JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 12 : 3011~3016 Dec. 2015

내부 알파탄소간 거리와 비네-코시 거리를 사용한 대규모 단백질 조각 라이브러리 구성

지상문*

Construction of Large Library of Protein Fragments Using Inter Alpha-carbon Distance and Binet-Cauchy Distance

Sang-mun Chi*

Department of Computer Science and Engineering, Kyungsung University, Busan 608-736, Korea

요 약

단백질의 삼차원 구조를 단백질의 국부적 구조인 단백질 조각의 일차원적 나열로 표현하면, 단백질 구조의 분석, 모델링, 탐색, 예측 등에 효과적으로 응용될 수 있다. 본 논문에서는 자연 상태의 단백질 구조를 정확하게 나타낼 수 있는 단백질 조각 라이브러리를 구성하기 위하여, 대규모 단백질 구조 자료를 이용 할 수 있는 거리 척도들의 효과적인 조합을 조사하였다. 단백질 조각 라이브러리를 구성하기 위해 군집화를 사용하였다. 초기 군집화 단계에서는 가장 계산량이 작은 내부 알파탄소간 거리를 사용하였고, 군집의 확장단계에서는 내부 알파탄소간 거리, 비네-코시거리와 평균 제곱근 오차를 조합하여 사용하였다. 제안한 거리 척도의 조합으로 대규모 자료를 이용하여 단백질 조각 라이브러리를 구성하였다. 구성된 라이브러리를 사용하여 단백질 구조를 나타내는 실험에서 작은 평균 제곱근 오차가 발생함을 확인하였다.

ABSTRACT

Representing protein three-dimensional structure by concatenating a sequence of protein fragments gives an efficient application in analysis, modeling, search, and prediction of protein structures. This paper investigated the effective combination of distance measures, which can exploit large protein structure database, in order to construct a protein fragment library representing native protein structures accurately. Clustering method was used to construct a protein fragment library. Initial clustering stage used inter alpha-carbon distance having low time complexity, and cluster extension stage used the combination of inter alpha-carbon distance, Binet-Cauchy distance, and root mean square deviation. Protein fragment library was constructed by leveraging large protein structure database using the proposed combination of distance measures. This library gives low root mean square deviation in the experiments representing protein structures with protein fragments.

키워드 : 단백질 구조, 단백질 조각 라이브러리, 내부 알파탄소간 거리, 비네-코시 거리, 평균 제곱근 오차

Key word : Protein structure, Protein fragment library, Inter alpha-carbon distance, Binet-Cauchy distance, Root mean square deviation.

Received 21 August 2015, **Revised** 08 September 2015, **Accepted** 22 September 2015

*** Corresponding Author** Sang-Mun Chi (E-mail:smchiks@ks.ac.kr,Tel:+82-51-663-5146)
Department of Computer Science and Engineering, Kyungsung University, Busan 608-736, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.12.3011 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

대용량 단백질 구조 분석 기술의 발전으로 단백질 구조 자료가 크게 증가하고 있다. 이러한 대규모 자료를 단백질 구조 모델링에 효과적으로 이용하는 방법으로, 단백질 구조 자료에 자주 나타나는 국부적 구조인 단백질 조각을 모델링 단위로 사용하려는 연구가 활발하다[1-6]. 단백질 조각으로 단백질 구조를 나타내는 방법은 단백질의 국부적 구조를 알파-나선, 베타-평판과 코일 구조로 간략화 하는 기존의 방법에 비하여 보다 정밀한 표현이 가능하며, 삼차원의 단백질 구조를 단백질 조각들의 일차원적인 나열로 나타내어 단백질 구조의 분석, 모델링, 탐색, 예측 등에 효과적으로 응용할 수 있기 때문이다[7].

자연 상태의 단백질 구조를 정확하게 나타낼 수 있는 대표적 단백질 조각들을 얻기 위해서는 단백질 조각들 사이의 유사성을 효과적으로 측정할 수 있는 거리 척도가 필요하다. 거리척도로서 널리 사용되는 방법은 원자들 사이의 평균 제곱근 오차 (RMSD: root mean square deviation), 단백질 구조의 (Φ,Ψ) 비틀림 각도와 내부 알파탄소간의 거리 등이 있다. 가장 널리 쓰이는 평균 제곱근 오차는 비교하려는 두 단백질 구조를 중첩하는 평행이동과 회전변환을 수행한 후에 대응되는 원자들 사이의 거리를 계산한다[8]. 평균 제곱근 오차는 비교 대상이 되는 단백질의 길이에 종속적이며, 상동 단백질 간에도 큰 거리가 나올 수 있으며, 평균 제곱근 오차가 중간인 범위에서 거리척도의 성능이 저하된다고 알려져 있다[9]. 특히, 거리척도의 계산에 많은 시간이 필요하므로 본 논문에서 사용하려는 대용량 자료의 분석에 적용하기 어렵다. 단백질 구조의 (Φ,Ψ) 비틀림 각도를 사용할 경우는 거리척도의 값에 너무 잡음이 많고, 작은 각도의 변화에도 큰 변화가 발생한다. 내부 알파탄소간의 거리는 단백질을 이루는 각 아미노산의 알파탄소간의 거리를 사용하여 단백질의 구조를 표현하므로, 단백질간의 거리의 계산이 간편하나, 거울상 대칭을 구별하지 못하는 단점이 있다. 최근에 제안된 비네-코시 점수[9]는 비교하는 단백질 구조의 모양의 유사성을 측정하는 방법으로, 비교적 계산시간이 적고 거울상 대칭의 구별이 가능하지만, 평균 제곱근 오차와 상관관계가 작은 단점이 있다.

본 논문에서는 대규모 단백질 구조 자료로부터 단백질 조각 라이브러리의 구성이 가능하도록 계산시간이 적게 소요되는 거리 척도를 이용한다. 이를 위하여 기존의 거리척도들인 평균 제곱근 오차, 알파탄소간 거리, 비네-코시 점수의 장점을 효과적으로 조합하여 대용량 자료의 분석에 적용하였다.

Ⅱ. 단백질 구조간의 거리척도

본 논문에서 사용한 단백질 구조 사이의 거리를 측정하는 거리척도에 대하여 알아본다. 가장 널리 쓰이는 원자 간의 평균 제곱근 오차는 비교하려는 두 단백질 구조를 중첩하기 위하여 평행 이동과 회전 변환을 수행한다. 이러한 변환은 각각의 단백질 구조간의 내부 거리는 변하지 않으면서 단백질의 위치와 방향을 바꾸어, 비교 하려는 두 단백질 구조를 최적으로 중첩되게 한다. 중첩된 두 단백질 구조에서 대응되는 원자들 사이의 거리의 제곱을 평균하여 제곱근을 한다.

비교하려는 두 단백질 구조에서 원자들의 좌표집합을 $N\times 3$ 행렬인 X와 Y로 나타난다고 하자. 행렬 X의 원소인 X_{ij}는 i번째 원자의 j좌표이고, 두 구조의 거리는 $\sum_i\sum_j (X_{ij}-Y_{ij})^2$이고, 간략히 $\| X-Y\|^2$로 표시한다. 평균 제곱근 오차 방법의 첫 단계는 두 구조의 중심을 원점으로 이동하여 $\sum_i X_{ij}=\sum_i Y_{ij}=0$ $(j,1\le j\le 3)$이 되게 한다. 두 번째 단계에서는 회전 변환을 수행한 후에 식 (1)을 최소화시키는 최적 회전 행렬 R을 찾는다.

$$\min_R \| XR-Y\|^2 \tag{1}$$

여기서, 회전행렬은 R^TR가 항등행렬이고 행렬식 $\det(R)=1$인 정규 직교행렬의 조건을 만족해야 한다. 최종적으로 평균제곱근 오차, $RMSD\ (X,Y)$는

$$\sqrt{\frac{1}{N}\| XR-Y\|^2}. \tag{2}$$

이러한 회전행렬을 찾는데, 가장 널리 사용되는 캅쉬 알고리즘[8]은 다음을 최소화하는 회전행렬을 찾는다.

$$\frac{1}{N}(\| X\|^2+\quad \| Y\|^2-2(s\sigma_1+\sigma_2+\sigma_3)) \tag{3}$$

여기서 σ_i는 3×3 행렬 X^TY의 세 개의 특이값으로 $0 \le \sigma_1 \le \sigma_2 \le \sigma_3$이고, s는 X^TY의 행렬식의 부호이다.

본 논문에서 두 번째로 사용하는 거리척도인 내부 알파탄소간의 거리는 단백질을 구성하는 아미노산 서열들에서 각 아미노산의 중심 원자인 알파탄소들 사이의 거리로서 단백질 구조를 나타낸다. 즉, m개의 아미노산으로 이루어진 단백질의 구조는 $m(m-1)/2$개의 알파탄소간의 거리를 원소로 가지는 벡터로 표현된다. 이렇게 표현된 벡터들 X, Y간의 유클리디언 거리로서 단백질 구조간의 거리를 나타내고, 이를 내부 알파 탄소간 거리 $IDC(X, Y)$로 나타낸다. 이 방법은 두 단백질 구조를 중첩하기 위한 변환이 필요하지 않으므로 고속 계산이 가능하나, 단백질 구조를 표현하는 벡터가 거울상 구조에 대해서 동일하므로 거울상 대칭을 구별하지 못하는 단점이 있다[2].

최근에 제안된 단백질 구조간의 거리인 비네-코시 점수는 비네-코시 항등식에 기초한다[9]. 두 개의 행렬 $X, Y \in R^{n\times m}$ $(m \le n)$에 대하여 다음의 비네-코시 항등식이 성립한다.

$$\det(X^TY) = \sum_{S/|S|=m} \det(X_S)\det(Y_S) \quad (4)$$

여기서, $S/|S|=m$는 $\{1,2,\ldots,n\}$의 숫자에서 중복없이 m개를 선택되어 만들어진 집합으로 X_S 는 X에서 이러한 m개의 행을 선택하여 얻은 부분행렬이다.

행렬 X가 단백질 구조를 나타낼 때에는 $m=3$이고, $\det(X_S)$는 원점과 X_S의 각 행에 해당하는 세 개의 점이 이루는 평행육면체의 체적이다. 따라서, 비네-코시 항등식의 우변은 단백질 구조 X, Y에서 각각 선택된 세 개의 점들과 원점으로 이루어지는 평행육면체들의 체적간의 상관관계를 나타내고, 이는 두 단백질 구조의 유사성을 나타낸다. 우변을 직접 계산하면 n개의 행에서 세 개의 행을 선택하는 경우의 수가 매우 많지만, 좌변을 이용할 경우에는 적은 연산으로 계산이 가능하다.

본 논문에서는 식 (4)를 정규화한 비네-코시 점수[9]를 사용한다.

$$BC(X, Y) = \frac{\det(X^TY)}{\sqrt{\det(X^TX)\det(Y^TY)}} \quad (5)$$

식 (5)는 행렬식의 특성으로부터 회전행렬 R에 대하여 불변이다. 즉, $BC(X, Y) = BC(X, YR)$이므로, 단백질 구조의 비교에 식 (2)의 *RMSD* 방법에서 필요한 최적 회전 행렬을 고려할 필요가 없으므로 고속 계산이 가능하다. 본 논문에서는 BC를 식 (6)의 간단한 방법으로 거리로 변환한 비네-코시 거리를 사용하였다.

$$BCD(X, Y) = 1 - BC(X, Y) \quad (6)$$

Ⅲ. 거리척도들의 조합을 사용한 단백질 조각 라이브러리 구성 방법

본 논문에서는 여러 거리 척도의 장점을 조합하여 대용량 단백질 자료로부터 단백질 구조를 정확하게 나타낼 수 있는 단백질 조각 라이브러리를 구성한다. 단백질 구조 자료에 군집화를 수행하고, 각 군집의 중심 단백질 조각을 대표적인 단백질 조각들을 선택하였다. 군집화를 초기화와 확장 단계로 나누고, 각 단계마다 여러 거리 척도의 조합을 사용한다.

초기화 단계에서는 미리 정해진 개수의 군집을 갖도록 군집화를 수행한다. 군집의 중심의 초기화는 미리 정해진 군집의 수만큼의 단백질 조각들을 단백질 구조 자료에서 임의로 선택하여, 각 군집의 중심으로 하였다. 내부 알파탄소간 거리를 사용하여 군집의 중심을 갱신하였는데, 학습 자료의 모든 단백질 조각과 이것에 가장 가까운 군집의 중심과의 거리를 구한 합들이 더 이상 작아지지 않을 때까지 갱신을 반복하였다. 알파탄소간 거리는 거울상 이미지를 구별하지 못하는 단점이 있으나, IV장의 실험결과에서 보듯이 평균제곱근 오차보다 1000배 이상의 계산 속도를 가지므로 대규모 자료에 대한 군집화를 고속으로 수행할 수 있다. 또한, 내부 알파탄소간 거리로 단백질 구조를 표현한 벡터는 추가적인 변환이나 계산 없이 군집화에 이용되며, 매우 안정적으로 수렴하는 장점이 있다.

확장 단계에서는 초기화 단계에서 얻은 군집을 기반으로 초기화 단계에서 사용한 내부 알파탄소간 거리의 단점을 보완할 수 있도록 약간의 군집을 추가한다. 즉, 알파탄소간 거리는 단백질 조각의 거울상 이미지를 구별하지 못하므로, 확장단계에서는 이를 구별할 수 있는 거리 척도를 사용하여, 현재 군집들의 중심과 가장 먼

거리를 가지면서 이상치(outlier)가 아닌 단백질 조각을 찾아서 군집의 개수를 증가시킨다.

비네-코시 거리 *BCD*는 거울상 이미지를 구별할 수 있으면서 단백질 구조의 유사성을 측정할 수 있으므로, 이미 만들어진 군집의 중심들과 가장 먼 거리를 가지는 학습 자료내의 단백질 조각을 찾는데 사용하였다. 이러한 군집의 개수를 증가시키는 과정에서 만들어지는 새로운 중심들과 가장 먼 거리를 가지는 학습 자료내의 단백질 조각을 새로이 탐색하려면 많은 시간이 소요되므로, 초기화 단계에서 얻은 군집에 대하여 거리가 먼 단백질 조각들을 구하고, 이들 중에서 다음 두 가지 조건을 만족하는 단백질 조각을 단백질 조각 라이브러리에 추가하는 간략한 방법을 사용하였다. 모든 학습 자료의 단백질 조각 X_k에 대하여 초기화 단계에서 얻은 군집의 중심들과 가장 가까운 비네-코시 거리를 이 단백질 조각과 군집간의 거리 BCD_k로 정의한다. 이러한 BCD_k를 내림차순으로 정렬하여, 큰 값에 해당하는 단백질 조각부터 다음의 두 조건을 만족하면 군집의 중심으로 추가한다. 첫 번째 조건은 이상치가 아닌지를 평가하는 식 (7)과 (8),

$$BCD(X_k, Y) < BCD_k/2 \tag{7}$$

$$ID(X_k, Y) < \mu(ID) \tag{8}$$

을 만족하는 학습 자료의 단백질 조각 Y가 3개 이상 존재하는 경우이다. 여기서 $\mu(ID)$는 학습 자료내의 단백질조각과 이것이 속하는 군집의 중심간의 알파탄소간 거리들의 평균이다. 첫 번째 조건의 계산에는 모든 학습 자료를 탐색해야 하므로, 원자간 평균 제곱근 오차 *RMSD* 보다 훨씬 계산량이 적은 비네-코시 거리 *BCD*와 내부 알파탄소간 거리 *ID*를 사용하였다. *BCD*는 가장 널리 사용되는 *RMSD*와의 상관관계가 크지 않으므로[9], *ID*를 동시에 고려한 거리로서 이상치를 찾았다.

두 번째 조건은 군집의 중심으로 추가되는 단백질 조각은 기존의 군집의 중심과 가깝지 않아야 하는 조건으로, 현재 군집의 중심들인 모든 C_i에 대하여

$$\mathrm{RMSD}(X_k, C_i) > TH_{\mathrm{rmsd}} \tag{9}$$

일 경우만 새로운 중심으로 추가한다. 여기서 TH_{rmsd}는 초기화 단계에서 얻은 각각의 군집의 중심에 대하여 가장 가까운 거리의 다른 중심과의 거리를 구하고, 이들 거리값 들에서 하위 사분위수(quantile)를 사용하였다. 두 번째 조건의 계산에는 *RMSD*를 사용하였다. 이는 군집의 개수는 학습 자료의 개수에 비하여 매우 작으므로 계산시간이 크게 소요되지 않기 때문에 계산량이 크더라도 거리척도로서 가장 널리 사용되는 *RMSD*를 사용하였다.

Ⅳ. 실험 및 결과

이 장에서는 제안한 방법으로 구성한 단백질 조각 라이브러리를 사용하여 단백질 구조를 표현하는 실험을 하고, 단백질 구조의 비교에 사용되는 거리척도들의 계산 속도를 조사한다.

실험에는 Astral SCOPe 2.04[10] 중에서 단백질 구조를 대표할 수 있고 단백질 서열간의 동일성이 40%이하인 중복성이 적은 단백질 구조 자료를 사용하였다. 본 연구에서는 단백질 구조의 중간에 절단이 없고 모든 삼차원 좌표가 존재하는 8204개의 단백질 자료를 추출하여 실험에 사용하였는데, 무작위로 선택한 80%는 단백질 조각 라이브러리를 구성하는 자료로 사용하고, 나머지 20%의 자료로서 단백질 조각 라이브러리의 성능을 평가하였다.

본 논문에서 사용한 8204개의 단백질 자료는 기존 연구[1-6]에서 각각 사용한 342 단백질, 200 도메인, 1429 단백질, 1428 도메인, 1020 체인, 4824 체인에 비교하면 대규모이다. 여기서, 도메인은 구조적으로 의미있는 독립적 구조이고, 체인은 1개 이상 모여서 단백질 하나를 구성하므로, 도메인과 체인은 단백질 구조의 일부분이다. 본 논문의 실험에 사용한 대규모 자료를 처리하기 위해서는 효과적이고 적은 계산량을 가지는 방법이 적용되어야 한다.

본 논문의 실험에서 가장 많은 시간을 소비하는 단백질 구조간의 거리를 계산하는 방법들의 수행시간을 조사하였다. 주어진 단백질 조각과 가장 거리가 가까운 단백질 조각 라이브러리내의 단백질 조각을 찾는 과정의 시간을 측정하였다. 이러한 과정은 단백질 조각 라이브러리를 구성하기 위한 군집화 과정에서 대부분의 계산에 해당되고, 단백질 라이브러리의 성능을 측정하

는 과정의 계산과정의 대부분을 차지한다. 실험은 리눅스 운영체제와 3.4GHz 클락 속도의 CPU를 사용하는 컴퓨터에서 수행되었다.

표 1에 길이가 4인 2,637,312개의 단백질 조각에 대하여 단백질 라이브러리를 구성하는 군집의 개수가 각각 100, 200, 300, 400, 500개인 경우의 수행시간을 모두 합한 후에 5로 나눈 시간을 나타내었고, 길이가 5인 2,630,749개의 단백질 조각은 군집의 개수가 각각 500, 1000, 2000, 3000, 4000개의 경우이다. 비네-코시 거리는 평균 제곱근 오차의 계산에 비하여 수행속도가 단백질 조각 길이가 각각 4와 5인 경우에 51배와 26배 빠르다.

이는 비네-코시 거리를 제안한 논문[9]의 결과와도 일치한다. 또 다른 거리척도인 내부 알파탄소간 거리는 평균 제곱근 오차보다 1000배 이상 고속으로 계산이 가능하다. 따라서 본 논문에서는 대규모 자료를 처리하기 위해서, 평균 제곱근 오차대신에 고속인 비네-코시 거리와 알파탄소간 거리를 적용하였다.

Table. 1 Average execution time of three distance measures for fragment length 4 and 5 (sec)

distance measures / fragment lengths	RMSD	Binet-Cauchy distance	inter alpha-carbon distance
4	7822.38	152.77	2.45
5	32960.88	1277.51	23.25

본 논문의 단백질 조각 라이브러리는 군집화를 수행한 결과로서 얻은 군집의 중심으로 이루어져 있고, 이 라이브러리의 유효성은 단백질 구조를 근사하는 정확도로 평가할 수 있다. 학습 자료와는 독립적인 나머지 20%의 1640개 단백질들을 단백질 조각으로 나눈 후에, 이와 가장 가까운 단백질 조각 라이브러리의 원소와의 평균제곱근 오차를 계산하였다. 표 2에는 이러한 *RMSD*의 평균으로서 단백질 조각 라이브러의 성능을 나타내었다. 단백질 조각의 길이가 4인 경우에는 각각 100, 200, 300, 400, 500의 초기화 군집 개수에서 시작하여, III장의 확장단계에서 20%까지 군집의 개수를 증가하거나, 20%이전에 식 (7)-(9)의 조건을 만족시키지 않는 경우에 확장단계를 종료하였다. 단백질 조각의 길이가 5인 경우에는 같은 방법을 사용하였고, 초기화 군집의 개수를 500, 1000, 2000, 3000, 4000을 사용하였다.

표 2에서 보듯이, 모두 0.5 Å 이하의 높은 정확도록 근사할 수 있었고, 같은 단백질 조각의 길이에서는 군집의 크기가 클수록 다양한 단백질 조각을 라이브러리에 포함하고 있으므로, 평균 제곱근 오차가 감소하였다. 또한 단백질 조각의 길이가 4일 때가 5일 때보다 적은 수의 단백질 조각을 근사하므로, 라이브러리의 크기가 작아도 더 정확한 근사가 가능하다.

Table. 2 RMSD for fragment length 4 and 5

fragment length / library size	RMSD average
4 / 120	0.200
4 / 240	0.164
4 / 323	0.154
4 / 478	0.139
4 / 591	0.129
5 / 600	0.306
5 / 1200	0.268
5 / 2400	0.242
5 / 3600	0.227
5 / 4800	0.217

기존 단백질 라이브러리[2]를 다운로드 하여, 본 논문에서 사용한 평가 자료로 그 성능을 측정하였다. 단백질 조각 14개와 단백질 조각의 길이가 4일 경우에 0.541 Å 이었고, 본 논문의 방법으로 구성한 크기가 14인 라이브러리의 경우에는 0.451 Å 으로 정확성이 증가하였다.

기존 단백질 라이브러리[2]에서 단백질 조각의 길이가 5이고, 라이브러리의 크기가 100, 224일 경우에 각각 0.641 Å , 0.592 Å 이었고, 본 논문의 방법으로 구성하였을 경우에는 0.449 Å , 0.366 Å 으로 정확성이 증가하였다. 기존의 라이브러리[2]는 200개의 도메인으로 구성된 작은 단백질 자료에서 최적화 된 것으로, 보다 대규모 자료에 대한 실험에서는 정확도가 감소함을 알 수 있다. 또한, 기존 라이브러리의 구성방법은 많은 계산시간 때문에 대규모 라이브러리를 제공하지 않으므로 본 논문에서 제안한 나머지 대규모의 라이브러리와의 비교를 수행할 수는 없었다.

실험을 통하여 알아본 바와 같이, 본 논문에서 제안한 계산량이 적은 거리 척도를 사용하여 대규모 자료에서 단백질 조각 라이브러리의 구성이 가능하였고, 기존의 단백질 조각 라이브러리보다 대규모인 단백질 조각 라이브러리가 평균 제곱근 오차를 감소시킴을 확인하였다.

V. 결 론

본 논문에서는 대규모 단백질 자료를 이용하여 단백질 조각 라이브러리를 구성하기 위해 계산량이 적은 거리 척도를 제안하였다. 기존의 거리 척도를 효과적으로 조합하였는데, 초기의 군집화 과정에서는 가장 계산량이 작은 내부 알파탄소간 거리를 사용하였고, 확장단계에서는 내부 알파탄소간 거리, 비네-코시거리와 평균 제곱근 오차거리를 조합하여 사용하였다.

제안한 거리척도를 사용한 단백질 조각 라이브러리의 구성은 기존의 방법에 비하여 계산 시간이 크게 감소되므로 대용량 분석기술의 발전에 따라 크게 증가하는 단백질 구조 자료에 적용이 가능하다. 제안한 대규모 단백질 조각 라이브러리로 단백질 구조를 표현하였을 경우에 평균 제곱근 오차가 더욱 감소하므로, 이를 여러 단백질 구조 분석에 적용할 수 있다. 하지만, 제안한 방법을 단백질 구조의 분석, 모델링, 탐색, 예측 등에 효과적으로 적용하기 위해서는 단백질 서열을 단백질 조각 라이브러리로 변환하는 방법과 단백질 조각 라이브러리로 표현된 단백질 구조들을 비교하는 방법의 개발이 필요하다.

REFERENCES

[1] A. G. de Brevern, C. Etchebest, and S. Hazout, “Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks,” *Proteins*, vol. 41, pp. 271-287, 2000.

[2] R. Kolodny, P. Koehl, L. Guibas, and M. Levitt, “Small libraries of protein fragments model native protein structures accurately,” *Journal of Molecular Biology*, vol. 323, pp. 297-307, 2002.

[3] A. C. Camproux, R. Gautier, and P. Tuffery, “A hidden markov model derived structural alphabet for proteins,” *Journal of Molecular Biology*, vol. 339, pp. 591-605, 2004.

[4] T. Hamelryck, J. T. Kent, and A. Krogh, “Sampling realistic protein conformations using local structural bias,” *PLoS Comput. Biol.* vol. 2, e131, pp. 1121-1133, 2006.

[5] S. C. Li, D. Bu, J. Xu, and M. Li, “Fragment-HMM: A new approach to protein structure prediction,”, *Protein Science*, vol. 17, pp. 1025-1934, 2008.

[6] I. Kalev and M. Habeck, “HHfrag: HMM-based fragment detection using HHpred,” *Bioinformatics*, vol. 27, no. 22, pp. 3110-3116, 2011.

[7] A. P. Joseph, et al., “A short survey on protein blocks,” *Biophys. Rev.* vol. 2, pp. 137-145, 2010.

[8] W. Kapsch, “A discussion of the solution for the best rotation to relate two sets of vectors” *Acta Crystallog. sect.*, vol. 34, pp. 827-828, 1978.

[9] F. Guyon and P. Tuffery, “Fast protein fragment similarity scoring using a Binet-Cauchy kernel,” *Bioinformatics*, vol. 30, no. 6, pp. 784-791, 2014.

[10] N. K. Fox, S. E. Brenner J. M. Chandonia, “SCOPe: Structural Classification of Proteins－extended, integrating SCOP and ASTRAL data and classification of new structures,” *Nucl. Acids Res*. 42(Database issue), D304-309, 2014.

지상문(Sang-Mun Chi)

1991년 서울대학교 수학교육학과 졸업(이학사)
1993년 한국과학기술원 수학과 졸업(이학사)
1998년 한국과학기술원 전산학과 졸업(공학박사)
1993년 ~ 2000년 삼성전자 무선사업부 선임연구원
2001년 ~ 현재 경성대학교 컴퓨터공학과 교수
※관심분야 : 생물정보학, 기계학습, 이산미분기하