A Bipartite Graph Based Model of Protein Domain Networks

J.C. Nacher[1], T. Ochiai[2], M. Hayashida[3], and T. Akutsu[3]

[1] Department of Complex Systems, Future University-Hakodate, Japan
[2] Faculty of Engineering, Toyama Prefectural University, Japan
[3] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan

Abstract. Proteins are essential molecules of life in the cell and are involved in multiple and highly specialized tasks encoded in the amino acid sequence. In particular, protein function is closely related to fundamental units of protein structure called *domains*. Here, we investigate the distribution of kinds of domains in human cells. Our findings show that while the number of domain types shared by k proteins follows a scale-free distribution, the number of proteins composed of k types of domains decays as an exponential distribution. In contrast, previous data analyses and mathematical modeling reported a scale-free distribution for the protein domain distribution because the relation between kinds of domains and the number of domains in a protein was not considered. Based on this finding, we have developed an evolutionary model based on (1) growth process and (2) copy mechanism that explains the emergence of this mixing of exponential and scale-free distributions.

Keywords: Growing networks, protein domains, scale-free networks.

1 Introduction

The complexity of a wide variety of systems as the metabolic pathways, protein interaction networks, social relationships or transportation systems, can be investigated in terms of networks where the elementary units of the system are represented by nodes and their interactions as edges. In recent years, empirical analyses and theoretical modeling of networks have rapidly become a highly-active research area, uncovering the existence of unexpected organizing principles and similarities in real systems, with sizes ranging from hundreds to billions of nodes [1,2,3,4]. Whereas at a global level, real complex networks deviate from predictions of random graph theory [5] and display a scale-free and hierarchical organization [6,7], a complementary perspective at a local level reveals a significant prevalence and variety of highly characteristic patterns of interactions, such as motifs, modules, cliques and communities with specific functional tasks [8,9,10].

Recent experimental efforts in proteomics have generated a massive amount of newly sequenced proteins, molecular structures, foldings mechanisms as well as interacting domains data. Using this information, protein interaction maps have

J. Zhou (Ed.): Complex 2009, Part I, LNICST 4, pp. 525–535, 2009.

been constructed and analyzed. Although these networks are still incomplete, it allows for the first time the study of the large-scale structure of functional interactions within a cell for a variety of organisms. These analyses have shown that cellular networks such as metabolic pathways, protein-protein interaction networks can be classified as scale-free networks [2,11,12].

A protein is a long chain of amino acids encoding important cellular functions. Each protein can be composed of one or more protein *domains* that represent fundamental building blocks with specific structural and functional features. However, a different classification allows the definition of *protein modules* considered as a more compact structural unit in a protein with a length in the range of 20-40 residues [13,14].

In this work, we will focus on proteins composed of domains as fundamental building blocks, in particular we have analyzed the empirical data corresponding to proteins and interacting domains using human proteome information collected from the UniProt[26] (UniProtKB/Swiss-Prot Release 56.0 of 22-Jul-2008) and Integr8[27] (Release 84 constructed from UniProt 14.0) databases. We then investigate the distribution of kinds of domains in human cells. Our findings show that while the number of a domain type shared by k proteins follows a scale-free distribution, the number of proteins composed of k types of domains decays as an exponential distribution. This finding has not been reported before, as previous analyses [15,16,17] did not study the relation between kinds of domains and the number of domains in a protein.

This problem can be investigated using a bipartite graph whose nodes can be classified into two disjoint sets N (proteins) and M (domains) such that each edge connects a node in N and one in M [18]. For example, N_k indicates the number of protein with k edges if the protein is composed of k domains. Similarly, M_k denotes the number of domains with k edges if this domain is shared by k proteins.

Based on our empirical findings on the dissimilar nature of N_k and M_k distributions, we have developed an evolutionary model using *the rate equation approach*, first suggested by Krapivsky et al.[19], that explains the emergence of this mixing of exponential and scale-free distributions. The model requires (1) growth process and (2) copy mechanism. We first use the rate equation approach for constructing the discrete mathematical equations corresponding to bipartite graphs. We then transform them into differential equations and solve them using the continuum limit.

2 Theoretical Model and Experimental Results

2.1 Theoretical Model

Let us consider a bipartite graph, whose nodes are divided into two disjoint sets N (proteins) and M (domains), and only connections between two nodes in different sets N and M are allowed as shown in Fig. 1. In what follows, N_k denotes the number of proteins (square) with k edges (domains). Similarly, M_k denotes the number of domains (circle) shared by k proteins. Furthermore, we consider that each domain represents a specific kind of domain. Therefore, two domains corresponding to the same type of domain are not allowed. This is a crucial point in our analysis. Then, we propose an algorithm that builds a power-law distribution for M_k and an exponential distribution for N_k.

1. The model is initialized with a same small number l of N-nodes and M-nodes. Each node from l_N and from l_M is connected by an edge, then the degree of all N-nodes and M-nodes is only one, where we have assumed $l{=}l_N{=}l_M$.
2. At time $t = 1$, with probability α_N, a randomly selected N-node is copied. Otherwise, with probability β_N, a new N-node is added. We then connect this new N-node to n_0 randomly selected M-nodes. In this process, $\alpha_N + \beta_N = 1$.
3. At the same time step, with probability α_M, a randomly selected M-node is copied. Otherwise, with probability β_M, a new M-node is added. We then connect this new M-node to m_0 randomly selected N-nodes. As in the above process, $\alpha_M + \beta_M = 1$.
4. Steps (2) and (3) are iterated t times until a desired number of nodes is generated. At the end, the network will consist of the same number $t{+}l$ of N-nodes and M-nodes.

Therefore, our model of growing bipartite networks is composed of two main ingredients: (1) growth process and (2) copy mechanism. Fig. 1 illustrates these mechanisms for both sets of nodes. From this algorithm, we construct the rate equation for the bipartite network. The rate equation approach was first introduced in network science by Krapivsky et al., [19] and applied to the study of percolation [20], protein evolution networks [21] and citation networks as well as used in extensive theoretical analyses [22]. Furthermore, it has also been applied to the computation of the node degree correlations [23]. On the other hand, models applied to bipartite graphs are much less numerous and only a very few works have addressed the issue [25]. See also the review on rate equation approach for further information [24]. By following our algorithm, the rate equation for the time evolution of the number of nodes with degree k in both sets of nodes N_k and M_k can be written as:

$$\frac{dN_k}{dt} = \alpha_M\left(\frac{k-1}{M(t)}N_{k-1} - \frac{k}{M(t)}N_k\right) + \beta_M\left(\frac{m_0}{N(t)}N_{k-1} - \frac{m_0}{N(t)}N_k\right) + \alpha_N\frac{N_k}{N(t)} + \beta_N\delta_{kn_0} \quad (1)$$

$$\frac{dM_k}{dt} = \alpha_N\left(\frac{k-1}{N(t)}M_{k-1} - \frac{k}{N(t)}M_k\right) + \beta_N\left(\frac{n_0}{M(t)}M_{k-1} - \frac{n_0}{M(t)}M_k\right) + \alpha_M\frac{M_k}{M(t)} + \beta_M\delta_{km_0} \quad (2)$$

where $N(t) = t + l$ and $M(t) = t + l$ are the total number of N-nodes and M-nodes at time t, respectively. In these equations, δ_{kn_0} and δ_{km_0} indicate the

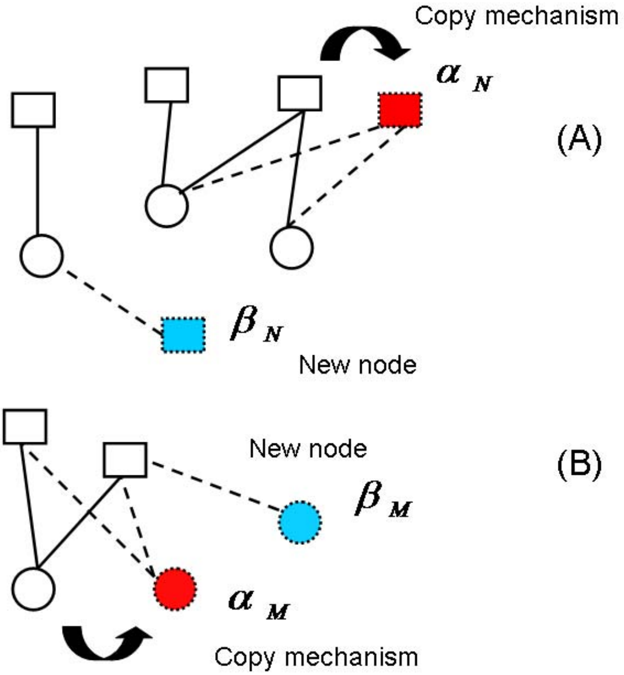

Fig. 1. Description of growth and copy mechanisms in our model for bipartite graphs. Squares (proteins) (A) and circles (kinds of domains) (B) can be added and copied. One protein connected to one (two) kind of domains indicates that this protein consists of one (two) kinds of domains.

contribution of a new node connected to already existing n_0 and m_0 nodes. Next, by introducing the probability distribution $n_k = N_k/N(t)$ and $m_k = M_k/M(t)$, we obtain

$$\frac{d((t+l)n_k)}{dt} = \alpha_M((k-1)n_{k-1} - kn_k) + \beta_M m_0(n_{k-1} - n_k) \\ + \alpha_N n_k + \beta_N \delta_{kn_0} \qquad (3)$$

$$\frac{d((t+l)m_k)}{dt} = \alpha_N((k-1)m_{k-1} - m_k) + \beta_N n_0(m_{k-1} - m_k) \\ + \alpha_M m_k + \beta_M \delta_{km_0} \qquad (4)$$

In the limit $t \to \infty$, we obtain the equation for the stationary distribution:

$$n_k = \alpha_M((k-1)n_{k-1} - kn_k) + \beta_M m_0(n_{k-1} - n_k) \\ + \alpha_N n_k + \beta_N \delta_{kn_0} \qquad (5)$$

$$m_k = \alpha_N((k-1)m_{k-1} - km_k) + \beta_N n_0(m_{k-1} - m_k) \\ + \alpha_M m_k + \beta_M \delta_{km_0} \qquad (6)$$

In the continuum k limit, these equations take the following form:

$$n_k = -\frac{d}{dk}\{(\alpha_M k + \beta_M m_0)n_k\} + \alpha_N n_k \tag{7}$$

$$m_k = -\frac{d}{dk}\{(\alpha_N k + \beta_N n_0)m_k\} + \alpha_M m_k \tag{8}$$

Then, from the last equation, we obtain

$$m_k \propto (\alpha_N k + \beta_N n_0)^{-\frac{1-\alpha_M+\alpha_N}{\alpha_N}} \tag{9}$$

In the limit for large k ($k \to \infty$),

$$m_k \propto k^{-\frac{1-\alpha_M+\alpha_N}{\alpha_N}} \tag{10}$$

$$\sim k^{-\frac{1+\alpha_N}{\alpha_N}} \tag{11}$$

where we have used $\alpha_M \sim 0$ in the last equation. Therefore, the degree distribution for M-nodes (number of domains shared by k proteins) obeys a power-law.

On the other hand, from Eq. (7), we can write

$$n_k \propto (\alpha_M k + \beta_M m_0)^{-\frac{1-\alpha_N+\alpha_M}{\alpha_M}} \tag{12}$$

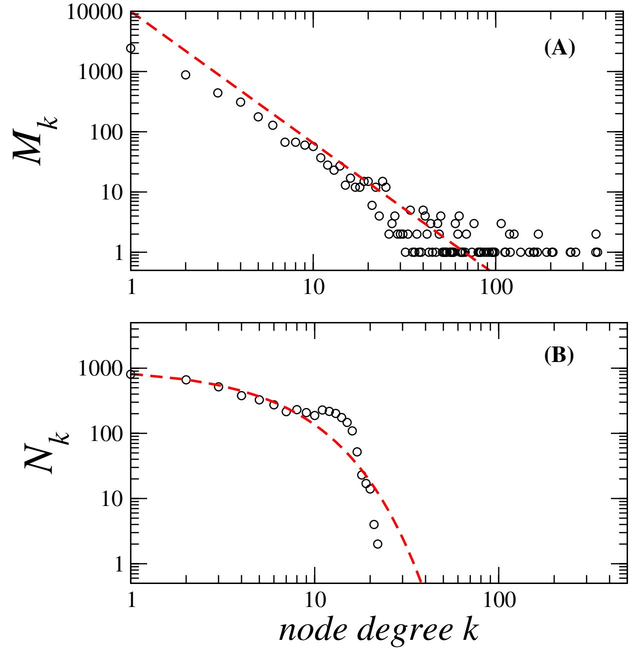

Fig. 2. Theoretical results (red dashed line) of the model and computational simulation (black circles) with $\alpha_N = 0.8$ and $\alpha_M = 0.05$, (A) Power-law distribution with degree exponent 2.18. (B) Exponential decay.

In particular, in the limit $\alpha_M \to 0$,

$$n_k \propto e^{-\frac{\beta_N}{m_0}k} \tag{13}$$

Therefore, we obtain that the degree distribution for N-nodes (number of proteins composed of k types of domains) obeys a exponential decay. We highlight main features of the model as follows:

1. By using a bipartite growing network model composed of copy and random attachment processes with supression of copy of M-nodes (types of domains) ($\alpha_M \sim 0$), we reproduce the observed distributions of power-law and exponential decay of several real networks composed of two types of nodes.
2. $\alpha_M \sim 0$ implies that M-nodes (kinds of domains) are unlikely copied, if compared to N-nodes. This is meaningful because kinds of domains are unique and cannot be duplicated by definition. This asymmetry in the growing mechanisms is fundamental to derive the observed mixing distributions.

2.2 Model Simulation

When both parameters α_N, α_M take values close to one simultaneously, a so-called "*giant fluctuation*" occurs [20]. It indicates that a model that only includes the copy mechanism (i.e., a model configuration with α_N, α_M close to one) does not behave well and the resulting distribution is singular and resembles the sum of delta functions in the large k region. Therefore, the contribution of a "*noise*" term is needed. While in Krapivsky et al. [20], the noise effect is introduced through a mutation-like mechanism, in our model the noise contribution comes from the random attachment mechanism when at least one of the parameters β_N, β_M is non zero.

Thus, with the exception of the case α_N, α_M close to one, we show the computational simulation of our model in the following three figures. Fig. 2 shows the degree distribution when the copy mechanism of M-nodes is supressed and copies of N-nodes are allowed. The simulated distribution M_k obeys a power-law, while the other distribution N_k obeys an exponential decay. This copy mechanism supression of M-nodes (domains) is meaningful because we are considering kinds of domains in our problem, and a kind of domain should be unique by definition. Next, Fig. 3 shows the case when both N-nodes and M-nodes are allowed to be copied. Then, both simulated distribution N_k and M_k obey a power-law. Finally, we consider the case when N-nodes and M-nodes have the copy mechanism supressed. As shown in Fig. 4, both simulated distribution N_k and M_k follow an exponential decay. Here we note that simulation results show the degree distribution N_k and M_k, instead of probability distribution n_k and m_k.

2.3 Experimental Results

We have performed an empirical analysis using human proteins collected from the UniProt[26] (UniProtKB/Swiss-Prot Release 56.0 of 22-Jul-2008) and Integr8[27]

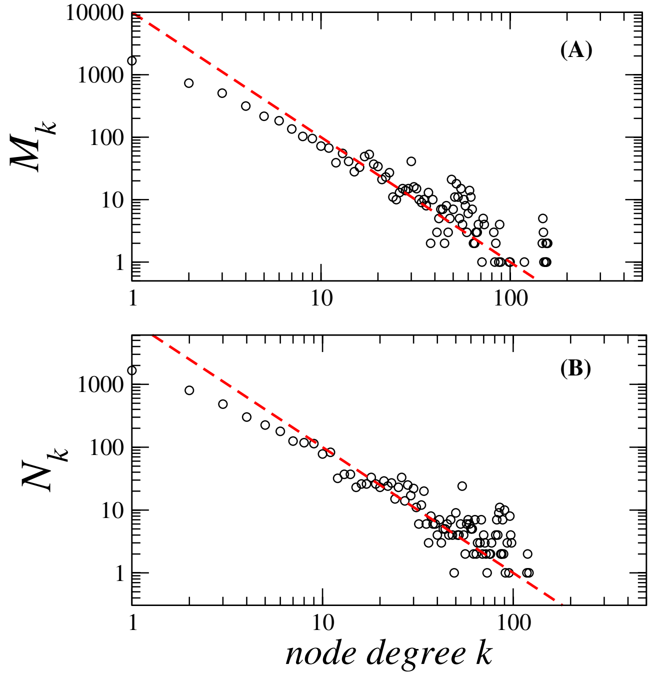

Fig. 3. Theoretical results (red dashed line) of the model and computational simulation (black circles) with $\alpha_N = 0.5, \alpha_M = 0.5$ Both figures (A) and (B) show a power-law distribution with degree exponent 2

(Release 84 constructed from UniProt 14.0) databases. Integr8 database provides non-redundant set of UniProt entries representing each complete proteome. We have obtained Pfam[28] domains for each protein from the DR line of UniProt format.

Fig. 5 shows an exponential distribution for the number of kinds of domains in a protein. Human proteins were downloaded from the UniProt and Integr8 databases. Next, Fig.6 shows the distribution of the number of domain types shared by k human proteins in the UniProt and Integr8 databases. In this case, we can observed that the distribution follows a scale-free distribution. These results are in agreement with the predictions of our evolutionary model shown in Fig. 2. It is worth noticing that our model generates the same number of domains as proteins because the number of M-nodes and N-nodes is the same by construction. However, we have also analyzed and computed this case of asymmetric growth in the number of nodes. Although we omit the main derivation for space reasons, our results show that not only the mixing of scale-free and exponential distributions is conserved but also the exponent degree of power-law is kept invariant under the asymmetric growth. To be precise, only the exponent of the exponential decay distribution depends on the asymmetric growth.

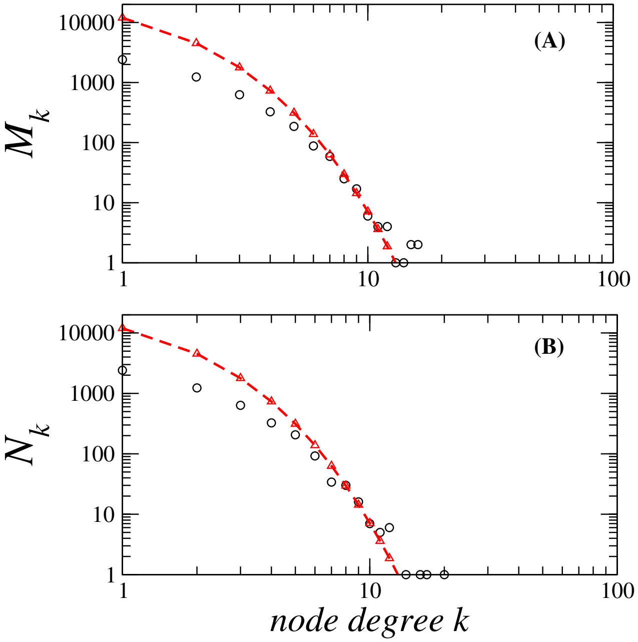

Fig. 4. Theoretical results of the model (red dashed line) and computational simulation (black circles) with $\alpha_N = 0.05, \alpha_M = 0.05$ Both (A) and (B) distributions show an exponential decay

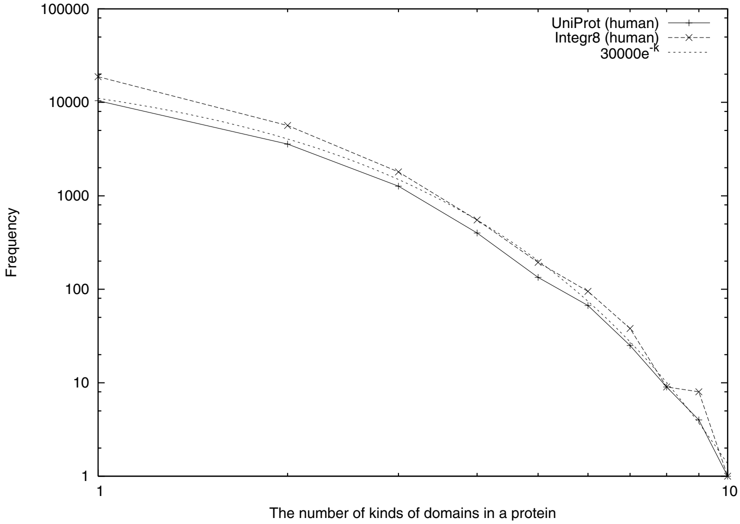

Fig. 5. The distribution of the number of kinds of domains in a protein for human proteome space. Data collected from UniProt and Integr8 databases.

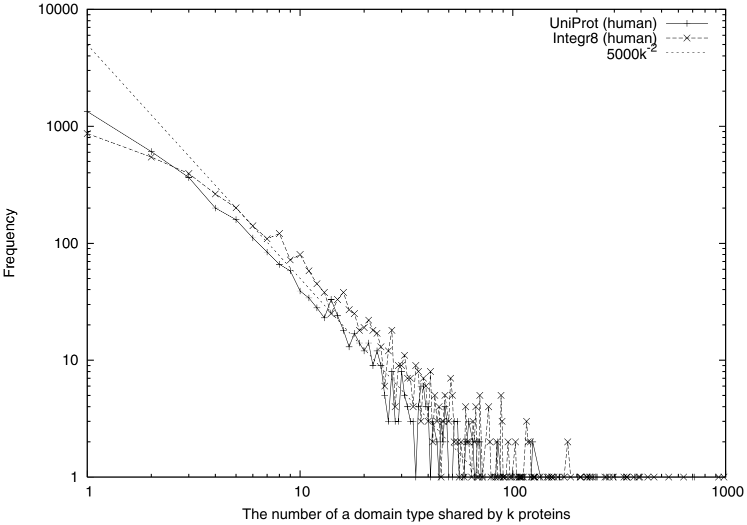

Fig. 6. The distribution of the number of domain types shared by k proteins for human proteome space. Data collected from UniProt and Integr8 databases.

3 Conclusion

In summary, we have investigated the distribution of protein and kinds of domains in human cells. Our results indicate that while the number of a domain type shared by k proteins follows a scale-free distribution, the number of proteins composed of k types of domains decays as an exponential distribution. It is worth noticing that previous data analyses and mathematical modeling reported a scale-free distribution for the protein domain distribution because the relation between kinds of domains and the number of domains in a protein was not considered.

Based on this finding, we have developed a simple evolutionary model based on (1) growth process and (2) copy mechanism. This model based on the rate equation approach for computing bipartite graphs does not only predict the observed asymmetry in the distribution of protein composed of k unique domains and number of domains shared by k proteins but also predicts the degree exponent for the power-law in the vicinity of value 2.

Furthermore, the model elucidates that the supression of copy mechanisms in one set of nodes is enough to create the mixture distribution and the symmetry breaking. This copy mechanism supression of M-nodes (domains) is reasonable because we are considering kinds of domains in this problem, and a kind of domain can be considered unique by definition.

Of particular interest for future work will be to extend the current model of bipartite graphs to be applied to other biological systems like gene regulatory networks where nodes represent operons encoding transcriptional factors

(TFs) and the target genes [29]. As this transcriptional network also exhibits an assymetric distribution for outgoing and incoming degrees, similar ideas shown in the current model could be helpful for its investigation.

References

1. Dorogovtsev, S.N., Mendes, J.F.F.: Evolution of Networks: From Biological Nets to the Internet and WWW. Oxford University Press, Oxford (2003)
2. Barabási, A.-L., Oltvai, Z.N.: Network Biology: Understanding the Cells's Functional Organization. Nature Reviews Genetics 5, 101–113 (2004)
3. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University Press, Cambridge (2004)
4. Newman, M., Barabási, A.-L., Watts, D.J.: The Structure and Dynamics of Networks. Princeton University Press, Princeton (2007)
5. Erdös, P., Rényi, A.: On the Evolution of Random Graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17–61 (1960)
6. Barabási, A.-L., Albert, R.: Emergence of Scaling in Random Networks. Science 286, 509–512 (1999)
7. Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L.: Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551–1555 (2002)
8. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U.: Network Motifs: Simple Building Blocks of Complex Networks. Science 298, 824–827 (2002)
9. Shen-Orr, S., Milo, R., Mangan, S., Alon, U.: Network Motifs in the Transcriptional Regulation Network of Escherichia coli. Nat. Genetics 31, 64–68 (2002)
10. Palla, G., Derenyi, I., Farkas, I., Vicsek, T.: Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature 435, 814–818 (2005)
11. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.-L.: The Large-scale Organization of Metabolic Networks. Nature 407, 651–654 (2000)
12. Jeong, H., Mason, S., Barabási, A.-L., Oltvai, Z.N.: Lethality and Centrality in Protein Networks. Nature 411, 41–42 (2001)
13. Go, M.: Modular Structural Units, Exons, and Function in Chicken Lysozyme. Proc. Natl. Acad. Sci. USA 80, 1964–1968 (1983)
14. Go, M.: Correlation of DNA Exonic Regions with Protein Structural Units in Haemoglobin. Nature 291, 90–92 (1981)
15. Wuchty, S.: Scale-free Behavior in Protein Domain Networks. Mol. Biol. Evo. 18, 1694–1702 (2001)
16. Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, E.V.: Birth and Death of Protein Domains: A Simple Model of Evolution Explains Power Law Behavior. BMC Evo. Biol. 2, 18 (2002)
17. Nacher, J.C., Hayashida, M., Akutsu, T.: Protein Domain Networks: Scale-free Mixing of Positive and Negative Exponents. Physica A 367, 538–552 (2006)
18. Newman, M.E.J., Strogatz, S.H., Watts, D.J.: Random Graphs with Arbitrary Degree Distributions and Their Applications. Phys. Rev. E 64, 026118 (2001)
19. Krapivsky, P.L., Redner, S., Leyvraz, F.: Connectivity of Growing Random Networks. Phys. Rev. Lett. 85, 4629 (2000)
20. Kim, J., Krapivsky, P.L., Kahng, B., Redner, S.: Infinite-order Percolation and Giant Fluctuations in a Protein Interaction Network. Phys. Rev. E. 66, 055101 (2002)

21. Ispolatov, I., Krapivsky, P.L., Yuryev, A.: Duplication-divergence Model of Protein Interaction Network. Phys. Rev. E. 71, 061911 (2005)
22. Krapivsky, P.L., Redner, S.: Organization of Growing Random Networks. Phys. Rev. E. 63, 066123 (2001)
23. Barrat, A., Pastor-Satorras, R.: Rate Equation Approach for Correlations in Growing Network Models. Phys. Rev. E 71, 036127 (2005)
24. Krapivsky, P.L., Redner, S.: Rate Equation Approach for Growing Networks. Lecture Notes in Physics 625, 3–22 (2003)
25. Ergün, G.: Human Sexual Contact Network as a Bipartite Graph. Physica A 308, 483–488 (2002)
26. The UniProt Consortium: The Universal Protein Resource (UniProt). Nucleic Acids Research 36, D190–D195 (2008)
27. Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., Apweiler, R.: Integr8 and Genome Reviews: Integrated Views of Complete Genomes and Proteomes. Nucleic Acids Research 33, D297–D302 (2005)
28. Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, J.S., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L., Bateman, A.: The Pfam protein families database. Nucleic Acids Research 36, D281–D288 (2008)
29. Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., Gerstein, M.: Genomic Analysis of Regulatory Network Dynamics Reveals Large Topological Changes. Nature 431, 308–312 (2004)