상호작용 네트웍 사전 구축을 이용한 단백질 기능 예측

진희정° 조환규
부산대학교 정보컴퓨터공학부
{hjjin°, hgcho}@pusan.ac.kr

Protein Function Prediction by Constructing Interaction Network Dictionary

HeeJeong Jin° HwanGue Cho
Dept. of Computer Science and Engineering, Pusan National University

요 약

단백체는 세포가 처해있는 환경에 따라, 그리고 각 조직 별로 유동적으로 존재하며, 세포의 실제적인 기능을 표현해준다. 이러한 이유로 세포 내에서 일어나는 실제적인 현상들을 전체 단백질 단계에서 통합적으로 파악하고자 하는 단백체학 연구가 활발하게 진행되고 있다. 미지의 단백질의 기능을 밝혀내는 연구는 단백체학의 가장 기본적이면서 중요한 부분이라고 할 수 있다. 본 논문에서는 "단백질 상호작용 네트웍 사전(PIND)"을 구축함으로써 단백질의 기능을 예측하는 새로운 방법론을 소개한다.

1. 서 론

단백질들은 생물체를 구성하는 중요한 성분인 동시에, 생물이 살아가는데 필요 요소인 효소, 항체 호르몬 등으로 작용한다. 세포 내에서 일어나는 신호전달, 세포 주기, 분화, DNA 복제 및 전사, 번역, 대사등 거의 모든 반응들은 수많은 단백질의 상호작용을 통해 수행되고 조절되어진다. 따라서 서로 상호작용하는 단백질들의 관계를 조사하면 생체 내에서 일어나는 현상의 메카니즘을 알아낼 수 있고, 이를 통하여 여러 가지 병을 예방·치료 할 수 있다. 많은 연구자들이 이에 대한 연구를 진행하고 있는데, 실험 결과로 얻어진 단백질 상호작용 데이터를 단백질-단백질 상호작용(Protein-Protein Interaction) 데이터라고 한다. 전 세계적으로 이미 구축되어 있는 PPI 데이터베이스로는 YPD[1], DIP[2],MIPS[3]등이 존재한다. 각 데이터베이스는 실험 방법론이나, 검증에 따라 데이터의 종류나 양의 차이가 있다.

현재까지 연구되어 온 단백질 기능 예측 알고리즘에는 PPI 데이터를 이용한 Majority Rule[4]이나 Random Markov Model[6], Topological Structure Abalysis[6,7], 단백질 서열을 이용하여 분석하는 방법[8]이 있다. Hui et al[9]의 논문에서는 같은 클러스터에 포함된 단백질들과 서로 다른 클러스터에 포함된 단백질들간의 상관 계수를 알아보았다. 이러한 연구들은 "유사한 기능을 하는 단백질들이 상호작용한다"는 개념을 바탕으로 한 것이다.

Majority Rule을 이용한 단백질 기능 예측 방법은 PPI 데이터에서 특정 노드(단백질) p_i의 기능을 예측할 때, 그 단백질과 상호작용하는 단백질들 중, 알고 있는 기능들을 모두 모아서 그 중 가장 많이 나타나는 기능을 p_i의 기능으로 알려주는 것이다. 이러한 방법론은 현재 노드와 직접적으로 상호 작용하는 단백질만으로 기능을 예측한다는 문제점이 존재한다. 따라서 항상 예측하는 단백질의 기능이 이웃 단백질에 포함되어 있어야한다는 큰 단점이 존재한다.

Random Markov Model은 단백질 상호작용 데이터를 Gibbs 분포로 정의하고, 단백질 상호 작용 데이터의 기능을 알고자 하는 단백질과 패스가 설정될 수 있는 모든 단백질들의 기능을 고려한다. 따라서 기능을 알고자 하는 단백질과 직접 상호작용하는 단백질들의 기능을 모두 알지 못하더라도 Majority Rule과는 달리 기능을 예측할 수 있는 장점이 있다. 하지만 현재 단백질에서 기능이 모두 알려져 있는 경우 이들 기능이 가장 많은 영향을 주고, 이웃 단백질들이 포함하고 있는 기능들이 가장 높은 확률값을 가지기 때문에, 기능을 알고자 하는 단백질의 기능은 이웃하는 단백질들의 기능으로 예측된다. 하지만, 실제 단백질 상호작용 데이터에서는 상호작용하는 단백질들의 기능과는 전혀 다른 기능을 가지는 단백질들이 존재한다. 이러한 경우, 낮은 확률값을 고려하지 않으면, 기능 예측에 실패할 수 있다.

Topological Structure Analysis 방법론은 서로 유사한 기능을 하는 단백질들은 상호 작용하는 경우가 많고, 이들은 함께 모여 있다는 점을 기초로 하여 HCS(Highly Connected Subgraph)와 같은 영역을 단백질 상호 작용 데이터에서 찾아낸 후, HCS 그래프 안에서 Majority Rule을 사용하여 기능을 유추하는 방법이다. Topological Structure Analysis 방법론은 예측율은 높지만, HCS와 같은 특정 구조에 포함되지 않은 단백질들은 기능을 예측하는 단백질들의 범위에 벗어나므로 예측을 할 수 없다는 단점이 있다.

2. PPI 데이터의 분석

현재 개발된 PPI 데이터를 이용한 단백질 기능 예측 방법론들은 "유사한 기능을 하는 단백질들이 상호작용한다"는 개념을 바탕으로 한 것이다. 하지만, 특정 단백질들은 그 단백질과 상호작용하는 단백질들이 가지고 있는 기능들과는 전혀 다른 기

능을 하는 경우가 있다. PPI 데이터에서 단백질을 p_i, p_i의 기능을 $F(p_i)$, p_i와 상호작용하는 단백질들을 $N(p_i)$, $N(p_i)$의 기능을 $F(N(p_i))$라고 하자. 그림 1은 MIPS에서의 단백질 *YKL143w*와 $N(YKL143w)$를 나타낸다. 각 단백질의 기능은 하나의 문자로 표현되어있다. 그림 1에서 $F(YKL143w)$과 $F(N(YKL143w))$ 사이에는 공통되는 기능이 없음을 알 수 있다. 따라서 기존의 단백질 기능 예측 방법론으로는 예측되어질 수 없다.

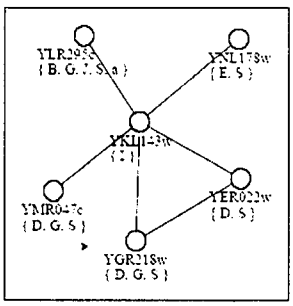
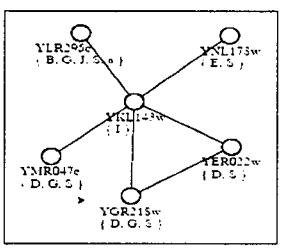


그림 1. MIPS의 한 단백질과 그 상호작용하는 단백질들의 예 : *YKL143w* 단백질의 $F(YKL143w)=I$ 이지만, $I \notin F(N(YKL143w))$이다.

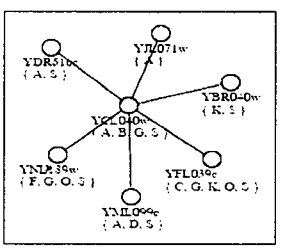
본 논문에서는 PPI 데이터에서의 각 단백질들을 자신의 기능과 상호작용하는 단백질들의 기능사이의 관계에 따라 세 가지로 분류하였다.

- *D-Protein* : 단백질 p_i는 $F(p_i) \cap F(N(p_i)) = \varnothing$이면, D-Protein이다.
- *O-Protein* : 단백질 p_i는 $F(p_i) \cap F(N(p_i)) \neq \varnothing$이고, $F(p_i) \not\subset F(N(p_i))$이면, *O-Protein*이다.
- *S-Protein* : 단백질 p_i는 $F(p_i) \subset F(N(p_i))$이면, *S-Protein*이다.

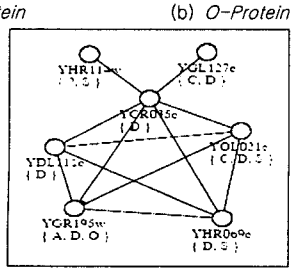
그림 2는 MIPS에서의 *D-Protein, O-Protein, S-Protein* 단백질의 예를 나타낸다.



(a) *D-Protein*



(b) *O-Protein*



(c) *S-Protein*

그림 2. *D-Protein, O-Protein S-Protein* 단백질의 예 : (a) $F(YKL143c) \cap ?(N(YKL143c)) = \{H, G, S, D, S, D, G, S, E, S, B, G, J, S, a\}$ 이므로, *YKL143c*은 D protein 이다. (b) $\{H, G, S, D, S, D, G, S, E, S, B, G, J, S, a\} \cap \phi$. (b) $F(Y\ CL040w) \cap ?(N(Y\ CL040w)) = \{A, B, G, S\} \cap \{K, S, A, D, S, F, G, O, S, A, S, C, G, K, O, S\} = \{A, G, S\}$이므로, *YCL040w*은 *O-Protein*이다. (c) $F(Y\ CR035c) \subset ?(N(Y\ CR035c))$ 이므로, *Y CR035c*는 *S-Protein* 이다.

표 1은 MIPS 데이터에서의 *D-Protein, O-Protein, S-Protein* 단백질을 분류한 것이다. 표 1에서 알 수 있듯이, MIPS 데이터에 Majority Rule을 적용한 경우, *S-Protein*인 57.45%의 단백질들만을 정확하게 예측할 수 있다. 따라서 *D-Protein*과 *O-Protein*들을 예측하기 위해서 새로운 방법이 필요하다.

표 1. MIPS에서의 단백질의 분류

Total	*D-Protein*	*O-Protein*	*S-Protein*
3,387	387(11.05%)	1,067(31.50%)	1,946(57.45%)

3. PIND를 이용한 단백질 기능 예측

표 2는 MIPS의 모든 단백질들의 $F(N(p_i))$의 맞추어본것이다. 표 2에서 단백질 p_i, p_j쌍의 $F(N(p_i)) = F(N(p_j))$인 비율이 아주 높음을 알 수 있다. $F(N(p_i)) = F(N(p_j))$이고 $F(N(p_i)) = F(N(p_j))$인 단백질은 전체 MIPS에서 34.8%이며, $F(p_i)) = F(p_j)$ 데이터에서는 73.3%이다. 따라서 MIPS 전체 데이터에서 $F(N(p_i)) = F(N(p_j))$인 단백질들을 이용하여 기능을 예측하여도 전체의 34.8%의 단백질은 정확하게 기능을 예측할 수 있다.

표 2. MIPS에서의 단백질 p_i, p_j쌍의 $F(N(p_i)) = F(N(p_j))$과 $F(p_i)) = F(p_j)$인 비율

Total	$F(N(p_i)) = F(N(p_j))$	$F(N(p_i)) = F(N(p_j)) \cap F(p_i) = F(p_j)$
3,387	1,609(47.50%)	1,180(34.80%, 73.30%)

*D-Protein*과 *O-Protein*들을 예측하기 위해서, 본 논문에서는 PIND를 구축한다[10]. 이를 위해 PPI의 각 기능을 알파벳 문자 하나로 표현하고, 각 단백질의 기능을 알파벳 문자로 이루어진 $F(p_i)$을 만들었다. *PIND*는 각 단백질 p_i에 대하여, $F(p_i)$와 $F(N(p_i))$의 문자열을 저장하는 사전형식의 구조를 갖는다. 표 2는 MIPS의 *PIND*의 예를 나타낸다.

표 3. MIPS의 PIND

Index	p_i	$F(p_i)$	$F(N(p_i))$
1	Q0045	{B,S}	{B,S}
2	Q0085	{B,G,J,S,a}	{B,G,J,S,a}
..	..	..	..
9	TY1A_PI.	{N}	{D,S,S}
..	..	..	..

단백질 p_i의 기능을 예측하기 위해서 전체 *PIND*에서 $F(N(p_i))$가 유사한 단백질들 *k*개를 선택하고, 선택된 단백질들의 기능

을 p_i에 할당한다. 본 논문에서는 $F(N(p_i))$의 유사성을 계산하기 위해서 *Czekanovski-Dice* measure를 사용한다. 이를 위해 집합 F를 $F(p_i)$라 하고, $r(x)=\{F(N(p_i))|F(p_i)=x \text{ and } p_i \in MIPS\}$, $R=\{r(x)|x \in F\}$라 하자. 모든 $r(i)$와 $r(j)$의 쌍에 대하여 *Czekanovski-Dice* measure를 적용한다. a를 $r(i)$와 $r(j)$에 공통적인 기능의 수, b를 $r(i)$의 기능의 수, c를 $r(j)$의 기능의 수라 하면, 유사정도 $Simil(i,j)$는 다음과 같이 계산되어 진다.

$$Simil(i,j)=\frac{2 \cdot a}{b+c}$$

4. 실험결과

본 논문에서는 기능 예측의 정확도를 위해서 leave-one-out 방법을 사용하였다. leave-one-out 방법은 기능을 알고 있는 하나의 단백질을 기능을 모르는 단백질로 표시한 다음 그 기능을 예측하는 방법이다. 본 논문에서는 기능을 아는 모든 단백질에 대해서 leave-one-out 방법으로 기능을 예측하고, 그 정확도를 측정하였다. 그림 3은 단백질의 degree에 따른 PIND를 이용한 기능 예측의 정확도를 나타낸다. degree가 1이상인 경우(전체 단백질)에서는 SN이 가장 높지만, 반면에 SP는 가장 낮다. degree가 5이상인 단백질들만을 예측했을 경우에는 SP가 가장 높지만, 반면에 SN은 가장 낮다.

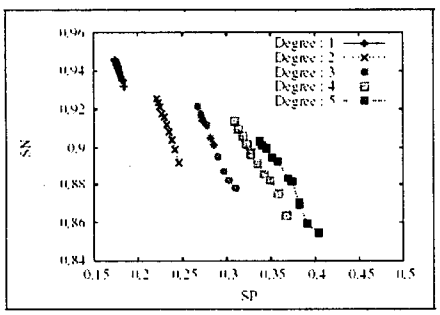
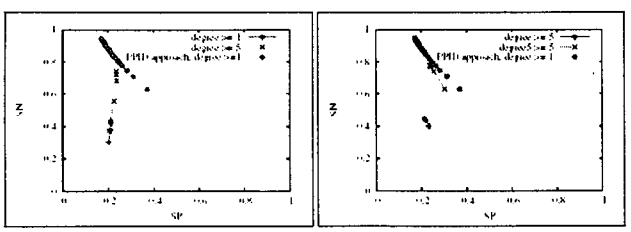


그림 3. MIPS 데이터의 기능 예측 결과. "degree : d"는 d보다 큰 degree를 가진 단백질들만을 이용하여 기능을 예측한 결과라는 뜻이다.

그림 4는 Majority Rule과 Chi-Square 방법을 본 논문에서 사용한 PIND를 이용한 방법의 기능 예측 정확도를 비교한 것이다. Majority Rule에서는 SN=0.77(SP=0.23이고 degree는 5이상)일 때 가장 높고, Chi-Square는 SN=0.74(SP=0.23이고 degree는 5이상)일 때 가장 높다. 그림 3와 4를 통해서 본 논문에서 제시한 방법이 Majority Rule과 Chi-Square 방법론에 비하여 신뢰성이 높은 방법이라 할 수 있다. 이전에 개발된 Random Markov Model과는 직접적인 비교를 할 수 없었지만, Markov Model의 Majority Rule과의 비교 결과를 통하여 본 방법이 경쟁력이 있다고 볼 수 있다.



(a)Chi-Square와 PIND 방법론 (b) Majority Rule와 PIND 방법론

그림 4. (a) Chi-Square와 PIND 방법론의 비교, (b) Majority Rule와 PIND 방법론의 비교

5. 결론

미지의 단백질의 기능을 밝혀내는 연구는 단백체학의 가장 기본적인 문제이다. 현재까지 "유사한 기능을 하는 단백질들이 상호작용한다"는 개념을 바탕으로 한 여러 방법론들이 제시되었지만, 이러한 개념에 맞지 않은 단백질들로 인하여 높은 예측 정확도를 기대할 수 없었다. 본 논문에서 제시한 PIND를 이용한 방법은 실험 결과를 통해서 Majority Rule과 Chi-Square 방법론에 비하여 신뢰성이 높은 방법이며, 가장 최근에 개발된 Random Markov Model과도 경쟁력이 있다. PIND 모델의 주요 내용은 다음과 같다.

- PIND를 이용한 방법은 이전에 제시된 이웃 단백질들의 기능을 세는 Majority Rule과 Chi-Square과 같은 방법론들에 비하여 높은 sensitivity와 specificity를 제공한다.
- PIND를 이용한 방법은 이전에 개발된 방법론들에 비하여 특히 *D-Protein, O-Protein*에서 높은 sensitivity와 specificity를 제공한다.

6. 참고 문헌

[1] Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI, "The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data," *Nucleic Acids Research*, 27, 1999

[2] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D, "The Database of Interacting Proteins", *Nucleic Acids Research*, 32, 2004

[3] Mewes HW, Frishman D, Guldener U, Mannhaupt G, Mayer K, Mokrejs M, Morgenstern B, Munsterkotter M, Rudd S, Weil B., "MIPS: a database for genomes and protein sequences", *Nucleic Acids Research*, 30, 2002

[4] Alexei Vazquez, Alessandro Flammini, Amos Maritan and Alessandro Vespoignani, "Global protein function prodiction from protein-protein interaction networks", *Nature Biotechnology*, 21, 2003

[5] Minghua Deng, Kui Zhang, Shipra Mehta, Ting Chen and Fengzhu Sun, "Prediction of Protein Function Using Protein-Protein Interaction Data", *Journal of Computational Biology*, 10, 2003

[6] Dongbo Bu, Yi Zhao, Lun Cai, Hong Xue, Xiaopeng Zhu, Hongchao Lu, Jingfen Zhang, Shiwei Sun, Lunjiang Ling, Nan Zhang, Guoji Li and Runsheng Chen, "Topological structure analysis of the protein-protein interaction network in budding yeast", *Nucleic Acids Research*, 31, 2003

[7] N.Przulj, D.A.Wigle and I.Jurisica, "Functional topology in a network of protein interactions", *Bioinformaics*, 20, 2004

[8] S Wuchty, A N Oltvai and A-L Barabasi, "Evolutionary conservation of motif constituents in the yeast protein interaction netwok", *Nature Genetics*, 35, 2003

[9] Hui Ge, Zhihua Liu, George M.Church, and Marc Vida, "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae", *Nature Genetics*, 33, 2001

[11] Hee-Jeong Jin and Hwan-Gue Cho, "Computational Method for Protein Function Prediction by Constructing Protein Interaction Network Dictionary", *INTERNATIONAL JOURNAL OF PATTERN RECOGNITION AND ARTIFICIAL*, to be appear