단백질 상호작용 네트워크에서 구조적 특징과 필수 단백질의 연관성 분석

An Analysis of Association for Essential Proteins in Protein-Protein Interaction Network

강태호, 류제운*, 이윤경*, 여명호, 정영수*, 권미형*, 유재수, 김학용*
충북대학교 정보통신공학과, 충북대학교 생화학과*

Kang tae-ho, Ryu jae-woon*, Lee yoon-kyoung*, Yeo myung-ho, Jung young-su*, Kwon mi-hyeong*, Yoo jae-soo, Kim hak-yong*
Department of Computer and Communication Engineering Chungbuk National University, Department of Biochemistry Chungbuk National University*

요약

단백질 상호작용 네트워크는 허브(Hub)라 할 수 있는 상호작용 수가 많은 소수의 단백질과 상호작용 수가 적은 다수의 단백질들로 구성된다. 최근 들어 여러 연구들에서 허브 단백질이 비 허브(Non-hub) 단백질보다 상호작용 네트워크에 필수적인 단백질일 가능성이 높다고 설명하고 있다. 이러한 현상을 중심-치명 룰(Centrality-lethality Rule)이라 하는데, 이는 복잡계 네트워크에서 허브단백질의 중요성 및 네트워크 구조의 중요성을 설명하기 위한 방법으로 폭넓게 신뢰받고 있다. 이에 본 논문에서는 중심-치명 룰이 항상 옳게 적용되는지를 확인하기 위해 Uetz, Ito, MIPS, DIP, SGD, BioGRID와 같은 효모에 관한 공개된 모든 단백질 상호작용 데이터베이스들을 분석하였다. 흥미롭게도, 상호작용 데이터가 적은 데이터베이스들(Uetz, Ito, DIP)에서는 중심-치명 룰을 잘 나타냈지만 상호작용 데이터가 대용량인 데이터베이스들(SGD, BioGRID)에서는 중심-치명 룰이 잘 맞지 않음을 확인하였다. 이에 따라 SGD와 BioGRID 데이터베이스로 부터 얻은 상호작용 네트워크의 특징을 분석하고 DIP 데이터베이스의 상호작용 네트워크와 비교해보았다.

Abstract

The protein interaction network contains a small number of highly connected protein, denoted hub and many destitutely connected proteins. Recently, several studies described that a hub protein is more likely to be essential than a non-hub protein. This phenomenon called as the centrality-lethality rule. This rule is widely credited to exhibit the importance of hub proteins in the complex network and the significance of network architecture as well. To confirm whether the rule is accurate, we investigated all protein interaction DBs of yeast in the public sites such as Uetz, Ito, MIPS, DIP, SGD, and BioGRID. Interestingly, the protein network shows that the rule is correct in lower scale DBs (e.g., Uetz, Ito, and DIP) but is not correct in higher scale DBs (e.g., SGD and BioGRID). We are now analyzing the features of networks obtained from the SGD and BioGRD and comparing those of network from the DIP.

I. 서 론

단백질은 인체를 이루고 있는 세포 구성은 물론 생명현상을 주관하는 각종 화학반응들에 단백질이 관여하고 있다. 단백질은 일정한 질서에 따라 서로 조립되기도 하고, 기능적으로 연관돼 네트워크를 이루고 있다. 이 네트워크는 생명체의 상황에 따라 분자들 간의 조합과 연관 관계를 달리하여 유동적으로 생명체를 구성한다. 단백질 분자는 각각 서로 다른 분자를 인지할 수 있고, 적절한 상황에서 서로 결합해 공동 작업을 펼친다. 따라서 이들이 어떻게 생명체의 역동적인 기능을 하게 되는지, 이들이 질병과는 어떻게 관련돼 있는지를 밝히려면 단백질 사이의 상호작용 네트워크를 규명해야 한다.

단백질이 이루는 네트워크는 현재로서는 극히 일부의 연관성에 대한 정보만 밝혀져 있어 정확한 모습을 이야기하기는 어렵다. 그렇지만 현재까지의 연구결과를 정리해보면, 단백질 사이의 분자 네트워크는 불균일한 모습이며 네트워크가 척도 없는(Scale-free) 네트워크로서 다른 많은 복잡계 네트워크와 공통된 성질을 보인다[1]. 척도 없는 네트워크는 링크수가 많은 소수의 허브 노드와 링크수가 적은 다수의 노드들이 공존하는 구조를 취하고 있다. 네트워크에서 일반적으로 다른 단백질에 비해 상호작용이 많은 단백질을 허브(Hub) 단백질이라

한다. 허브 단백질은 네트워크에서 경로를 단축시키는데 큰 역할을 함으로서 중요한 역할을 담당한다고 추측할 수 있다. 하지만 허브 단백질의 높은 접근성은 네트워크에서 지름길 역할을 하는 동시에 네트워크의 견고성에 대해서는 약점이 된다. 네트워크의 견고성은 네트워크가 공격을 받거나 일부분의 기능정지가 발생하더라도 온전하게 기능을 수행할 수 있는 능력을 의미하며 이를 측정하는 대표적인 방법으로 위상 견고성(Topological Robustness)이 있다[4, 5].

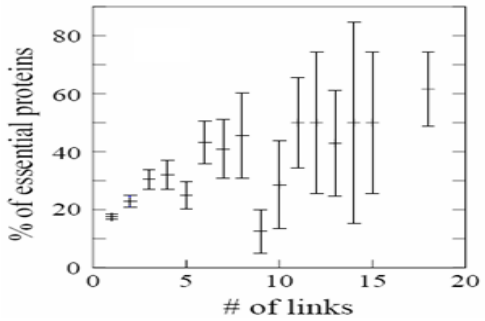
세포의 생존에 중요하게 관여하는 단백질을 필수(essential) 단백질 혹은 치사(lethal) 단백질이라 한다[4]. 필수 단백질은 세포의 성장속도에 관여하거나 특정 항생제에 감수성에 영향을 주는 단백질이다. 그리고 치사 단백질은 필수 단백질과 그 의미는 유사하나, 이 단백질을 시스템적으로 제거했을 때 세포의 생존 유무를 결정하는 단백질을 말한다. 그렇지만 기존 논문들이 필수 단백질과 치사 단백질을 대부분 같은 의미로 사용하고 있다. 따라서 본 논문에서도 치사 단백질과 필수 단백질의 같은 의미로 사용한다.

네트워크에서 허브 단백질은 필수 단백질일 가능성이 높다는 것을 쉽게 추측할 수 있다. 하지만 모든 허브 단백질이 필수 단백질인 것은 아니다. 다만 허브 단백질이 비허브 단백질보다 필수 단백질일 확률이 높다는 것이다. 허브 단백질이면서 필수 단백질은 생명체 내에서 아주 중요한 역할을 하는 것 중 하나이다. 이처럼 시스템생물학자들은 네트워크의 위상기하학적 특성과 단백질 기능간의 연관성을 찾기 위해, 허브 단백질 연구에 초점을 두고 있다. 따라서 본 논문에서는 단백질 상호작용 네트워크를 통해 허브 단백질과 필수 단백질 사이의 상관관계를 밝히고자 한다.

II. 관련 연구

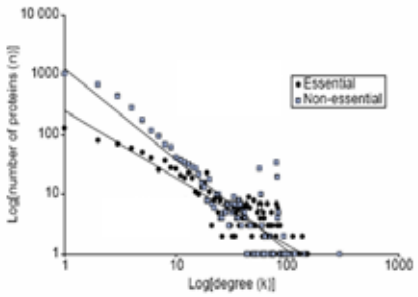
효모의 단백질 상호작용 네트워크는 척도없는 네트워크이다. 척도 없는 네트워크는 많은 수의 노드를 무작위로 제거했을 때에도 노드들 간의 연결성(connectivity)은 대체로 유지가 잘되며 전체 네트워크 모양이 크게 깨지지 않는다. 네트워크 구조가 매우 견고하다. 그러나 허브 단백질이 제거되었을 경우에는 PPI 네트워크의 평균 단계(Average degree) 는 감소하고 지름(diameter) 은 급속히 증가한다. 이는 허브 단백질이 생존에 많은 영향을 하는 중요한 기능을 할 것이라 쉽게 이해할 수 있다.

필수 단백질은 일반적으로 세포에서 특정 유전자를 돌연변이(mutation) 시키거나 제거 했을 때 세포가 살아있는지의 유무를 통해 구분한다[2, 3]. 그리고 기존연구들에서는 세포 생존에 중요한 역할을 하는 필수 단백질은 허브 단백질과 밀접한 관계를 가질 것이라 여기고 있다 [6, 7, 8]. 이를 증명해 주는 결과는 그림 1과 같다[4]. 그림 1은 링크수가 증가함에 따라 포함하는 필수 단백질의 수를 백분율로 나타낸 것인데, 링크수가 많은 단백질들이 필수 단백질일 확률이 크다는 것을 확인할 수 있다. 그러나 실제 링크수가 많은 단백질은 그 수가 비교적 매우 적기 때문에 그 중 일부가 필수 단백질일 경우에는 높은 비율을 차지할 수밖에 없다는 문제가 있다.



▶▶ 그림 1. 링크수별 필수 단백질 분포

Yu방식에서는[6] 단백질들을 필수 단백질과 비필수 단백질로 구분하고 모듈 내에서 링크수가 증가함에 따라 이에 속하는 단백질의 수로 그래프를 그린다. 그 결과는 그림 2와 같다. 링크수가 작을 때에는 비필수 단백질의 수가 필수 단백질 수보다 많지만 링크수가 점점 증가함에 따라 필수 단백질의 수가 역전하는 것을 볼 수가 있다. 이를 통해 두 그룹의 상호작용이 모두 멱함수 분포를 가지고 있으며, 상호작용의 수가 많은 단백질에는 비필수 단백질보다는 필수 단백질의 수가 많은 것을 확인하였다. 기존에는 허브를 정의할 수 있는 기준이 사실 명확하지 않았으나, Yu 방법은 수식적으로 허브에 대한 정의를 가능하게 했다. 즉, 링크수가 증가함에 따라 필수 단백질의 수가 비필수 단백질의 수보다 많아지는 그 순간 이후를 허브 단백질이라 정의하고 있다.



▶▶ 그림 2. 필수 단백질과 비필수 단백질의 링크수별 필수단백질 분포(로그비율)

III. 연구 결과

본 논문에서는 효모의 단백질 상호작용 정보를 제공해주는 여러 데이터베이스를 분석하여 하나의 데이터베이스로부터 잘못 이해할 수 있는 네트워크의 구조들을 비교·분석해 보고 어느 데이터베이스가 네트워크를 구축하는데 있어 정확한 데이터를 제공해주는지를 분석하였다. 이를 위해 Y2H로 얻은 실험 데이터 Uetz 데이터베이스[9], Ito 데이터베이스[10]와 여러 실험 데이터들을 통합한 MIPS 데이터베이스[11], DIP 데이터베이스[12] 그리고 알려진 문헌에 대해 텍스트마이닝(Text-mining)을 하고 정제 과정을 거친 SGD 데이터베이스[13], BioGRID 데이터베이스[14]를 분석하였다.

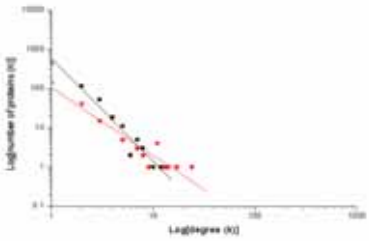
필수 단백질의 정보는 MIPS와 SGD에서 제공하는 데이터를 추출하여 사용하였다. 그리고 각각의 데이터베이스에서 추출한 상호작용 데이터는 정확성을 높이기 위해 내포하고 있는 중복 데이터들을 제거하고, 단백질의 세포내 위치 정보를 이용한 정제과정을 통해 False-positive 데이터를 제거하였다. 이들 각 데이터베이스에서 추출한 단백질수와 상호작용의 수는 표 1과 같다. 표 1에서의 허브단백질 기준은 그림 3과 같이 Yu방법[6]을 이용하여 얻은 결과이다.

[표 1] 데이터 요약

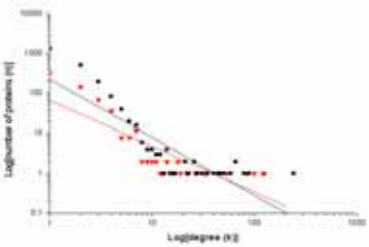
	단백질	상호작용	허브 기준
Uetz	1,032	967	7
Ito	3,187	4,193	54
MIPS	4,469	11,912	1,123
DIP	4,870	16,604	22
SGD	5,154	69,109	195
GRID	5,170	69,196	181

허브 단백질에 대한 정의 기준은 데이터베이스마다 다르다. MIPS 데이터베이스는 허브 기준이 1,123개로 이는 실제로 존재하지 않는다. 따라서 MIPS는 허브를 정의할 수 없었다. 특이한 것은 MIPS를 제외한 데이터베이스에서는 상호작용의 수가 증가한 만큼 허브 기준도 증가함을 확인할 수 있다(그림 3). 그리고 각 데이터베이스마다 다르게 정의된 허브를 가지고 필수 단백질의 분포가 허브 단백질에 주로 분포하는지를 살펴보았다. 결과적으로 비교적 상호작용 데이터가 적은 Uetz , Ito, DIP 데이터베이스는 허브 단백질에서 필수 단백질을 많이 포함하고 있는 반면, 상호작용 데이터가 다른 데이터베이스들에 비해 급격하게 증가한 SGD, BioGRID 데이터베이스는 필수 단백질이 허브에 오히려 적게 포함되어 있음을 확인하였다(그림 4)

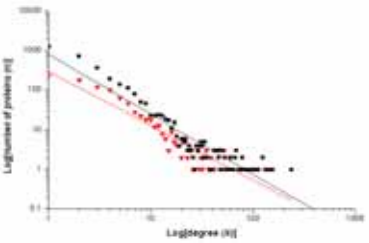
A. Filtered Ito



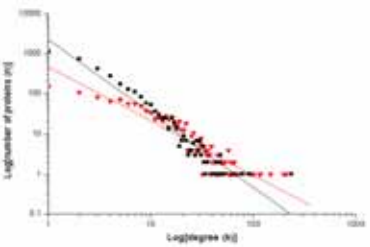
B. Filtered Uetz



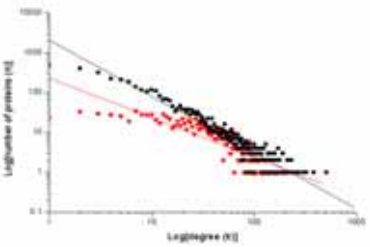
C. Filtered MIPS



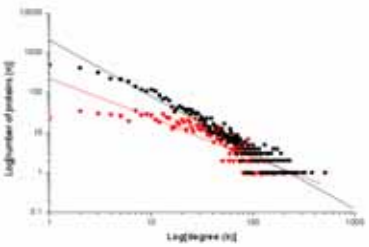
D. Filtered DIP



E. Filtered SGD

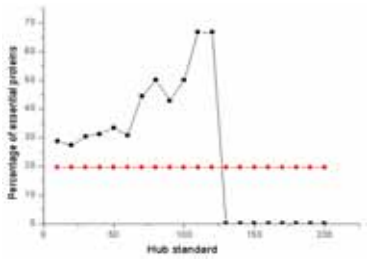


F. Filtered SGID

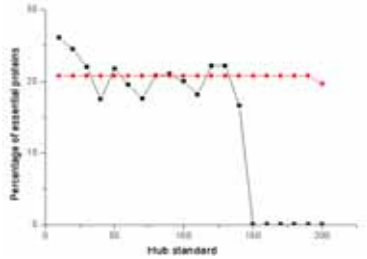


▶▶ 그림 3. 데이터베이스 구조 및 허브

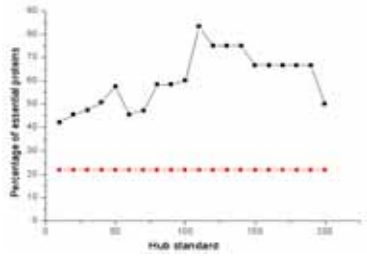
A. Filtered Ito



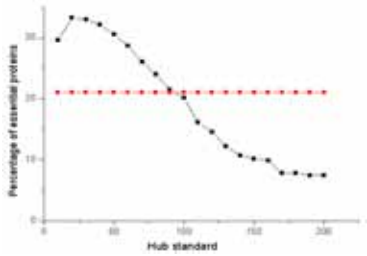
B. Filtered MIPS



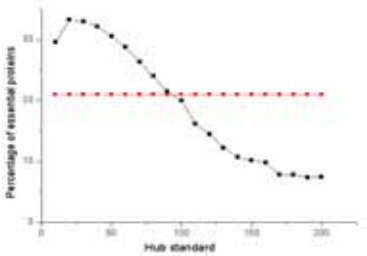
C. Filtered DIP



D. Filtered SGD



E. Filtered GRID



▶▶ 그림 4. 허브기준에 따른 필수 단백질의 분포

IV. 결 론

본 논문에서는 효모의 PPI 정보를 제공해주는 Uetz, Ito, MIPS, DIP, SGD, BioGRID 등의 많은 데이터베이스들이 제공하는 PPI 데이터를 이용해 네트워크의 허브와 필수 단백질의 연관성을 분석하였다. 이를 통해 데이터베이스가 비교적 작은 경우 이전연구들의 결과와 같이 상호작용 수가 많은 단백질 즉 허브 단백질에 필수 단백질이 많이 포함되어있는 것을 확인하였다. 반면 SGD, BioGRID와 같이 PPI 데이터 크기가 큰 데이터베이스의 경우에는 허브단백질과 필수단백질 사이의 관계가 이전 결과들과 달라짐을 확인하였다. 결과적으로 고효율성 기법을 통해 대량의 데이터가 추출되고있는 현 시점에서는 이전 연구들에서 제시된 허브 단백질과 필수단백질 사이의 관계가 재조명 되어야 할 필요가 있다.

▌참 고 문 헌▐

[1] Ravasz E, Barabasi AL . Hierarchical organization in complex networks. Phys. Rev. E67, 026122(2003)
[2] Macdonald PR, et al. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature 402, 413 - 418(1999)
[3] Winzeler EA, et al. . Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 285, 901 - 906(1999)
[4] Jeong H, Mason SP, Barabasi AL, Oltvai ZN . Lethality and centrality in protein networks. Nature 411, 41(2001)
[5] Proulx SR, Nuzhdin S, Promislow DE. Direct selection on genetic robustness revealed in the yeast transcriptome. PLos ONE. 2(9), e911(2007)
[6] Yu H, Greenbaum D, Lu HX, Zhu X, Gerstein M. Genomic analysis of essentiality within protein networks. TRENDS in Genomics 20, 227-231(2004)
[7] He X, Zhang J. Why Do Hubs Tend to Be Essential in Protein Networks? PLos Genet. 2:6, 0826-0834 (2006)
[8] Aragues R, Sali A, Bonet J, Marti-Renom MA, Oliva B. Characterization of protein hubs by inferring interacting motifs from protein interactions. PLoS Comput Biol. 3(9), 1761-71(2007)
[9] Uetz P, et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623-627(2001)
[10] Ito T, Chiba T, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome PNAS 98, 4569-4574(2001)
[11] Gűldener U, et al. MPact: the MIPS protein interaction resorce on yeast. Nucleic Acids Research 34, D436-D441 (2006)
[12] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of interacting proteins: 2004 update. Nucleic Acids Res., 32, D449 - D451(2004)
[13] Hirschman JE, et al.. Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. Nucleic Acids Res. 34, D442-445 (2006)
[14] Stark C, et al. . BioGRID: a general repository for interaction datasets. Nucleic Acids Res. 34, D535-D539 (2006)