2008－MPS－72 (31)
2008－BIO－15 (31)
2008／12／18

タンパク質間相互作用ネットワークにおける相互作用ドメイン対の確率的選択に基づくべき乗分布のモデル化

ホセ・ナチェル[†]　林田 守広[‡]　阿久津 達也[‡]

[†]はこだて未来大学 システム情報科学部 複雑系科学科

[‡]京都大学 化学研究所 バイオインフォマティクスセンター

概要 タンパク質間相互作用ネットワークの次数分布は、多くの生物種に対して負のべき乗則に従うことが知られている。また相互作用のモデルとしては、タンパク質の部分構造であるドメインを用いて、二つのタンパク質が相互作用するとき、かつそのときに限り、二つのタンパク質に含まれているドメイン対のうち少なくとも一つが相互作用するというモデルが提案されている。本研究ではある生物種について各タンパク質に対するドメイン組成が与えられたもとで、相互作用するドメイン対を確率的に一様選択するモデルを考察する。相互作用するドメイン対の数が少ないうちは、タンパク質間相互作用ネットワークの次数分布はべき乗則に従うが、多くなると正規分布に近い分布となることが数理的に導かれる。UniProt タンパク質データベースのヒトのドメイン組成データに適用することで、実際のデータに対してもこの現象が観測されることを確認した。

Modeling Power-law Distribution in Protein-protein Interaction Networks based on Random Selection of Interacting Domain Pairs

J.C. NACHER[†], Morihiro HAYASHIDA[‡], and Tatsuya AKUTSU[‡]

[†]*Department of Complex Systems, Future University-Hakodate*

[‡]*Bioinformatics Center, Institute for Chemical Research, Kyoto University*

Abstract We propose a model for protein-protein interaction networks that reveals the emergence of two possible topologies. We show that depending on the number of randomly selected interacting domain pairs, the connectivity distribution follows either a scale-free distribution, even in the absence of the preferential attachment, or a normal distribution. This new approach only requires an evolutionary model of proteins (nodes) but not for the interactions (edges). The edges are added by means of random interaction of domain pairs. As a result, this model offers a new mechanistic explanation for understanding complex networks with a direct biological interpretation because only protein structures and their functions evolved through genetic modifications of amino acid sequences. These findings are supported by numerical simulations using *H. sapiens* protein domain data from UniProt database.

1 Introduction

Understanding of complex interactions at scales from molecular level to large ecosystems is a key challenge in life science. Recently, the development of new technologies together with high-throughput experiments in DNA microarrays, proteomics and metabolomics has led to a massive accumulation of biological data in an effort to analyze and unravel the complex biological phenomena that take place in a cell. Interestingly, most biological networks such as metabolic networks and protein-protein interaction networks were also classified as fat-tail, scale-free-like networks [1].

Proteins are molecules assembled from amino-acids using information present in genes and perform many critical functions in cells. The high

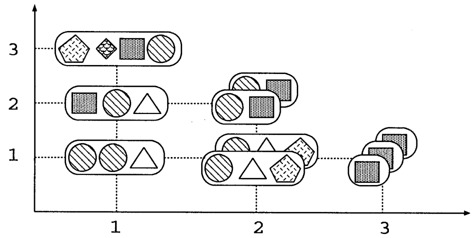

Figure 1: Domain pattern distribution. Horizontal axis: the number of proteins with the same domain pattern. Vertical axis: the number of different domain patterns.

complexity of the protein structure allows a hierarchical classification composed of fundamental interacting units defined as domains [2]. A protein domain can be defined as a building block of the entire protein molecule that is functionally and structurally independent. Proteins consist of one or more domains with different properties [2].

In this work, we propose an alternative construction of scale-free networks inspired from biological systems, where the preferential attachment is not explicitly required. In particular, by considering the PPI networks, the relationship between the emergence of the scale-free topology and the number of interacting domains is investigated. As a main result we show that, depending on the number of interacting domain pairs, the PPI networks can develop two fundamentally different topologies. In the first regime, we found that when a relatively small number of interacting domain pairs was selected with uniformly random probability, the degree distribution of the PPI network followed a power-law distribution with the exponent in agreement with the observed experimental data in several organisms. In the second regime, when the number of interacting domain pairs was relatively large the degree distribution approximately followed the normal distribution. This model offers a new mechanistic explanation for understanding complex networks with a direct biological interpretation [3].

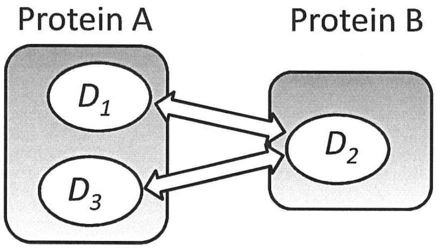

Figure 2: Protein interaction model. Protein A (B) consists of domains D_1 and D_3 (D_2). If proteins A and B interact, among pairs of domains, (D_1, D_2) and (D_3, D_2), at least a pair interacts.

2 Theoretical Analysis

Protein functions are closely related to the interaction between proteins. However, the elemental units responsible for the interaction are the protein domains which physically interact with one another to execute the cell functions. Most proteins comprise one or two domains, however some proteins may contain up to several domains.

Several researches have recently investigated the PPI networks in different organisms [4] and a variety of models for rebuilding their scale-free topology have been suggested [1, 5]. In addition, a few studies reported that power-law distributions are also present in protein domain networks [5]. We assume that each protein consists of one domain. Let n_D denote the number of proteins having domain D. From previous works [6], we can claim that the distribution of protein domains follows:

$$P(n_D = k) \quad \propto \quad k^{-\gamma} \tag{1}$$

with exponent $\gamma \approx 2$. This distribution indicates the probability to find a specific domain pattern k copies of which exist in the organism (see Fig. 1).

We consider a simple model for protein-protein interaction [7]. The main concept behind this model is that if two domains D_1 and D_2 interact, proteins having domain D_1 and proteins having domain D_2 should interact (see Fig. 2).

We select an interacting domain pair with uniformly random probability among all the possible domain pairs, and connect the corresponding protein pairs by an edge. The process is repeated N times. Therefore, the total number of interacting domain pairs in the network is equal to N.

We first describe in detail the case of $N = 1$. Let L be the total number of different domains. We assume that the domain distribution follows a power law $k^{-\gamma}$. To be precise, we assume that there exist $ak^{-\gamma}$ types of domains each of which has k paralogous proteins. Furthermore, we assume that the maximum k is bounded by K. This assumption is reasonable since the number of paralogous proteins is finite. Then, the total number of different domains L is approximated as follows,

$$\begin{aligned} L = \int_1^K ak^{-\gamma}dk &= \frac{a}{1-\gamma}\left[k^{1-\gamma}\right]_1^K \\ &= a\left[\frac{K^{1-\gamma}-1}{1-\gamma}\right]. \end{aligned} \tag{2}$$

It is noted that we approximated $\sum_1^K ak^{-\gamma}$ by $\int_1^K ak^{-\gamma}dk$ in order to obtain a simple analytical form. Thus, we have

$$\frac{a}{L} \approx \frac{1-\gamma}{K^{1-\gamma}-1}. \tag{3}$$

Suppose that domains A and B are selected as an interacting domain pair. The probability that such a pair is selected is

$$P(n_A = l)\cdot P(n_B = m) = \left(\frac{al^{-\gamma}}{L}\right)\left(\frac{am^{-\gamma}}{L}\right). \tag{4}$$

Then, we will have l proteins with degree m, and m proteins with degree l. Thus, the expected number of proteins having degree m is approximated by

$$\begin{aligned} P(n_B = m)\cdot E[n_A] &\approx \frac{am^{-\gamma}}{L}\int_1^K k\cdot\frac{al^{-\gamma}}{L}dk \\ &= \left(\frac{a}{L}\right)^2\cdot m^{-\gamma}\cdot\frac{1}{2-\gamma}\left[k^{2-\gamma}\right]_1^K \\ &= \left(\frac{1-\gamma}{K^{1-\gamma}-1}\right)^2\cdot\left(\frac{K^{2-\gamma}-1}{2-\gamma}\right)\cdot m^{-\gamma} \end{aligned} \tag{5}$$

where $\gamma \neq 2$, and $E[Z]$ means the expected value of Z. To be precise, we should calculate $2\cdot P(n_B = m)\cdot E[n_A]$ since we should also consider the case of $n_A = m$. However, it is still an approximation because the case of $n_A = n_B = m$ is counted twice. Since we are interested in asymptotic behavior, we use the above form.

For the case of $\gamma = 2$, the expected number is approximated by

$$\begin{aligned} P(n_B = m)\cdot E[n_A] &\approx \frac{am^{-\gamma}}{L}\int_1^K \frac{al^{-1}}{L}dk \\ &= \left(\frac{1-\gamma}{K^{1-\gamma}-1}\right)^2\cdot(\ln(K))\cdot m^{-\gamma}. \end{aligned} \tag{6}$$

If N is small compared with the total number of domains, it is expected that each domain interacts with at most one other domain. Therefore, we can still approximate the expected number of proteins having degree m by

$$N\cdot\left(\frac{1-\gamma}{K^{1-\gamma}-1}\right)^2\cdot\left(\frac{K^{2-\gamma}-1}{2-\gamma}\right)\cdot m^{-\gamma}. \tag{7}$$

If N is large, the situation changes. In such a case, one domain A interacts with domains $B_1, B_2, \cdots, B_M$. Then, the degree of each protein consisting of A will be $n_{B_1}+n_{B_2}+\cdots+n_{B_M}$ where n_{B_i} denotes the number of proteins consisting of domain B_i and we assume that $B_i \neq B_j$ for $i \neq j$. Then, the expected number of proteins having degree m is approximated by

$$\begin{aligned} \sum_{m_1+\cdots+m_M=m} & P(n_{B_1} = m_1)\cdot P(n_{B_2} = m_2) \\ & \cdots P(n_{B_M} = m_M)\cdot\int_1^K k\cdot\frac{ak^{-\gamma}}{L}dk \end{aligned} \tag{8}$$

In addition, it should be noted that if N is very large, then M is large. For large M, m follows the normal distribution (regardless of distributions of n_{B_i}) by the *central limit theorem.* Since $\int_1^K k\cdot\frac{ak^{-\gamma}}{L}dk$ can be considered as a constant (depending on K and γ), the degree distribution is expected to follow the normal distribution.

3 Experimental Results

We have used the Database of Interacting Proteins (DIP) for constructing the PPI networks of several organisms. We report here the results for *C. elegans, S. cerevisiae, D. melanogaster, E. coli, H. pylori, M. musculus* and *H. sapiens* organisms. In all cases, a power-law distribution was found with exponent close to 2 (see Fig. 3).

Moreover, we have investigated the threshold of interacting domain pairs by using

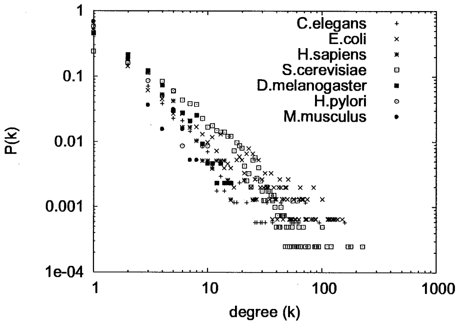

Figure 3: Degree distribution $P(k)$ of PPI networks for several organisms from DIP database.

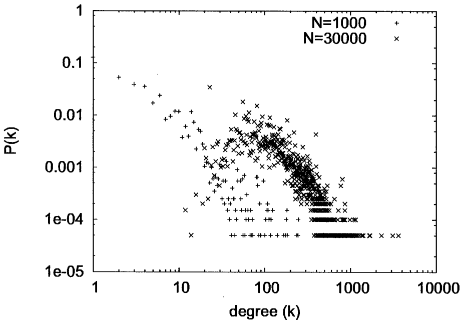

Figure 4: Degree distribution $P(k)$ of PPI networks for the number of interacting domain pairs $N = 1000, 30000$ simulated using *H. sapiens* protein domains from UniProt database.

UniProt database for protein sequences and Pfam database for domains. We constructed proteins with real composition of domains. Next, we selected two interacting domains with uniformly random probability, and connected the corresponding protein pair by an edge. We repeated the process for different values of N. The results are shown in Fig. 4.

In order to assess that the actual PPI network is consistent with the assumption of few interactions per domain, we computed the ratio between the number of *H. sapiens* domains obtained from Pfam database and the number of known protein-protein interactions according to the DIP database. This ratio gives a small value of 0.4, which is compatible with our assumption.

4 Concluding remarks

We have proposed a new model for protein-protein interaction networks that generates scale-free networks even in the absence of the preferential connectivity. We have shown that the PPI networks can exhibit two fundamentally different topologies according to the number of interacting domain pairs. These findings were consistent with the results of biological experimental data. As a result, this approach offers a new mechanistic explanation for understanding PPI networks with a direct biological interpretation. It only requires an evolutionary model of proteins (nodes) and information of domain structure but not for the interactions (edges). The concept behind this approach is that only protein structures and their functions evolved through genetic modifications of amino acid sequences and interactions are just a consequence of the evolved domain structure. In summary, our results give new conceptual insights into the origin of the observed scale-free topology in PPI networks and may open new directions for understanding the emergence of power-laws in different biological contexts.

References

[1] H. Jeong, S. Mason, A.-L. Barabási, Z.N. Oltvai. *Nature*, 411, 41–42, 2001.

[2] R.F. Doolittle. *Ann. Rev. Biochem.*, 64, 287–314, 1995.

[3] J.C. Nacher, M. Hayashida, T. Akutsu. Emergence of scale-free distribution in protein-protein interaction networks based on random selection of interacting domain pairs. *BioSystems*, in press.

[4] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori and Y. Sakaki. *Proc. Nat. Acad. Sci USA*, 98, 4569–4574, 2001.

[5] J. Qian, N.M. Luscombe and M.B. Gerstein. *J. Mol. Biol.*, 313, 673–681, 2001.

[6] J.C. Nacher, M. Hayashida and T. Akutsu. *Physica A*, 367, 538–562, 2006.

[7] E. Sprinzak and H. Margalit. *J. Mol. Biol.*, 311, 681–692, 2001.