A mathematical model for generating bipartite graphs and its application to protein networks

2009 J. Phys. A: Math. Theor. 42 485005

(http://iopscience.iop.org/1751-8121/42/48/485005)

J. Phys. A: Math. Theor. **42** (2009) 485005 (10pp) doi:10.1088/1751-8113/42/48/485005

A mathematical model for generating bipartite graphs and its application to protein 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

Received 14 August 2009, in final form 17 September 2009
Published 17 November 2009
Online at stacks.iop.org/JPhysA/42/485005

Abstract

Complex systems arise in many different contexts from large communication systems and transportation infrastructures to molecular biology. Most of these systems can be organized into networks composed of nodes and interacting edges. Here, we present a theoretical model that constructs bipartite networks with the particular feature that the degree distribution can be tuned depending on the probability rate of fundamental processes. We then use this model to investigate protein-domain networks. A protein can be composed of up to hundreds of domains. Each domain represents a conserved sequence segment with specific functional tasks. We analyze the distribution of domains in *Homo sapiens* and *Arabidopsis thaliana* organisms and the statistical analysis shows that while (*a*) the number of domain types shared by k proteins exhibits a power-law distribution, (*b*) the number of proteins composed of k types of domains decays as an exponential distribution. The proposed mathematical model generates bipartite graphs and predicts the emergence of this mixing of (*a*) power-law and (*b*) exponential distributions. Our theoretical and computational results show that this model requires (1) growth process and (2) copy mechanism.

PACS numbers: 87.14Gc, 89.75.K, 87.23.Kg, 87.15Aa

1. Introduction

Many complex systems, ranging from protein interaction networks and social relationships to transportation infrastructures, can be modeled as networks where the individuals or elementary units of the system are represented by nodes and their interactions as edges. Recent empirical and theoretical studies on complex networks have shown that many disparate systems, with sizes ranging from hundreds to billions of nodes [1–4], have common characteristics and are governed by similar organizing principles. In particular, these findings show that complex networks deviate from predictions of random graph theory [5] made several decades ago and

display a scale-free and hierarchical organization [6, 7]. Furthermore, local-level interaction analyses indicate a significant prevalence and variety of highly characteristic patterns of interactions, such as motifs, modules, cliques and communities with specific functional tasks [8–10].

Proteome analyses of any organism represent an extraordinary challenge and have already generated a massive amount of newly sequenced proteins, molecular structures, folding mechanisms as well as interacting domains data. Based on this information, genome-scale protein domain statistics as well as protein interaction maps can be investigated. Although these networks are still incomplete, it allows for the first time a systems biology approach to the cells. In this context, network biology is promising to provide an understanding of the large-scale structure of functional interactions of any organism. Network analyses have shown that several kinds of cellular networks such as metabolic pathways and protein–protein interaction networks can be classified as scale-free networks [2, 11, 12].

Proteins are long chains of amino acids encoding a large variety of essential functions in the cells of living organisms. Each protein can be composed of one or more conserved sequence segments with specific functional tasks, which are called structural modules or domains. These protein domains represent fundamental building blocks with structural and functional features. However, it is worth noticing that 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 only 20–40 residues [13, 14]. It is worth noticing that while the proteins are individual molecules, protein domains represent just a part of protein sequences and, therefore, are not visible independent units.

In this work, we investigate proteins composed of domains as fundamental building blocks. In particular, we have analyzed the empirical data corresponding to proteins and domains using *Homo sapiens* and *Arabidopsis thaliana* proteome information collected from the UniProt [15] (UniProt/Swiss-Prot Release 56.0) and Integr8 [16] (Release 84 constructed from UniProt 14.0) databases. We then investigate the distribution of kinds of domains in *H. sapiens* and *A. thaliana* cells. The data analysis shows 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. Here, we aim to develop an evolutionary model that rebuilds these asymmetric distributions. In contrast, most previous theoretical models [17–19] did not focus on the relation between the number of kinds of domains in proteins and that of domain types shared by k proteins.

This problem can be investigated by means of 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 [20]. For example, N_k indicates the number of proteins composed of k domains. Similarly, M_k denotes the number of domains shared by k proteins.

Inspired by the empirical findings on the dissimilar nature of N_k and M_k distributions, we have constructed a growing network model using *the rate equation approach*, first suggested by Krapivsky *et al* [21], that explains the emergence of this mixing of exponential and scale-free distributions. The model requires two evolutionary related ingredients: (1) growth process and (2) copy mechanism. We first use the rate equation approach to construct the discrete equations corresponding to the bipartite graph. We then transform them into differential equations and solve them using the continuum limit.

2. Mathematical 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

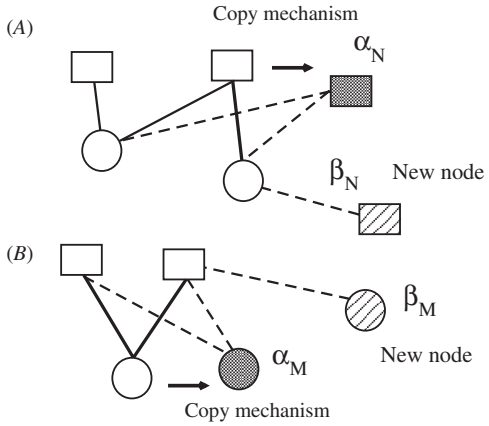

Figure 1. Description of growth and copy mechanisms in our model for bipartite graphs. Squares (proteins) (*A*) and circles (kinds of domains) (*B*) can be added and copied. One protein connected to one (two) kind of domains indicates that this protein consists of one (two) kind of domains.

allowed as shown in figure 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.

(i) The model is initialized with the same small number l of N-nodes and M-nodes. Each node from l_N is connected to a different node in l_M, and then the degree of all N-nodes and M-nodes is only 1, where we have assumed $l = l_N = l_M$.

(ii) 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, $\alpha_N + \beta_N = 1$.

(iii) 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, $\alpha_M + \beta_M = 1$.

(iv) 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. Figure 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* [21] and applied to the study of percolation [22], protein evolution networks [23] and citation networks as well as used in extensive theoretical analyses [24]. Furthermore, it has also been applied to the computation of the node degree correlations [25]. On the other hand, models applied to bipartite graphs are much less numerous and only a very few works have addressed the issue [27]. See also the review on rate equation approach for further information [26]. By following our algorithm, the rate equation for the time evolution of the number of nodes with degree k

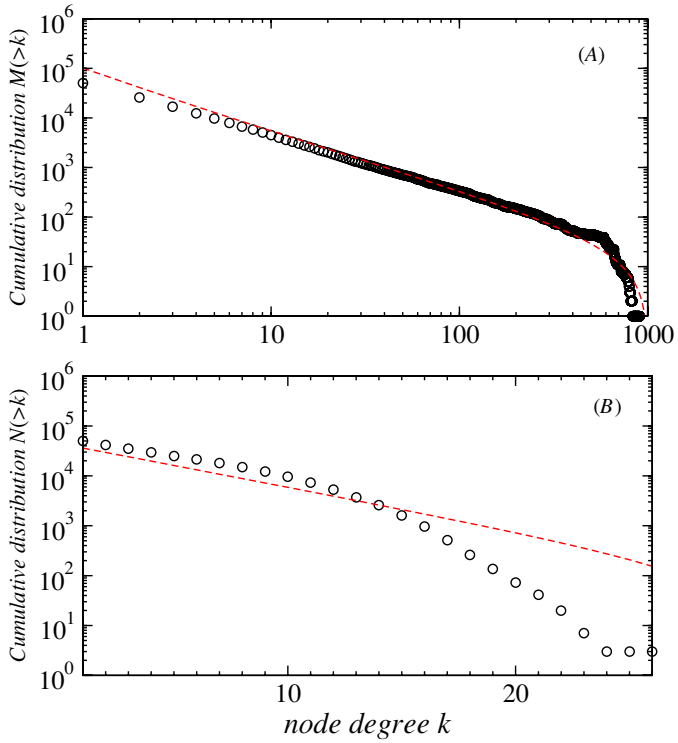

Figure 2. Fraction of nodes with degree greater than k. Computational simulation (circles) and theoretical results (red dashed line) of the model using (A) equation (7) and (B) equation (10) with $\alpha_N = 0.8$ and $\alpha_M = 0.05$. (A) Power-law decay with $\gamma' = 1.25$. (B) Theoretical results show an exponential decay or even faster than exponential for a small fraction of nodes with higher degree. (This figure is in colour only in the electronic version)

in both sets of nodes N_k and M_k can be written as

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

$$\frac{\mathrm{d}M_k}{\mathrm{d}t} = \alpha_N \left(\frac{k-1}{N(t)} M_{k-1} - \frac{k}{N(t)} M_k \right) + \beta_N \left(\frac{n_0}{M(t)} M_{k-1} - \frac{n_0}{M(t)} M_k \right) + \alpha_M \frac{M_k}{M(t)} + \beta_M \delta_{km_0}, \tag{2}$$

where $N(t) = t + l$ and $M(t) = t + l$ are the total numbers of N-nodes and M-nodes at time t, respectively. In these equations, δ_{kn_0} and δ_{km_0} indicate the contribution of a new node connected to already existing n_0 and m_0 nodes. It is to be noted that these two equations have symmetric forms. From now, we focus on the construction of the equations for N-nodes. Next, by introducing the probability distribution $n_k = N_k/N(t)$, we obtain

$$\frac{\mathrm{d}((t+l)n_k)}{\mathrm{d}t} = \alpha_M[(k-1)n_{k-1} - kn_k] + \beta_M m_0(n_{k-1} - n_k) + \alpha_N n_k + \beta_N \delta_{kn_0}. \tag{3}$$

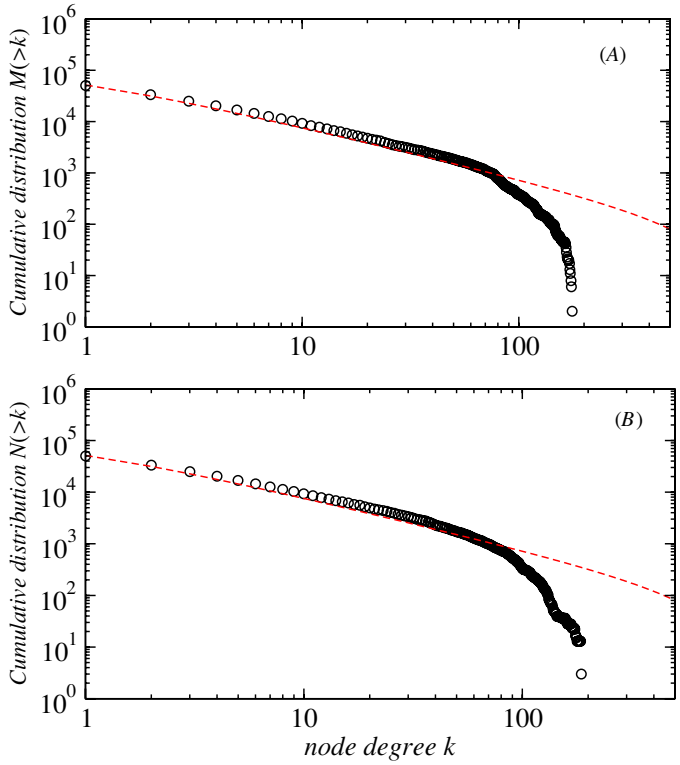

Figure 3. Fraction of nodes with degree greater than k. Computational simulation (circles) and theoretical results (red dashed line) of the model using (A) equation (7) and (B) equation (10) with $\alpha_N = 0.5, \alpha_M = 0.5$. Both figures ($A$) and ($B$) show a power-law distribution with $\gamma' = 1$.
(This figure is in colour only in the electronic version)

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

$$n_k = \alpha_M[(k-1)n_{k-1} - kn_k] + \beta_M m_0(n_{k-1} - n_k) + \alpha_N n_k + \beta_N \delta_{kn_0}. \tag{4}$$

In the continuum k limit, this equation takes the following form:

$$n_k = -\frac{\mathrm{d}}{\mathrm{d}k}[(\alpha_M k + \beta_M m_0)n_k] + \alpha_N n_k. \tag{5}$$

Similarly, the equation for M-nodes reads as

$$m_k = -\frac{\mathrm{d}}{\mathrm{d}k}[(\alpha_N k + \beta_N n_0)m_k] + \alpha_M m_k. \tag{6}$$

Then, from the last equation, we obtain

$$m_k \propto (\alpha_N k + \beta_N n_0)^{-\frac{1-\alpha_M+\alpha_N}{\alpha_N}}. \tag{7}$$

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

$$m_k \propto k^{-\frac{1-\alpha_M+\alpha_N}{\alpha_N}} \tag{8}$$

$$\sim -k^{-\frac{1+\alpha_N}{\alpha_N}}, \tag{9}$$

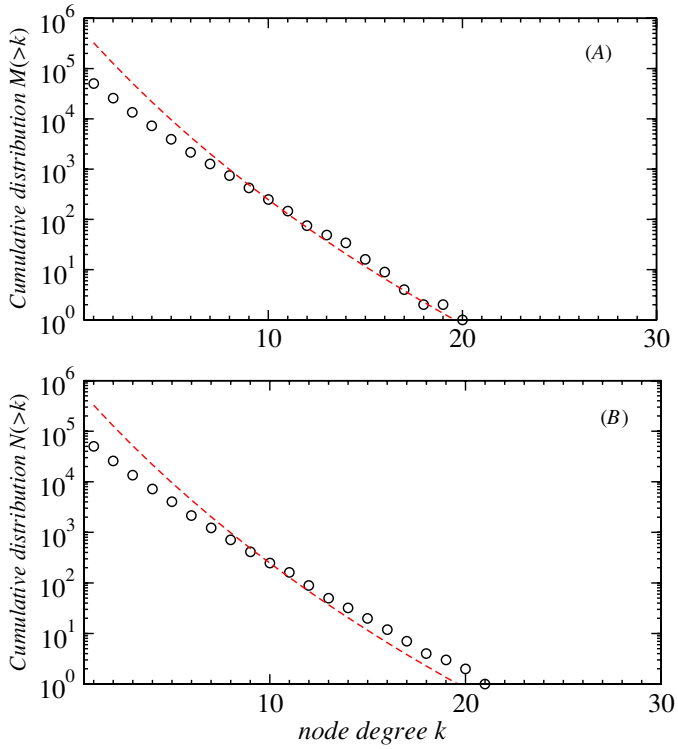

Figure 4. Fraction of nodes with degree greater than k. Computational simulation (circles) and theoretical results of the model (red dashed line) of the model using (A) equation (7) and (B) equation (10) with $\alpha_N = 0.05, \alpha_M = 0.05$. Both ($A$) and ($B$) distributions show an exponential decay.

(This figure is in colour only in the electronic version)

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 equation (5), we can write

$$n_k \propto (\alpha_M k + \beta_M m_0)^{-\frac{1-\alpha_N+\alpha_M}{\alpha_M}}. \tag{10}$$

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

$$n_k \propto \mathrm{e}^{-\frac{\beta_N}{m_0}k}. \tag{11}$$

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

(i) By using a bipartite growing network model composed of copy and random attachment processes with suppression 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.

(ii) $\alpha_M \sim 0$ implies that M-nodes (kinds of domains) are unlikely to be 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.

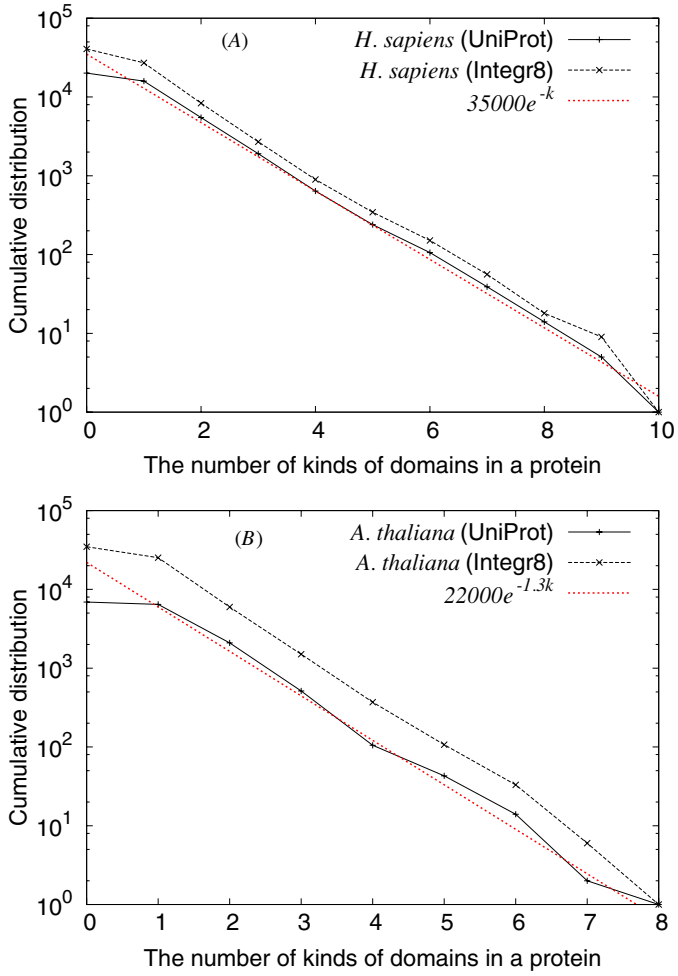

Figure 5. Fraction of proteins with at least that number of kinds of domains for (*A*) *H. sapiens* and (B) *A. thaliana* organisms. Data collected from UniProt and Integr8 databases. The data are consistent with an exponential decay.

3. Model simulation

When both parameters α_N, α_M take values close to 1 simultaneously, a so-called giant fluctuation occurs [22]. It indicates that a model that only includes the copy mechanism (i.e. a model configuration with α_N, α_M close to 1) 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* [22], 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 1, we show the computational simulation of our model in the following three figures. Figure 2 shows the cumulative distributions $N(> k)$ and $M(> k)$ of connectivities when the copy mechanism of M-nodes is suppressed

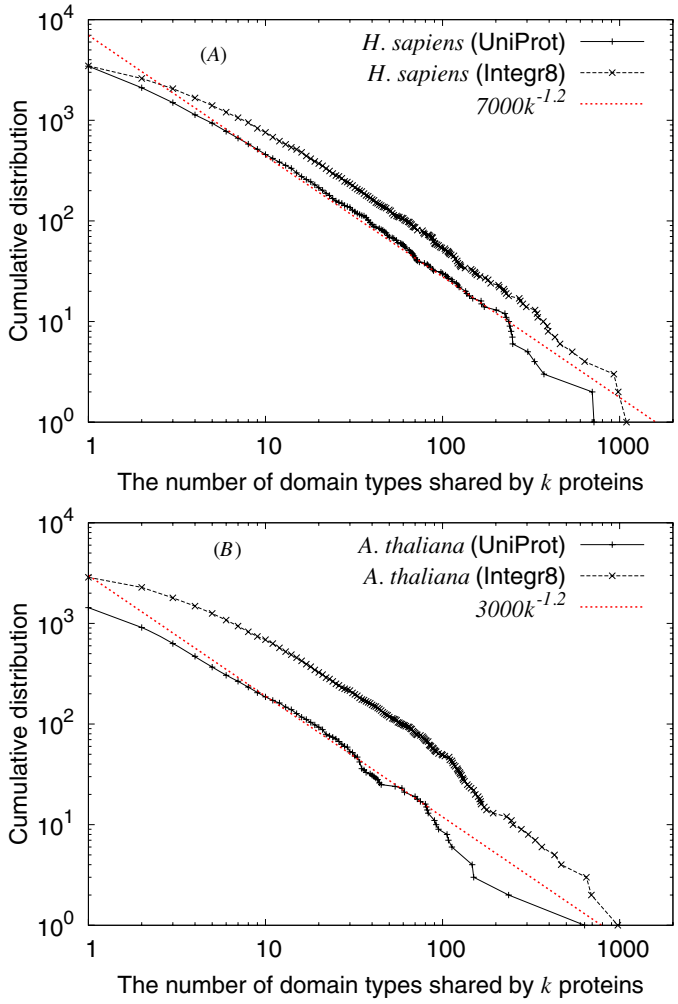

Figure 6. Fraction of domains shared by at least that number of proteins: (*A*) *H. sapiens* and (*B*) *A. thaliana* organisms. Data collected from UniProt and Integr8 databases. The data show a power-law decay with $\gamma' = 1.2$.

and copies of *N*-nodes are allowed. It is well known that if the cumulative distribution $M(> k)$ obeys a power law with exponent γ', the degree distribution $M(k)$ (or probability distribution m_k) also follows a power law with exponent $\gamma = \gamma' + 1$. The simulation results show that while $M(> k)$ obeys a power law, $N(> k)$ follows a distribution that decays exponentially or even faster for higher degrees. This copy mechanism suppression 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, figure 3 shows the case when both *N*-nodes and *M*-nodes are allowed to be copied. Then, both distributions obey a power law. Finally, we consider the case when *N*-nodes and *M*-nodes have the copy mechanism suppressed. As shown in figure 4, both distributions follow an exponential decay. Here we remark that simulation results show the cumulative degree distribution $N(> k)$ and $M(> k)$, instead of probability distribution n_k and m_k.

4. Experimental results

We have performed an empirical analysis using human proteins collected from the UniProt [15] (UniProtKB/Swiss-Prot Release 56.0 of 22 July 2008) and Integr8 [16] (Release 84 constructed from UniProt 14.0) databases. Integr8 database provides a 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.

Figure 5 shows an exponential distribution for the number of kinds of domains in a protein. *H. sapiens* and *A. thaliana* proteins were downloaded from the UniProt and Integr8 databases. Next, figure 6 shows the cumulative distribution of the number of domain types shared by k *H. sapiens* and *A. thaliana* proteins in the UniProt and Integr8 databases. In this case, we can observe that the distribution follows a power-law decay. These results are in agreement with the analytical predictions of our evolutionary model shown in figure 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 and the results suggest that the mixing of scale-free and exponential distributions is conserved. Moreover, a natural extension of the present model could be to consider a noise term for partial rewiring of duplicated nodes, which would lead to a more realistic modeling of protein duplication processes containing divergence feature.

Mutation is an important process in biological evolution. While the duplication process essentially generates the scale-freeness, mutation-related mechanisms play an important role by avoiding the giant-fluctuation effect. In this work, the addition of new nodes could be interpreted as a large (or complete) mutation, where all the links are rewired. However, an analytical derivation of the model with a more precise mutation process that explicitly includes a selective rewiring of edges would be an important future extension of this work.

5. Conclusion

In this work, we have developed a mathematical model for constructing bipartite networks with the particular feature that the degree distribution can be tuned. We have then applied it to the distribution of protein domains in *H. sapiens* and *A. thaliana* cells. The statistical analyses 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.

Based on this empirical observation, the proposed evolutionary model requires at least the following ingredients: (1) growth process and (2) copy mechanism. The rate equation approach was used for constructing the discrete equations of growing bipartite graphs and for deriving the analytical results. This model does not only explain 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 $\gamma = 2$.

Furthermore, the model elucidates that the suppression of copy mechanisms in one set of nodes is enough to create the mixture distribution and the symmetry breaking. This copy mechanism of suppression 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.

It is worth noticing that genome evolution consisting of domain shuffling events has also been investigated using more complex models based on diffusion approximation of birth-and-death processes [18, 29]. While the emergence of the power law of domain families was investigated using these models, the nature of the exponential decay for the number of

kinds of domains in a protein as well as for protein complexes was only shown using data analyses and dynamic simulations [30]. In contrast, here we used a rate equation approach for developing a bipartite graph-based model that simultaneously explains the emergence of these asymmetric distributions. Furthermore, this model notably simplifies the computation and easily allows the incorporation of additional evolutionary mechanisms. Similar evolutionary rules shown in protein family network analysis [31] could be incorporated in the proposed bipartite framework. A natural extension and further analysis could be to consider a noise term for partial rewiring of duplicated nodes and a precise analytical derivation and simulation of the model when mutation process is incorporated.

Finally, the present bipartite network construction is very general which makes it suitable for its extension and possible application to many other non-biological and biological systems. Of particular interest would be the study of the movie-actor and scientific collaborations [32] and the music recommendation networks [33], which also exhibit an asymmetric distribution for both sets of nodes.

References

[1] Dorogovtsev S N and Mendes J F F 2003 *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford: Oxford University Press)
[2] Barabási A-L and Oltvai Z N 2004 *Nat. Rev. Genet.* **5** 101
[3] Pastor-Satorras R and Vespignani A 2004 *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge: Cambridge University Press)
[4] Newman M E J, Barabási A-L and Watts D J 2007 *The Structure and Dynamics of Networks* (Princeton, NJ: Princeton University Press)
[5] Erdös P R and Rényi A 1960 *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17
[6] Barabási A-L and Albert R 1999 *Science* **286** 509
[7] Ravasz E, Somera A-L, Mongru D A, Oltvai Z N and Barabási A-L 2002 *Science* **297** 1551
[8] Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D and Alon U 2002 *Science* **298** 824
[9] Shen-Orr S, Milo R, Mangan S and Alon U 2002 *Nat. Genet.* **31** 64
[10] Palla G, Derenyi I, Farkas I and Vicsek T 2005 *Nature* **435** 814
[11] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A-L 2000 *Nature* **407** 651
[12] Jeong H, Mason S, Barabási A-L and Oltvai Z N 2001 *Nature* **411** 41
[13] Go M 1983 *Proc. Natl Acad. Sci. USA* **80** 1964
[14] Go M 1981 *Nature* **291** 90
[15] The UniProt Consortium: The Universal Protein Resource (UniProt) 2008 *Nucleic Acids Res.* **36** D190
[16] Kersey P *et al* 2005 *Nucleic Acids Res.* **33** D297
[17] Wuchty S 2001 *Mol. Biol. Evol.* **18** 1694
[18] Karev G P, Wolf Y I, Rzhetsky A Y, Berezovskaya F S and Koonin E V 2002 *BMC Evol. Biol.* **2** 18
[19] Nacher J C, Hayashida M and Akutsu T 2006 *Physica* A **367** 538
[20] Newman M E J, Strogatz S H and Watts D J 2001 *Phys. Rev.* E **64** 026118
[21] Krapivsky P L, Redner S and Leyvraz F 2000 *Phys. Rev. Lett.* **85** 4629
[22] Kim J, Krapivsky P L, Kahng B and Redner S 2002 *Phys. Rev.* E **66** 055101
[23] Ispolatov I, Krapivsky P L and Yuryev A 2005 *Phys. Rev.* E **71** 061911
[24] Krapivsky P L and Redner S 2001 *Phys. Rev.* E **63** 066123
[25] Barrat A and Pastor-Satorras R 2005 *Phys. Rev.* E **71** 036127
[26] Krapivsky P L and Redner S 2003 *Lecture Notes in Physics* **625** 3
[27] Ergün G 2002 *Physica* A **308** 483
[28] Finn R D 2008 *Nucleic Acids Res.* **36** D281
[29] Karev G P, Berezovska F S and Koonin E V 2005 *Bioinformatics* **21** 12
[30] Beyer A and Wilhelm T 2005 *Bioinformatics* **21** 1610
[31] Goh K I, Kahng B and Kim D 2005 *J. Korean Phys. Soc.* **46** 551
[32] Ramasco J J, Dorogovtsev S N and Pastor-Satorras R 2004 *Phys. Rev.* E **70** 036106
[33] Cano P, Celma O and Koppenberger M 2006 *Chaos* **16** 013107