제36회 한국정보처리학회 추계학술발표대회 논문집 제18권 제2호 (2011. 11)

단백질 상호작용 네트워크에서의 단백질 기능예측을 위한 패턴 마이닝*

김태욱*, 이미정*, 이패패*, 류근호*
*충북대학교 데이터베이스/바이오인포메틱스 연구실
{twkim, mjlee, khryu, lipeipei}@dblab.chungbuk.ac.kr

Prediction of Protein Function using Pattern Mining in Protein-Protein Interaction Network

Taewook Kim*, Meijing Li*, Peipei Li*, Keun Ho Ryu*
*Database/Bioinformatics Laboratory, Chungbuk National University

요 약

단백질 사이의 상호작용 네트워크(PPI network: Protein-Protein Interaction network)를 이용하여 단백질 기능을 예측 하는 것은 단백질 기능 예측 기법들 중에서 중요한 작용을 한다. 하지만 PPI를 이용한 단백질 기능 예측은 기능의 복잡도와 다양성으로 인해 제한적인 결과를 나타내 왔다. 따라서 본 논문에서는 기존의 연구들 보다 높은 정확도로 단백질 기능을 예측하기 위해 기능 예측을 하려는 단백질과 상호작용 하는 단백질들에 그래프 마이닝 기법을 적용하여 빈발 2-노드 상호작용 패턴을 찾고, 그 패턴을 이용하여 단백질 기능을 예측하는 접근법을 제안하였다. 실험데이터로 DIP(Database of Interacting Proteins)에서 제공하는 단백질 상호작용 데이터를 사용하였으며, 다른 기존의 단백질 기능 예측 기법들보다 높은 정확도를 보여주었다.

1. 서론

최근 의학, 생물학 등 다양한 분야에서 단백질 기능 예측이 새로운 연구 과제로 떠오르고 있다. 단백질의 기능을 예측하고 단백질간의 상호작용을 연구함으로써 질병을 예방하고 신약을 개발하는 것이 가능하기 때문에 그 중요성은 더욱 커지고 있다. 이러한 단백질의 기능을 알기 위해 수많은 생물학적 실험이 이루어지고 있지만 실험 대상 단백질의 수가 너무 많기 때문에 많은 비용과 시간이 소요된다. 따라서 시간과 비용을 줄이고 좀 더 정확한 단백질 기능을 예측하기 위한 수많은 연구가 진행되고 있다.[1] 단백질 상호작용 네트워크를 통한 단백질의 기능 예측은 단백질의 기능적 관계의 복잡성과 다양성으로 인해 도전적인 과제로 남아 있으며[2], 최근에는 단백질 상호작용 네트워크에서 그래프 기반 마이닝 방법과 통계학적 방법 등을 이용하여 단백질의 기능을 규명하기 위한 다양한 연구가 시도되고 있다.

기존의 단백질 상호작용 네트워크를 이용한 단백질 기능 예측 연구는 대부분 기능을 알지 못하는 단백질과 상호작용하는 인접 이웃 단백질의 기능 또는 그 경로를 통해 기능을 예측하였다[3]. Chuan Lin은 단백질 기능 예측을 위하여 기능이 알려지지 않은 단백질과 공통 이웃을 가지는 단백질을 이용하여 기능을 예측하였다[4]. 이는 기존의 상호작용하는 두 단백질은 비슷한 기능을 가진다는 개념을 확장하여 두 단백질간의 직접적인 상호작용이 없어도 공통 이웃을 가질 경우 비슷한 기능을 가진다는 이론에서 비롯된 것이다. 이와 비슷한 개념으로 Alexei Vazquez는 단백질의 기능을 분류하고, 기능을 알 수 없는 하나의 단백질과 상호작용하는 이웃 단백질들이 공통된 기능을 가진 경우 그 공통된 기능과 비슷한 기능을 가질 것이라 예측하였다[5]. 이밖에도 기능이 알려지지 않은 단백질과 상호작용하는 이웃 단백질을 이용한 방법들이 단백질 기능 예측을 위해 다양한 방법으로 시도되고 있다. 하지만 기존의 이런 방법들은 하나의 단백질이 다양한 기능을 갖는 경우, 또는 상호작용하는 두 개의 단백질이 서로 전혀 다른 기능을 가진 경우 예측된 기능과 실제 단백질의 기능이 전혀 달라질 수 있기 때문에 정확도에 있어 제한된 결과를 나타냈다. 이러한 문제점을 해결하기 위해 본 논문에서는 단백질 상호작용 네트워크상에 그래프 마이닝을 적용하여 빈발 패턴을 추출하였다. 그래프는 각 노드를 단백질로, 노드 사이의 간선은 두 단백질 사이의 상호작용을 표현하였다[6]. 단백질의 기능들을 분류하여 각 노드에 레이블 하였으며, 레이블을 이용하여 그래프 상에서 2개의 노드를 가진 빈발 패턴을 찾아 기능 예측을 실

* 이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원 (No. 한국연구 2011-0001044)과 질병관리본부 인체자원은행 프로젝트(4851-307)의 지원을 받아 수행된 연구임

행 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 단백질 기능의 분류를 위해 사용한 카테고리에 대한 설명과 그래프 마이닝을 적용하여 네트워크에서의 빈발 패턴을 추출하는 과정을 기술 하였다. 3장에서는 추출된 빈발 패턴을 이용하여 각 패턴의 정확도를 계산하고 가장 높은 정확도를 가진 패턴으로 기능을 예측하는 과정을 기술 하였으며 4장은 모델 평가를 위한 실험 및 결과 분석을 기술 하였다. 끝으로 5장에서는 이 논문의 결론을 맺는다.

2. 그래프 마이닝 기법 기반의 단백질 상호작용 빈발 패턴 추출

본 논문의 궁극적인 목표는 주어진 PPI 데이터를 이용하여 그래프 마이닝 기법을 적용하고 빈발 패턴을 추출하여 높은 정확도로 단백질의 기능을 예측하는 것이다. 또한 단백질 기능의 분류를 위해 MIPS에서 제공하는 카테고리를 사용하였다.

2.1 단백질 기능 카테고리

본 논문에서는 단백질 기능의 포괄적 분류를 위해 MIPS에서 제공하는 Funcat(Functional Catalogue Database) 카테고리를 사용하였다.[7] MIPS는 독일의 Max-Planck-Institute Bioinformatics 그룹에서 생성한 단백질 데이터베이스로 유전체 서열의 기능 분석 및 분류에 중점을 두고 있다[8]. Funcat은 총 27개의 카테고리를 사용하였으며 각 카테고리는 세부 카테고리로 나누어져 있다. 또한 다양한 생물의 단백질 기능을 이해하기 쉽게 서술하였다.

실험을 위해 DIP에서 제공하는 Saccharomyces cerevisiae 단백질 상호작용 데이터를 사용하였으며[9], 이를 위해 Funcat의 전체 카테고리 중 16개의 카테고리를 사용하였다. 또한 하나의 단백질이 다수의 기능을 가질 경우, 하나의 카테고리가 여러 단백질에 중복 되는 것을 허용하였다.

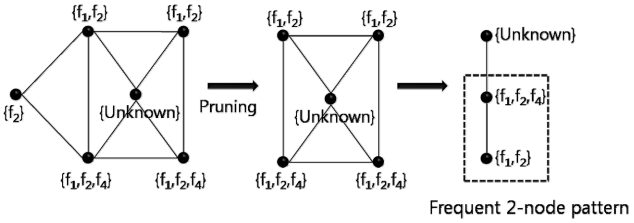
<표 1> Funcat 단백질 기능 카테고리

ID	Functional Catalogue
01	Metabolism
02	Energy
10	Cell Cycle and DNA Processing
11	Transcription
12	Protein Synthesis
14	Protein Fate
16	Protein with binding Function or Cofactor Requirment
18	Regulation of Metabolism and Protein Function
20	Cellular Transport
31	Cell Rescue
34	Interaction with the Environment
38	Transposable Elements, Viral and Plasmid Proteins
40	Cell Fate
41	Development
42	Biogenesis of Cellular Components
43	Cell Type Differentiation

2.2 빈발 패턴 추출

본 논문에서는 방향성과 가중치가 없는 그래프로 단백질 상호작용 네트워크를 표현하였다.[6] 그래프는 G(V,E)로 표현된다. $V(v_1...v_n)$는 각 단백질을 나타내는 노드의 집합이며, $E(e_1...e_n)$는 노드간의 연결, 즉 단백질간의 상호작용을 나타내는 간선들의 집합이다. 단백질 기능 카테고리 집합 $F(f_1...f_n)$은 각 노드 단백질이 가진 기능에 따라 레이블 하였으며 기능을 알지 못하는 단백질은 {Unknown}으로 레이블 하였다.

단백질 상호작용 네트워크 안에서 빈발 패턴을 추출하기 위해 aprior 알고리즘을 사용하였다[10.] 알고리즘은 그래프안의 기능을 알지 못하는 단백질과 상호작용하는 이웃단백질들의 레이블을 이용하여 모든 2-노드 패턴을 탐색하고, 사용자가 정한 최소 지지도 값을 넘지 못하는 패턴을 가지치기 단계에서 제거하여 빈발 패턴을 추출한다. 여기서 2-노드 패턴은 단백질 기능을 예측하려는 단백질 노드와 상호작용하는 두 이웃 단백질의 기능들을 말한다. 그림1은 최소 지지도를 2로 정했을 때 그래프 안에서 2-노드 빈발 패턴을 탐색하는 예를 보여준다. 기능을 알지 못하는 단백질 노드는 Unknown으로 레이블 하였으며, 그것의 이웃 단백질들은 각각의 단백질이 가진 기능에 따라 레이블 하였다.



2-node pattern	Frequency
{f1,f2}-{f2}	1
{f1,f2}-{f1,f2}	1
{f1,f2,f4}-{f2}	1
{f1,f2,f4}-{f1,f2}	2
{f1,f2,f4}-{f1, f2,f4}	1
Minimum support : 2	

(그림 1) 2-노드 빈발패턴 추출

3. 단백질 기능 예측

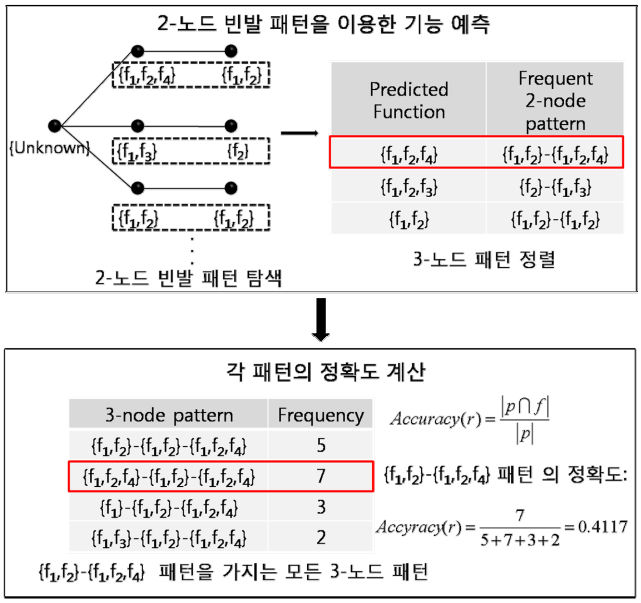
기능을 알지 못하는 단백질의 기능을 예측하기 위해, 위 단계에서 추출한 2-노드 빈발 패턴들을 검색하고, 패턴들을 이용하여 기능을 예측하였다. 여기서 2-노드 빈발 패턴은 알고리즘에 의해 발견된 패턴 중 최소 지지도와 같거나 그 이상의 빈도수를 가진 패턴을 말한다. 단백질 기능 예측 단계는 다음과 같다.

- 각각의 빈발 패턴이 클래스 예측을 위해 가지는 정확도 (Accuracy)를 계산한다.

$$Accuracy(r) = \frac{|p \bigcap f|}{|p|} \quad \text{식 (3.1)}$$

여기서 r은 단백질 패턴으로부터 단백질 기능 레이블의 rule을 말하고, p는 예측하려는 단백질 P와 상호작용하는 2-노드 빈발 패턴을 말하며, f는 단백질 기능 레이블을 말한다.

- 각 후보 패턴을 패턴이 가지는 정확도에 따라 내림차순으로 정렬하여 정확도가 가장 높은 패턴이 가지는 기능을 Unknown의 기능으로 예측한다.



(그림 2) 패턴의 정확도를 이용한 기능예측

(그림 2)는 발견된 2-노드 빈발 패턴을 이용하여 패턴 각각의 정확도를 계산하는 과정의 예를 보여준다. 먼저 Unknown으로 레이블된 노드와 상호작용하는 2-노드 빈발 패턴을 이용하여 Unknown의 기능을 예측한다. 예측된 기능에 따라 3-노드 패턴을 정렬하고, 전체 네트워크에서 예측된 3-노드 패턴의 빈발도를 카운팅 하여 정확도를 계산한다. 예측된 3-노드 패턴 중 가장 정확도가 높은 패턴을 이용하여 단백질의 기능을 예측한다. (그림2)에서는 2-노드 빈발 패턴 $\{f_1,f_2\}$-$\{f_1,f_2,f_4\}$을 이용하여 기능을 $\{f_1,f_2,f_4\}$으로 예측하고, 전체 네트워크에서 $\{f_1,f_2\}$-$\{f_1,f_2,f_4\}$ 패턴을 포함하는 3-노드 패턴을 찾아 그 중 예측된 $\{f_1,f_2,f_4\}$-$\{f_1,f_2\}$-$\{f_1,f_2,f_4\}$패턴의 비율을 계산하여 정확도를 측정한다.

4. 실험 평가

4.1 데이터

논문에서는 DIP(Database of Interacting Proteins)에서 제공하는 단백질 상호작용 데이터를 사용하였다. DIP는 생물학적 실험을 통해서 얻어진 단백질 상호작용을 데이터를 가진 대표적인 웹사이트 중 하나이다. 우리는 DIP가 가지는 몇몇 생물 종의 데이터 중 가장 많은 데이터를 가진 Saccharomyces cerevisiae 데이터를 사용하였다. DIP에서 제공하는 Saccharomyces cerevisiae 데이터는 1274개의 단백질 노드와 3222개의 상호작용을 포함한다.

4.2 기존 단백질 기능 예측 기법과의 비교

본 논문에서 제안한 패턴 마이닝 기법의 성능을 평가하기 위해 기존의 Neighbor Counting 기법과 정확도를 비교하였다. 정확한 비교를 위해 본 논문에서 사용한 Saccharomyces cerevisiae 데이터를 적용하였다. 정확도는 단백질 기능이 정확히 일치하는 경우와 부분적으로 일치하는 경우를 구분하여 <표 2>에 나타내었다. 두 경우 모두에서 Pattern-Mining Method가 Neighbor Counting Method 보다 높게 측정되어 더 나은 성능을 보여주었다.

<표 2> Pattern-based Method 와 Neighbor Counting Method의 정확도 비교

Method	Exact match	Inclusive match
Pattern-Mining Method	0.688	0.820
Neighbor Counting Method	0.440	0.758

5. 결론

본 논문에서는 단백질 상호작용 네트워크를 이용하여 단백질 기능을 예측하는데 중점을 두었다. 기존의 방법과 달리 그래프 마이닝 기법을 적용하여 기능을 모르는 단백질과 상호작용하는 빈발한 2노드-패턴을 찾고, 발견된 패턴들이 가지는 정확도를 계산하여 가장 높은 정확도를 가지는 패턴의 기능을 예측에 사용하였다. 본 논문에서 제안한 방법을 평가하기 위해 DIP에서 제공하는 Saccharomyces cerevisiae 데이터를 사용하여 정확도를 측정하였다. 정확도는 예측된 단백질 기능이 실제 단백질 기능과 정확히 일치하는 경우와 부분적로 일치하는 경우를 구분하였으며, 각각 0.688과 0.820의 정확도를 나타내었다.

참고문헌

[1] Hee-Jeong Jin, J.H Yoon and H.G Cho, "An analysis System for Protein-Protein Interaction Data Based on Graph Theory" Korean Institute of Information Scientists and Engineers vol 33, p267, 2006

[2] Y.-R. Cho, W. Hwang, M. Ramanathan, and A. Zhang, "Semantic integration to identify overlapping functional modules in protein interactionnetworks," BMC Bioinf., vol. 8, p. 265, 2007.

[3] Y. Cho and A. Zhang, "Predicting protein funtion by frequent functional association pattern mining in protein interaction networks", Vol. 14, No. 1, pp. 30-36,

January 2010
[4] C. Lin, D. Jiang, and A. Zhang. Prediction of protein function using common-neighbors in protein-protein interaction networks. In Proc. IEEE Symposium on BIBE 2006.
[5] Vazquez,A., Flammini,A., Maritan,A. and Vespignani,A. (2003) Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol., 21, 697±700
[6] P. Lee, C. Huang, J. Fang, J.J.P. Tsai and K. Ng, "Study of the protein-protein interaction networks via random graph approach", Fourth IEEE International Conference on Cognitive Informatics, pp.110-119, July 2005
[7] A. Ruepp, A. Zollner, D. Maier, K. Albermann, J. Hani, M. Mokrejs, I. Tetko, U. Guldener, G. Mannhanupt, M. Munsterkotter and H.W. Mewes, "The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes", Nucleic Acids Research, Vol. 32, No. 18, pp. 5539-5545, 2004
[8] H.W. Mewes, S. Dietmann, D. Frishman, R. Gregory, G. Mannhaupt, K.F. X. Mayer, M. M¨unsterk¨otter, A. Ruepp, M. Spannagl, V. St¨umpflen, and T. Rattei, "MIPS: Analysis and annotation of genome information in 2007." Nucleic Acid Res., vol. 36, pp. D196 - -D201, 2008.
[9] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie and D. Eisenberg, "The database of interacting proteins: 2004 update", Nucleic Acids Research, Vol. 32, pp. D449-D451, 2004
[10] R. Agrawal and R. Srikant, ""Fast algorithm for mining association rules,""in Proc. 20th Int. Conf. Very Large Databases (VLDB), 1994, pp. 487 - -499.
[11] B. Liu, Y. Ma and C.K. Wong, Improving an association rule based classifier Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, Lyon, France (September 2000), pp. 504 - 509.