2006－BIO－4（10）
2006／2／10

タンパク質ドメインネットワークにおける混合スケールフリー次数分布

Jose C. Nacher　　林田　守広　　阿久津　達也

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　本稿においては、頂点がタンパク質に対応し、対応するタンパク質が同じドメインを持つ場合に頂点間に辺を引くという方法により構成した「タンパク質ドメインネットワーク」の次数分布について研究を行った。大腸菌からヒトまで６種類の生物のタンパク質データについて調べたところ、いずれの生物においても、指数-1 と指数+1 の二種類のべき乗分布を混合した分布が得られた。そこで、この混合分布を説明するために、突然変異とドメイン重複に基づく、タンパク質ドメイン構造の進化モデルを考え、その次数分布を解析した。その結果、指数-1 と指数+1 の混合分布が得られることを理論的に説明することにも成功した。

Scale-free mixing in protein domain networks

Jose C. Nacher, Morihiro HAYASHIDA and Tatsuya AKUTSU

Bioinformatics Center, Institute for Chemical Research, Kyoto University

Abstract　In this article, we present a theoretical model for studying the protein domain networks, where one node of the network corresponds to one protein and two proteins are connected if they contain the same domain. The resulting distribution of nodes with a given degree, k, shows not only a power-law with negative exponent $\gamma = -1$, but it resembles the superposition of two power-law functions, one with a negative exponent and another with a positive exponent $\beta = 1$. We call this distribution pattern "*scale-free mixing*", which is conserved among organisms from *Escherichia coli* to *Homo sapiens*. To explain the emergence of this superposition of power-laws, we propose a basic model with two main components: (1) mutation and (2) duplication of domains. Precisely, duplication gives rise to complete subgraphs (i.e., cliques) on the network, thus for several values of k a large number of nodes with degree k is produced, which explains the positive power-law branch of the degree distribution.

1　Introduction

Graph theory described the real-life networks with random distribution of nodes and edges ([1]) since the second half of the twentieth century. The degree distribution (i.e.,the probability $P(k)$ that a randomly chosen node has a number of edges (degree of node) k) of such random graph is a Poisson distribution, characterized by a peak at the average degree value $< k >$ and an exponential tail. However, at the dawn of twenty-first century, the availability of large databases allowed to uncover some organizational principles hidden in the large networks. In particular, the discovery of the small-world effect ([2]) and the emergence of scaling in random networks (i.e., scale-free distribution) ([3]) showed us that the topology of real large networks substantially differs from the network topology predicted by the random graph theory. While the small-world property shows that the mean distance between nodes increases not faster than logarithmically with the total number of nodes, the scaling in networks emerges from a statistical abundance of nodes with a large number of edges k (i.e.,

hub nodes) compared with the average degree value $< k >$. The resulting distribution is a scale-free distribution $P(k) \sim k^{\gamma}$, where γ is usually a negative real value between -1 and -4, and it is called the exponent of the power-law.

There are many examples of scale-free topology in non-biological systems as the Internet, electrical power grids, the World Wide Web, and airline routes ([4, 5]). On the biological side, the degree distribution of chemical compounds and chemical reactions (metabolic pathways) ([6, 7]), and the distribution of the gene expression level in cells (transcriptional organization) ([8, 9, 10]) are examples of scale-free organization. Furthermore, some hierarchical scale-free topologies ([4, 11, 12, 13]) were recently proposed for describing the organizational levels in such biological systems. Concerning proteins, some works have recently studied the protein interaction networks, revealing the scale-free behaviour of these networks ([14, 15]). In addition, a few works carried out some analyses about protein domains ([16, 17, 18]), where the scale-free organization was also found.

In the present work, we focus on the proteins, which play crucial roles in many subcellular processes. In particular, we study the organization and evolutionary origin of protein domain networks. A protein domain can be defined as a well-defined region within a protein that either performs a specific function or constitutes a stable structural unit (also known as building block) ([19]). In particular, our study contains the following main points:

First, we developed a model of protein domain networks based on (1) mutation of domains and (2) duplication of domains. In our network, each node is one protein and two proteins are connected if they contain the same domain. When the degree distribution $P(k)$ of the network is measured, we found that $P(k)$ resembles the superposition of two power-law functions, one with a negative exponent $\gamma = -1$ and another with a positive exponent $\beta = 1$. We call this distribution pattern "*scale-free mixing*". As we will explain later in detail, the origin of the power-law branch with positive exponent $\beta = 1$ emerges from the existence of complete subgraphs (i.e., cliques) of d proteins, where the degree of each protein will be $d-1$. Therefore, a scale-free distribution with positive exponent is found as $P(d-1) = d$. The emergence of this positive power-law branch of the degree distribution is our main result.

Second, we compare the results of our model with experimental data [21]. The results reveal the existence of the *scale-free mixing* pattern in the experimental data, and furthermore, it is conserved among organisms as *Escherichia coli, Saccharomyces cerevisiae, Arabidopsis thaliana, Drosophila melanogaster, Mus musculus, and Homo sapiens.*

The article is organized as follows. In Section 2, a theoretical model is presented to study the protein domain networks, and the scale-free mixing is found. In Section 3, the experimental evidence for the positive and negative power-laws in protein domain networks is shown. Section 4 discusses our results and the final section presents the conclusions. A complete version of this work can be found in [20].

2 Theoretical model

We propose a model of protein domain networks based on (1) mutation of domains and (2) duplication of domains. We will first generate the distribution of domains $P_D(k)$, and by using this distribution we will compute the value of the degree distribution of the protein domain network $P(k)$. Therefore, we remark that we only have one network in our model with degree distribution $P(k)$. The proposed model is as follows.

First, we assume that there are N proteins, each of which consists of only one domain. In addition, these domains are different to each other. Second, we repeat T times the following

steps: (1) with probability $(1-a)$, we create a new protein with a new domain. (2) Otherwise, we randomly select one protein and make a copy of it. In this model, the step (1) corresponds to mutation of one protein and the step (2) corresponds to protein duplication. This model is quite similar to the Barabási-Albert (BA) model [3] by considering the following: (1) A type of domain in our proposed model corresponds to a node in the BA network model. (2) The number of proteins with the identical domains in our proposed model corresponds to the degree of a vertex in the BA network model. (3) Creation of a new protein corresponds to the addition of a new node. Let i, k_i denote some domain (i-th domain) and the number of proteins consisting of i, respectively. Then, we can obtain

$$\frac{dk_i}{dt} = a\frac{k_i}{t} \tag{1}$$

because a copy of a protein is created with probability a at each time step, and there are t proteins at time step t. From this, the number of proteins consisting of i reads as:

$$k_i = c\left(\frac{t}{t_i}\right)^a \tag{2}$$

where c is an appropriate constant and t_i is the time when the i-th domain was first created. Then, as in the BA model, we have:

$$P_D(k) \propto k^{[-1-(1/a)]} \tag{3}$$

This equation gives us the domain pattern distribution $P_D(k)$. Precisely, $P_D(k)$ indicates the number of proteins consisting of the same domains and made of exactly k copies of proteins. We will illustrate this distribution by using two simple examples: (1) Let us focus on the case that the proteins are single domains (hereafter protein A means a protein of single domain A). By following our proposed model, after T iterations we can find that the protein A has k copies. However, we can also find that proteins B and C have k copies. Therefore, $P_D(k)$ gives a value of 3. (2) Suppose we have one single domain protein with three copies (i.e., $k = 3$).

On the other hand, and by following Eq. (3), we can easily control the exponent of this power-law (i.e., $\gamma = [-1-(1/a)]$) by changing the mutation rate a. We can see that for any mutation rate the value of γ is above 2. However, if the mutation rate is very small (which is plausible in a realistic biochemical process) [19], γ is close to 2. Furthermore, we have computed the experimental distribution of domain pattern by using the *UNIPROT-Swissprot* database for protein sequences and *InterPro, Pfam, and Smart* for domain databases. Our results showed that, from *E.coli* to *H.Sapiens* organisms, the scale-free distribution was generated and the predicted exponent of value 2 was obtained.

Emergence of scale-free mixing. Here, we can consider a network of proteins, where one node of the network is one protein and two proteins are connected if they contain the same domain. Therefore, by using Eq. (3) the degree distribution of this network should read as:

$$P(k) \propto k^{[-(1/a)]} \tag{4}$$

In Fig. 1, we show the results of the computer simulation of our model by using $a \sim 0.9$. We see a power-law with negative exponent $\gamma = -1$, as it is predicted by Eq. (4). In addition, our results show a series of parallel power-law distribution with positive exponent with a value of $\beta = 1$.

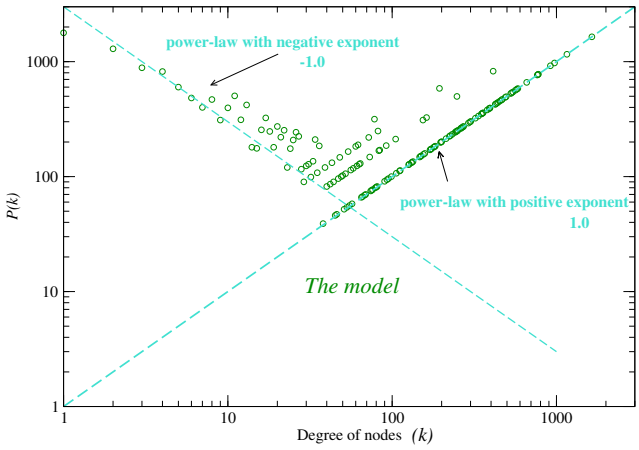

Figure 1: The results of our proposed model by considering only mutation and protein duplication mechanisms. The distribution of nodes with degree k resembles a superposition of two power-law distributions, one with positive exponent $\beta = 1$ and another negative exponent $\gamma = -1$. We call this type of distribution scale-free mixing. N=100 initial single domain proteins and up to T=50000 iterations.

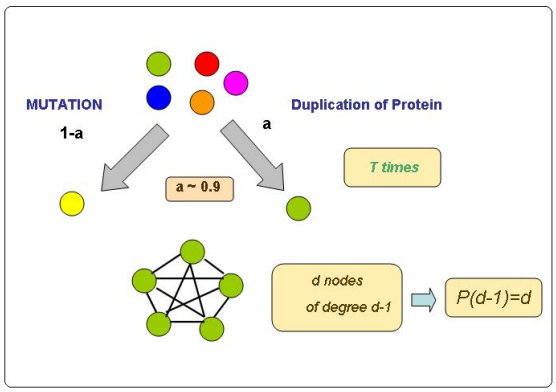

Figure 2: Scheme of the process that generates a scale-free mixing distribution. By starting with N different proteins, we repeat T times the following steps: 1) with probability (1-a), we create a new protein with new domain (mutation). 2) Otherwise, we randomly select one protein and make a copy of it (protein duplication). The parameter is set to the value of a=0.9, which is plausible in a realistic biochemical process.

In Fig. 2, we illustrate the origin of the power-law branch with positive exponent $\beta = 1$. After T iterations, one protein can consist of d copies with the same domain. Next, we connect two proteins if they consist of the same domain. Then, we can easily see that a cluster of d proteins is generated, where the degree of each protein will be $d - 1$. Therefore, a power-law distribution with positive exponent emerges as $P(d - 1) = d$. Interestingly, the signal of this distribution was also found for all the organisms as we will show in next section.

3 Experimental results

In order to compare our model with real data and to shed light on the evolutionary origin of the protein domains from a network perspective, we carried out an extensive study on protein and protein domain databases. We used the UNIPROT Knowledgebase database

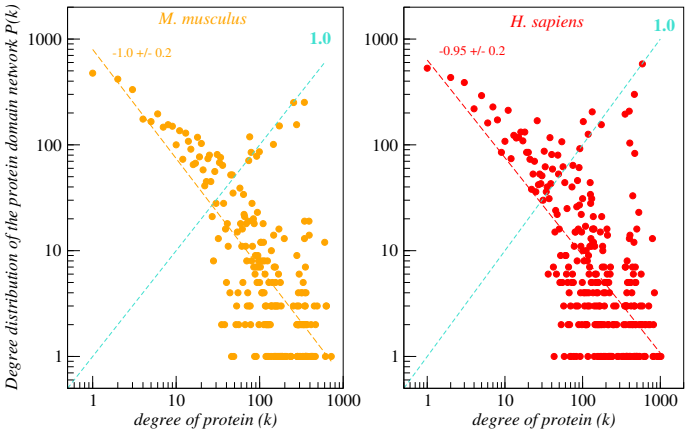

Figure 3: The degree distribution $P(k)$ of the protein domain network in *M. musculus and H. sapiens* organisms is very similar. The distribution resembles a superposition of two power-law functions, one with negative exponent $\gamma = -1$ and another with positive exponent $\beta = 1$ (dashed-lines). Data of protein sequences are from [21], and we used Pfam (circles) protein domain database.

for the individual protein sequences. In particular, we used the SwissProt part of version 2.5. In this database, the domains of each protein are referred to databases of domains by using their accession numbers. The referred domain databases are *InterPro, Pfam, ProDom, SMART, Prints and ProSite*. Although we investigated all these protein domain databases, here we only present the results obtained from *Pfam, InterPro and SMART* databases. It is also worth noticing that similar results were obtained by using the other databases.

Based on the information from the databases, we generate a protein domain network. The connectivity of the network can be specified by using an *adjacency matrix* a_{ij}. This is a square matrix of $N \times N$, where N is the total number of nodes in the network. For directed networks, its element a_{ij} is equal to 1 if there is an edge coming from the node i to the node j and 0 otherwise. The adjacency matrix of an undirected network is symmetrical: $a_{ij} = a_{ji}$. If self-loops are absent, the adjacency matrix have zeros on the diagonal $a_{ii} = 0$.

Here, we construct an undirected network by considering a protein as a node made of some domains, and an edge is connecting two proteins if they are sharing one or more domains. By using the adjacency matrix, the degree of the protein i reads as $k_i = \sum_{j=1}^{N} a_{ij}$. Therefore, we can define the degree distribution $P(k)$ as the number of proteins with specific degree k in the network.

We have evaluated the degree distribution $P(k)$ and the result is shown in Fig. 3. The degree distribution of nodes seems to be a superposition of two power-law functions, one with negative exponent $\gamma = -1$ and another one with a positive exponent $\beta = 1$. This distribution is observed in *E. coli, S. cerevisiae, A. thaliana, D. melanogaster, M. musculus and H. sapiens* organisms. In particular, the signal of the positive power-law branch is observed more clearly in higher organisms as *M. musculus* and *H. sapiens* shown in Fig. 3.

4 Discussion

This model has some advantages, and they are enumerated as follows: First, it may be reasonable from a viewpoint of evolution of proteins because it is based on well-known fundamental mechanisms as protein mutation and protein duplication. Second, the exponent of

the power-law in Eq. (4) is flexible and can be modified by tuning the parameter a. Third, and interestingly, the model does not explicitly require preferential attachment. In contrast, it is well-known that BA model requires growth and preferential attachment mechanisms to generate the scale-free distribution. By following our construction, we can see that the preferential attachment is not required.

Table 1: Specific domains in the vicinity of $P(k) \sim k$ in *H.sapiens* organism. First column k indicates the node degree of each protein, second column $P(k)$ indicates the number of nodes (proteins) with degree k, and the third column shows the accession number and the name of the most common domain in the connected proteins and between parenthesis (in bold letter) we show the number of times that this particular domain is shared. Note that "$A; B$" means domains A and B are contained in one protein, "A/B" means each domain A, B is contained in different proteins

k	$P(k)$	Domains in *H.sapiens*
588	582	PF00001(**582**) 7 transmembrane receptor (rhodopsin family)
464	300	PF00047 (**299**) Immunoglobulin domain
398	208	PF00096 (**206**) Zinc finger, C2H2 type
352	195	PF00069 (**195**) Protein kinase domain
174	155	PF00046 (**153**) Homeobox domain
133	205	PF00400 (**100**) / PF00036 (**79**) WD domain, G-beta repeat / EF hand
118	126	PF00076 (**98**) / PF00008 (**25**) RNA recognition motif / EGF-like domain
92	92	PF00071 (**88**) Ras family

It is also worth noticing that Fig. 1 shows up to three parallel positive power- law distributions. The bottom one follows $P(d-1) = d$ and holds only if there exists one kind of protein (i.e., consisting of the same domain) having d copies. The middle one corresponds to the presence of two kinds of proteins with d copies, therefore $P(d-1) = 2d$. The top one emerges when three kinds of proteins with d copies are present. In that case, we have $P(d-1) = 3d$.

Furthermore, we have analyzed the domains in the vicinity of the power-law branch of the distribution with positive exponent $\beta = 1$ shown in Fig. 3 by looking for commonalities, as for example similar protein function among organisms. We present the results for the human organism in Table 1.

The results reveal that some of the analyzed domains show similarities in a functional level. For example, let us focus on the *H.sapiens* (Table 1) and *M.musculus* (Table not shown) organisms, where the positive power-law is exhibited more clearly. In Table 1, we see that PF00047 (*immunoglobulin domain*)and PF00001 (7 $trans-membrane$ *receptor domain*) appear with high frequency. Concerning the PF0047, we can say that $Immunoglobulin$ proteins are usually present in large quantity and variety in cells in order to recognize potential viruses. On the other hand, receptors are also present in a relevant number and variety in order to recognize extra-cellular signals.

In both types of domains, we see that duplication (for increasing the number of proteins) and mutation (for increasing the variety in a functional level) mechanisms play a significant role. Interestingly, as we explain in Fig. 1 and Fig. 2, the positive power-law branch of the distribution emerges precisely from these two mechanisms.

Finally, we extended our model by including the (3) domain insertion feature (i.e., *shuffling domain mechanism*). Our results showed that by adding this mechanism (3), the tail of

the experimentally measured distribution is reproduced more accurately. However, we also found that (1) and (2) mechanisms are enough to generate (a) the observed scale-free mixing topology and (b) predict the relevant positive and negative exponents.

5 Conclusion

In this article, we have presented a model of protein domain networks based on (1) mutation of domains and (2) duplication of domains. The degree distribution of the network generated by our model resembles a superposition of two power-laws, one with negative exponent $\gamma = -1$ and another with positive exponent $\beta = 1$, and we called it scale-free mixing distribution. In particular, we found that the power-law branch with positive exponent $\beta = 1$ emerges from the duplication mechanism. Precisely, duplication generates cliques in the network, therefore for several values of k, a large number of nodes (proteins) with degree k are produced. The emergence of this positive power-law branch of the distribution is our main result.

The results of our model were compared with protein domain networks of six organisms generated with data from the UniProt Knowledgebase-Swissprot database for protein sequences and using InterPro, Pfam and Smart for domain databases. Our results indicate that the signal of this positive power-law branch of the measured distribution is also observed in the experimental data and it is conserved among organisms from *Escherichia coli* to *Homo sapiens*.

The emergence of the positive and negative exponent power-law distributions found in our study may represent a fingerprint of the mechanisms occurred during the protein domain evolution. In particular, it may indicate that, although the domain shuffling mechanisms occurred during the evolutionary stages, the evolutionary process was also governed by the genetic duplication of domains and mutations. Moreover, the scale-free mixing organization becomes manifest when at least these two mechanisms are considered.

It is also important to remark that while the branch of the distribution with negative exponent emerges from a statistical abundance of nodes with a large number of edges k (i.e., hub nodes) compared with the average degree value $< k >$, the positive branch of the distribution emerges in our model due to the existence of complete subgraphs.

Finally, it would be interesting to determine whether this *scale-free mixing* pattern is only a feature of the protein domain network or it could also be found in other biological systems or artificial and technical networks (for example, domain servers and Internet). In the latter case, we may be dealing with a universal property of the evolution of networks.

As a future work, we expect to extend our analysis of domains to elucidate the relationship between this scale-free mixing organization and specific biological functions.

Acknowledgement This work was partially supported by Grant-in-Aid Nr.17017019 from MEXT (JAPAN).

References

[1] P. Erdös and A. Rényi, Publ. Math. Inst. Hung. Acad. Sci. **5**, 17 (1960).

[2] D.J. Watts and S.H. Strogatz, Nature **393**, 440 (1998).

[3] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[4] S.N. Dorogovtsev and J.F.F. Mendes Evolution of Networks: From Biological Nets to the Internet and WWW, 2003, Oxford University Press, Oxford.

[5] A. Barrat, M. Barthelemy, R. Pastor-Satorras and A. Vespignani, Proc. Natl. Acad. Sci. USA **101**(11), 3747 (2004).

[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai and A.-L. Barabási, Nature **407**, 651 (2000).

[7] J.C. Nacher, T. Yamada, S. Goto, M. Kanehisa and T. Akutsu, Physica A **349**, 349 (2005).

[8] C. Furusawa and K. Kaneko, Phys. Rev. Lett. **90**, 008102 (2003).

[9] V.A. Kuznetsov, Genetics **161**, 1231 (2002).

[10] H.R. Ueda *et al.*, Proc. Natl. Acad. Sci. USA **101**(11), 3765 (2004).

[11] E. Ravasz, A.L. Somera, D.A. Mongru, Z.N. Oltvai and A.-L. Barabási, Science **297**, 1551 (2002).

[12] A.-L. Barabási and Z.N. Oltvai, Nature Reviews Genetics **5**, 101 (2004).

[13] J.C. Nacher, N. Ueda, M. Kanehisa, and T. Akutsu, Physical Review E **71**, 036132 (2005).

[14] H. Jeong, S. Mason, A.-L. Barabási, Z.N. Oltvai, Nature **411**, 41 (2001).

[15] S.H. Yook, Z.N. Oltvai and A.-L. Barabási, Proteomics **4**, 928 (2004).

[16] S. Wuchty, Mol. Biol. Evol. **18**(9), 1694 (2001).

[17] J. Qian, N.M. Luscombe and M. Gerstein, Journal of Molecular Biology **313**, 673 (2001).

[18] N.V. Dokholyan, Gene **347**, 199 (2005).

[19] R.F. Doolittle, Ann. Rev. Biochem. **64**, 287 (1995).

[20] J.C. Nacher, M. Hayashida and T. Akutsu, Physica A, in press.

[21] UniProt Knowledgebase database (Universal Protein Resource). We used the SwissProt part of version 2.5. The URL link is http://www.pir.uniprot.org/.