대용량 미세환경 정보처리를 통한 단백질 기능 예측 자동화

민혜영*, 윤성로**
*스탠퍼드대학교 통계학과
**고려대학교 전기전자전파공학부
e-mail : sryoon@korea.ac.kr

Automatic Protein Function Prediction Through Processing Large-Scale Protein Microenvironment Information

Hyeyoung Min*, Sungroh Yoon**
*Department of Statistics, Stanford University, California, USA
**School of Electrical Engineering, Korea University, Seoul, Korea (Corresponding author)

요 약

정보처리 기술의 발전에 따라 정보기술을 통한 생명과학 문제 해결을 연구하는 생명정보학 (bioinformatics) 분야에서도 보다 대용량의 바이오 정보를 처리하게 되었다. 특히 우리 몸을 이루는 핵심 요소인 단백질의 기능 예측 자동화는 다루어야 할 정보량이 매우 방대한 관계로 일찍부터 컴퓨터를 사용한 정보처리 기법이 중요하게 다루어져 왔다. 본 연구에서는 특정 단백질 주변의 미세환경 (microenvironment)에 관한 정보를 수집하고 분석하여 그 기능이 알려진 다른 종류의 단백질 주변의 미세환경과 비교함으로써 기능을 예측하는 방법에 대해 소개한다.

1. 서론

2000 년 초에 완결된 HGP (Human Genome Project)의 경우 인간 유전자의 완전 해독이라는 목표를 컴퓨터를 사용하여 보다 신속하게 달성할 수 있었다. 몇 가지 가설을 우선 세우고 이를 증명하기 위해 잘 준비된 실험을 수행하는 전통적인 생명과학의 패러다임에서 벗어나, 불특정 다수의 실험을 먼저 수행하고 이를 통해 수집된 대용량의 정보를 컴퓨터로 분석하여 많은 수의 생물학적 가설을 도출해내는 새로운 패러다임을 사용하였기 때문이다.

HGP 에 의해 인간 유전자의 종류가 밝혀진 이후, 개별 유전자의 기능 및 여러 유전자 간 상호작용을 설명하기 위한 연구가 활발히 수행되면서 방대한 양의 생물학 정보가 생성되고 있다. 특히 단백질 기능을 컴퓨터를 사용하여 밝혀내기 위한 연구도 단백질 기능 예측 자동화 (Automatic Protein Function Prediction) 라는 이름으로 활발히 연구되고 있다. 특히 그 종류와 기능이 무척 다양하기 때문에, 컴퓨터를 이용한 관련 정보의 분석이 핵심적이라고 할 수 있다.

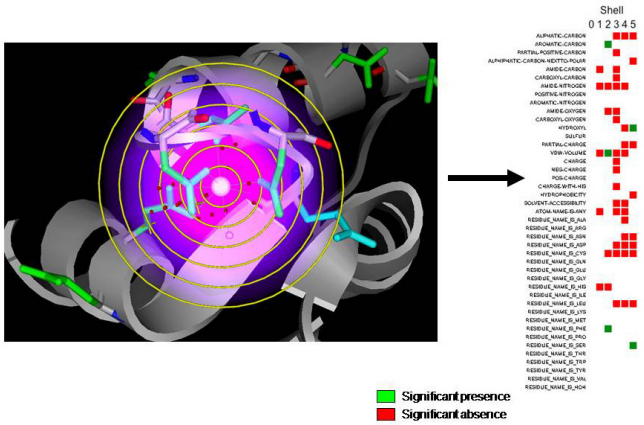
본 연구에서는 인터넷에 공개되어 있는 PDB (Protein Data Bank) [1] 상의 거의 모든 단백질 개개에 대해 그 주변 환경에 대한 각종 정보를 구축하고, 이를 분석함으로써 각종 단백질의 기능을 예측하고, 기존에 알려지지 않았던 새로운 기능을 발굴하는 것을 그 목적으로 한다.

구체적으로, 우선 개별 단백질에 대해 1 차원적인 아미노산 (amino acid) 배열 (sequence) 상에 중심 (center)을 설정하고, 이를 중심으로 다수의 동심 구 (co-centric sphere)를 형성한 뒤, 각 동심 구 상의 각종 미세 환경 변수 (micro-environmental parameters)를 추출해 낸다. 이 정보를 264 차원의 벡터 (vector) 형식으로 전체 단백질에 대해 얻어낸 뒤, 같은 클러스터에 속한 단백질은 유사한 기능을 가진다는 가설 아래, 단백질 미세 환경의 벡터들을 클러스터링 (clustering) 기법을 통해 군집화 한다.

이와 관련하여 이미 선행연구를 수행하였으나 [2], 선행연구에서는 미세 환경 벡터간 의미 있는 거리 메트릭 (metric)을 도출해 내는데 중점을 두었고, K-means clustering 이라는 비교적 간단한 클러스터링 방법을 사용하였다. K-means clustering 은 O(Kn) 알고르즘으로 그 수행 속도는 비교적 빠르나, 예측값 최대화 (Expectation Maximization)에 속하는 방법으로 최종 결과가 최초에 K 개의 클러스터 중심을 어떻게 정하느냐에 좌우된다는 한계를 가진다. 또한 현재 단백질의 기능이 총 몇 가지냐에 대한 명확한 해답이 없는 상태로, K 값을 정하는 과정에서도 다양한 문제가 발견되었다.

본 연구에서는 기존 선행연구에서 사용된 벡터간 거리 metric 을 근간으로 하여 새로운 군집화 방법을 사용한 단백질 기능 예측 자동화 기법 개발에 초점을 두었다.

2. 연구 방법 및 결과



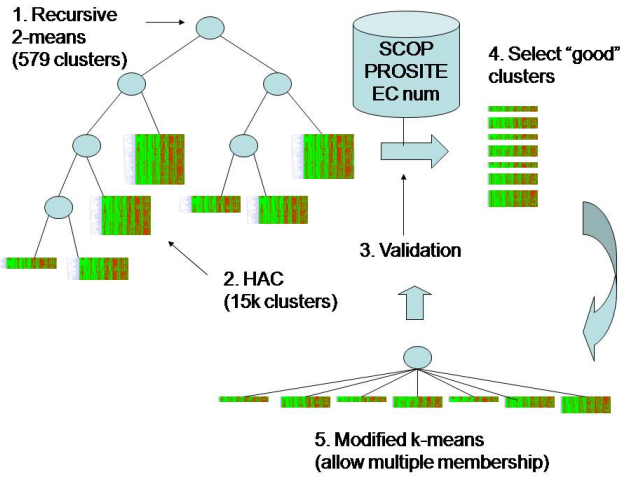
(그림 1) 단백질 미세환경 추출

우선 기존 연구와 동일하게 각 단백질 중심으로부터 동심 구를 설정하고 이를 중심으로 각종 환경 변수를 추출해내었다 (그림 1). 사용된 환경 변수는 <표 1>과 같다. (44 개의 변수를 1 angstrom 을 간격으로 한 6 개의 동심 구상에서 측정, 총 256 개 변수가 사용되었으나 아래 표에는 그 값이 언제나 일정한 2 개 변수를 제외한 42 개의 변수만 포함하였다.)

<표 1> 본 연구에 사용된 단백질 미세 환경 변수

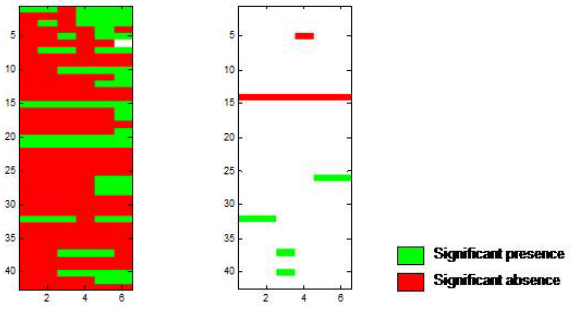
1 Aliphatic carbon	22 residue name is ala
2 aromatic carbon	23 residue name is arg
3 carbon w/partial positive charge	24 residue name is asn
4 aliphatic C next a polar atom	25 residue name is asp
5 amide carbon	26 residue name is cys
6 carboxyl carbon	27 residue name is gln
7 amide nitrogen	28 residue name is glu
8 positively charged nitrogen	29 residue name is gly
9 aromatic nitrogen	30 residue name is his
10 amide oxygen	31 residue name is ile
11 carboxyl oxygen	32 residue name is leu
12 hydroxyl oxygen	33 residue name is lys
13 sulfur	34 residue name is met
14 float partial charge	35 residue name is phe
15 +float Van der Waals volume, #atoms	36 residue name is pro
16 +float charge	37 residue name is ser
17 +float negative charge	38 residue name is thr
18 +float positive charge	39 residue name is trp
19 +float charge on histidines	40 residue name is tyr
20 float hydrophobicity	41 residue name is val
21 +int solvent accessibility	42 residue name is other

위 과정을 통해 약 1 천만 개의 264 차원 벡터를 생성하였으며, K-means clustering 방법 대신 recursive 2-means 방법과 hierarchical clustering 방법이 접목된 새로운 군집화 방법을 통한 단백질 기능 예측 자동화 기법을 사용하였다. Hierarchical clustering 방법은 $O(n^2)$ 알고리즘으로 그 수행 속도가 K-means clustering 방법보다 느릴 수 있지만, 본 연구에서는 우선적으로 2-means clustering 방법을 연속적으로 사용하여 데이터 크기를 줄인 후에 hierarchical clustering 방법을 적용함으로써 수행 시간에 대한 문제를 최소화 하였다. 또, recursive 2-means clustering 방법을 적용함으로 인해 미리 K 값을 정해야 한다는 K-means clustering 의 한계도 극복하고자 노력하였다. 또한 하나의 벡터를 복수의 클러스터에 할당함으로써, 하나의 단백질이 여러 가지 기능을 가질 수 있다는 생물학적 사실을 반영하도록 하였다. (그림 2) 에 본 연구에 사용된 군집화 기법이 요약되어있다.



(그림 2) 본 연구에 사용된 군집화 기법

위 방법을 사용하여 약 15,000 개의 클러스터를 얻었으며 이 중 인슐린 등 일부 중요 단백질의 미세 환경정보를 SCOP [3] 등 독립적인 별개의 데이터베이스를 통해 검증하였고, (그림 3)과 같이 visualization 하였다.



(그림 3) 인슐린 미세 환경

3. 결론

단백질 기능 예측 자동화를 위한 유용한 정보처리 기법을 고안하였으며 그 효용성을 중요 단백질에 대해 검증하였다. 후속 연구를 통해 적용분야 확대를 계획하고 있다.

참고문헌

[1] J. L. Sussman et al., "Protein Data Bank (PDB)," *Acta Crystallogr*. D 54, 1078-1084, 1998.

[2] S. Yoon et al., "Clustering protein environments for function prediction: finding PROSITE motifs in 3D," *BMC Bioinformatics*, 8(Suppl 4):S10, May 2007.

[3] A. G. Murzin et al., "SCOP: A structural classification of proteins database for the investigation of sequences and structures," *Journal of Molecular Biology*, vol. 247, no. 4, pp. 536-540, April 1995.