원자간 결합 분포를 이용한 단백질 구조 검색 시스템

박성희, 박수준, 이성훈, 박선희
*한국전자통신연구원 바이오정보연구팀
e-mail : sunghee@etri.re.kr

Protein Structure Retrieval System using Bond-line Histogram of Atoms

Sung Hee Park, Soo Jun Park, Seong Hun Lee, Seon Hee Park
Bioinformation Research Team, ETRI at Daejeon, Korea

요 약

현재 생물분자의 기능적 관점에서 단백질 구조에 관심이 많이 모아지고 있다. 단백질의 기능은 구조에서 기인하기 때문에 두 단백질의 구조간의 유사성을 측정할 수 있는 방법은 두 단백질의 기능의 유사성을 유추할 수 있다. 본 논문에서는 두 단백질의 원자간 결합선 분포의 유사성을 기반으로 한 웹 환경에서 동작하는 단백질 구조 검색 시스템을 설계 구현하였다. 두 단백질의 구조의 유사성을 측정하기 위한 단백질의 표현(representation)으로는 3 차원 에지 히스토그램을 사용하였다. 3 차원 에지 히스토그램, 즉, 3 차원 공간 상에서의 원자간 결합선 분포에 기반한 단백질 구조 검색 시스템은 많은 양의 단백질 구조 정보로부터 원하는 형태의 단백질 구조를 빠른 시간에 검색할 수 있는 장점을 가지므로 스크리닝의 전단계(pre-screening)에서 사용될 수 있다.

1. 서론

생체 내에서의 생화학작용들은 유전자 발현에 의해 생성된 생물분자(biomolecular)인 단백질의 작용에 의해서 대부분 이루어진다. 그리고, 그 기능은 단백질의 3 차원적 구조(모양)에 의해 결정된다. 따라서, 두 단백질의 구조간의 유사성을 측정할 수 있는 방법은 두 단백질의 기능의 유사성을 유추할 수 있다. 즉, 구조결정학자들이 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 유사척도(similarity measure)가 제안되어 왔다. 초기에는 단백질 원자의 위치와 원자들 간의 거리 비교에 따라 유사도 측정을 하였다. 이는 계산량이 너무 많고 에러에 민감한 단점이 있어 단백질 알파 탄소의 위치만을 가지고 유사도를 측정하였다[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 접근방법으로 단백질들을 그 단백질이 포함하는 2 차구조의 벡터형태로 표현하고 이들 벡터를 이용하여 유사도를 측정하는 방법에 대한 연구가 있었다[3].

본 논문에서는 또다른 접근으로서 단백질의 원자들이나 혹은 특정원자들의 위치에 의한 표현과는 달리 원자들 사이에 형성되는 결합(bond)들의 분포를 이용하여 두 단백질을 비교하는 기법을 제안하고 이를 구현하였다. 이 기법은 단백질의 원자들간의 결합을 선으로 표현한 막대모형(stick model)이나 선모형(wireframe model)같은 표현 모델에서 3 차원 공간 상에서 에지를 추출하여 이를 히스토그램화하고 두 히스토그램간의 유사도를 측정하는 방법이다.

다음 장에서는 원자간 결합선 분포를 표현한 3 차원 에지 히스토그램을 이용한 단백질 구조 비교 모델을 제시하고 3 장에서는 이를 적용하여 웹환경에서 단백질 구조를 비교 검색하는 시스템의 프로토타입을 구현하고 검색결과를 살펴보았고 4 장에서 결론을 맺는다..

2. 3D 에지 히스토그램을 이용한 단백질 구조 비교 모델

3D 에지히스토그램을 이용한 단백질 비교 순서도는 그림 1 과 같다.

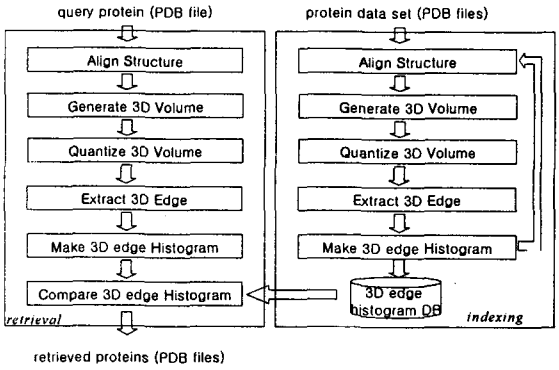


그림 1. Flowchart of protein structure retrieval system

A. *Align Structure(AS)*

3 차원 구조 정렬은 매우 어려운 문제 중의 하나이다. 본 논문에서는 단백질 3 차원 전체구조의 방향성(orientation)을 정렬시키기 위하여 주성분분석(Principal Component Analysis)를 이용한다. 주성분분석의 기하학적인 의미는 가장 길쭉하게 퍼진 축을 주축으로 삼고 그 주축으로 정렬을 할 수 있다는 데 있다.

그림 2 는 단백질 도메인 1a0r 의 G 체인의 주성분 분석 전과 주성분 분석 후의 구조 정보를 보여준다. 주성분 분석 후에 장축 순으로 변환된 것을 볼 수 있다.

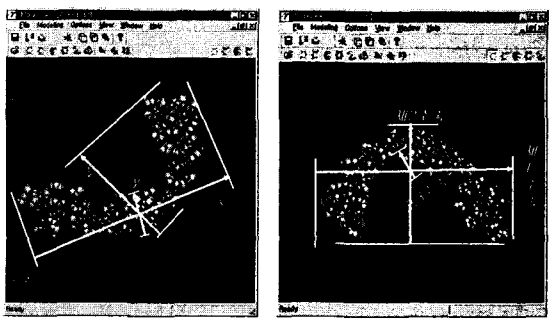


그림 2. 주성분 분석 전(a)와 주성분 분석 후(b)의 축 길이((a) x 축:64.865Å, y 축: 26.187 Å, z 축: 47.82 Å (b)제 1 주축: 69.492 Å, 제 2 주축: 39.55 Å, 제 3 주축: 22.24 Å)

B. *Generate 3D Volume(GV)*

원자간 결합선 분포를 구하기 위해서는 3 차원 공간을 일정한 크기로 자르거나(digitizing) 일정한 간격으로 취해야 한다(sampling). 주성분 분석에 의하여 변환된 단백질 구조정보로부터 원자들의 3 차원 위치정보를 읽어 들인다. 읽어 들인 위치정보로부터 결합(bond) 정보를 생성하고 이를 이용하여 3 차원 입체(volume)를 생성하기 위하여 공간적 샘플링을 수행한다.

C. *Quantize 3D Volume(QV)*

3 차원 양자화 과정에서는 단백질 3 차원 구조 공간을 잘게 복셀(voxel)로 나누고 결합선(bond)이 복셀을 지나는 경우 1 지나지 않는 경우 0 으로 표현한다. 이렇게 전체 3 차원 구조 공간을 이진화로 양자화한다. 그림 3 에서 결합선이 지나는 복셀의 경우 짙은 색으로 표현되고 그렇지 않는 복셀은 연한 색으로 표현되었다.

그림 3. Example of 3D structure space quantation

D. *Extract 3D Edge(EE)*

양자화 과정을 거치면 그림 3 와 같이 결합선이 지나는 부분과 지나지 않는 부분에 의해 경계선이 생긴다. 3 차원 에지 추출과정은 이들 경계선을 종류별로 추출하는 단계이다. 이 과정에서는 우리는 그림 4 과 같이 10 종류의 3 차원 에지를 정의하고 각 복셀에 대해서 10 종류의 edge 성분을 추출한다.

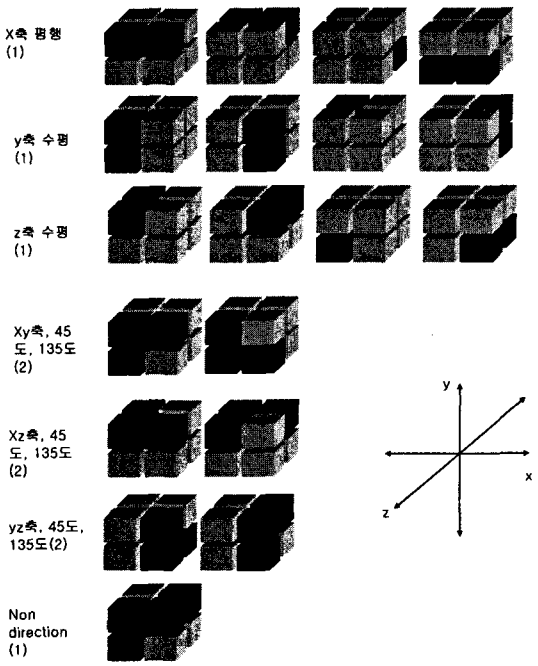


그림 4. Ten 3D-edge patterns

최상위의 에지 패턴은 x 축에 평행한 에지로 4 가지

가 생성될 수 있는데, 이들을 같은 x 축 평행 에지패턴으로 본다. y, z 축 평행 에지 패턴도 x 축 평행에지 패턴과 같이 4 가지가 생성될 수 있으나, 이를 각각 같은 에지 패턴으로 본다. xy 축, xz 축, yz 축에 각각 45 도, 135 도의 에지패턴이 가능하다. 그리고, 마지막으로 방향성을 정할 수 없는 비방향성 에지패턴을 정의하였다. 따라서, 총 10 가지의 에지패턴을 정의하여 에지를 추출한다. 그림 3 에서 괄호안의 숫자는 해당블럭에서 정의되는 에지의 수를 나타낸다.

E. Make 3D edge Histogram(MH)

3 차원 에지 추출과정(EE)을 통하여 추출된 3 차원 에지들의 분포, 즉, 3 차원 에지 히스토그램을 생성하기 위하여 3D 입체(volume)를 먼저 3 차원 구조 공간을 각 축에 대하여 4 X 4 X 4 로 나눈다(그림 5). 나누어진 구조 공간을 subblock 이라 하고 각 subblock 에 대하여 위에서 정의한 10 종류의 에지 패턴을 추출한다. 총 히스토그램 빈 수는 640 개이며 이들 각각을 3 차원 에지히스토그램이라 한다.

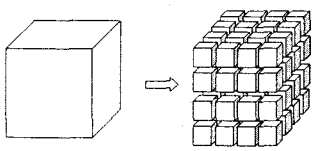


그림 5. Subblocks in 3D space

3 차원 에지 히스토그램 빈의 의미는 표 1 과 같다.

표 1. Semantics of 3D edge histogram bins

bins	semantics
3D_Edge[0]	X axis parallel edge of subblock(0,0,0)
3D_Edge[1]	Y axis parallel edge of subblock(0,0,0)
3D_Edge[2]	Z axis parallel edge of subblock(0,0,0)
3D_Edge[3]	Xy axis 45 degree edge of subblock(0,0,0)
3D_Edge[4]	Xy axis 135 degree edge of subblock(0,0,0)
3D_Edge[5]	Xz axis 45 degree edge of subblock(0,0,0)
3D_Edge[6]	Xz axis 45 degree edge of subblock(0,0,0)
3D_Edge[7]	Yz axis 45 degree edge of subblock(0,0,0)
3D_Edge[8]	Yz axis 45 degree edge of subblock(0,0,0)
3D_Edge[9]	Non-directional edge of subblock(0,0,0)
3D_Edge[10]	X axis parallel edge of subblock(0,0,1)
…	…
3D_Edge[638]	Yz axis 45 degree edge of subblock(3,3,3)
3D_Edge[639]	Non-directional edge of subblock(3,3,3)

3. 구현 및 결과

A. 시스템 구성도

본 시스템은 웹 환경에서 동작하는 서버-클라이언트 모델로서 웹에 접속하여 다중 사용자에 의해 서비스될 수 있는 형태의 프로토타입이다. 그림 6 는 이러한 시스템의 구성도를 대략적으로 보여준다.

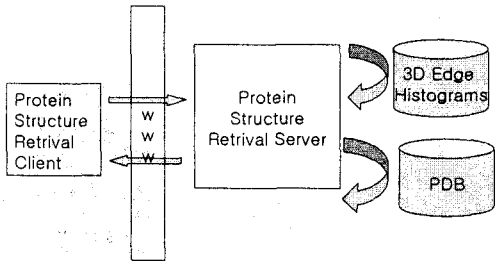


그림 6. Architecture of protein structure retrieval system

B. 색인

색인은 검색에 사용될 색인 파일을 만드는 과정으로 단백질들로부터 히스토그램을 추출한다. 실험에서는 100 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 오른 쪽 결과 프레임에 보여진다. 그림 7 는 질의 및 검색 인터페이스를 보여주며 단백질 1a5k 의 체인 B 를 질의단백질로 한 질의 결과를 보여주고 있다.

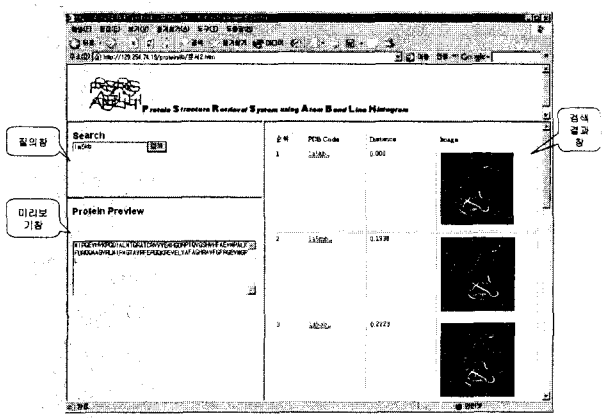
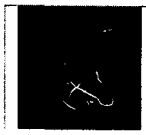
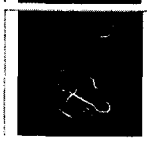


그림 7. Retrieval interface (query protein: chain b of 1a5kb)

D. 검색결과

단백질 1a5k 에 대한 단백질 구조 검색을 수행한 결과를 표 2 에서 보여주고 있다. 유사도(여기서는 히스토그램간의 "차이" 개념으로 값이 작을수록 유사정도가 커짐)가 0.5 이하의 5 개의 파일의 경우 유사한 모양의 단백질을 검색함을 볼 수 있다(표 2).

표 2. 검색결과 (질의단백질: 1a5kb)

순위	PDB Code	Distance	Secondary Structure Image
1	1a5kb_	0.000	
2	1a5mb_	0.1938	
3	1a5ob_	0.2223	
4	1a5nb_	0.2243	
5	1a5lb_	0.3268	
6	1a4pa	1.0198	
7	1a1ua_	1.0503	
8	1a1uc_	1.0589	
9	1a3p__	1.0739	
10	1a1ka3	1.0748	

4. 결 론

본 논문에서는 기존의 단백질 원자의 위치에 기반 단백질 구조비교에서 단백질 원자들간의 결합선의 분포를 이용한 단백질 구조비교기법을 제시하였다. 단백질 결합선 분포를 이용한 단백질 구조 비교를 위해 단백질 결합선의 종류를 10 가지의 3D 에지로 정의하였고, 이를 이용하여 웹 상에서 단백질 구조 검색을 할 수 있는 검색시스템을 구현하였다. 먼저, 비교하고자 하는 단백질 데이터 베이스의 단백질을 기하학적 정렬을 위하여 주성분분석(PCA)를 하고 이들 결합선 분포를 이용한 단백질 구조 비교가 단백질 전체의 모양과 원자들간의 결합각도를 세밀하게 비교하고 있지는 않지만 전체적인 분포를 고려함으로 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