제26회 한국정보처리학회 추계학술발표대회 논문집 제13권 제2호 (2006. 11)

잔기 위치 예측을 위한 단백질 기하학적 특징 추출 기법

유기진, 정광수, 류근호
충북대학교 데이터베이스/바이오인포메틱스 연구실
e-mail:{heyu4580,ksjung,khryu}@dblab.chungbuk.ac.kr

An Extraction Technique of Protein Geometric Features for Predicition of Residue Location

Ki Jin Yu, Kwang Su Jung, Keun Ho Ryu
Database/Bioinformatics Laboratory
Chungbuk National University

요 약

생명현상을 이해하기 위해서는 단백질의 기능 규명이 이루어져야한다. 단백질 기능 규명을 위한 서열 분석 방법은 서열 상동성이 현저히 낮은 경우 단백질 기능 예측이 불가능하고, 과거의 전체적인 단백질 구조 분석을 통한 기능 예측의 문제점이 보고되고 있다. 이 논문에서는 기능상 중요한 의미를 가지고 있는 단백질의 특정하위구조의 기하학적 특징을 추출하여 이 특징과 잔기의 위치와의 관계를 규명하였다. 또한 NaiveBayes, SVM, C4.5의 분류알고리즘을 이용하여 각 알고리즘별 분류성능을 평가하였다. 기능상 중요한 의미를 가지고 있는 특정하위구조를 비교함으로써 모르는 단백질의 기능을 예측할 수 있다.

1. 서론

생명체의 생물학적 기능과 과정을 이해하는데 있어서 생명체를 이루고 있는 단백질 기능의 규명이 중요하다. 알려지지 않은 단백질의 기능을 규명하는 방법으로 단백질 서열을 분석하는 방법과 단백질의 3차 구조 접근방법이 있다. 서열분석은 알려지지 않은 단백질 서열을 기존의 알려진 서열과 비교함으로써 서열의 기능과 진화정도를 알 수 있지만 서열 상동성이 현저히 낮은 경우는 예측이 불가능하다.

서열 상동성이 낮은 경우, 단백질 폴드 구조의 유사도를 통해 기능 유추가 가능하다[1-4]. 또한 단백질 폴드의 전체적인 구조가 다르더라도 단백질 표면의 리간드가 결합하는 영역의 구조가 유사하면 서로 유사한 기능을 수행한다[5]. 따라서 전체 서열과 복잡한 단백질 3차 구조 전체를 필요로 하지 않고 기능상 중요한 의미를 가지고 있는 단백질의 특정하위구조를 분석하여 단백질의 기능 예측이 가능하다.

단백질의 표면에서는 다른 물질과 결합하여 상호작용이 일어난다. 표면의 활성사이트는 특정 리간드와 결합하는 성질을 갖고, 그 결합에 의한 상호작용으로 단백질 기능이 수행하게 된다.

이 논문에서는 잔기의 기하학적 특성을 이용하여, 어떤 잔기가 활성사이트에 관여하는지를 알아보기 위해 단백질의 모든 잔기의 위치를 예측하는 기법을 제시하였다. 잔기의 3차 구조 좌표 값을 이용하여 각각의 잔기에서 다른 잔기가 이루는 거리를 측정하였다. 또한 거리를 이용하여 잔기간의 거리를 한눈에 알아볼 수 있는 ContactMap을 생성하고 ContactMap에 나타난 잔기간의 거리를 이용하여 새로운 기하학적인 특징을 추출하였다.

추출된 특징에 NaiveBayes, SVM, C4.5 분류기를 이용하여 잔기의 위치를 예측하고, 보다 정확한 잔기 위치 예측에 기여한 속성을 분석하였다. 활성사이트에 존재하는 잔기와 다른 영역에 존재하는 잔기의 위치 예측 결과를 통해 활성사이트의 정보를 추출해 낼 수 있고, 이를 단백질 기능 예측에 활용하여 또한 신약개발을 위한 시발점이 된다.

이 논문은 2006년도 교육인적자원부 지방연구중심대학 육성사업의 지원에 의하여 연구되었음

2. 관련연구

단백질 구조에서 유사한 폴드(fold)나 서열을 가지고 있다고 하더라도 완전히 서로 다른 기능을 가질 수 있음이 확인되었다[6]. 또한 단백질 상호간 분명한 기능적 관계를 가지고 있음에도 불구하고 구조 및 서열의 유사도와는 관련성이 없는 경우가 밝혀졌다[7]. 이런 문제로 인하여, 단백질의 상호작용이 직접적으로 일어나는 단백질 표면으로부터 기능관계를 탐지하려는 연구들이 최근에 활발히 진행되고 있다.

Schmitt등[8]은 단백질 서열이나 폴드 상동성과는 독립적으로 단백질 사이의 기능적 관계를 탐지하기 위한 새로운 방법을 개발하였다. 두 활성사이트(binding site)의 매칭정도를 측정하여 점수를 매기는 안을 고안하였다. 각각의 활성사이트에 있는 모든 원자를 고려하는 것이 아니고, 움푹한 구멍(cavity)을 정의하는 수도센터(pseudocenter)를 설정하여 단백질의 특정모양을 비교하였다. 활성사이트 안에 있는 하위그룹은 클릭 알고리즘(clique algorithm)에 의해 탐지되었다.

BinKowski[9]은 단백질표면의 활성사이트에 해당하는 아미노산서열과 공간적 패턴을 탐지하여 단백질 기능관계를 유추하는 새로운 접근방법을 설명하였다. 두 단백질에서 활성사이트를 비교하면 서열유사도(sequence identity)가 51%에 이르지만 단백질 전체서열 비교에서는 16%의 유사도를 갖는 것을 확인하였다. 또한 pvSOAR 프로그램을 사용하여 단백질표면의 유사도를 구조적으로 측정하였다.

3. 잔기 위치 예측을 위한 특징 추출

3.1 Residue Center 정의

단백질 전체 폴드를 비교하여 기능 유추하는 기법에서는 단백질의 백본을 이용한다. 단백질 백본은 각 잔기의 Cα의 좌표를 추출하여 계산한다. 그러나 단백질 표면의 경우, 구조의 정확한 형태를 결정하는 것은 Side Chain의 영향력이 크다. 따라서 우리는 PDB에서 제공하는 잔기의 원자 좌표 값 중 수소를 제외한 Side Chain을 이루는 원자를 추가하여 가상의 중심을 구하고, 표면의 기하학적 특징을 추출한다. 이렇게 함으로써 표면의 구조가 Cα에 치우치지 않고 실질적인 표면구조에 근접하게 된다. 이렇게 구한 가상의 중심점을 우리는 Residue Center라 정의한다.

식(1)은 Residue Center의 x좌표를 계산하는 공식이다. 잔기의 Cα와 수소를 제외한 Side Chain의 n개 원자의 각 x좌표의 값을 더하여 Cα와 원자의 수로 나눈다. 같은 방법으로 y, z 좌표 값을 구한다.

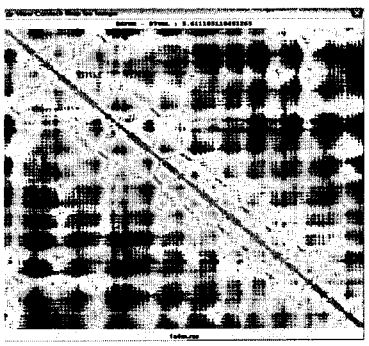
$$Rx = \frac{\sum_{i=1}^{n} x_n + x_{c\alpha}}{n+1} \quad (1)$$

3.2 ContactMap 생성

(그림 1)은 단백질의 잔기간의 거리를 x, y 좌표 평면에 나타낸 ContactMap이다. 식(1)을 이용하여 Residue Center를 구하고, 식(2)를 이용하여 잔기간의 거리(Euclidean Distance)를 구한다. 식(2)에서 단백질의 한 잔기 R^i의 Residue Center를 (R^ix, R^iy, R^iz)로 표현하고 같은 방법으로 잔기 R^j의 Residue Center를 (R^jx, R^jy, R^jz)로 표현한다.

$$Distance(R^i, R^j) = \sqrt{(R^jx - R^ix)^2 + (R^jy - R^iy)^2 + (R^jz - R^iz)^2} \quad (2)$$

ContactMap은 잔기간 거리의 범위를 설정하여 색깔별로 표현한 것이다. 거리 0-10Å를 가지는 잔기를 녹색으로 표현하고, 10-20Å는 노란색, 20-30Å는 분홍색, 30-40Å는 빨간색, 40Å이상의 잔기는 검정색으로 나타내어 각 잔기와 다른 잔기가 이루는 거리를 관찰하여 전체적으로 어떤 잔기가 근접하게 위치하는지 비교할 수 있다. 가로축과 세로축의 값이 같기 때문에 대각선을 중심으로 대칭을 이룬다.



(그림 1) ContactMap

3.3 ContactMap으로부터 기하학적 특징 추출

특정 잔기를 기준으로 범위를 설정하여 그 범위내에 위치하는 잔기의 수와 그 잔기들의 거리의 합을 추출한다. 우리는 이렇게 측정된 기하학적 특징을 이용하여 잔기의 구조적인 위치(표면, 내부, 패치)의 연관성을 밝히고자 한다. <표 2>에서처럼 다

양한 범위(Å)를 설정하고 그 범위 내에 위치하는 잔기의 수를 C, 범위 내 모든 잔기의 거리 합을 S로 설정한다. 이렇게 총 24개의 기하학적인 특징을 추출하였고, 실험을 통하여 이 특징들이 잔기의 위치 예측에 얼마나 기여하는지를 알아본다.

<표 2> 추출된 기하학적 특징

범위 (Å)	속성		
0-5, 5-10, 10-15, 15-20, 20-25, 25-30, 30-	잔기 수	C_{tab}	7개
	잔기 거리	S_{tab}	7개
0-10, 0-15, 0-20, 0-25, 0-30	잔기 수	C_{accum}	5개
	잔기 거리	S_{accum}	5개

4. 추출된 특징에 대한 성능평가

4.1 Data Set 생성

추출된 속성이 잔기 위치 예측에 얼마나 기여하는지를 실험하기 위해 우리는 Enzyme-Ligand Data Set[10]을 사용하였고, 이 Data Set은 Enzyme Classification(E.C.) number를 기준으로 Hydrorases, Isomerases, Ligases, Lyases, Oxidoreductases, Transferases로 여섯 개의 카테고리로 이루어져 있다. Data Set에서 명시한 단백질의 PDB CODE를 이용하여 잔기의 좌표 값을 추출하고, SURFNET[11]을 이용하여 단백질의 PATCH 정보를 추출하였다. 이렇게 만들어진 Data Set에는 활성사이트와 관련 있는 영역의 잔기(PATCH class)와, 활성사이트와 관련이 없는 잔기(NOT_PATCH class) 정보를 포함하고 있다.

<표 2>에서 언급한 속성(C_{tab}, S_{tab}, C_{accum}, S_{accum}) 중 분류에 중요한 영향을 미치는 속성을 알아보기 위해 네 가지 속성을 각각 조합하여 다시 <표 3>의 Data Set을 구성하였다. 다음절에서는 만들어진 14가지 속성 SET을 이용하여 성능 평가를 한다.

<표 3> 조합 속성 SET

① S_{tab}	② S_{accum}	③ C_{tab}	④ C_{accum}
⑤ S_{tab},S_{accum}	⑥ S_{tab},C_{tab}	⑦ S_{tab},C_{accum}	⑧ S_{accum}, C_{tab}
⑨S_{accum},C_{accum}	⑩ C_{tab},C_{accum}	⑪ S_{tab}, S_{accum},C_{tab}	⑫ S_{tab}, S_{accum},C_{accum}
⑬ S_{accum} C_{tab} C_{accum}		⑭ S_{tab},S_{accum},C_{tab},C_{accum}	

4.2 성능평가

우리가 제시한 분류 방법을 검증하기 위해 WEKA version 3.4.8을 사용하였고 NaiveBayes, Support Vector Machine, 의사결정트리 C4.5 알고리즘에 적용하였다. 평가를 위해 식(3)의 정확률, 식(4)의 제곱근평균제곱오차(RMSE), 식(5)의 TP(True Positive) Rate를 사용하였다. <표 4>를 보면, 전체 정확률은 C4.5와 SVM이 비슷하게 좋은 성능을 보여주고, PATCH의 TP Rate는 NaiveBayes가 가장 높은 점수를 보여준다.

<표 4> 알고리즘 평가

알고리즘	정확률 (%)	제곱근평균제곱오차	TP Rate	
			NOT_PATCH	PATCH
Naive Bayes	64.59	0.53	0.72	0.49
SVM	71.06	0.53	0.96	0.14
C4.5	71.17	0.45	0.88	0.32

$$\text{정확률} = \frac{TP+TN}{TP+FP+FN+TN} \quad (3)$$

$$\text{제곱근평균제곱오차(RMSE)} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(P_{(ij)} - T_j)^2}$$

$P_{(ij)}$: 결과값, T_j : 예측값 (4)

$$TP\,Rate = \frac{TP}{TP+FP} \quad (5)$$

알고리즘에 따라 가장 높은 정확률을 가지는 속성은 다르게 측정되었다. <표 5>에서, NaiveBayes 알고리즘은 ③번 속성 SET이 가장 우수하였고, C4.5는 ⑫번 속성 SET이 대체적으로 높은 정확률을 산출하였다.

<표 5> 정확률이 가장 높은 속성SET과 정확률(%)

알고리즘	Hydrorases	Isomerases	Ligases	Lyases	Oxidoreductases	Transferases	ALL
Naive Bayes	③ 71.39	① 68.00	④ 67.80	③ 64.05	③ 67.99	③ 66.77	③ 67.37
SVM	74.59	⑬ 72.52	⑭ 78.08	⑬ 68.28	⑥ 72.89	⑭ 69.18	68.25
C4.5	⑫ 75.99	⑫ 70.68	⑨ 83.21	① 67.43	⑩ 71.80	⑫ 69.80	⑤ 70.66

<표 6>은 Enzyme의 6개 그룹과 모든 Enzyme Data Set을 분류기에 적용하였을 때, 최대 TP Rate를 산출하는 속성 SET과 TP Rate이다. PATCH의 TP Rate를 기준으로 NaiveBayes 알고리즘은 전체 속성을 조합한 속성 SET ⑭번, C4.5 알고리즘은 ⑥번이 가장 좋은 결과를 산출하였고 평균적으로 NaiveBayes의 TP Rate가 더 높다. NOT_PATCH의 분류에서는 C4.5에 의한 TP Rate 결과가 다른 알고

리즘에 비해 높은 결과를 얻었다.

<표 6> TP Rate가 가장 높은 속성SET과 TP Rate

	알고리즘	Hydrorases	Isomerases	Ligases	Lyases	Oxidoreductases	Transferases	ALL
PATCH	Naive Bayes	⑭ 0.34	⑬ 0.58	⑥⑨ 0.73	⑭ 0.53	⑭ 0.67	⑬⑭ 0.50	⑭ 0.47
	SVM	0	⑥ 0.35	⑭ 0.03	⑬ 0.31	⑥ 0.37	⑭ 0.21	0
	C4.5	⑭ 0.31	② 0.47	⑨ 0.33	⑥ 0.42	⑩ 0.47	⑫ 0.43	⑥ 0.31
NOT_PATCH	Naive Bayes	③ 0.88	② 0.78	④ 0.71	③ 0.74	① 0.70	③ 0.8	③ 0.81
	SVM	1	② 0.99	1	1	⑨ 0.94	② 0.96	1
	C4.5	② 0.98	① 0.87	②⑨ 0.97	④ 0.88	⑨ 0.89	⑭ 0.95	②④ 0.94

분류기에 적용한 결과 정확률 68.94%, 제곱근평균제곱오차 0.5077, NOT_PATCH의 TP rate 0.85, PATCH의 TP rate 0.32의 평균을 얻었다. 그러나 정확률은 높지만 Data Set의 15,152개 잔기 중 NOT_PATCH의 수는 10,342개로 68%를 차지하고 NOT_PATCH의 TP Rate가 높은 점수의 0.85이기 때문에, 4,810개로 Data Set의 32%를 차지하는 PATCH 분류의 정확률이라고 할 수 없다. 따라서 PATCH의 TP Rate와 그 외의 성능 평가를 통해 NaiveBayes을 이용한 분류가 더 정확함을 알 수 있고 모든 속성을 조합한 속성 SET ⑭번을 training set으로 설정할 경우 좀 더 정확한 분류 결과를 산출하였다.

5. 결론

단백질의 기능을 예측하기 위한 방법으로 단백질의 서열과 3차 구조의 분석이 많이 실행되어 왔다. 그러나 단백질 서열의 상동성이 낮은 경우는 단백질의 기능을 정확히 예측할 수 없다. 그리고 단백질의 전체적인 구조가 다르더라도 활성사이트의 구조가 같다면 같은 기능을 수행한다.

이러한 문제점을 보완하기 위해서 우리는 모든 단백질 잔기의 위치를 파악하여 위치에 따른 잔기의 특성을 밝히고 이에 따른 단백질의 영역별 특성을 이용하여 단백질 기능을 예측하기 위한 방법의 첫 단계를 제시하였다. C4.5 알고리즘은 가장 높은 정확률을 보여주지만, 전체 잔기의 68%를 차지하는 NOT_PATCH가 정확하게 분류되는 가능성이 크다.

단백질 표면의 중요한 정보를 제공하는 PATCH는 전체 잔기의 32%에 불과하기 때문에 이를 기준으로 하였을 때 NaiveBayes 알고리즘이 PATCH를 분류하는데 SVM이나 C4.5 보다 더 정확한 결과를 산출하였다.

활성사이트의 위치예측에서 PATCH의 TP Rate를 높이기 위해 친수성, 소수성, 전하 등의 물리화학적 특징을 이용하여 그에 상응하는 리간드의 구조적 특징과 물리화학적 성질을 좀 더 추가적으로 보완할 계획이다. 또한 이런 특징에 잔기의 보존정보를 추가하여 좀 더 정확한 단백질 기능을 예측하고, 이는 신약개발을 위한 기반연구로 활용될 수 있다.

참고문헌

[1] L. Holm, and C. Sander, "Protein structure comparison by alignment of distance matrices," J. Mol. Biol., Vol. 233, pp.123-138, 1993.

[2] J. F. Gibrat, T. Madej, and S. H. Bryant, "Surprising similarities in structure comparison," Curr. Opin. Struct. Biol., Vol. 6, pp.377-385, 1996.

[3] C. A. Orengo, and W. R. Taylor, "SSAP: sequential structure alignment program for protein structure comparison," Methods Enzy-mol., Vol. 266, pp.617-635, 1996.

[4] R. B. Russell, and G. J .Barton, "Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels," Proteins, Vol. 14, pp.309-323, 1992.

[5] P. C. Babbitt, "Definition of enzyme function for the structural genomics era," Curr. Opin, Chem. Biol., Vol. 7, pp.230-237, 2003.

[6] L. M. Kauvar and H. O. Villar, "Deciphering cryptic similarities in protein binding sites," Curr. Opin. Biotechanol., Vol. 9, pp.390-394, 1998.

[7] A. Via, F. Ferre, B. Brannetti, A. Valencia, and M. Helmer-Citterich, "Three-dimensional view of the surface motif associated with the P-loop structure: *cis* and *trans* cases of convergent evolution," J. Mol. Biol., Vol. 303, pp.455-465, 2000.

[8] S. Schmitt, D. Kuhn, and G. Klebe, "A new method to detect related function among proteins independent of sequence and fold homology," J. Mol. Biol., Vol. 323, pp.387-406, 2002.

[9] T. A. Binkowski, L. Adamian, and J. Liang, "Inferring functional relationships of proteins from local sequence and spatial surface patterns," J.Mol.Biol., Vol. 332, pp.505-526, 2003.

[10] R. A. Laskowski, N. M. Luscombe, M. B. Swindells and J. M. Thornton, "Protein clefts in molecular recognition and function," Protein Sci., Vol. 5, pp.2438-2452, 1996.

[11] R. A. Laskowski, "SURFNET: A program for visualizing molecular surfaces, cavities, and intermolecular interactions," J. Mol. Graph, Vol. 13, pp.307-308, 323-330, 1995.