \cdot a}{||a||^2} a$$

위에서 구해진 법선 벡터 N을 a라고 두고, 벡터 E_i의 시작점을 시작점, 벡터 E_j의 시작점을 끝점으로 하는 벡터를 u라고 두면 앞의 공식에 의해서 벡터 w_2를 구할 수 있다. 이 벡터 w_2는 벡터 E_i의 시작점에서 평면 P에 투영된 벡터 E_j의 시작점까지의 변위(displacement)를 나타낸다. 또한 벡터 $\overrightarrow{P_1P_3}$은 벡터 E_j의 시작점에서 끝점까지의 변위를 나타내므로 평면 P에 투영된 벡터 E_j의 시작점과 끝점은 다음과 같이 구할 수 있다.

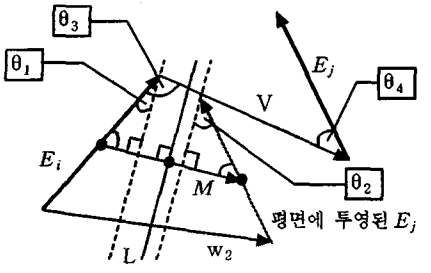
E_i의 시작점 +	w_2 =	평면에 투영된 E_j의 시작점
E_j의 시작점 +	$\overrightarrow{P_1P_3}$ =	평면에 투영된 E_j의 끝점

위와 같은 방법으로 벡터 E_j를 두 벡터 E_i와 E_j에 평행한 평면 P에 투영하였다.

3.2 각도 상관관계의 계산

투영된 두 벡터의 중점을 연결한 벡터를 M이라 하면 평행선과 직각삼각형의 성질을 이용하여 가상의 중심선 L과 투영된 두 벡터 사이의 두 각 θ_1과 θ_2는 <그림 3>과 같은 방법으로 구할 수 있는

데 두 벡터가 이루는 각은 아래에 나오는 벡터의 내적이 이용된다. 그리고 나머지 두 각 θ_3과 θ_4는 평행선의 성질을 이용하여 쉽게 구할 수 있다.



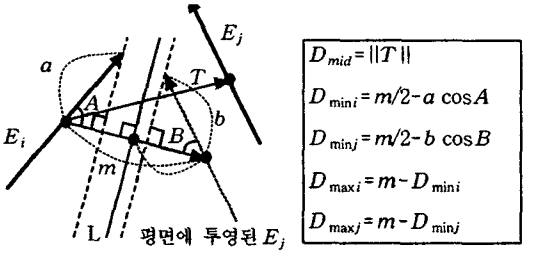
<그림 3> E_i와 E_j사이의 각도관계 계산

두 벡터가 이루는 각 θ는 벡터의 유클리드 내적(Euclidean inner product)을 이용하여 구할 수 있다. u, v를 2차원 또는 3차원 공간상의 벡터, θ를 이들이 이루는 각이라 할 때 u, v가 영이 아닌 벡터이면 다음과 같이 표기될 수 있으며, 이로부터 각 θ를 구할 수 있다[4].

$$\cos\theta = \frac{a \cdot b}{\|a\|\|b\|} = \theta_1 \ , \quad \text{acos}(\theta_1) = \theta$$

3.3 거리 상관관계의 계산

투영된 두 벡터의 중점을 연결한 벡터 M의 길이를 m이라 하면 가상의 중심선 L과 투영된 두 벡터 사이의 다섯 가지 거리는 <그림 4>에 나와있는 식으로 그 값을 구할 수 있다. 그런데 여기서 벡터의 방향과 위치에 따라 최대거리와 최소거리가 뒤바뀔 수도 있으며 벡터 E_i나 벡터 E_j가 중심선 L과 만날 경우 최소거리를 0으로 처리한다.

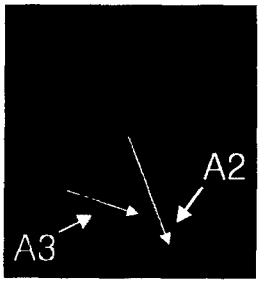


<그림 4> E_i와 E_j사이의 거리관계 계산

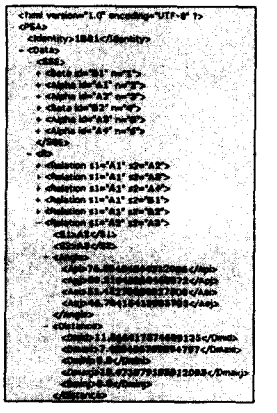
3.4 실험 및 결과

앞 절에서 기술한 방법을 JAVA 프로그램으로 구현하여 그 결과를 살펴보았다.

1B01이라는 단백질은 네 개의 α-나선과 두 개의 β-관상조각으로 이루어져 있는데, 그 구조는 <그림 5>와 같다. 여기서 α-나선은 원통으로 표현되었으며, 원통 속의 화살표는 두 개의 α-나선 A2와 A3의 방향을 나타내고 있다. 이 두 이차구조 요소 사이의 각도 및 거리 관계를 생성된 1B01의 PSAML 문서(<그림 6>)에서 보여준다. 이 문서에서, <그림 5>에서 보는 바와 같이 투영된 두 이차구조 요소는 서로 교차하게 되는데 가상의 중심선 L을 생각해 보았을 때, 두 이차구조 요소 사이의 관계 중 최소거리가 0인 것을 확인할 수 있으며 나머지 각도 및 거리 관계에 대한 수치를 얻을 수 있다.



<그림 5> 1B01 단백질 구조



<그림 6> 1B01의 PSAML

4. 결론

본 논문에서는 PSA에서 정의한 단백질의 이차구조 간의 각도와 거리 관계를 계산하는 방법에 대해서 기술하고 그 결과를 살펴보았다. PSA에서 제공되는 이차구조의 상관관계들은 단백질의 구조에 관한 공간상의 여러 가지 각도 및 거리 정보를 제공함으로써 단백질 구조를 비교하는 여러 형태의 시스템[5]을 개발하는데 유용하게 이용될 수 있다.

앞으로는 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 변환 도구 개발," 추계학술발표대회 논문집(하), 한국정보처리학회, 제 9권, 제 2호, 2002.

[4] 이장우, "알기쉬운 선형대수," 범한서적, 1998.

[5] Jin-Hong Kim, Geon-Tae Ahn, Min-Su Cho, Su-Hyun Lee, and Myung-Joon Lee, "An XML Representation of Protein Data for Structure Comparison and Its Application," Proc. of 2002 KSBI annual meeting, 2002.