단백질 상호작용 네트워크에서 필수 단백질의 견고성 분석

Analysis of Essential Proteins in Protein-Protein Interaction Networks

류제운*, 강태호, 유재수**, 김학용***

충북대학교 생화학과*, 충북대학교 정보통신공학과**

Jae Woon Ryu(jwryu@kribb.re.kr)*, Tae-Ho Kang(thkang@netdb.cbnu.ac.kr)**,

Jae-Soo Yoo(yjs@chungbuk.ac.kr)**, Hak Yong Kim(hykim@chungbuk.ac.kr)*

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 보고되고 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 비교하였다.

■ 중심어 : | 단백질-단백질 상호작용 네트워크 | 필수 단백질 | 허브 단백질 |

Abstract

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 a 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.

■ keyword : | Protein-Protein Interaction Network | Essential Protein | Hub Protein |

* 본 논문은 2006학년도 충북대학교 학술연구지원사업의 연구비지원에 의하여 연구되었습니다.

접수번호 : #080603-002

접수일자 : 2008년 06월 03일

심사완료일 : 2008년 06월 10일

교신저자 : 김학용, e-mail : hykim@cbnu.ac.kr

I. 서 론

단백질은 세포를 구성하는 거대분자들 중의 하나로 생체 화학반응을 포함한 생명현상을 조절하고 수행하는 역할을 하고 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 척도없는(scale-free) 네트워크로서 다른 많은 복잡계 네트워크와 공통된 성질을 보인다[1]. 척도없는 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[2]. 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[3]. 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 논문에서는 치사 단백질과 필수 단백질을 같은 의미로 사용하되 필수 단백질로 통일시켜 사용하기로 한다.

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[6]. 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[4]. 척도 없는 네트워크는 많은 수의 노드를 무작위로 제거했을 때에도 노드들 간의 연결성(connectivity)은 대체로 유지가 잘되며 전체 네트워크 모양이 크게 깨지지 않는다. 다시 말해, 네트워크 구조가 매우 견고하다. 그러나 허브 단백질이 제거되었을 경우에는 단백질 네트워크의 평균 단계(average degree) 는 감소하고 지름(diameter) 은 급속히 증가한다. 이는 허브 단백질이 생존에 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[7][8]. 그리고 기존연구들에서는 세포 생존에 중요한 역할을 하는 필수 단백질은 허브 단백질과 밀접한 관계를 가질 것이라 여기고 있다[6][9][10]. Jeong은[4] 링크수가 증가함에 따라 포함하는 필수 단백질의 수를 백분율로 나타내면서 링크

수가 많은 단백질들이 필수 단백질일 확률이 크다는 것을 밝혔다. 그러나 실제 링크수가 많은 단백질은 그 수가 비교적 매우 적기 때문에 그 중 일부가 필수 단백질일 경우에는 높은 비율을 차지할 수밖에 없다는 문제가 있다.

Yu은[6] 단백질들을 필수 단백질과 비필수 단백질로 구분하고 모듈 내에서 링크수가 증가함에 따라 이에 속하는 단백질의 수로 그래프를 그린다. 그 결과 링크수가 작을 때에는 비필수 단백질의 수가 필수 단백질 수보다 많지만 링크수가 점점 증가함에 따라 필수 단백질의 수가 역전되는 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.

III. 연구 결과

본 연구에서는 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 yeast two hybrid (Y2H) 방법으로 얻은 실험 데이터로부터 구축한 Uetz 데이터베이스[11], Ito 데이터베이스[12]와 여러 실험 데이터들을 통합한 MIPS 데이터베이스[13], DIP 데이터베이스[14] 그리고 알려진 문헌에 대해 텍스트마이닝(text-mining)을 하고 정제 과정을 거친 SGD 데이터베이스[15], BioGRID 데이터베이스[16]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[17]. 이들 각 데이터베이스에서 추출한 단백질수와 상호작용의 수는 [표 1]과 같다. [표 1]에서의 허브단백질 기준은 Yu방법[6]을 이용하여 얻은 결과이다.

표 1. 데이터 요약

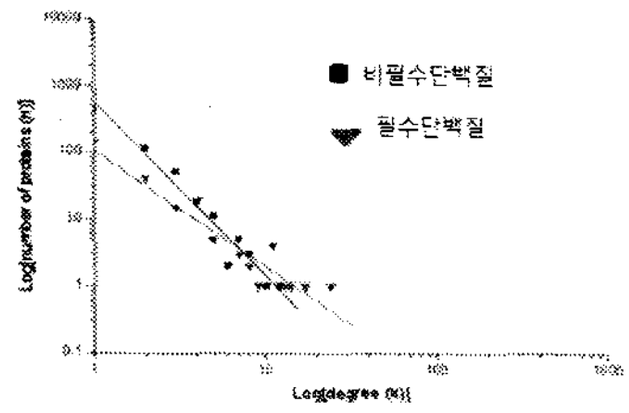
	단백질	상호작용	허브 기준
Uetz	1,032	967	7
Ito	3,187	4,193	54
MIPS	4,469	11,912	1,123
DIP	4,870	16,604	22
SGD	5,154	69,109	195
GRID	5,170	69,196	181

허브 단백질에 대한 정의 기준은 데이터베이스마다 다르다. 일반적으로 MIPS를 제외한 데이터베이스에서는 상호작용의 수가 증가한 만큼 허브 기준치도 증가함을 확인할 수 있다[표 1]. MIPS의 경우는 hub의 기준이 1,123으로 높게 나왔는데, 이는 실재 효모 단백질에는 1,123개의 링크를 가지고 있는 단백질은 존재하지 않으며 이론에 의한 추정치이기 때문에 허브 기준을 정의할 수 없다. 따라서 MIPS는 허브단백질의 견고성이나 필수성을 연구하기에는 적절한 데이터베이스라고 할 수 없기 때문에 앞으로의 연구에서는 MIPS로부터 얻은 연구 결과에 대해서는 언급하지 않기로 한다[표 1][그림 1D].

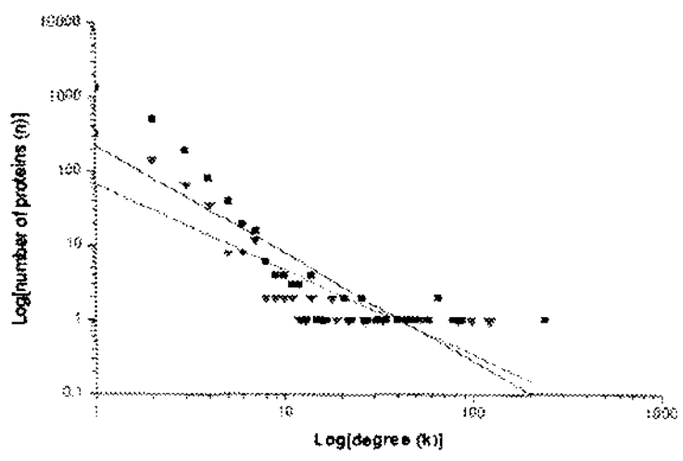
Y2H실험으로 얻은 Uetz와 Ito 데이터베이스는 상호작용수가 비교적 적으며 허브단백질이 필수단백질이라는 다른 연구결과와 비교적 잘 일치하고 있다[그림 1A]와 [그림 1B]. 여러 실험결과를 통합한 DIP 데이터베이스는 상호작용수가 Ito의 결과에 비해 약 4배 정도 증가하였는데, 이 역시 허브 단백질이 필수단백질이라는 가설이 성립되는 것으로 나타났다[그림 1C]. 그러나 텍스터마이닝을 하고 정제한 SGD 데이터베이스와 BioGRID 데이터베이스는 허브 기준이 각각 195와 181로 나타났는데, 실재 허브 단백질이 필수 단백질이 되는 경향성을 나타내지 못했다[그림 1E]와 [그림 1F],

[그림 2D]와 [그림 2E].

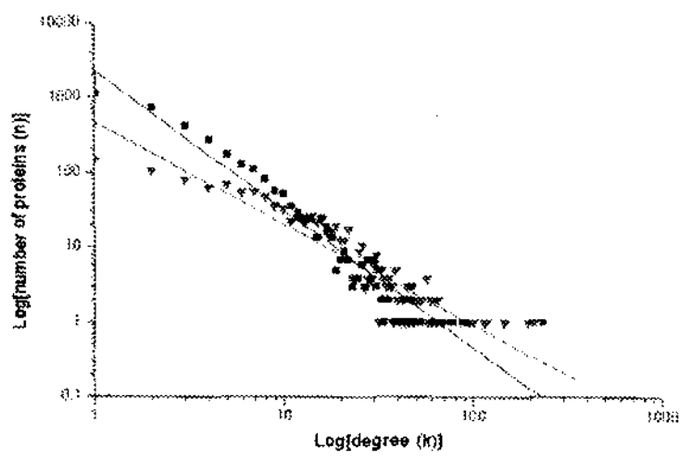
A. Filtered Uetz DB



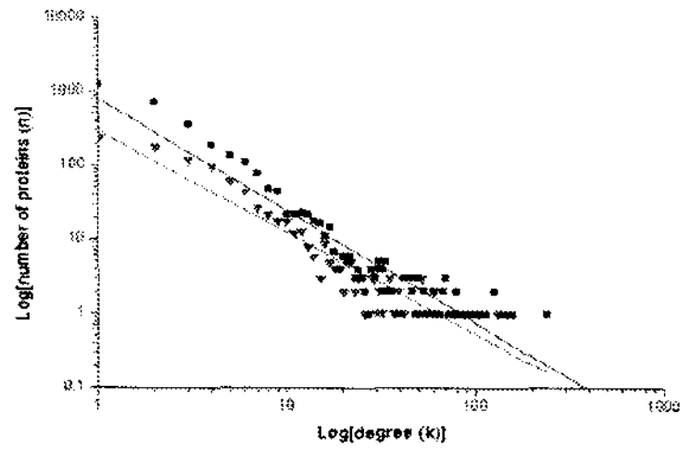
B. Filtered Ito DB



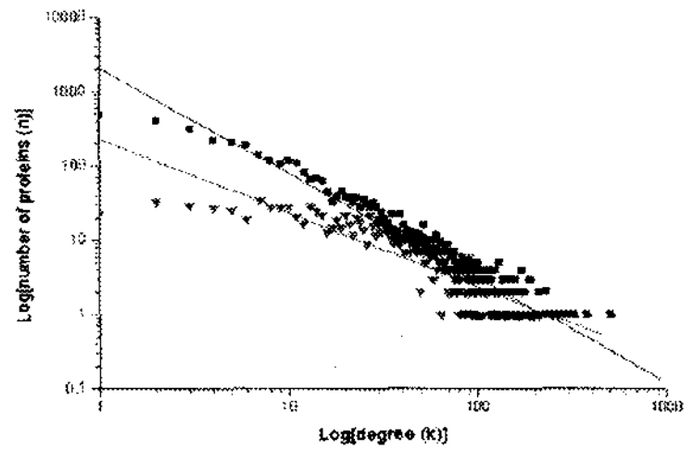
C. Filtered DIP DB



D. Filtered MIPS DB



E. Filtered SGD DB



F. Filtered BioGRID DB

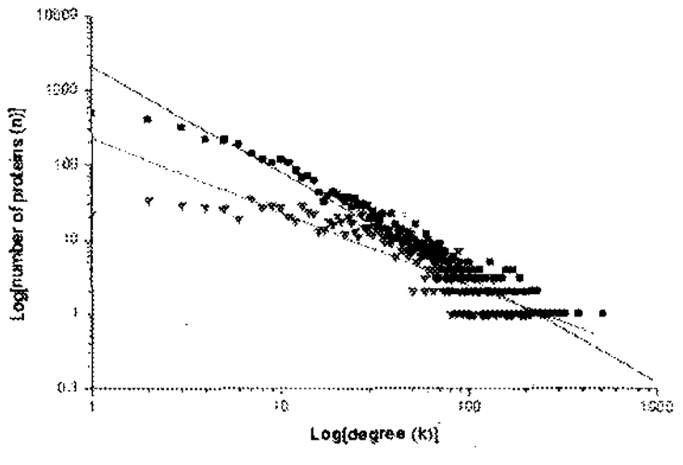
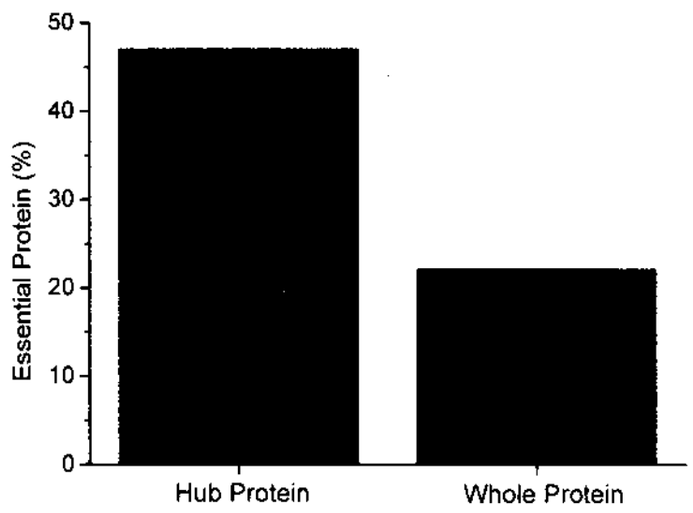


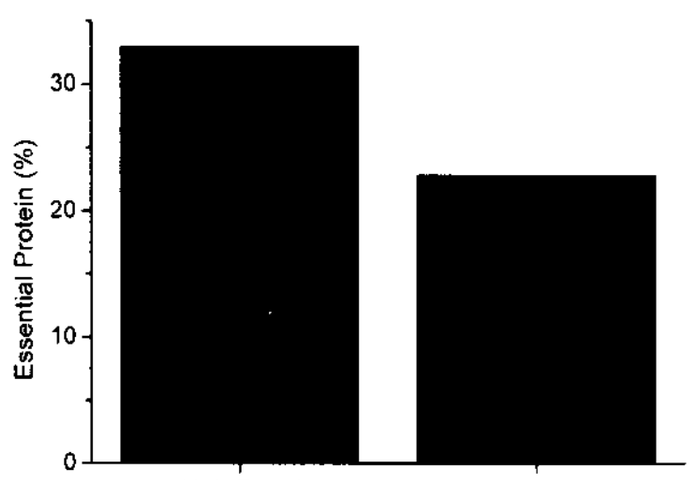
그림 1. 단백질 상호작용 데이터베이스의 멱함수 분포와 허브

각 데이터베이스마다 수치가 다른 허브 기준을 가지고 필수 단백질이 허브 단백질에 주로 분포되어 있는지를 살펴보았다. Uetz 및 DIP 데이터베이스는 허브 단백질의 약 47%가 필수 단백질이며 전체 단백질의 약 22%가 필수 단백질인 것과 비교해볼 때 허브 단백질이 필수 단백질인 경향이 높게 나타났다 [그림 2A]와[그림 2C]. Ito 데이터베이스는 다소 낮은 허브 단백질의 약 33%가 필수 단백질이다[그림 2B]. 이에 반해 SGD와 BioGRID 데이터베이스는 전체 단백질의 약 22%가 필수 단백질이었으며 허브 단백질의 약 8%만이 필수 단백질인 것으로 나타났다. 비교적 Uetz, Ito, 및 DIP 데이터베이스의 결과는 허브 단백질이 필수 단백질일 경향성이 더 높다는 기존에 보고된 결과와 일치하는데 반해, SGD와 BioGRID의 경우는 반대로 나타났다. 이는 데이터베이스 구축 과정에서 택스터마이닝을 통해 대용량이 축적되면서 단백질 수의 증가에 비해 단백질-단백질 상호작용 수가 급격히 증가한 결과로 풀이 된다 [표 1].

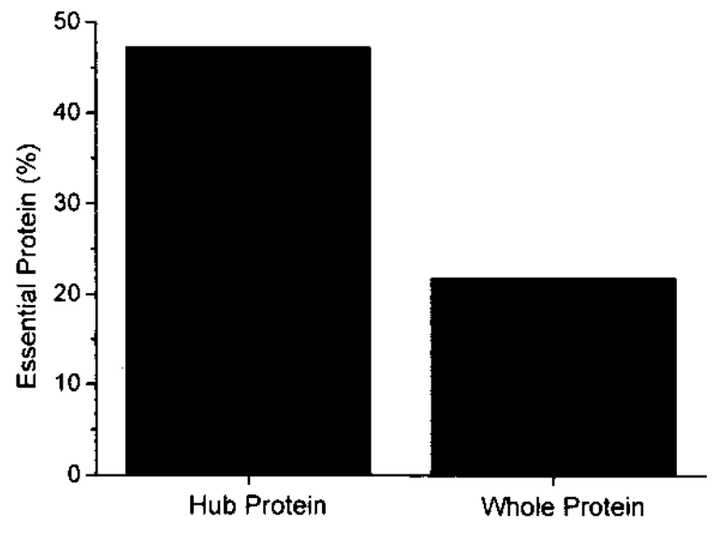
A. Filtered Uetz DB



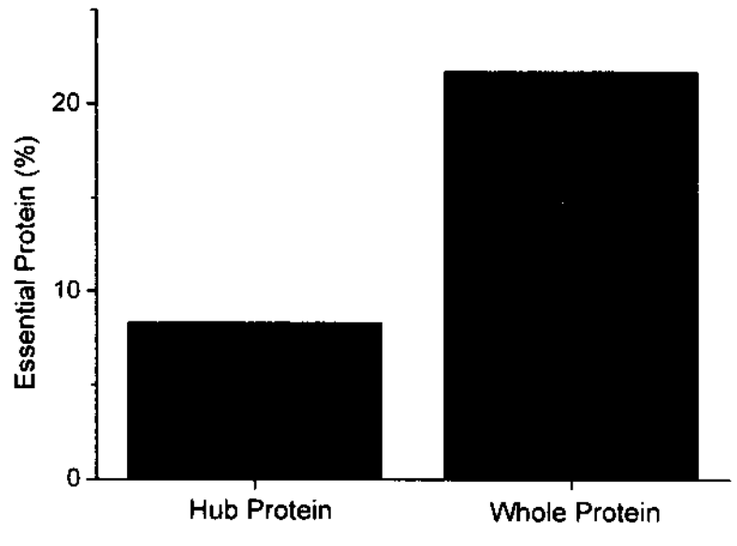
B. Filtered Ito DB



C. Filtered DIP DB



D. SGD



E. BioGRID

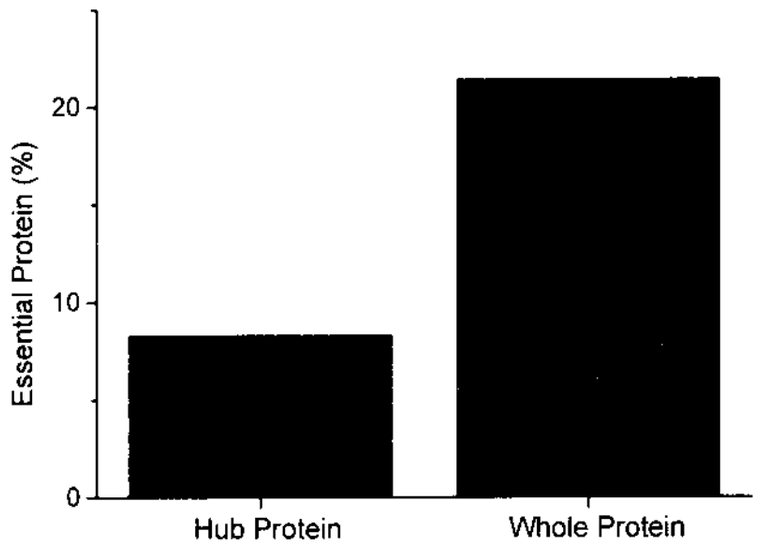


그림 2. 데이터베이스 구조 및 허브

SGD와 BioGRID 데이터베이스에서 급격히 증가한 상호작용이 어떤 형태로 분포되어 있는지를 DIP의 그것과 비교하여 차이를 분석하고자 하였다. 어떤 단백질의 상호작용 수를 Ki라 칭하고, 그 단백질과 이웃하는 단백질의 상호작용 수를 Kj로 칭하였다. 예를 들면, 단백질 A와 상호작용하는 단백질 수는 B를 포함하여 15개라면 Ki 값은 15이며 단백질 B와 상호작용하는 단백질 수가 20이라면 Kj 값은 20이 된다. 이러한 값을 가지는 단백질의 경우 수를 Kn으로 표시하여 3차원 그래프로 분포도를 분석하였다[그림 3].

상대적으로 상호작용 수가 적은 DIP의 단백질 상호작용 분포는 축 주위에 분포되어 있는데 반해, SGD와 BioGRID의 분포는 0에서 300까지 고르게 분포되어 있다. 이는 허브 단백질이 증가했을 뿐만 아니라 허브 단백질과 허브 단백질 상호작용 수가 크게 증가했음을 의미한다. 필수 단백질 수는 같은데도 불구하고 허브 단백질의 상호작용 수가 증가한 것이 허브 단백질의 필수 단백질 퍼센트를 상대적으로 낮추는 결과를 가져오게 하였다.

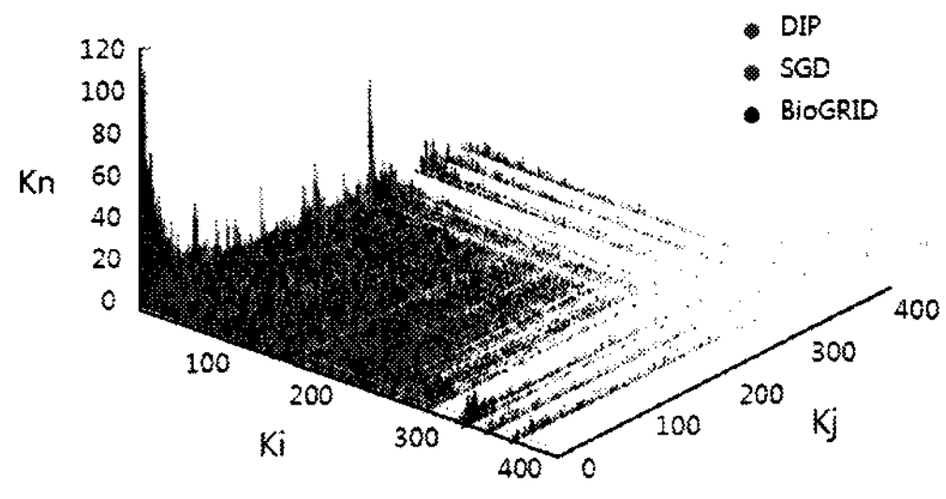


그림 3. DIP, SGD, BioGRID DB의 상호작용 수의 분포도 비교

IV. 결론

본 논문에서는 효모의 단백질-단백질 상호작용 정보를 제공해주는 Uetz, Ito, MIPS, DIP, SGD, BioGRID 등의 많은 데이터베이스들이 제공하는 PPI 데이터를 이용해 네트워크의 허브와 필수 단백질의 연관성을 분석하였다. Uetz, Ito, DIP과 같이 단백질 상호작용 단백질 수가 비교적 작은 경우는 이전 연구 결과들과 일치하였다[4][6]. 다시 말해 비교적 상호작용 수가 많은 단백질 즉 허브 단백질에 필수 단백질이 많이 포함되어있는 것을 확인하였다. 반면 SGD, BioGRID와 같이 단백질 상호작용 수가 큰 데이터베이스의 경우에는 허브단백질과 필수단백질 사이의 관계가 이전 결과들과 다르게 나타났음을 확인하였다.

결과적으로 고 효율성 기법을 통해 대량의 데이터가 추출되고 있는 현 시점에서는 이전 연구들에서 제시된 허브 단백질과 필수단백질 사이의 관계가 재조명 되어야 할 필요가 있다. 텍스터마이닝 기법 도입으로 대용량 데이터를 확보한 SGD나 BioGRID의 경우, 여러 실험 데이터를 통합한 DIP 보다 오히려 false positive 데이터를 상당량 제거하였다고 보고하였는데 [15][16], 실제 세포에서는 허브 단백질이 필수적인 경향을 나타내는 것이 아니라 오히려 필수 단백질들이 상호작용 수가 작은 단백질들로 흩어져 있기 때문에 오랜 세월 동안 박테리아와 같은 생명체가 적응과 진화를 거쳐 생존했을 가능성도 배제할 수 없다. 앞으로 실제 생명체에서는 허브 단백질이 어떤 성향을 띄는지에 대한 연구가 지속적으로 이루어져야할 것이다.

참 고 문 헌

[1] E. Ravasz and A. L. Barabasi, "Hierarchical organization in complex networks," Phys. Rev. E, Vol.67, No.2, pp.026122-0261228, 2003.

[2] A. L. Barabasi and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," Nat. Rev. Genet., Vol.5, No.2, pp.101-113, 2004.

[3] N. N. Batada, L. D. Hurst, and M. Tyers, "Evolutionary and physiologucal important of hub proteins," Plos Comput. Biol., Vol.2, No.7, pp.e88, 2006.

[4] H. Jeong, S. P. Mason, A. L. Barabasi, and Z. N. Oltvai. "Lethality and centrality in protein networks," Nature, Vol.411, No.6833, pp.41-42, 2001.

[5] S. R. Proulx, S. Nuzhdin, and D. E. Promislow, "Direct selection on genetic robustness revealed in the yeast transcriptome," PLos ONE., Vol.2, No.9, pp.e911, 2007.

[6] H. Yu, D. Greenbaum, H. X. Lu, X. Zhu, and M. Gerstein, "Genomic analysis of essentiality within protein networks," TRENDS in Genomics, Vol.20, No.6, pp.227-231, 2004.

[7] P. Ross-Macdonald, P. S. Coelho, T. Roemer, G. S. Roeder, and M. Snyder, "Large-scale analysis of the yeast genome by transposon tagging and gene disruption," Nature, Vol.402, No.6759, pp.413-418, 1999.

[8] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, M. Johnston, and R. W. Davis, "Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis," Science, Vol.285, No.5429, pp.901-906, 1999.

[9] X. He and J. Zhang, "Why Do Hubs Tend to Be Essential in Protein Networks?" Plos Genet, Vo.2, No.6, pp.0826-0834, 2006.

[10] R. Aragues, A. Sali, J. Bonet, M. A. Marti-Renom, and B. Oliva, "Characterization of protein hubs by inferring interacting motifs from protein interactions," PLoS Comput Biol., Vol.3, No.9, pp.1761-1771, 2007.

[11] P. Uetz, L. Giot, G. Caqney, T. A. Mansfield, R. S. Judson, S. Fields, and J. M. Rothberg, "A comprehensive analysis of protein-protein

interactions in *Saccharomyces cerevisiae*," Nature, Vol.403, No.6770, pp.623-627, 2000.

[12] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome," PNAS, Vol.98, No8. pp.4569-4574, 2001.

[13] U. Gűldener, M. Műnsterkotter, M. Oesterheld, P. Paqel, A. Ruepp, H. W. Mewes, and V. Stűmpflen, "MPact: the MIPS protein interaction resorce on yeast," Nucleic Acids Research, Vol.34, No.1, pp.D436-D441, 2006.

[14] 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 Res, Vol.32, No.1, pp.D449 - D451, 2004.

[15] J. E. Hirschman, R. Balakrishnan, K. R. Christie, D. Botstein, and J. M. Cherry, "Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome," Nucleic Acids Res, Vol.34, No.1, pp.D442-445, 2006.

[16] C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers, "BioGRID: a general repository for interaction datasets," Nucleic Acids Res., pp.34, No.1, pp.D535-D539, 2006.

[17] J. W. Ryu, H. Y. Kim, T. H. Kang, J. S. Yoo, and J. S. Chung, "Prediction of unannotated proteins from a protein interaction network filtered by using localization and domains in yeast," J. Kor. Phys. Soc, Vol.51, No.5, pp.1805-1811, 2007.

저 자 소 개

류 제 운(Jea Woon Ryu) 정회원

- 2006년 2월 : 충북대학교 생화학과(이학사)
- 2002년 8월 ~ 현재 : 충북대학교 생화학과(이학석사)

<관심분야> : 생물정보학, 신호전이 네트워크, 시스템 바이오

강 태 호(Tae-Ho Kang) 정회원

- 1999년 2월 : 호원대학교 정보통신공학과(공학사)
- 2002년 8월 : 충북대학교 정보산업공학과(공학석사)
- 2007년 8월 : 충북대학교 정보통신공학과(공학박사)
- 2007년 9월 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 Post-doc.

<관심분야> : 데이터베이스 시스템, 데이터 마이닝, 생물정보학, 시스템 바이오

유 재 수(Jae-Soo Yoo) 종신회원

- 1989년 2월 : 전북대학교 컴퓨터공학과(공학사)
- 1991년 2월 : 한국과학기술연구원 전산학과(공학석사)
- 2007년 8월 : 한국과학기술연구원 전산학과(공학박사)
- 1996년 ~ 현재 : 충북대학교 전기전자컴퓨터공학부 교수

<관심분야> : 데이터베이스 시스템, 정보검색, 멀티미디어 데이터베이스, 분산객체 컴퓨팅, 생물정보학

김 학 용(Hak Yong Kim) 정회원

- 1985년 2월 : 충북대학교 농화학과(농학사)
- 2002년 8월 : 충북대학교 화학과(이학석사)
- 2007년 8월 : 미국 코네티컷대학교, 분자 및 세포생물학과(이학박사)
- 1998년 ~ 현재 : 충북대학교 생명과학부 교수

<관심분야> : 시스템 바이오, 신호 전이, 단백질 네트워크, 생체동역학