제22회 한국정보처리학회 추계학술발표대회 논문집 제11권 제2호 (2004. 11)

단백질 구조 비교를 위한 전처리 기법으로서의 주성분 분석

박성희, 박찬용, 김대희, 박수준, 박선희
한국전자통신연구원 바이오정보연구팀
e-mail : {sunghee, cypark, dhkim98, psj, shp}@etri.re.kr

Principal Component Analysis as a Preprocessing Method for Protein Structure Comparison

Sung Hee Park, Chan Yong Park, Dae Hee Kim, Soo-Jun Park, Seon Hee Park
Bioinformatics Research Team, ETRI at Daejeon, Korea

요 약

본 논문에서는 두 단백질의 구조적 유사성을 기반으로 한 단백질 비교를 위해서 전처리 기법으로서의 주성분분석기법을 소개한다. 기존의 백본 및 알파탄소 간의 거리행렬(distance matrix), 2 차 구조 비교기법, 구역(segment)단위의 비교 기법과 같은 단백질 비교 기법들은 위치이동(translation)와 회전(rotation)에 불변한(invariant) 차이를 구하기 위하여 거리행렬을 이용하였다. 그리고, 난 다음 이들의 최적화 과정을 거쳤다. 그러나, 본 논문에서 제시하는 전처리 기법으로서의 주성분분석기법은 단백질 구조를 전체적인 구조 관점에서 위치를 정렬시킨 후에 단백질 간의 구조를 비교하는 방식이다. 단백질의 구조의 방향성(Orientation)을 맞춘 다음에는 다양한 단백질 표현으로 구조를 비교할 수 있다. 본 논문에서는 두 단백질의 구조의 유사성을 측정하기 위한 간결한 단백질 표현(representation)으로 3 차원 에지 히스토그램을 사용하였다. 이 기법은 방향성을 정렬하기 위하여 기존의 방법에서 사용되었던 반복적인 거리계산을 통한 최적화하는 과정을 없앰으로써 단백질 구조 비교 시간을 단축할 수 있는 새로운 단백질 구조 비교 패러다임을 가능하게 한다. 따라서, 이 패러다임을 통하여 적절한 단백질 구조 방향성 정렬과 단백질 구조 표현을 이용한 단백질 구조 비교 검색 시스템은 많은 양의 단백질 구조 정보로부터 원하는 형태의 단백질 구조를 빠른 시간에 검색할 수 있는 장점을 가질 수 있다.

1. 서론

생체 내에서의 생화학작용들은 유전자 발현에 의해 생성된 생물분자(biomolecule)인 단백질의 작용에 의해서 대부분 이루어진다. 그리고, 그 기능은 단백질의 3 차원적 구조(모양)에 의해 결정된다. 따라서, 두 단백질의 구조간의 유사성을 측정하는 방법은 두 단백질의 기능의 유사성을 유추할 수 있다. 즉, 구조결정학자들은 새롭게 밝혀낸 단백질 구조와 기존에 기능이 알려진 단백질의 구조와 비교를 통하여 새로운 단백질의 기능을 예측하려 하였다.

이를 위해서 지금까지 단백질 구조 비교를 위한 많은 단백질 표현(representation) 혹은 기술자와 그에 따르는 유사척도(similarity measure)가 제안되어 왔다. 초기에는 단백질 구조는 가장 평범하게 단백질을 구성하는 원자 전체, 단백질의 백본(backbone)을 구성하는 원자들, 알파 탄소 원자들 등으로 표현되어 왔다. 이는 계산량이 너무 많고 에러에 민감한 단점을 보완하려는 시도로 발전되어 왔다[1]. 또한, 최근에는 단백질을 일정한 아미노산 수 만큼씩 잘라서 그 잘라진 아미노산의 알파탄소의 위치의 평균값을 가지고 위와 같은 유사도를 측정하여 속도도 줄이면서 에러에 민감한 단점을 보완하는 연구가 있었다[2]. 다른 접근 방법으로 단백질들을 그 단백질이 포함하는 2 차 구조의 벡터형태로 표현하고 이들 벡터를 이용하여 유사도를 측정하는 방법에 대한 연구가 있었다[3]. 이러한 단백질 표현들은 이동 및 회전에 불변한 유사도 계산을 위하여 거리행렬(distance matrix)를 활용한다. 이 방

법은 최소의 거리를 찾기 위하여 최적화 과정을 거친다.

본 논문에서는 또다른 단백질 비교 알고리즘의 접근으로 단백질 구조의 이동 및 회전에 영향을 받은 단백질 구조 비교 알고리즘을 위하여 단백질 구조를 전체적으로 방향성(orientation)을 정렬한 후 간단한 단백질 표현에 의해 유사도를 측정하는 단백질 비교 방식을 제안한다.

여기서 단백질 구조의 방향성을 일치하기 위한 방법으로 컴퓨터 비젼 및 디지털 이미지 처리 분야에서 물체의 방향성을 정렬하는데 많이 사용된 주성분분석 및 호텔링 변환(Hotelling Transform)을 이용한다.

그리고, 변환된 후의 단백질 구조의 유사도 측정을 위하여 단백질 표현 기법은 원자들 사이에 형성되는 결합(bond)들의 3 차원 공간 상에서의 지역적 분포를 이용한 3D 에지 히스토그램을 사용하였다[4].

다음 장에서는 기존의 단백질 구조 비교 알고리즘과 제안된 단백질 구조 비교 알고리즘을 비교하고 3 장에서 주성분 분석을 통한 단백질 구조 방향성 정렬 결과를 알아 본다. 4 장에서는 이들을 이용한 단백질 비교 검색 시스템을 알아본다. 5 장에서 결론을 맺는다.

2. 기존의 알고리즘과 제안된 단백질 구조 비교 알고리즘과의 비교

기존의 알고리즘과 제안된 단백질 구조 비교 알고리즘의 순서도를 그림 1 과 같이 비교하였다.

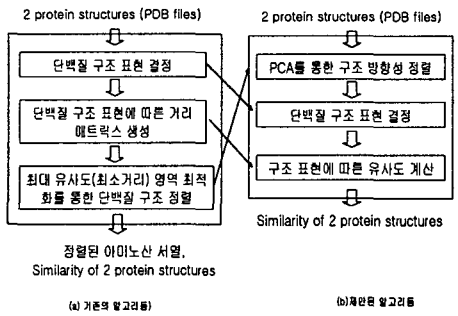


그림 1 기존(a)과 제안(b)된 단백질 구조 비교 순서도 비교

그림 1 (a)에서 볼 수 있듯이, 단백질 구조 비교를 위해서는 단백질의 거리 비교에 사용될 단백질 구조의 표현을 결정해야 한다. 기존에 사용되어 왔던 단백질 표현으로는 백본을 구성하는 원자들의 3 차원 위치, 알파탄소들만의 위치, 아미노산 서열을 몇 개의 구역(segment)로 나눠서 그 구역에 포함된 알파탄소들의 무게 중심 등이 있었다. 이들은 모두 원자 수준에서의 표현들이고, 다른 수준의 표현으로 2 차 구조들로 단백질을 표현하기도 하였다.

이들 단백질 표현을 이용하여 거리행렬를 생성한다. 이 행렬은 행은 비교하는 두 단백질 중 한 단백질 A 의 구조 표현 요소(예를 들어, 알파탄소)들을 의미하고 열은 다른 한 단백질 B 의 구조 표현 요소(앞 단백질과 동일한 표현)들을 의미한다. 따라서 행렬의 (i, j) 의 값은 A 단백질의 단백질 표현 i 번째 요소와 B 단백질의 단백질 표현 j 번째 요소간 거리를 유사도 혹은 거리계산척도를 이용해 계산한 수치들이다. 이 거리행렬을 이용하여 요소들간의 거리가 가장 최소인 조합을 찾아냈을 때, 단백질이 구조적으로 최적 정렬이 된 것이다. 그러나, 최소의 조합을 찾아내는 최적화 과정은 시간이 많이 걸리는 작업이다.

본 논문에서 제안한 단백질 구조 비교 알고리즘은 기존의 방식의 단백질 구조 비교에서 구조 정렬을 위한 최적화 과정을 단백질 구조의 고려할 사항을 통한 단백질 구조 정렬 과정을 먼저 행한 후에 단백질 구조 거리(유사도)를 계산하는 방식이다. 이 방식은 최적화 과정이 필요하지 않으므로 시간을 단축할 수 있는 장점이 있다.

컴퓨터 비젼 및 디지털 이미지 처리과 같은 분야에서는 2 차원 및 3 차원의 물체를 추출 및 인식하기 위한 기법들을 많이 연구해 왔다. 그 중의 고전적인 방법 중의 하나로 주성분 분석을 이용한 물체의 방향성 정렬 기법이 있다.

본 논문에서는 이러한 주성분 분석을 통한 단백질 구조의 방향성 정렬을 수행한다. 주성분 분석에 의한 단백질 구조의 방향성 정렬의 결과를 다음 장에서 살펴본다.

3. PCA 를 이용한 단백질 구조 방향성 정렬

논문에서는 단백질의 구조의 특징을 이용하여 먼저 단백질 구조를 정렬한 후 단백질의 유사도를 계산하는 방식을 제안한다. 그 방법의 효용성을 보이기 위하여, 본 논문에서는 단백질을 구성하는 원자들의 3 차원 좌표를 3 차원 자료로 보고 주성분 분석을 하여 단백질의 방향성을 정렬하는 방법을 제시한다.

단백질은 X-ray crystallography 방식이든 핵자기공명(NMR: Nuclear Magnetic Resonance)방식으로 단백질 구조가 결정되더라도 단백질 구조의 3 차원적 위치를 동일하게 부여할 기준이 존재하지 않는다. 즉, 동일한 단백질 구조를 가지고 구조를 결정하더라도, 이동 및 회전이 3 차원적 위치에 포함되게 된다. 이러한 단백질 3 차원 구조의 이동 및 회전은 단백질을 구조적으로 비교하는 데, 해결해야 할 문제를 제시한다.

이와 같이 단백질의 회전과 이동에 따른 동일 단백질의 다양한 3 차원 구조 표현을 해결할 수 있는 방법은 방향성 정렬을 수행하는 것이다.

방향성 정렬 방법으로 본 논문에서는 주성분 분석 기법을 사용한다.

예를 들어, 그림 2 에서 보듯이 단백질 1gp2:G 의 단백질 구조 정보 데이터베이스(PDB: Protein Data Bank)내에 들어 있는 정보는 y 축으로 긴 모양을 나타내고 있고(a) 1a0r:G 는 x 와 z 축으로 길쭉한 모양을 나타내고 있다(b).

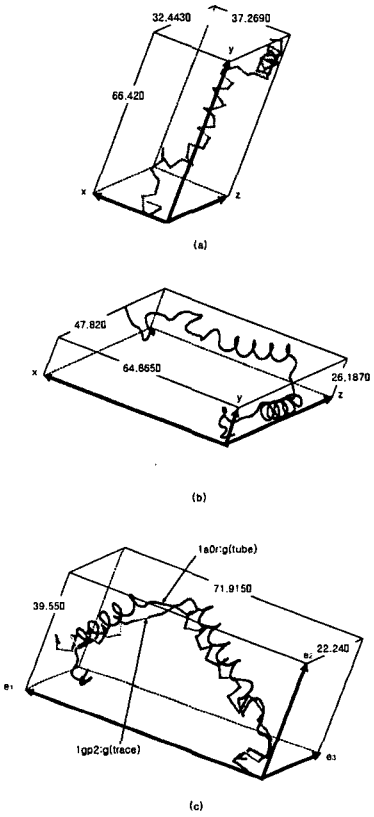


그림 2 단백질 1gp2:g 와 1a0r:g 의 주성분 분석 전과 후의 3 차원적 위치. (a) 주성분 분석 전 1gp2:G (b) 주성분 분석 전 1a0r:g (c) 주성분 분석 후 1gp2:g 와 1a0r:g 의 겹쳐진 위치

이들을 주성분 분석을 통하여 좌표변환을 하면 두 단백질이 새로운 좌표계에 의하여 주성분에 따라 정렬이 됨을 볼 수가 있다(c). (a), (b) 는 1gp2:g (y>>x>z 여기서 x: 37.269 Å, y: 66.42Å, z: 32.443 Å) 와 1a0r:g (x>z>y 여기서 x: 64.865 Å, y: 26.187 Å, z: 47.82 Å)의 PDB 파일의 위치, (c) 는 두 단백질 1gp2:g (x > y > z 여기서 x: 71.915 Å, y: 33.785 Å, z: 14.029 Å) and 1a0r:g (x > y > z 여기서 x: 69.492 Å, y: 39.55 Å, z: 22.24 Å)의 을 호텔링 변환 후 두 단백질의 위치를 겹쳐진 그림으로 표현하였다.

4. 단백질 구조 비교 검색 시스템 설계 및 구현

방향성이 정렬된 두 단백질의 구조적 유사성을 위하여 본 논문에서는 단백질 구조 표현 중의 하나인 3D 에지 히스토그램을 이용하였다. 주성분 분석을 통하여 먼저 방향성 정렬을 수행하고 3D 에지 히스토그램으로 단백질을 표현한 후 유사도 계산을 통한 단백질 구조 비교 및 검색 시스템을 설계하고 구현하였다.

A. 시스템 구성도

구현된 시스템의 구성도는 그림3과 같다.

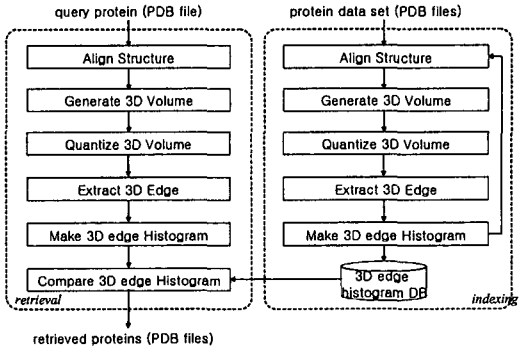


그림 3 제안된 방법을 이용한 시스템 구성도

B. 색인

색인은 검색에 사용될 색인 파일을 만드는 과정으로 단백질들을 모두 호텔링 변환을 한 후 3D 에지 히스토그램을 추출한다. 실험에서는 9700 여 개의 단백질 도메인 PDB 파일을 사용하였다.

C. 검색

검색인터페이스(그림 7)에서 질의형태로 비교하고자 하는 PDB 코드를 질의 창에 입력하고 검색(Retrieve)버튼을 누르면 검색결과가 보여진다. 그림 7 는 질의 및 결과 인터페이스를 보여주며 단백질 1a5k 의 체인 B 를 질의단백질로 한 질의 결과를 보여주고 있다.

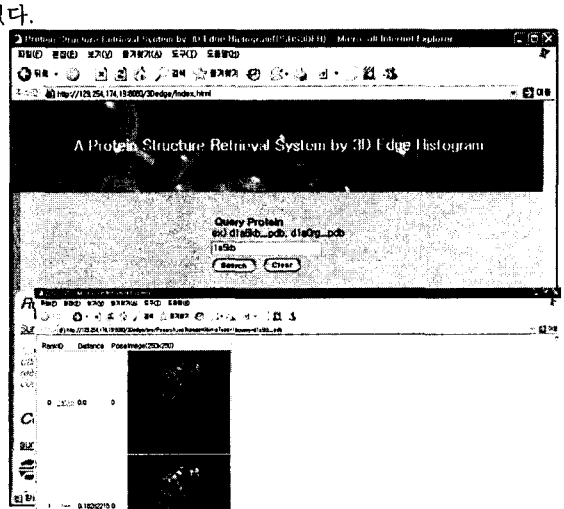
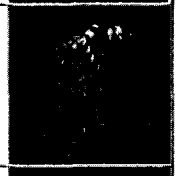


그림 4. Retrieval & Result interface (query protein: chain b of 1a5kb)

D. 검색결과

단백질 1a5k 의 B 체인에 대한 단백질 구조 검색을 수행한 결과를 표 2 에서 보여주고 있다. 유사도(여기서는 히스토그램간의 "차이" 개념으로 값이 작을수록 유사 정도가 커짐)가 0.5 이하의 파일의 경우 유사한 모양의 단백질을 검색함을 볼 수 있다(표 2).

표 1. 검색결과 (질의단백질: 1a5k:b)

Rank	ID	Distance	Pose	Image(250x250)
0	1a5kb	0.0	0	
1	1fwab	0.18282215	0	
2	1fwcb	0.18573979	0	
3	1a5mb	0.19384164	0	
4	1fwbb	0.20823689	0	
5	1fwjb	0.21349344	0	

4. 결론

본 논문에서는 기존의 단백질 구조 비교 방식과는 다른 구조의 방향성을 먼저 정렬한 후 단백질 표현에 의한 단백질 구조 유사도 계산 방식으로 반복적이지 않고 1 회의 유사도 계산 방식을 제안하였다. 이를 위해 먼저, 비교하고자 하는 단백질 데이터 베이스의 단백질을 기하학적 정렬을 위하여 주성분분석(PCA)를 하고 이들 결합선 분포를 이용한 3D 에지 히스토그램으로 표현한 단백질 구조를 비교 하였다. 이는 1 회의 비교를 통해 단백질의 유사정도를 계산해 낼 수가 있어 기존의 단백질 구조 비교 시스템의 처리 시간을 단축하는 장점이 있다. 빠른 검색을 통하여 스크리닝 전처리(prescreening) 단계에서 사용될 경우 더 정밀한 구조비교에 앞서 매우 효율적일 것으로 보여진다.

참고 문헌

[1] Lholm and C.Sander, "Protein Structure Comparison by alignment of distance matrices", Jounal of Molecular Biology, Vol. 233, pp. 123-138, 1993

[2] Rabian Schwarzer and Itay Lotan, "Approximation of Protein Structure for Fast Similarity Measures", Proc. 7th Annual International Conference on Research in Computational Molecular Biology(RECOMB), pp. 267-276, 2003

[3] Amit P. Singh and Douglas L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representation", Proc. Intelligent Systems for Molecular Biology, 1993

[4] S.H. Park, S.J. Park, S.H.Park, "A Protein Structure Retrieval System Using 3D Edge Histogram," Key Engineering Material Vol 277-279, pp. 324-330, 2004