Enzyme의 활성 사이트 표면 비교기법 설계

유남희*, 정광수*, 류근호*, 정용제
*충북대학교 데이터베이스/바이오인포메택스연구실, 충북대 생명광학부
e-mail:{nami,ksjung,khryu}@dblab.chungbuk.ac.kr, chungyj@chungbuk.ac.kr

Designing of Surface Comparison Method on Active Site of Enzyme

Nam Hee Yu*, Kwang Su Jung*, Keun Ho Ryu*, Yong Je Chung
*Database/Bioinformatics Laboratory,
Division of Life Science Chungbuk National University

요 약

단백질의 구조는 그 기능과 밀접히 연관되어 있기 때문에 구조에 조금이라도 변화가 생기면 바로 생체기능에 이상이 생긴다. 그래서 단백질 구조연구는 필수적이고 구조의 유사성 검색을 이용하여 단백질 기능을 예측한다. 그러나 전체적인 구조가 유사한 단백질이라도 기능에 중요한 특정구조가 다르게 되면 다른 기능을 수행 할 수 있고 구조가 다른 단백질이라도 핵심 영역의 구조가 유사하다면 유사한 기능을 수행할 수 있다. 이는 단백질의 기능이 특정 하위구조의 잘 보존된 활성 사이트에 따라 결정되기 때문이다. 이 논문은 단백질의 3차원 공간정보를 matrix로 표현 할 수 있는 가장 작은 평면도형인 삼각형을 이용하여 단백질 표면에 대한 상세한 형태비교를 제공한다. 단백질 표면에서 활성 사이트 아미노산 잔기의 side chain은 일반적으로 바깥을 향하여 표면의 형태를 결정짓기 때문에 단백질 표면을 비교하기 위해 side chain 정보가 필수적이다. 우리는 아미노산 잔기의 Cα원자에 side chain을 포함하여 Cα삼각형과 side chain 삼각형 2개를 하나의 특정하위구조 set으로 정의하고 이 하위구조로 distance matrix를 구축한다. 만들어진 distance matrix에 RMSD를 이용하여 활성 사이트의 표면을 비교한다. 제시한 기법은 단백질의 전체적인 서열과 구조 정보를 이용하지 않고, 활성 사이트의 특정하위 영역만을 고려함으로써 더욱 효과적이고 빠른 시간 내에 상세한 비교를 수행할 수 있다.

1. 서론

분자구조는 생명의 기능을 조정하는데 가장 구체적인 역할을 가지고 있다. 단백질의 구조 연구가 중요한 이유는 이들이 생체 내 기능 수행 시 각종 물질들이나 단백질들끼리 상호결합을 해야 하는데 이런 결합성을 결정짓는 것들이 3차원 구조이기 때문이다. 가령 후천성면역결핍증(AIDS) 치료약물의 표적인 단백질 분해 효소의 3차원구조를 알면 저해제를 보다 잘 디자인 할 수 있어 치료제 개발도 쉬워질 것이다. 단백질 구조연구는 구조의 유사성 검색을 이용하여 단백질 기능을 예측한다. 그러나 전체적인 구조가 유사한 단백질이라도 다른 기능을 수행 할 수 있고 구조가 다른 단백질이라도 핵심 구조가 유사하다면 유사한 기능을 수행할 수 있다.[1] 이는 단백질의 기능이 특정 구조적으로 잘 보존된 모티프에 따라 결정되기 때문이다. 진화적 관점에서 모티프는 기능을 유지하는 핵심적인 부분이 잘 보존되었다는 점에서 활성 사이트와 관련이 있다. 즉 활성 사이트가 위치해 있는 부분이 모티프부분으로 볼 수 있다. 효소의 단백질 표면에서 활성 사이트에 결합하는 이온, 작은 분자, 거대분자 등의 리간드(ligand)와의 상호작용에 의해 단백질이 기능을 수행한다.

이 논문은 활성 사이트 표면을 비교하기 위해 단백질의 3차원 공간정보를 matrix로 표현 할 수 있는 가장 작은 평면도형인 삼각형을 이용하여 상세한 형태비교를 제공한다. 단백질의 전체구조가 아닌 표면비교를 위해 단백질 표면형태에 영향을 주는 side chain의 정보를 고려한다. 먼저 Protein Data Bank를 통해 Cα와 side chain의 3차 좌표정보를 이용해 세 개의 Cα와 각 Cα에 대응하는 side chain으로 두 개의 다른 삼각형을 만든다. 이렇게 만들어진 두 개의 삼각형을 하나의 특정하위구조 set으로 정한다. 이 하위구조set을 이용하여 모든 경우의 수로 나올 수 있는 distance를 계산하여 하나의 distance matrix를 구축한다. 각 단백질의 distance matrix를 분석하여 활성 사이트에 대해 분류 및 기능을 예측한다. 제시한 기법은 단백질의 전체적인 서열과 구조 정보를 이용하지 않고서, 단백질 기능을 결정하는 활성 사이트를 효과적으로 비교함으로써 더욱 빠른 시간 내에 상세한 분석을 수행할 수 있다.

"이 논문은 2008년도 정부(교육과학기술부)의 재원으로 한국과학재단의 지원을 받아 수행된 연구(No. R11-2008-014-02002-0)이며 2008년 정부(교육과학기술부)의 지원을 받아 수행된 연구임" (지역거점연구단육성사업/충북BIT연구중심대학육성사업단)

2. 관련연구

단백질 구조분석은 구조의 유사성 검색을 이용하여 단백질 기능을 예측한다. 그러나 전체적인 구조가 유사한 단백질이라도 특정 구조가 다르면 다른 기능을 수행 할 수 있고 구조가 다른 단백질이라도 특정구조가 유사하다면 유사한 기능을 수행할 수 있다. 따라서 전체적인 구조를 비교하지 않고 기능을 결정하는 활성 사이트의 분석을 통하여 단백질 기능을 예측할 필요가 있다.[2] 따라서 이 절에서는 단백질의 하위구조를 비교 분석하는 기법들에 대해 소개한다.

Local feature frequency profile 방법을 제시한 In-Geol cho [3]는 구조적으로 유사성을 빠르게 접근하기 위해 단백질의 모든 Cα의 좌표들의 diatance를 구성하는 matrix를 구축한다. diatance matrix에서는 많은 하위구조특징이 추출되는데 medoid analysis을 통해 submatrix의 k-cluster 사용하여 대표 패턴 matrix를 생성한다. 미지의 단백질에 대해 구조적으로 유사한 matrix를 빠른 시간 내로 계산이 가능하다.

Craih T.Porter [4]는 PDB구조 데이터를 이용하여 활성사이트의 잔기를 식별하고 annotation하는 Catalytic site atlas 데이터베이스를 소개한다. CSA 데이터베이스는 Swiss-prot과 EC number의 식별자로도 검색이 가능하기 때문에 효소 촉매 반응의 타입에 따라 단백질을 분류하는 EC Number를 우리 논문에서 실험평가로 적용할 수 있다.

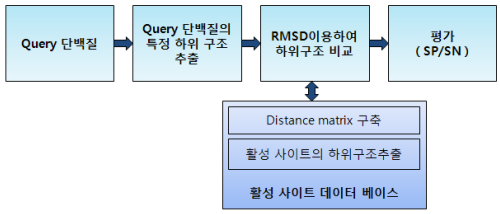
Fabrizio Ferre [5,6]는 바인딩 영역을 구성하는 아미노산 잔기를 벡터로 표현하고 벡터를 비교하는 알고리즘을 제안하였다. 벡터는 아미노산 잔기의 Cα원자 좌표와 side chain 중심점의 좌표를 이용하여 생성된다. 두 바인딩 영역을 mapping 시킨 후, 치환 메트릭스와 RMSD를 계산하여 가장 유사한 벡터를 seed벡터라고 정의한다. seed 벡터의 주변에 있는 벡터를 선택해 나가면서 바인딩 영역 아미노산 잔기의 벡터와 비교하고 치환 메트릭스의 상동성과 RMSD의 cut-off를 설정하여 유사한 벡터를 검색한다.

T. Andrew Binkowski [7]는 CASTp(Computed Atlas of Surface Topography of Protein) 데이터베이스에서 바인딩 영역의 정보를 추출하여 분석한다. 논문에서 실험한 데이터의 전체 서열을 비교하였을 때 16%의 유사성을 나타냈지만 바인딩영역의 서열은 51%정도의 유사함을 보였다. 또한 바인딩 영역의 형태를 비교하기 위해 cRMSD와 oRMSD를 계산한다. cRMSD는 단백질을 mapping시킨 후 아미노산 잔기의 좌표를 이용하여 아미노산 잔기들 간의 거리를 분석하고 oRMSD는 바인딩 영역의 아미노산 잔기의 중심점을 향한 벡터를 각각 생성하고 이 벡터들의 거리를 분석한다.

우리 논문에서는 In-Geol choi가 제시한 모든 Cα의 좌표들의 diatance를 구성하여 matrix를 구축하는 방법에 Cα 좌표로 삼각형 하나를 만들고 side chain좌표로도 표현될 수 있는 삼각형 하나를 만들어 두 개의 삼각형을 하나의 하위구조 set으로 정의하여 diatance matrix를 구축한다. 그리고 RMSD를 이용하여 바인딩 영역의 형태를 비교한 T. Andrew Binkowski처럼 구축된 distance matrix를 비교하는데 RMSD를 이용한다.

3. 활성싸이트 영역 비교설계

그림 1은 활성 사이트의 표면비교를 통해 단백질을 분류하고 기능을 예측하는 프레임워크를 나타낸다.

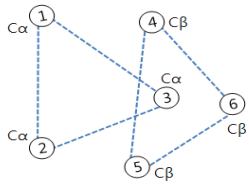


(그림 1) Framework

3.1 하위구조추출

우리 연구는 활성 사이트 표면을 비교하기 위해 단백질의 3차원 공간정보를 matrix로 표현 할 수 있는 가장 작은 평면도형인 삼각형을 이용하여 상세한 형태를 비교한다. 단백질의 기본구조인 아미노산이 한 개의 탄소원자에 한 개의 수소원자, 한 개의 아미노기, 한 개의 카르복실기, 한 개의 side chain(R group)이 결합하고 있다는 것을 고려한다.

우리 논문에서는 각 아미노산 잔기의 Cα 원자 좌표를 이용하여 한 단백질에서 표현할 수 있는 모든 삼각형을 만든다. 그리고 단백질 표면의 활성 사이트를 비교하기 위해서는 일반적으로 바깥을 향하여 표면의 형태를 결정짓는 아미노산 잔기의 side chain의 정보가 필수적이다. 따라서 우리는 아미노산 잔기의 Cα원자에 side chain을 포함하여 Cα삼각형과 side chain 삼각형 2개를 하나의 특정하위구조 set으로 정의한다.(그림 2)



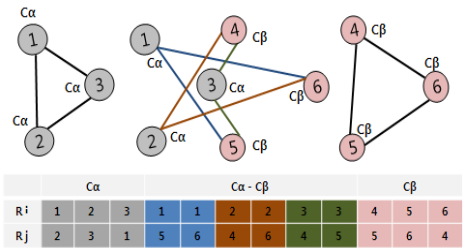
(그림 2) Cα와 side chain의 하위구조

먼저 활성 사이트 정보는 4절에 설명될 Catalytic Site Atlas 데이터베이스의 최신 버전을 이용한다. CSA에서 얻은 활성 사이트정보를 통해 PDB에서 Cα와side chain의 위치정보를 추출한다. 각 단백질의 활성 사이트의 위치정보에서 세 개의 Cα 원자 각 x,y,z좌표를 하나의 꼭지점으로 보고 삼각형으로 표현할 수 있는 모든 경우의 수를 고

려하되 삼각형 각 변의 길이에 threshold값을 주어 특정하위구조 즉, 표현할 수 있는 삼각형 set의 수에 제한을 준다. Cα 원자로 만들어진 삼각형에 대응하는 side chain도 같은 방법으로 삼각형으로 표현한다.

3.2 distance matrix 생성

활성 사이트의 특정하위구조set에 있는 삼각형 두 개의 꼭지점 6개를 이용해 모든 경우의 수를 고려하여 distance를 구한다. (그림 3) 이때 구조의 기여도에 따라 Cα끼리의 거리, Cα와 Cα와 부합하지 않는 side chain의 거리, side chain끼리의 거리 순으로 거리를 구한다. Cα와 부합하는 side chain들의 거리는 거의 유사하기 때문에 고려하지 않는다.



(그림 3) 각 Cα와 side의 distance

두 개의 삼각형 점들의 거리(Euclidean Distance)는 단백질에서 뽑을 수 있는 residue 3개의 Cα원자 좌표x,y,z와 그에 대응하는 3개의 side chain의 원자 좌표x,y,z를 이용하여 식1를 통해 계산한다.

$$\text{Distance}=\sqrt{\sum_{i,j=1}^{n}(Rix-Rjx)^2+(Riy-Rjy)^2+(Riz-Rjz)^2}$$

<식 1>

식 1에서 i,j는 Cα와 side chain의 원자좌표로 만들어진 삼각형 꼭지점 6개를 말한다. 하나의 특정하위구조set에서 계산된 각 아미노산 잔기의 Cα와 side chain의 거리 값을 matrix에 표현하고 이것을 distance matrix라고 정의한다.

	1Cα_a	2Cα_b	3Cα_c	4Cβ_a	5Cβ_b	6Cβ_c
1Cα_a	0	d21	d31	d41	d51	d61
2Cα_b	d21	0	d32	d42	d52	d62
3Cα_c	d31	d32	0	d43	d53	d63
4Cβ_a	d41	d42	d43	0	d54	d64
5Cβ_b	d51	d52	d53	d54	0	d65
6Cβ_c	d61	d62	d63	d64	d65	0

(그림 4) distance matrix

그림 4처럼 distance matrix는 6x6 matrix로 구성되고 하나의 matrix에는 Cα와 side chain으로 구성된 두 개의 삼각형 set인 하위구조의 distance로 되어있다.

3.3 비교기법

단백질의 활성 사이트 비교는 두 단백질 사이의 구조 차이를 표현해 주는 환경변수인 RMSD(root mean square deviation)를 이용한다. RMSD는 두 단백질 구조에 있어서 원자의 위치 사이의 제곱근 평균 제곱 편차 값으로 한 단백질의 모든 원자의 위치를 함수 값으로 하여 계산을 해둔 것이다. 즉, 단백질의 백본이나 알파탄소의 위치등 원자의 일부 그룹만의 위치를 함수 값으로 하여 계산해둔 값이다. 이것은 두 가지 구조를 가능한 겹쳐보이게 하는 superposition 알고리즘이다. 우리는 이 알고리즘을 이용해 distance matrix에 표현된 총 distance 12개를 i,j로 표현하고 q는 query 단백질의 하위구조 diatance matrix, db는 데이터베이스에 있는 하위구조 diatance matrix로 표현한다.

$$\text{RMSD}=\sqrt{\frac{1}{n}\sum_{i,j=1}^{n}(q_{di}-db_{dj})^2}$$

<식 2>

Query 단백질을 입력하면 단백질에서 표현될 수 있는 특정하위구조 set을 추출하여 그 중에 하나의 하위구조를 선택하여 distance matrix를 생성하고 데이터베이스에 있는 활성 사이트의 diatance matrix와 비교를 한다. 이때 RMSD를 이용하여 비교하는 두 matrix의 차이를 알아본다. 데이터베이스에서의 matrix와 적은 RMSD값을 갖는 하위구조를 query단백질에서 선택한 하위구조와 같은 단백질이라 분류한다.

4. 실험계획

4.1 data set

data set은 동물의 조직이나 침, 눈물 속에 들어 있는 항균성 효소이며 가수분해반응에 작용하는 hydrolase로 분류되는 lysozyme으로 구성한다. RCSB protein data bank에서 단백질의 PDB파일을 얻고 단백질의 표면에 위치하는 활성 사이트를분석하기 위해 단백질 3차원 구조 시뮬레이션 툴인 Rasmol (http:// www. umass. edu/microbio/rasmol/)을 이용하여 표면 아미노산 잔기를 추출한다. 그리고 활성사이트의 정보는 효소의 촉매기작과 관련된 아미노산 잔기의 정보를 제공하는 catalytic site atlas 데이터베이스(http://www. ebi.ac.uk/ thornton-srv/databases/CSA/)에서 추출한다. CSA파일은 데이터베이스에 있는 모든 단백질의 PDB ID와 활성사이트 아미노산 잔기의 이름,번호,체인등의 정보를 제공한다. PDB파일과 lysozyme 활성 사이트의 특정하위구조정보와 distance matrix정보는 데이터베이스를 구축하여 저장한다.

4.2 실험평가

제안한 방법에 의해 단백질의 하위구조가 정확하게 검색되었는지를 검증하기 위해 data set을 기능에 따라 클래스로 다시 분류한다. 효소 단백질은 효소 촉매 반응의 타입에 따라 단백질을 분류하는 EC Number (enzyme commission number) 가 주어진다. 따라서 단백질 기능을 기준으로 분류된 클래스는 EC Number를 참조하여 생성한다.

이 논문에서 제시한 기법은 구조를 기반으로 비교하여 기능을 예측하기 때문에 검색된 단백질이 동일한 클래스에 속하는지를 평가한다. 일반적으로 널리 사용되는 특이성(specificity: Sp)와 민감성(Sensitivity: Sn)을 평가 기준으로 사용된다.

<표 1> SP와 SN의 관계식

	Predictive Positive	Predictive Negative
Actual Positive	True Positive	False Negative
Actual Negative	False Positive	True Negative

TP, FP, FN은 표 1에서 보여주듯이 실제 클래스와 예측된 클래스를 비교하여 수치로 나타낸 것이다. TP는 A단백질을 정확하게 예측한 단백질의 수를 나타내고 ,FP는 B단백질을 A 단백질로 잘못 예측한 수이며 FN은 A단백질을 B 단백질로 잘못 예측한 수를 의미한다. 즉, TP는 True Positive, FN은 False Negative, FP는 False Positive를 의미한다.

$$Sn = \frac{TP}{TP+FN},\ Sp = \frac{TP}{TN+FP}$$

<식 3>

SN는 실제 A단백질 중 정확하게 A단백질로 예측된 결과로 실제 단백질의 제현률을 나타낸다. 이 식을 통해 B단백질로 잘못 예측된 A 단백질의 비율도 확인 할 수 있다. SP는 A단백질로 예측된 단백질 중 실제 A 단백질로 예측된 결과의 정확률을 나타낸다. 이 식을 통해 A 단백질로 잘못 예측된 B 단백질의 비율도 확인 할 수 있다.

5. 결론 및 향후계획

단백질 구조분석은 구조의 유사성 검색을 이용하여 단백질 기능을 예측한다. 그러나 전체적인 구조가 유사한 단백질이라도 다른 기능을 수행 할 수 있고 구조가 다른 단백질이라도 핵심 영역의 구조가 유사하다면 유사한 기능을 수행할 수 있다. 따라서 전체적인 구조를 비교하지 않고 기능을 결정하는 특정하위구조인 활성 사이트 분석을 통하여 단백질 기능을 예측할 필요가 있다.

이 논문에서 활성 사이트 표면을 비교하기 위해 단백질의 3차원 공간정보를 matix로 표현 할 수 있는 가장 작은 평면도형인 삼각형을 이용하여 활성 사이트의 표면에 대해 상세한 비교를 한다. 단백질 표면의 활성 사이트를 비교하기위해서는 일반적으로 바깥을 향하여 단백질 표면의 형태를 결정짓는 side chain 정보가 필수적이다. 따라서 우리는 아미노산 잔기의 Cα원자에 side chain을 포함하여 Cα삼각형과 side chain 삼각형 2개를 하나의 특정하위구조 set으로 정의한다. 이 하위구조set을 적용하여 distance matrix를 구축하여 RMSD를 이용하여 활성 사이트의 하위구조를 비교하고 구조정보를 체계적으로 분류한다.

이 논문에서 제시한 활성 사이트의 표면 비교 기법은 단백질의 아미노산 서열과 구조의 전체적인 정보를 이용하지 않고 활성 사이트의 정보만을 분석하고 비교한다. 이 연구를 통해 신약개발 분야에서는 질병의 치료약물의 표적인 효소의 3차원구조를 알게 되어 저해제를 보다 잘 디자인 할 수 있어 치료제 개발이 쉬워질 것이다. 향후과제로는 특정단백질의 3차 구조 모티프 정보들을 분석하여 특정구조가 가지는 규칙을 발견하고 데이터베이스화함으로써 특정 기능을 가지는 단백질에 대한 검색 및 특정 하위구조 검색에 효과적인 방법을 제공하고자 한다.

참고문헌

[1] P. C. Babbitt, "Definition of Enzyme Function for the Structural Genomics era," Curr. Opin. Chem. Biol., Vol. 7, pp.230-237, Apr. 2003.

[2] Kwang Su Jung, Ki Jin Yu, Keun Ho Ryu, Yong Je Chung, "Predicting Ligand Binding Site using Protein Surface Features," PACIFIC SYMPOSIUM ON BIOCOMPUTING, pp.72, Jan. 2007.

[3] In-Geol Choi "Local feature frequency profile : A method to measure structural similarity in proteins" PNAS, Vol.101, pp.3797-3802, 2004

[4] Craig T. Porter, Gail J. Bartlett, and Janet M. Thornton, "The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data," Nucl. Acids. Res., Vol. 32, pp.129-133, Jan. 2004.

[5] Fabrizio Ferre, Gabriele Ausiello, Andreas Zanzoni and Manuela Helmer-Citterich, "SURFACE : a Database of Protein Surface Regions for Functional Annotation," Nucleic Acids Research, Vol. 32, pp.240-244, Jan. 2004.

[6] Fabrizio Ferre, Gabriele Ausiello, Andreas Zanzoni and Manuela Helmer-Citterich "Functional annotation by identification of local surface similarities: a novel tool for structural genomics", BMC Bioinformatics, 2005

[7] T. Andrew Binkowski, Larisa Adamian and Jie Liang, "Inferring Functional Relationships of Proteins from Local Sequece and Spatial Surface Patterns," J.Mol.Biol., Vol. 332, pp.505-526, Sep. 2003.