Identifying the biological and physical essence of protein-protein network for yeast proteome : Eigenvalue and perturbation analysis of Laplacian matrix

이스트 프로테옴에 대한 단백질-단백질 네트워크의 생물학적 및 물리학적 정보인식 : 라플라스 행렬에 대한 고유치와 섭동분석

Iksoo Chang[1], Mookyung Cheon[1], Eun-Joung Moon[1], and Choongrak Kim[2]

[1] National Research Lab. For Computational Protomics and Biophysics, Dept. of Physics, Pusan Nat'l Univ. Busan 609-735, Korea
[2] Dept. of Statistics, Pusan Nat'l Univ. Busan 609-735, Korea

*To whom correspondence should be addressed. E -mail: chang@random.phys.pusan.ac.kr

Abstract

The interaction network of protein -protein plays an important role to understand the various biological functions of cells. Currently, the high -throughput experimental techniques (two -dimensional gel electrophoresis, mass spectroscopy, yeast two -hybrid assay) provide us with the vast amount of data for protein-protein interaction at the proteome scale. In order to rec ognize the role of each protein in their network, the efficient bioinformatic al and computational analysis methods are required.

We propose a systematic and mathematical method which can analyze the protein -protein interaction network rigorously and enable us to capture the biological and physical essence of a topological character and stability of protein -protein network, and sensitivity of each protein along the biological pathway of their network. We set up a Laplacian matrix of spectral graph theory bas ed on the protein -protein network of yeast proteome, and perform an eigenvalue analysis and apply a perturbation method on a Laplacian matrix, which result in recognizing the center of protein cluster, the identity of hub proteins around it and their relative sensitivities. Identifying the topology of protein -protein network via a Laplacian matrix, we can recognize the important relation between the biological pathway of yeast proteome and the formalism of master equation. The results of our systematic and mathematical analysis agree well with the experimental findings of yeast proteome. The biological function and meaning of each protein cluster can be explained easily. Our rigorous analysis method is robust for understanding various kind s of networks

whether they are biological, social, economical...etc

Introduction

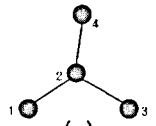
단백질-단백질 상호작용 네트워크는 세포의 다양한 활동(세포골격유지, 세포증식, 세포사멸, 단백질 합성 및 분해, 각종 대사작용 등)을 가능하게 한다. 현재 high-throughput experimental techniques (two-dimensional gel electrophoresis, mass spectrometry, yeast two-hybrid assay)를 통해서 proteome scale의 단백질-단백질 상호작용에 관한 방대한 데이터를 얻을 수 있다. 이런 방대한 데이터를 체계적으로 분석하고 그들로부터 또 다른 새로운 정보를 얻기 위해서는 computational 분석 (bioinformatical analysis)가 요구 되어진다. 이런 computational 분석은 미지의 단백질들을 동정하고 그 기능을 찾는데 드는 시간과 비용을 줄여준다. 또한 proteome scale의 단백질-단백질 상호작용 데이터의 computational 분석을 통해 proteome 전체의 topological 특성이나 stability, sensitivity 등을 조사함으로써 global system을 이해하려는 시도가 이루어지고 있으며, 이와 더불어 새로운 computational 분석 방법들도 제기되고 있다.

본 연구에서는 이러한 생물학적 네트워크의 연구에 대한 Laplacian matrix와 perturbation 방법이라는 새로운 접근을 제안하고, 이러한 방법을 통하여 yeast에서의 단백질-단백질 상호작용 네트워크에 적용시켜 이 네트워크의 특성들을 조사하여 기존의 연구들과 비교하였다

This work is supported by the National Research Laboratory program of the Ministry of Science and Technology, Korea/ Korea Institute of Science and Technology Evaluation and Planning.

Laplacian Matrix

Laplacian matrix는 spectral graph theory[1]에서 사용되는 방법으로, 어떤 주어진 그래프(graph)의 특성을 이해하기 위해 사용된다. Node와 edge들로 이루어진 그래프가 주어진다면, 먼저 각 node들과의 연결 상태를 나타내는 edge들을 사용해서 adjacency matrix를 그리고 각 node마다 몇 개의 다른 node들과의 연결되어 있는지를 알려주는 degree matrix를 세울 수 있다. 이 때 Laplacian matrix는 degree matrix에서 adjacency matrix를 뺀 값으로 정의되는데, 대각 element들은 각 node들의 degree로 주어지며, off-diagonal element들은 두 node를 연결하는 edge가 있으면 두 node에 대응되는 행렬의 element 값이 -1 또는 weighted 된 값으로 할당된다. 그리고 off-diagonal element 중에서 edge가 없는 두 node사이의 행렬 element값은 0으로 할당된다. 그림 1(a)는 node 개수가 4이고 edge 개수가 3인 그래프의 한 예이다, 이 그래프에 해당되는 각각의 adjacency, degree, 그리고 Laplacian matrix는 그림 1(b)와 같이 주어진다. Node의 개수가 4이므로 각각의 matrix는 4X4 matrix가 된다. 이러한 접근은 일반적으로 그림1과 같은 한 개의 그래프가 아닌 여러 개의 분리된 그래프에도 동일하게 적용시킬 수 있으며, 이 때 나타나게 되는 matrix들은 block diagonalized matrix 형태가 된다.



(a)

$$A=\begin{pmatrix}0&1&0&0\\1&0&1&1\\0&1&0&0\\0&1&0&0\end{pmatrix},\quad D=\begin{pmatrix}1&0&0&0\\0&3&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}\Rightarrow L=\begin{pmatrix}1&-1&0&0\\-1&3&-1&-1\\0&-1&1&0\\0&-1&0&1\end{pmatrix}$$

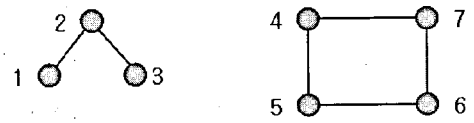
(b)

그림 1 : (a) Node 개수가 4이고 edge개수가 3인 그래프. (b) 이 그래프에 해당되는 각각의 4X4 Adjacency matrix(A), Degree matrix(D), 그리고 Laplacian matrix(L)

이러한 Graph Spectral Method를 응용한 최근 연구[2,3] 중 하나는 주어진 단백질에서 각 아미노산들의 그룹화(clustering)에 응용되어졌는데[2] 이러한 방식으로 단백질의 중요한 domain이나 active site에 대한 정보들을 구하는데 사용되어졌다. 이러한 응용은 Laplacian matrix의 eigenvector들의 정보로부터 그룹화와 각각의 클러스터(cluster)에 대한 center들을 구할 수 있다는 사실로부터 가능하게 되었다. 즉 주어진 그래프로부터 Laplacian matrix를 만들면, 이 matrix로부터 eigenvalue와 eigenvector들을 구하면 각 eigenvector에 그 그래프의 고유한 특성들이 포함되어 있다. 가령 N개의 node를 가지고 M개의 분리된 그래프로 이루어진 network이라면 NXN Laplacian matrix는 M개의 block diagonalized matrix이며 이 때 M개의 eigenvalue값이 0인 eigenvector들이 구해진다. Eigenvalue값이 0인 eigenvector들은 각각의 분리된 그래프들을 의미하며, eigenvector들의 element들의 값은 해당되는 그래프의 node에만 값이 주어지고 나머지는 0이다. 이런 식으로 M개의 eigenvector들을 통하여 각각의 분리된 그래프들은 자신에 속한 node들을 알 수 있게 된다. 또한 각 분리된 그래프에서 가장 중요한 hub node도 알 수 있는데, 모든 eigenvector들은 여러 분리된 그래프 중에서 한 그래프에 해당되는 node들에게만 0이 아닌 특별한 값들을 가진다. 이 때 어떤 한 그래프에 해당되는 node들이 0이 아닌 것들을 다 모으면 eigenvector들의 개수는 node개수와 정확하게 일치하고, 여기서 가장 큰 eigenvalue에 해당되는 eigenvector의 값들에는 가장 중요한 hub node와 주위의 node들에 대한 정보를 포함하고 있다. 이런 주어진 그래프에서 가장 큰 eigenvalue에 해당되는 eigenvector를 top eigenvector라 한다. 이 top eigenvector에서 element들의 값이 가장 큰 것이 hub node에 해당된다. 그리고 음수이면서 절대값이 큰 element들은 이 hub node에 일차적으로 연결된 node들이다. 이런 식으로 두 번째로 큰 eigenvalue에 해당되는 eigenvector는 두 번째 hub node에 대한 정보를 준다. 이러한 기본 개념을 이용하면 무수히 큰 network에 대해서 단 한 번의 eigenvalue analysis를 통하여서 전 network의 특성들을 빠르게 수집할 수 있다.

하지만 여전히 각 분리된 그래프들에 해당되는 node들을 알기 위해서는 적어도 M개의 eigenvector들을 분석해야 한다. 이러한 Laplacian matrix의 특성을 잘 살리면서도 여러 분리된 그래프들의 각각의 node에 대한 정보 즉 그룹화에 대해 좀 더 효율적인 방법이 제안되었는데, adjacency matrix를 만들 때 연결되지 않은 node들 사이의 값들에 0대신 작은 값인 Δ를 사용하면 원

래 분리된 모든 그래프들은 하나의 연결된 그래프가 된다. 그러므로 이렇게 adjacency를 만들고 degree matrix도 Δ만큼 보정을 주어 Laplacian matrix를 세운다. 그림 2는 두 개의 분리된 그래프로 이루어진 네트워크와 Δ를 사용한 Laplacian matrix의 예이다.



(a)

Laplacian Matrix L' (D-A) with small Δ value to make one cluster in total.

$$\begin{pmatrix} 1+5\Delta & -1 & -\Delta & -\Delta & -\Delta & -\Delta & -\Delta \\ -1 & 2+4\Delta & -1 & -\Delta & -\Delta & -\Delta & -\Delta \\ -\Delta & -1 & 1+5\Delta & -\Delta & -\Delta & -\Delta & -\Delta \\ -\Delta & -\Delta & -\Delta & 2+4\Delta & -1 & -\Delta & -1 \\ -\Delta & -\Delta & -\Delta & -1 & 2+4\Delta & -1 & -\Delta \\ -\Delta & -\Delta & -\Delta & -\Delta & -1 & 2+4\Delta & -1 \\ -\Delta & -\Delta & -\Delta & -1 & -\Delta & -1 & 2+4\Delta \end{pmatrix}$$

(b)

그림 2: (a) 2개의 분리된 그래프로 이루어진 네트워크. (b) 이에 해당되는 Δ로 보정한 Laplacian matrix.

이와 같이 Δ로 보정한 Laplacian matrix를 eigenvalue analysis를 하면 eigenvalue가 0인 eigenvector는 하나가 나오고 M-1개의 0보다 크지만 작은 값의 eigenvalue를 가진 eigenvector들이 만들어진다. 이 때 second lowest eigenvalue에 해당되는 eigenvector의 값들을 살펴보면 원래 분리된 그래프를 가진 네트워크에서 같은 그래프에 속하는 node들은 같은 값을 가지게 된다. 즉 M개의 그래프에 각각의 그래프마다 해당 node들이 동일한 값을 가지게 되어 이 eigenvector의 분석만을 통하여서도 각 node들의 그룹화가 가능하게 된다.

이처럼 Laplacian matrix의 eigenvalue analysis를 통하여 주어진 네트워크에서 각 그래프들의 node들을 알 수 있을뿐더러 각 그래프들의 중심이 되는 hub node와 그 이웃하는 node들을 단번에 찾을 수 있다. 즉 광대한 네트워크에서 그룹화와 중요한 node를 구별하는데 유용하게 사용될 수 있다.

Analogy to a Master equation

이러한 Laplacian matrix는 통계물리학에서 사용되는 master equation에서도 사용되어지는데, 한 입자가 격자나 어떤 클러스터 상에서 움직인다고 생각하자. 가령 그림 1(a)와 같은 그래프 상에서 사이트 i에서 입자를 발견할 확률을 P_i라고 하면, 이 계의 동역학 및 평형상태를 기술하는 master equation은 다음과 같이 주어지게 된다.

$$\begin{aligned} dP_1/dt &= -P_1 + P_2 \\ dP_2/dt &= P_1 - 3P_2 + P_3 + P_4 \\ dP_3/dt &= +P_2 - P_3 \\ dP_4/dt &= +P_2 - P_4 \end{aligned}$$

이 때 각각의 사이트에서의 에너지는 동일하다고 가정하게 된다. 이 식을 좀 더 간결하게 표현하면

$$\frac{d\vec{P}}{dt} = -M \cdot \vec{P}, \quad \text{where} M = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1 & 3 & -1 & -1 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{pmatrix}$$

여기서 M은 바로 앞에서 언급한 Laplacian matrix L과 동일하게 된다. 그러므로 Laplacian matrix를 사용하여 cluster들의 특성들을 알아내는 작업은 master equation을 이용한 미분방정식의 풀이와 동일하게

이 network가 가지는 계를 설명해 줄 것이다.

또한 master equation을 사용한 통계물리학적 방법에서 이미 알려진 여러 특성들도 동일하게 Laplacian matrix에 적용가능 할 것이다. 가령 n개의 node 또는 사이트로 이루어진 그래프 또는 클러스터가 있다면, master equation을 eigenvalue analysis를 했을 때 0인 eigenvalue는 한 개가 나올 것이며 이 eigenvalue에 해당되는 eigenvector의 element들의 합은 0이 아닌 어떤 값이 될 것이다. 하지만 n-1개의 0이 아닌 eigenvalue를 가지는 eigenvector들의 element합은 0이 된다. 이러한 eigenvalue값들은 각각 입자가 클러스터 위에서 움직임을 나타내는 동역학적 특성들을 내포하고 있는데, 0인 eigenvalue는 평형상태에서의 입자가 있을 때의 확률을 나타내며, second lowest eigenvalue는 비평형상태에 있던 입자가 평형상태로 찾아갈 때 가장 늦은 완화를 나타내는 모드가 된다. 물론 가장 큰 eigenvalue는 평형상태로 가장 빨리 가는 모드를 의미한다

Yeast Network Matrix

Laplacian matrix를 이용한 연구 중 단백질내 아미노산들의 네트워크에 적용한 연구는 역시 second lowest eigenvector를 분석해서 각 아미노산들을 그룹화하여 active site나 domain에서 아미노산들을 그룹을 형성시킬 수 있었으며, 각 cluster에 해당되는 top eigenvector로부터 그 클러스터의 가장 중요한 center 아미노산을 찾을 수 있었다. 이러한 접근 방식은 단백질내의 아미노산의 network뿐만 아니라 좀 더 거시적인 단백질-단백질 상호작용에도 응용될 수 있을 것이라 예상되며, Yeast proteome 의 단백질-단백질 network에도 적용하여 그 network의 특성을 이해하고자 한다.

단백질-단백질 상호작용의 데이터 중 본 연구에서 사용한 것은 Ito의 yeast two-hybrid 데이터[4] 중 core data를 사용하였다 (http://genome.c.kanazawa-u.ac.jp/Y2H). 여기에는 총 786개의 단백질이 754개의 상호작용을 가진다. 이러한 데이터로 단백질을 각 node로 상호작용을 edge로 정하여 786×786 adjacency matrix와 degree matrix, 그리고 Laplacian matrix를 만들었다. Laplacian matrix를 만들 때 Δ를 0과 0.01을 사용하여 만들었으며 각각에 대해 eigenvalue analysis를 수행하였다. Δ가 0인 경우 총 132개의 0인 eigenvalue를 얻었다. 이는 총 132개의 분리된 클러스터들이 존재함을 의미하며 각각의 클러스터에서 hub protein을 top eigenvector를 분석하여 구하였다. Δ가 0.01인 경우는 0인 eigenvalue가 1개 존재하고 131개의 0보다 약간 크지만 작은 값의 eigenvalue를 얻었으며, second lowest eigenvalue의 eigenvector에서 같은 값을 가지는 node들로 그룹화하여 분석하면 총 132개의 cluster가 구해진다. 이 경우도 마찬가지로 hub 단백질을 각각의 top eigenvector를 분석해서 구할 수 있었다. 132개의 클러스터 중 131개의 클러스터는 2에서 14개로 구성된 단백질들의 클러스터로 크기가 무척 작은데 비해 나머지 한 개는 총 417개의 단백질을 포함하고 있는 무척 큰 클러스터를 형성하고 있다. 이 큰 클러스터의 첫 번째 hub에서 다섯 번째 hub는 가장 큰 eigenvalue부터 다섯 번째 큰 eigenvalue에 해당되는 eigenvector들의 element로부터 구해지는데, 각각 SRP1, APG17, JSN1, TEM1, BZZ1이다.

Perturbation method

Laplacian Matrix를 사용하여 각 클러스터의 가장 중요한 center를 각 클러스터에 대응되는 top eigenvector를 통하여 얻을 수 있지만, 그 클러스터에서 다음으로 중요한 node를 알기 위해서는 top eigenvector가 아닌 그 다음 eigenvector를 살펴보고 찾아야 한다. 그래서 우리는 다른 방법으로 각 클러스터의 중요한 node를 순차적으로 찾을 수 있는 방법을 고안하였는데, 각 cluster에 대응되는 node와 edge들만으로 이루어진 Laplacian Matrix를 만들어 이 matrix의 eigenvalue의 값과 여기서 한 node씩 perturbation을 시켜서 생기는 matrix의 eigenvalue의 차이를 구하였고, 이 차이들의 합이 클수록 그 network에서 중요한 node가 된다는 것을 확인하였다. 이러한 접근 방법을 통해 yeast network의 각 cluster들에 대해 중요한 역할을 하는 단백질들을 순차적으로 얻을 수 있었다.

417개 단백질의 클러스터에 대해 이렇게 순차적으로 얻은 hub 단백질도 역시 앞에서 구한 순서와 동일한데 SRP1, APG17, JSN1, TEM1, BZZ1, NUP116, YDL239C, YDL100C 순이었다. 이 순서는 각 단백질이 상호작용하는 다른 단백질의 수와 비례한다. 그림 3은 가장 큰 cluster의 network 모습을 표1은 hub 단백질을 나타낸다

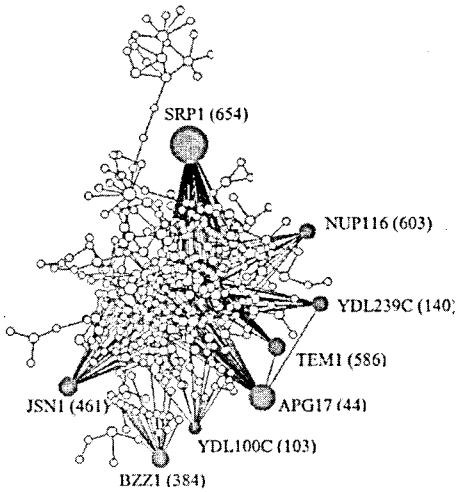


그림 3: 417개의 단백질로 이루어진 가장 큰 클러스터.

Hubs of the largest cluster from perturbation of matrix (from SD1 & SD2)

Protein name (no.)	First connected proteins
SRP1 (654)	**55**
APG17 (44)	**33**
JSN1 (461)	**22**
TEM1 (586)	**20**
BZZ1 (384)	**18**
NUP116 (603)	**15**
YDL239C (140)	**15**
YDL100C (103)	**11**

표1: 가장 큰 클러스터에서 8개의 hub 단백질과 상호작용하는 다른 단백질의 수

표1에서 보듯이 앞에서 Laplacian matrix와 perturbation 방법을 이용하여 구한hub 단백질들은 상호작용하는 단백질의 수에 따른 순차적인 단백질들과 잘 일치한다. 이렇게 구한 hub 단백질들은 실제 생태 내에서 매우 중요한 역할을 하는 경우가 많다. 표2에서와 같이 SRP1, TEM1, NUP116등의 단백질은 제거될 때 yeast는 세포가 죽게 되는 phenotype이 inviable인 단백질이다.

Protein name (no.)	Experiment Type	Phenotype
SRP1 (654)	Systematic deletion	Inviable
APG17 (44)	Systematic deletion	Exhibits growth defect on a non-fermentable (respiratory) carbon source
JSN1 (451)	Systematic deletion	Viable
TEM1 (586)	Systematic deletion	Inviable
BZZ1 (384)	Systematic deletion	Viable
NUP116 (603)	Systematic deletion	Inviable
YDL239C (140)	Systematic deletion	Viable
YDL100C (103)	Systematic deletion	Exhibits sensitivity at 5 generations when grown in 10uM nystatin

표 2: 가장 큰 클러스터에서 hub 단백질들의 생물학적 특성

Summary and Prospect

Spectral graph theory의 Laplacian matrix는 network의 특성을 이해하는데 사용될 수 있다. 특히 이러한 Laplacian matrix는 입자가 클러스터상에서의 평형 및 동역학을 기술하는 master equation의 관점으로 조명될 수 있으며, 이러한 관점에서 평형상태 및 완하 모드에 대한 해법과 network의 특성에 대해 관계 지워 질 수 있으며, 이러한 관점에서의 연구는 network이 동역학적으로 움직이는 계이거나 각 상호작용에 weight가 주어지면 더욱 중요하게 부각될 것이다. 우리는 이러한 Laplacian matrix를 사용하여 기본적인 형태의 yeast network의 clustering과 hub 단백질들을 성공적으로 얻을 수 있었다. 이러한 연구를 통해 각 상호작용에 weight가 들어가는 경우와 동역학적으로 움직이는 계와 같은 좀 더 복잡한 관계가 있는 단백질-단백질 상호작용 연구에 좀 더 체계화된 접근이 될 수 있을 것이다. 물론 이러한 연구는 단백질-단백질 상호작용에 관한 연구뿐 아니라 사회현상부터 자연현상에 이르는 다양한 네트워크에 응용될 수 있을 것으로 예상된다

References

[1] N. Deo. *Graph theory with applications to engineering and computer science*, Prentice Hall 1974.

[2] S. Vishveshwara, K.V. Brinda, and N Kannan. *Protein structure: Insights from graph theory*, J. Theor. Comp. Chem., 1, 187-211 (2002)

[3] C. H. Q. Ding. *Unsupervised feature selection via two-way ordering in gene expression analysis*, Bioinformatics, 19, 1259-1266 (2003)

[4] T. Ito, T. Chiba, R. Ozawa. M. Yoshida.. M. Hattori, and S. Yoshiyuki. *A computer two-hybrid analysis to explore the yeast protein interactome*. PNAS 98, 4569-4574 (2001)