서열 유사도와 특징 기반 분류를 융합시킨 단백질 기능 예측 시스템

문지환*, 김유성
인하대학교 정보통신공학과
e-mail : pianoboom83@hanmail.net

A Hybrid Protein Function Prediction System Using Sequence Similarity and Feature-based Classification

Ji Hwan Moon*, Yoo-Sung Kim
Dept. of Information & Communication Engineering, Inha University

요 약

단백질의 서열 정보와 기능 정보의 양이 증가함에 따라 컴퓨터 실험을 통한 단백질의 기능 예측이 가능해졌으며 정확성이 높은 예측 시스템을 개발하려는 여러 연구가 시도되고 있다. 대표적인 방법으로 서열 유사도를 기반으로 기능 예측을 하는 시스템이 제안되었으나 단백질 중에는 서열이 유사하지만 기능이 다르거나 또는 서열은 다름에도 불구하고 기능이 같은 단백질이 존재하기 때문에 서열의 유사도 만을 이용해서는 단백질의 기능 예측을 어렵다. 이러한 유사도 방법의 단점을 극복하기 위해 단백질 서열로부터 추출한 특징을 기반으로 분류하는 방법도 제안되었다. 본 논문에서는 이러한 기존 방법들의 장점을 얻기 위하여 서열 유사도 방법과 특징 기반 방법을 융합한 단백질 기능 예측 시스템을 제안하고 예측 정확성 분석을 위한 실험을 실시하였다. 실험의 결과에 따르면 제안된 융합시스템이 서열 유사도만을 이용한 방법과 특징 기반 방법보다 좋은 예측 정확률을 갖는 것으로 분석되었다.

1. 서론

단백질은 우리 몸을 구성하는 필수 요소 중 하나이다. 이러한 단백질은 20 가지 아미노산의 서열로 이루어져 있으며 각 아미노산은 DNA 염기가 3 개씩 모여서 이루어지는 코돈에 의해 결정된다. 효소는 단백질의 일종으로 우리 몸에서 일어나는 각 화학 반응의 촉매 역할을 담당하며 이를 효소의 기능이라 한다. 과학의 발달과 함께 해마다 쏟아져 나오는 생물학 정보의 양이 기하급수적으로 증가함에 따라 단백질의 서열 정보의 양도 증가하고 있으나 각 단백질의 기능을 밝히는 작업은 그 증가속도를 따라가지 못하고 있는 것이 현실이다. 생명정보학은 이러한 방대한 양의 생물학 정보를 전산학의 힘을 빌려 빠르고 정확하게 처리하는 것을 목적으로 하는 학문으로, 단백질 기능 예측도 많은 연구가 시도되고 있는 중요한 분야이다.

단백질 기능 예측에 가장 기본이 되는 방법은 서열 유사도를 이용한 방법이다[1]. 서열 유사도를 이용한 방법은 아미노산의 서열이 서로 유사한 단백질은 그 기능이 같다는 가설을 기반으로 한다. 그러나 서열 유사도를 이용한 방법에는 다음과 같은 문제가 존재한다. 단백질 중에는 서열이 서로 유사하지만 기능이 다르거나, 서열은 유사하지 않지만 기능이 같은 단백질이 존재한다는 것이다[2].

이러한 서열 유사도 방법의 한계를 극복하기 위한 대안으로 여러 연구들이 진행되고 있는데, 그 중 하나가 특징 기반 기능 예측 방법이다[3,4,5]. 특징 기반 예측 방법은 단백질의 아미노산 서열에서 아미노산 잔기의 총 수, 각 아미노산 잔기의 수, 양전하 및 음전하를 띈 잔기의 수, 총 원자량, 분자 무게 등의 특징 외에 여러 물리화학적 특징을 추출하여 단백질의 기능을 예측하는 방법으로 단백질의 서열 유사도나 구조의 유사도에 상관 없이 기능을 예측하는 것이다[5].

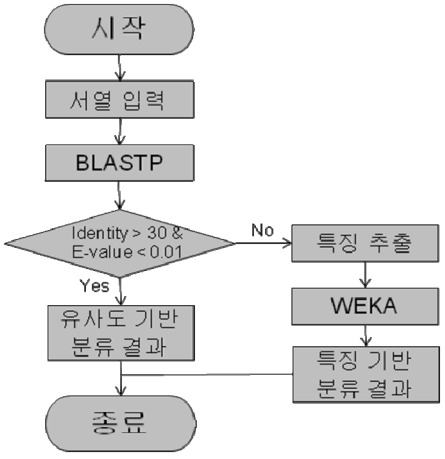
본 논문에서는 서열 유사도 방법과 특징 기반 방법을 직렬로 융합한 단백질 기능 예측 시스템을 제안한다. 서열 유사도 방법은 말 그대로 유사도를 이용하여 단백질의 서열을 예측하기 때문에 유사도가 낮은 단백질이 우연히 정확하게 예측되었다 하더라도 그 결과에 대한 신뢰도는 낮을 가능성이 있다. 따라서 제안하는 시스템은 서열 유사도 방법을 먼저 실행하여 1 차적인 예측을 하고 이 때, 임계값 이하의 유사도를 가진 단백질들은 특징 기반 방법을 적용하여 기능을 예측한다. 즉, 유사도가 높은 서열은 서열 유사도 방법을, 유사도가 낮은 서열은 특징 기반 분류 방법을 적용함으로써 서열 유사도 방법의 단점을 보완하고자 하는 것이다.

본 논문은 다음과 같이 구성되었다. 1 장은 서론이고 2 장에서는 제안하는 시스템의 구성을 설명한다. 3 장은 특징 기반 방법에서 사용하는 특징 집합의 구성이고 4 장은 실험에 사용된 데이터, 5 장은 실험 방법이다. 6 장에서는 실험 결과 및 분석을 다루며 마지막으로 7 장에서는 결론 및 향후 연구에 대해서 논한다.

2. 하이브리드 단백질 기능 예측 시스템의 구성

본 논문에서 제안한 하이브리드 단백질 기능 예측 시스템은 크게 서열 유사도 모듈과 특징 기반 분류 모듈로 구성되며 단백질 기능 예측 절차는 (그림 1)과 같다.

먼저 단백질의 서열을 입력 받으면 서열 유사도 모듈이 먼저 동작하여 1 차적 기능 예측을 한다. 서열 유사도 모듈에서 서열의 유사도를 검색하는 도구로 미국 NCBI(National Center for Biotechnology Information)의 BLAST(Basic Local Alignment Search Tool)중 단백질 서열 유사도를 찾아주는 BLASTP 를 사용하였다. 1 차적 기능 예측 후, 결과 파일을 탐색하여 입력된 단백질 서열이 데이터베이스 내의 서열과 임계값 이상의 유사도를 갖고 결과의 E-value 가 낮을 때, 그 결과를 실제로 예측된 결과로 인정하여 기능 예측을 종료한다. E-value 는 통계적인 값으로 높으면 높을수록 두 서열이 우연히 일치한다는 것을 나타낸다. 따라서 E-value 가 낮을수록 유사도의 신뢰도가 높다는 것을 의미하기 때문에 E-value 도 임계값의 하나로 적용하였다. 임계값 설정에 대한 내용은 5 장의 실험 방법에서 다룬다. 만약 유사도가 낮거나 E-value 가 높을 경우 특징 기반 분류 모듈을 이용하여 추가 기능 예측을 실시한다. 특징 기반 모듈에서는 먼저 입력된 서열로부터 특징을 추출하여 특징 집합을 만들어 WEKA[6]에서 사용 가능한 arff 파일로 저장한다. 그 다음, WEKA 를 이용하여 미리 구축된 모델을 통해 기능을 예측한다.



(그림 1) 단백질 기능 예측 절차

3. 특징 집합의 구성

단백질 기능 예측에 이용되는 특징 집합은 총 44 개의 특징으로 구성된다. <표 1>은 각 특징의 종류와 간단한 설명을 나타낸다.

<표 1> 특징의 종류와 설명

번호	특징	설명
1~20	아미노산 구성	전체 아미노산 개수에 대한 각 아미노산의 백분율
21~22	전하 아미노산	음전하와 양전하를 띈 아미노산의 개수
23~27	원자의 개수	탄소, 수소, 질소, 산소, 황
28	Aliphatic Index	지방족 곁사슬에 의한 상대적 크기
29	Extinction Coefficient	특정 파장에서의 빛 흡수도
30	GRAVY	상대적 소수성 수치의 평균
31	Instability Index	시험관에서의 안정성
32	Theoretical pI	이론적 등전점
33	분자 무게	단백질의 분자 무게
34	Polar	물리 화학적 특성
35	Aliphatic	
36	Aromatic	
37	Small	
38	Tiny	
39	Bulky	
40	Hydrophobic	
41	Acidic	
42	PPR	양전하를 띈 잔기 다음에 양전하를 띈 잔기가 오는 비율
43	NNR	음전하를 띈 잔기 다음에 음전하를 띈 잔기가 오는 비율
44	PNPR	양전하 다음에 음전하 또는 음전하 다음에 양전하를 띈 잔기가 오는 비율

특징 집합에서 사용된 특징들은 단백질 기능 예측에 관련된 기존 연구들([7-12])에서 사용된 특징들 중 2 개 이상의 관련 연구에서 사용된 특징들 중에서 선택하였다. 각 특징을 살펴보면, 1 번부터 20 번까지는 전체 아미노산의 개수에 대한 각 아미노산의 백분율을 나타낸다. 21 번과 22 번은 각각 음전하를 띈 아스파르트산과 글루탐산의 합과 양전하를 띈 아르기닌과 라이신의 합이다. 23 번부터 27 번은 탄소(C), 수소(H), 질소(N), 산소(O), 황(S) 원자의 개수이다. 28 번은 Aliphatic Index[7]로써 지방족 곁사슬에 의한 단백질의 상대적인 크기를 나타내고 29 번은 Extinction Coefficient[8]로 특정 파장에서의 빛 흡수도를 나타낸다. 본 시스템에서는 280nm 파장에서의 흡수도를 택하였다[9]. 30 번은 GRAVY[10]로 각 아미노산들의 상대적 소수성 수치의 합을 총 아미노산의 개수로 나눈 값이다. 31 번의 Instability Index[11]는 시험관에서 단백질이 얼마나 안정한가를 예측하는 수치이다. 32 번의 Theoretical pI[12]는 단백질의 구조를 배제하고 아미노산의 구성만으로 등전점을 구한 값이다. 33 번은 단백질의 분자 무게이다. 34 번부터 41 번은 물리화학적 특징으로 각 특징마다 해당하는 아미노산의 합으로 나타낸다. 42 번은 양전하를 띈 아미노산 다음에 또 양전하를 띈 아미노산이 오는

비율을, 43 번은 음전하를 띈 아미노산 다음에 또 음전하를 띈 아미노산이 오는 비율을, 마지막으로 44 번은 양전하를 띈 아미노산 다음에 음전하를 띈 아미노산 혹은 음전하를 띈 아미노산 다음에 양전하를 띈 아미노산이 오는 비율을 나타낸 특징이다.

4. 실험 데이터

효소는 단백질의 일종으로 제안하는 시스템에서 단백질의 기능을 예측한다는 것은 이 효소의 기능을 예측하는 것이다. 효소는 EC(Enzyme Commission) 번호 체계에 따라 각 단계별 혹은 그룹 별로 기능이 잘 분류되어 있기 때문에 기능 예측이 용의하다. 또한 효소의 기능을 예측해 봄으로써 인체 내 생명 반응의 가장 기본 단위에 대한 이해를 증진시키기 위해 효소의 기능을 예측해보고자 하였다. EC 번호는 4 단계의 숫자로 이루어져 있다. 각 숫자는 1.2.3.4 와 같이 '.' 로 구분하며 왼쪽에서 오른쪽으로 갈수록 더 하위 단계를 나타낸다. 제안하는 시스템은 질의와 데이터베이스 내 단백질 서열의 EC 번호를 비교하여 효소의 기능을 예측한다.

실험을 위해 사용한 데이터는 다음과 같다. 먼저 데이터베이스 구성을 위한 서열 정보는 KEGG(Kyoto Encyclopedia of Genes and Genomes) 43 버전을 사용하였으며 서열의 총 개수는 596,156 개이다. 학습 데이터 및 질의로 사용한 서열 정보는 KEGG 의 2010 년 9 월 7 일 버전이고 그 중, 하나의 EC 번호를 가지며 그 EC 번호가 후보군이 아닌 정보만 사용하였으며 서열의 총 개수는 39,937 개이다. 또한, 다른 연구와의 비교를 위해 다른 연구에서 사용하였던 상동성이 없는 35 개의 서열 정보[13]를 검증 데이터로 사용하였다.

5. 실험 방법

먼저 서열의 유사성 기준인 임계값을 정하기 위한 실험을 진행하였다. 유사도를 10%부터 90%까지 10%씩 증가시키면서 각 유사도 이하의 유사도를 갖는 서열을 분리시키고 각 유사도 이상의 유사도를 갖는 서열에 대한 서열 유사도 방법의 정확도를 측정하였다.

<표 2>는 단백질 분류 기준 중 유사도를 10%부터 50%까지의 변화시켰을 때의 정확도를 보여준다. 전체 서열에 대해서는 10%와 20%를 적용하였을 때가 가장 높지만, 상동성 없는 서열에 대해서는 30%일 때 가장 높은 정확도를 보였기 때문에 30%를 분류 기준으로 선택하였다.

E-value 의 임계 값은 다른 서열 유사도 방법에서 보편적으로 사용되는 값인 0.01 로 정하였다.

<표 2> 서열 정보에 대한 각 유사도별 정확도

	10%	20%	30%	40%	50%
전체 서열	95.93%	95.93%	95.22%	92.19%	87.45%
상동성 없음	60%	60%	62.86%	57.14%	48.57%

이렇게 정해진 유사도 임계 값 30%를 가지고 기존 방법들과의 예측 정확성 비교를 위해 다음과 같이 실험을 진행하였다.

먼저 BLAST 를 이용하여 전체 서열 정보에 대한 유사도 결과를 얻었다. 그 다음 E-value 가 0.01 이상, 유사도 30% 이하인 단백질을 분류하고 분류된 단백질을 모아서 특징을 추출하여 특징 집합을 만든 후, 그 집합으로 예측 모델을 구축하였다. 예측 모델을 구축하는 방법으로 데이터 마이닝의 Support Vector Machine 과 PART 를 이용하였다. 그 후, 검증을 위한 서열 정보인 상동성 없는 서열 정보를 마찬가지의 방법으로 분류하고 특징을 추출한 후 완성된 모델에 대해 검증을 하였다.

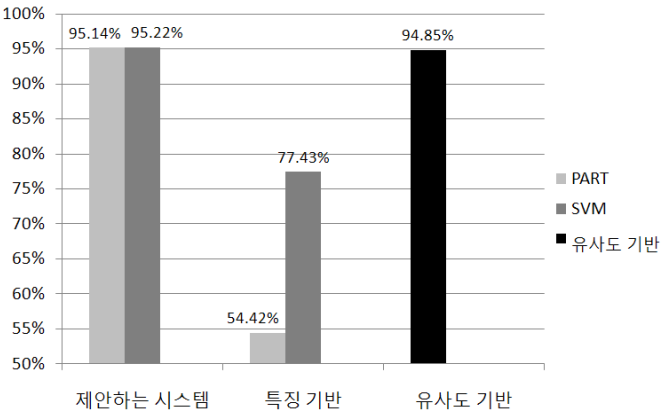
6. 실험 결과 및 분석

본 논문에서 제안하는 시스템을 이용한 실험 결과는 크게 둘로 나눌 수 있다. 하나는 실험에 사용된 전체 단백질 서열에 대한 예측 정확도이고, 다른 하나는 상동성이 없는 단백질 서열에 대한 예측 정확도이다.

<표 3>과 (그림 3)은 본 시스템과 서열 유사도 방법만을 이용한 경우, 그리고 특징 기반 방법만을 이용한 경우의 전체 서열 정보에 대한 예측 비교 결과를 보여준다.

<표 3> 각 기법 별 결과 비교표

	제안하는 시스템	특징 기반	유사도 기반
PART	95.14%	54.42%	94.85%
SVM	95.22%	77.43%	



(그림 3) 전체 서열에 대한 예측 결과 비교

특징 기반 방법은 총 27 개, 74 차원의 특징 집합에 대해 PART, Support Vector Machine 을 이용하 예측한 결과를 보여준다. 그 결과는 각각 PART 는 54.42%, Support Vector Machine 은 77.43%이다. 이 방법은 실험 데이터로 Swiss-Prot 의 데이터[14]를 사용하였기 때문에 직접적인 비교는 될 수 없으나, 전체적인 성능을 비교하기 위한 참고 자료로 사용하였다. 서열 유사도 방법만을 사용한 경우는 94.85%의 예측 정확도를 보여줬다. 여기서 서열 유사도 방법이라 함은 제안하는 시스템에서의 첫 번째 단계 즉, E-

value 는 0.01 이하, 유사도는 30% 이상인 단백질에 대해 BLAST 로 예측한 방법을 의미한다. 제안하는 시스템은 서열 유사도 방법과 특징 기반 방법을 융합한 방법 중 모델 구축에 PART 를 적용하였을 때 95.14%, Support Vector Machine 을 적용하였을 때 95.22%의 정확도를 보였다.

(그림 3)에서 나타난 것과 같이 특징 기반 방법은 어떠한 알고리즘을 사용하느냐에 따라 그 편차가 심하게 나타남을 볼 수 있다. 이에 반하여 제안하는 시스템은 사용하는 알고리즘에 따른 예측 정확도의 편차가 상대적으로 작음을 볼 수 있으며 또한, 예측의 정확도가 다른 두 방법에 비해 높음을 볼 수 있다.

다음으로 상동성이 없는 서열 정보에 대한 결과이다. <표 4>는 이러한 서열 정보를 이용한 단백질 기능 예측 연구인 SVMProt 과 제안하는 시스템의 예측 결과를 비교하여 나타낸다.

<표 4> 상동성이 없는 정보에 대한 결과 비교표

	SVMProt	제안하는 시스템
전체	86%	95.22%
상동성 없음	72%	62.86%

이러한 서열 정보는 Swiss-Prot 데이터베이스에서 키워드 검색을 통해 'novel', 'distinct', 'unrelated' 의 키워드를 키워드 'enzyme' 과 결합하여 검색한 결과이다. 위의 결과를 보면 전체적인 정확도 측면에서는 제안하는 시스템이 95.22%의 성능을 보이나 상동성이 없는 정보에 대해서는 SVMProt 이 더 정확함을 볼 수 있다. 이러한 문제를 해결하기 위해서는 서열 유사도 방법보다는 특징 기반 방법에서 좀 더 서열간의 독립적인 특징을 적용하는 것이 필요하다.

7. 결론 및 향후 연구

본 논문은 서열 유사도 방법과 특징 기반 방법의 융합을 통하여 단백질의 기능을 예측하는 시스템을 제안하였다. 전체적인 서열 정보에 대한 정확도의 결과는 PART 를 적용하였을 때 95.14%, Support Vector Machine 을 적용하였을 때 95.22%로 얻을 수 있었다. 그러나 상동성이 없는 정보에 대해서는 62.86%의 예측 정확도를 보여주었다.

제안하는 시스템이 좀 더 정확한 단백질 기능 예측 도구로 발전하기 위해서 특징 기반 방법의 개선이 필요하다. 특징 집합을 선정하는 데 있어서 서열 유사도에 독립적이면서 특정 기능을 갖는 단백질이 갖는 공통 특징을 찾아내어 특징 집합에 적용한다면 특징 기반 방법의 성능이 개선될 것이라 생각한다. 향후 이를 위한 연구를 진행할 예정이다.

참고문헌

[1] Stephen F. Altschul et al.(1990), "Basic Local Alignm-ent Search Tool", *J. Mol. Biol.*, **215**, pp.403-410.

[2] Damien Devos and Alfonso Valencia(2000), "Practical Limits of Function Prediction", *PROTEINS: Structure, Function, and Genetics*, **41**, pp.98-107.

[3] Paul D. Dobson and Andrew J. Doig(2004), "Predicting Enzymes Class From Protein Structure Without Alignments", *J. Mo. Biol.*, **345**, pp.187-199.

[4] Lars J. Jensen, Marie Skovgaard, and Soren Brunak(2002), *Protein Science*, **11(12)**, pp.2894-2898.

[5] 이범주(2009), "효소 기능 예측을 위한 아미노산 서열에서의 특징 추출", 박사학위 논문(충북대학교).

[6] WEKA, http://www.cs.waikato.ac.nz/ml/weka/

[7] A.J. Ikai(1980), "Thermostability and aliphatic index of globular proteins", *J. Biochem*, **88**, pp.1895-1898.

[8] Stanley C. Gill and Peter H. von Hippel(1989), "Calculation of Protein Extinction Coefficients from Amino Acid Sequence Data", *Analytical Biochemistry*, **182**, pp.319-326.

[9] E. Gasteiger et al.(2005), "Protein Identification and Analysis Tools on the ExPASy server", *The Proteomics Protocols Handbook*, pp.571-607.

[10] J. Kyte and R.F. Doolittle(1982), "A simple method for displaying the hydropathic character of a protein", *J. Mol. Biol.*, **157**, pp.105-132.

[11] Kunchur Guruprasad, B.V.Bhasker Reddy and Madhusudan W.Pandit(1990), "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence", *Protein Engineering*, **4(2)**, pp.151-161.

[12] B. Bjellqvist et al.(1994), "Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions", *Electrophoresis*, **15**, pp.529-539.

[13] L.Y. Han et al(2004), "Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach", *Nucleic Acids Research*, **32(21)**, pp.6437-6444.

[14] B. Boeckmann et al.(2003), "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003", *Nucleic Acids Research*, **31**, pp.365-370.