Sensitivity-dependent model of protein–protein interaction networks

Jingshan Zhang and Eugene I Shakhnovich

Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA

Received 30 May 2008
Accepted for publication 1 September 2008
Published 29 September 2008
Online at stacks.iop.org/PhysBio/5/036011

Abstract
The scale free structure \( p(k) \sim k^{-\gamma} \) of protein–protein interaction networks can be reproduced by a static physical model in simulation. We inspect the model theoretically, and find the key reason for the model generating apparent scale free degree distributions. This explanation provides a generic mechanism of ‘scale free’ networks. Moreover, we predict the dependence of \( \gamma \) on experimental protein concentrations or other sensitivity factors in detecting interactions, and find experimental evidence to support the prediction.

1. Introduction
‘Scale free’ networks have been observed in many areas of science [1] including social science, biology and the internet, where degree distributions follow (albeit noises) the power-law form \( p(k) \sim k^{-\gamma} \) within one or two orders of magnitude for \( k \). Here the degree \( k \) is the number of links a node has, and \( p(k) \) is the probability of a node to have degree \( k \). An important scale free network under experimental [2–5] and theoretical [1, 6–13] study is the protein–protein interaction (PPI) network, where a link between two proteins indicates a large enough binding energy between them. These studies have the goal that the topology of PPI networks could reflect how systems of various proteins have evolved in biological organisms.

It has been pointed out recently that scale free PPI networks could also result from variation of surface hydrophobicities of proteins. Starting from an approximately Gaussian distribution of surface ‘stickiness’ \( K \), and the binding free energy of two proteins is determined by the sum of their ‘stickiness’. In a more detailed description, there are \( K \) hydrophobic residues among the \( M \) surface residues on protein \( i \), and \( M = 100 \) is assumed to be a constant for all proteins. The probability of finding a protein with \( K \) hydrophobic surface residues is \( p_E(K) \)

\[
p_E(K) = \int dp \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(p - \overline{p})^2}{2\sigma^2}} \quad (1)
\]

with mean value \( \overline{p} \approx 0.2 \) and deviation \( \sigma \approx 0.05 \). This results in an approximately Gaussian distribution of the surface ‘stickiness’ \( K \), and the binding free energy of two proteins is determined by the sum of their ‘stickiness’. In a more detailed description, there are \( K \) hydrophobic residues among the \( M \) surface residues on protein \( i \), and \( M = 100 \) is assumed to be a constant for all proteins. The probability of finding a protein with \( K \) hydrophobic surface residues is

\[
p_E(K) = \frac{M!}{K!(M-K)!} \left( \frac{1}{2} \right)^K \left( 1 - \frac{1}{2} \right)^{M-K}. \quad (2)
\]

It can be seen that \( p_E(K) \) is close to a Gaussian distribution (figure 1(a)). The binding of proteins \( i \) and \( j \) is determined by the binding free energy

\[
\Delta G = -(K_i + K_j)F_0 + G_{(0)}, \quad (3)
\]

where \( \Delta G \) is negative for a strong binding, \( F_0 \) is the change of binding free energy upon burial of each hydrophobic residue and \( G_{(0)} \approx 6 kT \approx 10 kBT \) is a constant value determined by experiments [14, 15]. In support of this model, figure 3 of [14] showed that experimental result of binding...
energies can be described by the sum of stickiness terms and a constant term. If \( K_i + K_j > K_c \), the interaction is experimentally detectable, and the two proteins are labeled as linked in the PPI network.

3. Results and interpretations

3.1. Degree distributions

We calculate \( p(k) \) numerically (see numerical method for details) with given values of \( N \) and \( K_c \), where \( N \) is the total number of proteins in the network, and obtain apparent ‘scale free’ structure \( p(k) \propto k^{-\gamma} \) (figure 2). We set the default situation as \( N = 5000 \) and \( K_c = 83 \) to fit \( \gamma = 2 \). Figure 2 indicates that the apparent slope \( \gamma \) increases with \( K_c \), and increases as \( N \) decreases. More explicitly, the dependence of \( \gamma \) upon \( K_c \) is plotted in figure 3.

Let us interpret these results by analytical approaches. A protein with hydrophobicity \( K \) has a pass/fail line

\[
K_{\text{line}} = K_c - K.
\]  

Proteins with hydrophobicity above \( K_{\text{line}} \) are linked to it, while those with hydrophobicity below \( K_{\text{line}} \) do not. Therefore the protein with hydrophobicity \( K \) has an average degree

\[
\bar{k} \simeq N \int_{K_{\text{line}}}^{\infty} p_c(K') \, dK'.
\]  

In the mean-field approximation the degree of the protein \( k \) is just \( \bar{k} \), and the degree distribution is

\[
p(k) = p_c(K) \frac{dK}{dk} = \frac{p_c(K)}{N p_c(K_c - K)}.
\]  

Beyond mean-field approximation its degree fluctuates with deviation \( \sim \sqrt{\bar{k}} \), which will be addressed later.

Let us restrict the discussion to the mean-field approximation for the moment. We can note that the experimentally observable range \( 1 \lesssim \bar{k} \lesssim 100 \) only covers a small range of hydrophobicity (39 \( \lesssim K \lesssim 48 \) for the default situation), as indicated by the short red line in figure 1(b). In this range the hydrophobicity distribution \( p_c(K) \) is very close to exponential, since the short red line in figure 1(a) is nearly straight. So we can use linear approximation to produce the nearly straight lines in figure 2. Define

\[
a = \lim_{K \to K_{\text{line}}} \frac{d \ln p_c(K')}{dK'} \quad (7)
\]  

and

\[
b = \lim_{K \to K_{\text{line}}} \frac{d \ln p_c(K')}{dK'},
\]  

then equation (5) gives \( \bar{k} = e^{aK_{\text{line}}b} \). As a result we have \( p(k) \sim k^{-1+2b/a} \). This is a ‘scale free’ network with \( \gamma = 1 + 2b/a \).

To understand the undulations in \( p(k) \) at large \( k \) in figure 2, we must go beyond the mean-field approximation and deal with the fluctuation of degree with magnitude \( \sqrt{\bar{k}} \) for a given \( k \). Noting the \( K \) values are discrete integers, each \( K \) value produces a peak in \( p(k) \), centered at \( \bar{k} \) and with width \( \sqrt{\bar{k}} \).

Since \( \bar{k} \) grows with \( K \) almost exponentially, the distance between nearest-neighbor peaks \( \bar{k}(K+1) - \bar{k}(K) \)
grows linearly with $\bar{K}$. The undulations emerge at large enough $\bar{K}$, when the peak distance exceeds the peak width $\sqrt{\bar{K}}$.

Now we are ready to study the dependence of the slope $\gamma$ on parameter $K_c$ in figure 3. Approximating the hydrophobicity distribution as the Gaussian distribution $\ln p_x(K) \sim -(K - K_0)^2$, where $K_0$ is the most probable hydrophobicity value, we have

$$\gamma = 1 + \frac{b}{a} \approx 1 + \frac{K_c - K_{\text{line}} - K_0}{K_{\text{line}} - K_0}.$$  

(9)

We find $K_0 \simeq 20$ in equation (2), and $K_{\text{line}} \simeq 41.5$ is nearly a constant from equation (5) for a typical degree $k \simeq 5$, then $\gamma$ is a linear function of $K_c$ in equation (9), and forms a straight line (solid) in figure 3.

3.2. Dependence on experimental sensitivity

Different $\gamma$ values have been obtained in different PPI experiments, varying from $\gamma \approx 2.1$ to $\gamma \approx 2.5$ [3–5, 8–10]. To explain this variation, we note that different experiments might have different sensitivity in detecting PPI. Indeed, some interactions detectable in one experiment might be too weak to be detected in another experiment. An example of factors affecting experimental sensitivity is protein concentration/level, which is in turn controlled by gene expression and dependent upon the specific technique used to detect PPI. Even for the same experiment, the sensitivity in detecting interactions is actually reduced by setting a higher standard in identifying PPI, e.g., selecting only highly repeatable PPI data which effectively correspond to interactions with high affinity.

Let us study how $\gamma$ depends on these experimental sensitivity factors. In high-throughput experiments the concentration of protein–protein complex $C_{ij}$ must be high enough to be detected,

$$C_{ij} = \frac{C_i C_j}{C_0} \exp \left( \frac{-\Delta G}{k_b T} \right) \geq C_{\text{crit}},$$  

(10)

where the binding free energy $\Delta G$ is given by equation (3). $C_i$ and $C_j$ are the concentrations of proteins $i$ and $j$ in monomeric form, and the normalization concentration $C_0 = 1M$ is the convention. Rewriting this relationship in the form of association constant, the binding affinity should be strong enough to be detectable,

$$K_o = \frac{1}{C_0} \exp \left[ \frac{(K_i + K_j)F_0 - G(0)}{k_b T} \right] \geq \frac{1}{C_0} \exp \left[ \frac{K_iF_0 - G(0)}{k_b T} \right] = \frac{C_{\text{crit}}}{C_i C_j}.$$  

(11)

Thus the parameter $K_o$ of the model is determined by experimental concentration proteins

$$K_c = \left[ k_b T \ln \left( \frac{C_0C_{\text{crit}}}{C_i C_j} \right) + G(0) \right] / F_0.$$  

(12)

To estimate the only unknown parameter $F_0$ in this equation, we note that for the yeast two-hybrid screening technique the PPIs with binding affinity $K_o \geq \frac{C_{\text{crit}}}{C_i C_j} \simeq 1 \mu M^{-1}$ are detectable [16]. If we use $\gamma \approx 2.3$ and $K_c = 87$ for this threshold binding affinity, we can obtain an estimate $F_0 \approx 0.28k_b T$. With the help of this value we can use equation (12) to convert the $x$-axis of figure 3 from $K_c$ to the experimental variable $\frac{C_i}{C_j}$ (top of figure 3).

It can be seen from figure 3 that lower sensitivity, or lower $C_i, C_j$, leads to higher $\gamma$. This can be realized by lower protein concentrations through reduced gene expressions, or selecting only highly repeatable data of detected PPIs. This prediction is confirmed by figure 2(a) of [10], which clearly shows that the core data set of Ito et al. [4], containing only PPIs identified by at least three independent sequence tags, generates a steeper degree distribution than the full Ito data set does. Obviously the Ito core data correspond to relatively strong interactions, manifest in high $K_o$ and $K_c$. Note that the horizontal dots with $p(k) = 1/N$ at high $k$ in figure 2(a) of [10] should be excluded when fitting the slope $\gamma$, because they are actually in the $p(k) < 1/N$ region where a few nodes with arbitrary degree $k$ emerge occasionally. On the other hand, the protein concentrations in the yeast two hybrid experiments [3, 4] are not yet available, and the prediction about dependence of the slope $\gamma$ upon protein concentration needs verification from future experiments.

3.3. Clustering coefficient

We also study another important property of networks, clustering coefficient $C(k)$, and show the numerical result of the model in figure 4. If a protein is linked to $k$ proteins, the average number of links between the $k$ proteins, $t(k)$, cannot exceed $k(k-1)/2$. Here the averaging includes all possible realizations. The clustering coefficient is $C(k) \leq \frac{2t(k)}{k(k-1)} \leq 1$. Similarly to [17, 18], we obtain (figure 4) $C(k) \simeq 1$ at small $k$ and $C(k) \sim k^{-2}$ at large $k$. The experimental result [2–4, 10] has a similar shape with slope $\approx 2$ for large $k$, and $C(k)$ is smeared between 1 and $10^{-1}$ for small $k$. If we attribute the discrepancy between the model and experiment at small $k$ to false negatives, the model is in reasonable agreement with experiments. A physical picture is helpful for interpreting this result. As mentioned above, if there are the $k$ proteins linking to the same protein, their hydrophobicity exceeds $K_{\text{line}}$, while the hydrophobicity of all other proteins is below $K_{\text{line}}$. The mean-field relationship between $K_{\text{line}}$ and $k$ is equation (5). If we have $K_{\text{line}} \geq K_c/2$ at a small degree $k$, then the most hydrophobic $k$ proteins are all connected, and $C(k)$ is 1. At large enough $k$, however, $K_{\text{line}} < K_c/2$ and not all proteins
above $K_{\text{line}}$ are linked to each other. Then the clustering coefficient is determined by

$$C(k) \approx \frac{\sum_{K_{\text{line}}}^{M} AK_1 \int_{k_{\min}}^{K_{\text{line}}} dK_2 p_{E}(K_1)p_{E}(K_2)}{\sum_{K_{\text{line}}}^{M} dK_1 \int_{k_{\min}}^{K_{\text{line}}} dK_2 p_{E}(K_1)p_{E}(K_2)}. \quad (13)$$

The denominator is proportional to $K^{-2}$ according to equation (5). It corresponds to the square region between $K_{\text{line}}$ and $M$ in figure 5. The numerator, corresponding to the shadowed region in figure 5, is dominated by the region near the cutting line $K_1 + K_2 = K_2$, because $p_{E}(K)$ is nearly a sharp exponential function. Hence the numerator scales as the length of the cutting line, $K_2 = 2K_{\text{line}} \propto \ln K_2$ constant. Therefore, in agreement with Boguna et al [18], the numerator is a slow function of $k$ compared to the denominator, and the clustering coefficient scales as $C(k) \sim K^{-2}$ at large $K$. At small $K$, the square is totally in the shadow, leading to $C(K) \sim 1$. The step-like shape of $C(k)$, however, comes from the discreteness of integer $K$ values.

### 4. Numerical methods

We calculate $p(k)$ as an average of all possible realizations. The calculation is done with integer $K'$ and without the mean-field approximation. We ignore the unimportant difference between $N$ and $N + 1$ for large enough $N$. The exact form of equation (5) is

$$K = N \sum_{K'=\max(K_1,K_0)}^{M} p_{E}(K'). \quad (14)$$

and the degree distribution is

$$p(k) = \sum_{K'=0}^{M} p_{E}(K') \binom{N}{k} (\bar{K}/N)^k (1 - \bar{K}/N)^{N-k}. \quad (15)$$

Instead of the mean-field result, equation (13), the clustering coefficient is calculated as

$$C(k) = \sum_{K'=0}^{M-1} \left[ w(K') \left( \sum_{K_{1},K_{2}=K}^{M} p_{E}(K_{1})p_{E}(K_{2}) \theta(K_{1} + K_{2} - K_{c} + 1/2) \right) \right], \quad (16)$$

where

$$\theta(K) = \begin{cases} 1 & K > 0 \\ 0 & K < 0 \end{cases} \quad (17)$$

is the usual Heaviside step function, and

$$w(K) = \binom{N}{k} \left( \sum_{K'=0}^{M} p_{E}(K') \right)^k \left( \sum_{K'=0}^{K-1} p_{E}(K') \right)^{N-k} \quad (18)$$

is the probability that $k$ proteins have hydrophobicity $\geq K$ while the maximum hydrophobicity of the rest $N - k$ proteins is $K - 1$.

### 5. Conclusion and outlook

We study a static physical model to explain scale free PPI networks. We note that the experimentally observable part of degree distribution covers a limited range (from $k = 1$ to $k < 100$), and corresponds to a small range of hydrophobicity. The hydrophobicity distribution $p_{E}(K)$ in this small range is close enough to an exponential distribution. Therefore a linear approximation leads to the ‘scale free’ degree distribution $p(k) \sim K^{-\gamma}$, with $\gamma$ dependent on the threshold parameter $K_{c}$ and network size $N$. In experiments $K_{c}$ depends on the sensitivity factors, such as protein concentration, in detection of PPI. Our result provides a possible interpretation to the difference in experimental $\gamma$ values, and predicts the dependence of $\gamma$ on experimental sensitivity factors. This prediction is supported by the slope change [10] when comparing Ito data set and Ito core data set [4], and dependence of $\gamma$ on protein concentrations needs experimental verification in future. The distribution of another network property, clustering coefficient, produced in the model is also in reasonable agreement with that of experiment [10] and previous theoretical descriptions [17, 18].

The hydrophobicity distribution in the physical model has been arranged to reflect the reality in a simplified way. While the real distribution of protein ‘stickiness’ can be somewhat different from it, the generation of ‘scale free’ network will not be sensitive to the difference. More generally, ‘scale free’ degree distributions can also be produced by many smooth distributions of hydrophobicity, such as binomial, Gaussian, Poisson distributions and their modifications. This can be one of the reasons that scale free (in a limited range) networks are so widely observed.

A major part of PPI networks is obtained by the high-throughput yeast two-hybrid screening [3, 4]. This technique often produces a large fraction of false positives [19] which do not correspond to any real biological function, while real functional interactions presumably constitute a smaller portion of the detected result. While functional PPIs may involve formation of additional hydrogen bonds and salt bridges to obtain adequate binding affinity, these nonfunctional PPIs have not been evolutionarily selected and are formed primarily due to the hydrophobic effect. In this model we
show that a simple static network of nodes with different ‘stickiness’ can readily appear to be scale free. To this end, we use equation (3) because the nonfunctional PPIs are just random interfaces between two proteins without experiencing the evolutionary design of pairwise interface patterns. Moreover, this model could be used to extract information of nonfunctional interactions between unrelated proteins which randomly encounter in a real cell, and such information is in turn important in probing the general principles for cells to organize proteins in a cell. Namely, the stronger nonfunctional interactions, the more unrelated proteins interfere with each other, and the less protein types can coexist. Hence the nonfunctional interactions can limit the proteome size of a single cellular organism. Interesting related questions include the change of how much living cells have to do in constricting the nonfunctional interactions in the course of protein evolution, as well as the impact of higher temperature for thermophile organisms.

If the distribution of ‘stickiness’ is simply an exponential function, \( \ln p_i(K) \sim -K \), the model is simplified to \( a = b \) and \( \gamma = 2 \). This reduced simple situation would then be in complete agreement with one of the mathematical examples of networks briefly mentioned by Caldarelli et al [20], which has been applied to realistic networks such as gene regulation network [21]. Our finding indicates that this simple mathematical form [20] has more important impacts on systems in reality. Indeed, it is reasonable to expect distribution of many qualities, such as annual personal income and eagerness to learn knowledge, to be fitted by an approximately exponential distribution at least in some short range, and with suitable arrangements power-law distribution might emerge.

Masuda et al [17] followed the suggestion of Caldarelli et al [20] and studied essentially similar models to ours. But they did not relate the mathematical models to real systems. More importantly, they emphasized that the slope \( \gamma = 2 \) is universal, while the slope in our study not only deviates from 2 but also depends on experimental properties such as expression levels of proteins.

In contrast to this static model, most models of PPI networks focus on the development history of the network through gene duplications [11, 12], which is similar to ‘preferential attachment’ in growing networks [13]. It was found [12] that the network structure of the gene duplication model analytically approaches scale free [12] at \( k \to \infty \) if links of new nodes should be deleted by a probability larger than 1/2, and the degree distribution is comparable with experiments. Our approach serves as an alternative way to obtain ‘scale free’ PPI network. Further experiments, such as systematic study of dependence of apparent power \( \gamma \) on gene expression level, or other measures of protein concentration, will help clarify whether the static model or gene duplication mechanism is mainly responsible for the observed scale free nature of PPI networks.

**Acknowledgment**

This work is supported by NIH.

**Glossary**

**Protein–protein interaction network.** A network of many types of proteins of an organism; each type of protein is a node in the network. Two nodes are labeled as linked if the two types of proteins can interact with each other with sufficient affinity.

**Degree.** The number of links a node has in the network. If a node in the protein–protein interaction network has degree \( k \) then this protein can interact with \( k \) other types of proteins.

**Scale free network.** In such a network, the number of nodes with degree \( k \) decreases with \( k \), and the dependence is a power-law function.

**Yeast two hybrid.** A molecular biology technique used to discover protein–protein interactions by testing for physical interaction/binding between two proteins, respectively. This technique is able to test interactions between a large number of proteins rapidly (so-called high-throughput screening).

**Sensitivity in detecting interactions.** Only strong enough interactions between proteins are identified as ‘interacting’ pairs. If the sensitivity in detection becomes higher, slightly weaker interactions becomes detectable, and more interactions are detected.

**Surface hydrophobicity.** The fraction of hydrophobic amino acids among the amino acids on the surface of a protein. If hydrophobic amino acids are buried either in the formation of a protein or in the formation of a protein–protein complex, they are no longer in contact with water, and thus lower the total free energy. The hydrophobic effect is important in the interaction of proteins, especially in non-functional interactions.

**References**

[1] Albert R and Barabasi A-L 2002 Statistical mechanics of complex networks Rev. Mod. Phys. 74 47
[2] Xenarios I, Rice D W, Salwinski L, Baron M K, Marcotte E M and Eisenberg D 2000 DIP: the database of interacting proteins Nucl. Acids Res. 28 289
[3] Uetz P et al 2000 A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae Nature 403 623
[4] Ito T, Tashiro K, Muta S, Ozawa R, Chiba T, Nishizawa M, Yamamoto K, Kuhara S and Sakaki Y 2000 Toward a protein–protein interaction map of the budding yeast: a comprehensive system to examine two-hybrid interactions in all possible combinations between the yeast proteins Proc. Natl Acad. Sci. USA 97 1143
[5] Rual J F et al 2005 Towards a proteome-scale map of the human protein–protein interaction network Nature 437 1173
Lim J et al 2006 A protein–protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration Cell 125 801

[6] Deeds E J, Ashenberg O and Shakhnovich E I 2006 A simple physical model for scaling in protein–protein interaction networks Proc. Natl Acad. Sci. USA 103 311

[7] Shi Y, Miller G, Qian H and Bomsztyk K 2006 Free-energy distribution of binary protein–protein binding suggests cross-species interactome differences Proc. Natl Acad. Sci. USA 103 11527

[8] Jeong H, Mason S, Barabsi A-L and Oltvai Z N 2001 Lethality and centrality in protein networks Nature 411 41

[9] Wagner A 2001 The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes Mol. Biol. Evol. 18 1283

[10] Yook S Y, Oltvai Z N and Barabsi A-L 2004 Functional and topological characterization of protein interaction networks Proteomics 4 928

[11] Rzhetsky A and Gomez S M 2001 Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome Bioinformatics 17 988

Qian J, Luscombe N M and Gerstein M 2001 Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model J. Mol. Biol. 313 673

Bhan A, Galas D J and Dewey T G 2002 A duplication growth model of gene expression networks Bioinformatics 18 1486

Sole R V, Pastor-Satorras R, Smith E and Kepler T B 2002 A model of large-scale proteome evolution Adv. Complex Syst. 5 43

Vazquez A, Flammini A, Maritan A and Vespignani A 2003 Modeling of protein interaction networks ComPlexUs 1 38

[12] Kim J, Krapivsky P L, Kahng B and Redner S 2002 Infinite-order percolation and giant fluctuations in a protein interaction network Phys. Rev. E 66 055101

Pastor-Satorras R, Smith E and Sole R V 2003 Evolving protein interaction networks through gene duplication J. Theor. Biol. 222 199

Ispolatov I, Krapivsky P L and Yuryev A 2005 Duplication-divergence model of protein interaction network Phys. Rev. E 71 061911

[13] Barabsi A-L and Oltvai Z N 2004 Network biology: understanding the cell functional organization Nature Rev. Genetics 5 101

[14] Janin J 1995 Protein–protein recognition Prog. Biophys. Mol. Biol. 64 145

[15] Tamura A and Privalov P L 1997 The entropy cost of protein association J. Mol. Biol. 273 1048

[16] Estojak J, Brent R and Golemis E A 1995 Correlation of 2-hybrid affinity data with in-vitro measurements Mol. Cell. Biol. 15 5820

[17] Masuda N, Miwa H and Konno N 2004 Analysis of scale-free networks based on a threshold graph with intrinsic vertex weights Phys. Rev. E 68 036124

[18] Boguna M and Pastor-Satorras R 2003 Class of correlated random networks with hidden variables Phys. Rev. E 68 036112

[19] Deane C M, Salwinski L, Xenarios I and Eisenberg D 2002 Protein interactions—two methods for assessment of the reliability of high throughput observations Mol. Cell. Proteomics 1 349

[20] Caldarelli G, Capocci A, De Los Rios P and Muoz M A 2002 Scale-free networks from varying vertex intrinsic fitness Phys. Rev. Lett. 89 258702

[21] Balcan D and Erzan A 2004 Random model for RNA interference yields scale free network Euro. Phys. J. B 38 253