상호작용 맵에서 단백질 기능 예측

정재영° 최재훈 박종민 박선희
한국전자통신연구원
{jjy72°, jhchoi, jmpark93, shp}@etri.re.kr

A Protein Function Prediction in Interaction Maps

JaeYoung Jung° JaeHun Choi JongMin Park SeonHee Park
Bioinformatics Research Team, Computer System Research Department
Electronic Telecommunication Research Institute(ETRI)

요 약

단백질 상호작용 데이터는 현 생물정보학에서 기능이 알려지지 않은 단백질의 기능 예측에 높은 신뢰성이 있는 프로티오믹스의 계산 모델에 이용되고 있다. 일반적으로 이 단백질 기능 예측 알고리즘들은 대규모의 2차원 단백질-단백질 상호작용 맵에서 *Guilt-by-Association* 개념 기반으로 개발되고 있다. 본 논문에서는 단백질-단백질 상호작용 데이터를 이용한 그래프 기반 단백질 기능 예측 모델을 개발하였다. 특히, 이 모델은 대량의 상호작용 데이터에서 정확한 기능 예측을 수행할 수 있다는 장점을 가지고 있다. 이를 위해 Yeast에 대한 단백질 상호작용 맵, Homology 및 Interaction Generality를 이용하여 이 모델를 평가하였다.

1. 서 론

초기의 생물정보학(bioinformatics) 연구가 유전자 서열 정보 등 방대한 양의 새로운 생물학 데이터들을 저장 및 분석하기 위한 데이터베이스 개발에 초점을 맞추었다면, 현재는 미지의 단백질 기능 예측을 위한 다양한 접근 방법들에 더 많은 관심을 가질 때이다.

세포 내 여러 종류의 분자들 중에서 단백질 상호작용 데이터는 고 성능 실험 기법(high-throughput technology)들이 개발· 이용되어 실험으로부터 대규모 데이터를 생산하게 되었다. 이렇게 얻어진 방대한 실험 데이터로 구성된 2차원 단백질-단백질 상호작용 맵(interaction map)에 그래프 이론을 적용하면 기능이 알려지지 않은 단백질(unknown function proteins)의 기능에 대한 이해의 폭을 확장할 수 있다. 또한 단백질 상호작용 맵은 직· 간접 상호작용 데이터를 이용하여 기능이 알려지지 않은 단백질들의 기능 예측 및 단백질 복합체(protein complex)등의 주요한 분석 도구로 이용되고 있다.

본 논문에서는 단백질-단백질 상호작용 데이터를 이용한 기능이 알려지지 않은 단백질의 기능 예측을 위한 계산 모델에 대해 살펴보고, 전문가 지식 및 유전자의 유사성을 이용한 단백질 기능 예측 알고리즘을 제안한다. 이러한 단백질 상호작용 예측 모델의 개발은 상호작용의 데이터의 기능상 분류 및 새로운 실험의 비용을 줄여 줄 것으로 기대된다.

2. 단백질 기능 예측

어떤 단백질이 무슨 기능을 할 것이냐에 대한 단서는 이미 기능을 알고 있는 다른 단백질과 어떻게 상호작용을 하는지 여부를 살펴봄으로써 알 수 있으며, 이러한 원리를 *Guilt-by-Association*이라고 한다. [1]

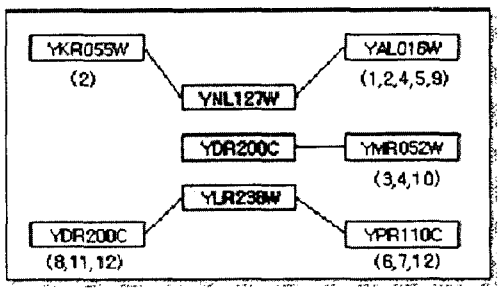

그림 1. *Guilt-by-Association* 방법

그림 1은 Yeast *Saccharomices Cerevisiae*의 단백질-단백질 상호작용의 일부분이다. 그림에서 기능이 밝혀지

지 않은 단백질들은 회색 박스로 표현이 되고, 나머지 단백질들은 기능이 밝혀졌으며 소괄호 내부 숫자의 기능들을 가진다. 단백질 YNL127W, YDR200C 그리고 YLR238W에 대하여 Guilt-by-Association을 이용하면 기능이 알려지지 않은 단백질에 대해 YNL127W (2), YDR200C (3, 4, 10) 그리고 YLR238W(12)의 기능을 부여할 수 있다.[2]

단백질-단백질 상호작용 데이터를 이용한 단백질 기능 예측 방법은 그림 2로 표현된다. 먼저 데이터베이스에 저장된 물리적인 단백질-단백질 상호작용 데이터로부터 그래프 알고리즘을 이용하여 단백질 상호작용 맵을 구성한다. 그런 다음 *Guilt-by-Association* 방법을 적용하여 기능이 밝혀진 인접한 단백질을 통하여 기능이 알려지지 않은 특정 단백질의 기능을 예측한다.

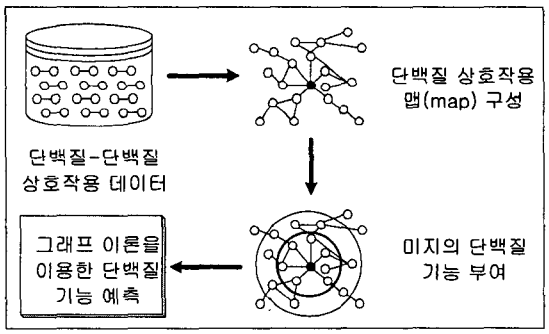

그림 2. 단백질-단백질 상호작용 데이터를 이용한 단백질 기능 예측 방법

2.1 단백질 상호작용 데이터

지금까지 밝혀진 Yeast 데이터는 약 6,000여 개의 유전인자가 있으며, 약 60%의 단백질에 대한 기능이 밝혀졌다.[2] 본 논문에서는 Yeast 단백질의 기능 예측을 위한 데이터 셋(set)으로 KDD Cup 데이터를 이용하였는데, 이는 MIPS 데이터베이스로부터 만들어졌으며 각 단백질은 6개의 속성(attribute) 및 13개의 기능(function)과 15개의 Localization, 그리고 단백질-단백질 상호작용(protein-protein interaction)의 데이터로 구성되어있다.[3]

단백질 기능 예측 실험에 사용된 데이터를 자세히 살펴보면 학습 데이터(training data)에 나타나는 단백질들이 862개 이며, 테스트 데이터(test data)는 380개이다. 그리고 단백질의 이름을 고려하지 않고 속성 및 기능을 갖는 학습 단백질이 4,346개이다.

2.2 예측 모델

본 논문에서 제안하는 모델은 실험상에서 나타나는 False-Positive 상호작용에 대한 정량화 및 Expression Correlation 을 이용하는 기능 예측모델이다.

Interaction-Generality Algorithm

Annotate_Function(*TrainSet, TestSet, T, gth, eth,* λ, *d*)

// *TrainSet* is a hash table where each entry is presented by a pair of (P_i, F_i)

// *TestSet* is a list entry of proteins P_j for protein of unknown function

// *T* is a symmetric interaction matrix

// *gth* is the threshold value for interaction generality

// *eth* is the threshold value for expression correlation

// λ is the value for choosing λ – largest values with *gth* and *eth*

// *d* is the interaction depth

while TestSet.size() > 0 do

1. Select a set of neighbors of P_j in *TestSet*
   - $Q=\langle q_1, q_2, \ldots, q_K \rangle$ where $q_m=0$ for all m
   - $F_j=\langle f_1, f_2, \ldots, f_K \rangle$ where $f_m=0$ for all m
   - compute $Ng^d(j)$
2. Compute the function frequency of P_j
   - for $P_i \in Ng^d(j)$ do
     if $G(i,j) \leq gth$ and $EC(i,j) > eth$ then
     increment q_m by 1 if P_i has $f_m=1$
3. Annotate functions to F_j
   - decide λ largest functions $f'_1, f'_2, \ldots, f'_\lambda$ using Q
   - assign $\langle f'_1, f'_2, \ldots, f'_K \rangle$ to F_j
4. Add (P_j, F_j) to TrainSet
5. Remove Pj from TestSet

그림 3. 단백질 기능 예측을 위한 알고리즘

그림 3은 본 논문에서 제안하는 Interaction Generality 알고리즘이다. TrainSet은 기능이 알려진 단백질과 기능의 쌍 (P_i, F_i)으로 P_i 는 단백질의 이름, F_i 는 단백질 P_i 의 밝혀진 기능 벡터이다. TestSet은 기능이 알려지지 않은 단백질과 기능의 쌍으로 $(P_j, ?)$으로 표현된다. T 는

단백질-단백질 상호작용을 나타내는 2 차원 인접 행렬, *gth* 는 Interaction Generality의 값, *eth*는 상호작용을 하는 두 단백질의 correlation coefficient는 나타내는 2 차원 인접 행렬이다. 여기서 λ은 λ-번째 높은(λ-largest) 빈도수를 위한 값이며, d는 기능 예측에 고려되는 직-간접 상호작용의 *d*-깊이(*d*-depth)내 속한 단백질을 포함시키기 위한 값이다. 제안된 알고리즘은 파라미터 *gth*, *eth*, λ, *d* 등의 변화에 따라 기능 예측에 고려되는 단백질의 수를 제한하게 된다. TestSet의 P_j 에 대해 *d*-깊이 내 상호작용을 하는 TrainSet의 단백질 $P_i \in Ng^d(j)$로부터 기능 빈도 벡터 F_j 를 계산한다. 만약, P_i 와 P_j 의 interaction generality가 *gth*보다 작고, Expression Correlation이 *eth*보다 작으며 F_j 의 계산에서 P_i는 제외된다. F_j로부터 λ-largest 기능만이 P_j의 기능으로 예측하고, (P_i, F_j)는 TrainSet으로 옮겨진 후 TestSet에서는 제거된다. 알고리즘의 수행은 TestSet에 속한 단백질이 없거나 예측되는 TestSet의 단백질이 존재하지 않으면 예측을 종료한다.

2.3 성능 평가

그래프 기반 기능 예측 모델을 이용하여 KDD Cup 상호작용 데이터를 시각화 및 미지의 단백질에 대하여 기능 예측 값을 부여한 결과는 그림 4로 표현된다.

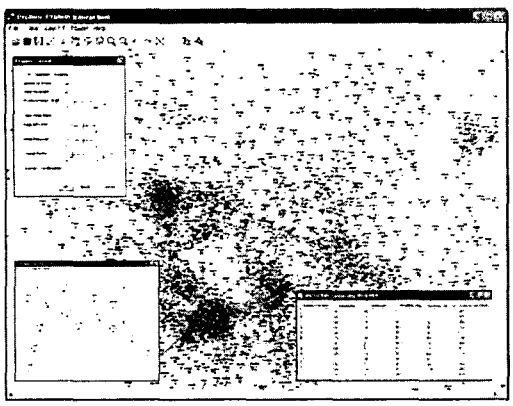

그림 4. 단백질 상호작용 시각화 및 기능 예측

예측모델에 대한 시뮬레이션은 IG의 값을 1부터 5까지 가변 시키면서 실험을 하였으며, 표 1은 IG가 5인 경우의 예측 성능을 보여준다. 단백질 기능예측 모델 결과는 Confusion Matrix로 된다. 또한 실험에서 Expression Correlation 값은 0.5 이상 되는 경우에만 예측을 하도록 하였다. 본 논문에서 시뮬레이션 결과 380개의 TestSet 데이터 중 304개의 단백질 기능 예측이 가능하였으며, 85% 이상의 높은 예측 성능 결과를 얻을 수 있었다.

λ	1	2	3	4	5	6	7
TP	543	602	644	652	654	654	654
TN	3006	2966	2859	2824	2812	2811	2811
FP	144	184	291	150	148	148	148
FN	259	200	158	150	148	148	148
Acc(%)	89.10	90.28	88.64	87.96	87.70	87.68	87.68

표 1. 단백질 기능예측 모델에 대한 결과치; True Positive(TP), True Negatives(TN), False Positives (FP), False Negatives(FN)

4. 결론 및 향후 연구

본 논문에서는 생물정보학 분야에서 그래프 이론을 기반으로 한 단백질 기능 예측 모델을 제안하였다. 이 모델은 *Guilt-by-Association*의 개념에 Expression Correlation 및 Interaction Generality 정보를 이용하고 있다. 또한, KDD Cup Yeast 단백질 데이터 셋에서 본 모델을 평가하였을 때, 높은 성능을 나타냄을 확인하였다.

향후 연구 과제로는 단백질 기능 예측에 필요한 풍부한 데이터 셋에 대한 연구가 필요하며, 생물정보학의 전문가들과 협력을 통하여 표준화된 데이터 셋을 준비함으로써 계산모델의 객관적 평가 방법에 대한 연구의 병행이 요구된다.

참고문헌

[1] Stephen Oliver, " Guilt-by-Association goes global," Nature (2002), 601-603

[2] Benno Schwikowski, Peter Uetz, and Stanley Fields, " A network of protein-protein interactions in yeast," Nature Biotechnology, Vol.18, No.3, Dec.2000, 1257-1261

[3] Jie Cheng, Christos Hatzis, Hisashi Hayashi, Mark-A. Krogel, Shinichi Morishita, David Page, and Jun Sese, " KDD cup 2001 report," SIGKDD Explorations, 3, Jan. 2002, 47-64