Journal of Proteomics & Bioinformatics - Open Access

Research Article JPB/Vol.1/May 2008

Protein Interaction Network Double Exponential Model

Piotr H. Pawlowski[1]*, Szymon Kaczanowski[1,2], Piotr Zielenkiewicz[1,3]

[1]Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warszawa, Poland
[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, USA
[3]Plant Molecular Biology Laboratory, Warsaw University, Warszawa, Poland

*Corresponding author: Piotr H. Pawlowski, Ph.D., Institute of Biochemistry and Biophysics,
Polish Academy of Sciences, Warszawa, Poland, Fax: (48) 39121623; E-mail: piotrp@ibb.waw.pl

Received April 13, 2008; Accepted May 14, 2008; Published May 20, 2008

Citation: Piotr HP, Szymon K, Piotr Z (2008) Protein Interaction Network. Double Exponential Model. J Proteomics Bioinform 1: 061-067. doi:10.4172/jpb.1000010

Abstract

The proper theoretical description of the distribution of the node degree for yeast protein-protein interaction network was investigated to deal with the observed discrepancy between usually proposed models and the existing data. The power law or the generalized power law with exponential cut-off were shown to be inaccurate within a wide range of degree values. Proposed linear-combination-of-exponential-decays-method exactly characterizing the distribution by the spectrum of decay constants revealed two separate parameter domains. A consequent hypothesis that the node degree distribution could follow the universal double exponential law was successfully verified by selected model comparison using the AIC criterion. BIND and DIP data for H. pylori, E. coli, *S. cerevisiae, D. melanogaster, C. elegans* and A. thaliana were used for this purpose. A linear change in the magnitude of the distribution components with proteome size was observed, manifesting the evolutional stability of the process of developing the protein interaction network. Proposed kinetic model of protein evolution, considering the two hypothetical protein classes, first, with a relatively rapid emerging rate and a short characteristic residence time, and the second one, with the opposite properties, analytically described the nature of bi-exponential pattern. The model presents a situation in which evolutionary conserved proteins increase their interactions due to specific kinetic conditions. Thus, we oppose the opinion that the majority of such interactions are biologically significant, and, therefore the older parts of interactome are more complex. We believe that our interactome results support the hypothesis of Stuart Kaufman, presented in his book "The Origin of Order", that random mutations and natural selection constitute the origin of order and complexity.

Keywords: Protein interaction network; Exponential model; Scale free network; Power law

Introduction

The degree of a node (or connectivity) is the number of edges that are adjacent to it. From the theoretical point of view, it is one of the basic measures characterizing the importance of the node in the network. Although the power law (PL) and the generalized power law supplemented with an exponential cut-off (GPL-EC) were widely popularized (Wagner, 2001; Jeong et al., 2001) as the rules describing the distribution of the node degrees in protein-protein interaction network, attempts at a more exact mathematical description are still being undertaken (Thomas et al., 2003; Berg et al., 2004). The reasons are both of practical and methodological nature. The first reason pertains to the still evolving databases, and the second one concerns the facts that the usually simple shape of arrangement of experimental points may be fitted in various manners giving at different theoretical assumptions quite similar results. According to the DIP data (see Materials and Methods) we could observe that the degree distribution of nodes of S. cerevisiae protein interaction network follows approximately a PL or a GPL-EC, but only for the degree values k smaller than 10. For higher values of k we saw a serious discrepancy between the theory and the experiment, already reported by others as an exponential decay (Wilhelm et al., 2003).

There are additional indications (Barabási and Oltvai, 2004; Pereira-Leal et al., 2005) that the biological network characteristics may contain an exponential component. The main aim of the present paper is to resolve whether by using a more complex exponential-type model one can better describe the distribution of node degree in the protein interaction network. Developing the above idea we proposed to consider a node distribution as a linear combination of exponential decays $A_i \exp(-\lambda_i k)$, with amplitudes A_i and decay constants λ_i being positive values. Our method applied to S. cerevisiae DIP data revealed two separate domains of λ_i, with two characteristic values of the parameters related to the relatively "fast", then "slow", tendency of a distribution to decay along k-axis. This led to the natural concept that a double exponential curve $a_1 \exp(-d_1 k) + a_2 \exp(-d_2 k)$ could be a better model of the node degree distribution than the standard or modified power law. This supposition was confirmed by using BIND or DIP data for 6 different organisms and the AIC criterion (see Materials and Methods). The obtained results led to analysis of the dependence of both exponential contributions to the total protein pool on proteome size, clearly indicating a linear trend. In consequence, this analysis helps us to better characterise the evolutionary mechanism leading to the observed double exponential distribution and points out its universal elements.

To explain the bi-exponential character of node degree distribution, the kinetic model of protein network evolution was proposed. It relates the searched distribution formula to the parameters describing the rate of some creation and disruption processes, postulated as being important in formation of the net. According to our model, two basic types of proteins, marked "1" and "2", with a different dynamics of evolutional behaviour were assumed. They were shown to be good candidates, from a statistical point of view, for the low-connected nodes and hubs, respectively.

The discussed results suggest that the process of evolution leads to a "biological" order in the interactome. Therefore, they support the hypothesis of Kaufman (1993) that the process of random mutation and selection always leads to complexity.

Materials and Methods

Protein interaction network data for H. pylori ($N_{k>0}$ = 724 nodes, N_e = 1403 edges) and S. cerevisiae (analogous values 4135 and 7839) were taken from Coevolution and Self-organisation in Dynamical Networks data sets (COSIN, http://www.cosin.org) derived from the Database of Interacting Proteins (DIP, http://dip.doe-mbi.ucla.edu/). Data for E. coli (399 and 312), D. melanogaster (7910 and 23128), C. elegans (3227 and 5026) and A. thaliana (487 and 959) were taken from Biomolecular Interaction Network Database (BIND, http://www.bind.ca/Action). Only single protein-protein interaction records (without self-interaction) were analyzed. No non-interacting proteins were reported.

According to our method of linear combination of exponential decays (LCED), a S. cerevisiae node degree distribution (histogram) was tentatively described by the sum:

$$n_k = \sum_{i=0}^{i\_\max} A_i \exp(-\lambda_i k) \qquad (1)$$

where n_k was a number of $k\_$ degree nodes and $i\_$ max was the maximal value of a sum index .Equation 1 was fitted to the experimental data, at i_max = 50 and gridded spectrum of decay constants λ_i = {0, 0.025, 0.050, 0.075… 1.250}. The fit had been repeated 20 times to find the sets of amplitudes A_i , and then the respective averages $<A_i>$ were analysed. As a fitting algorithm the NonlinearRegress procedure (NRP) from Mathematica 4.1 (http://www.wolfram.com) was applied, with substitution $A_i = (A'_i)^2$ to guarantee only the positive value of amplitude. Random starting conditions, A'_{0i} , were being selected within the range $0.5 < A'_{0i} > 1.5$.

In the final modelling with a double exponential law (DEL),

$$n_k = a_1 \exp(-d_1 k) + a_2 \exp(-d_2 k) \qquad (2)$$

In the alternative modelling with a PL,

$$n_k = A k^{-\gamma} \qquad (3)$$

and with a GPL-EC,

$$n_k = A(k + k_0)^{-\gamma} \exp(-k / k_c) \qquad (4)$$

The fits were performed in the range $1 \le k \le 15$, using NRP once (without squared substitution of amplitude), and at default starting conditions (1.0).

To rate the quality of the proposed models, corrected Akaike's Information Criterion (AICc) was adopted, defined as:

$$AICc = z \ln(\sigma^2) + 2m + \frac{2m(m+1)}{z-m-1} \qquad (5)$$

where σ^2 is the average squared residual for a given model, - the number of model $\boldsymbol{m}$ parameters, and $\boldsymbol{z}$.- the number of observations (Burnham and Anderson, 2004). In the case of PL, $\boldsymbol{m}$ = 2. For GPL-EC and DEL, $\boldsymbol{m}$ = 4. The number of analysed points was $\boldsymbol{z}$= 15 in each competing model. Models with a smaller AICc value were being favoured.

In the theoretical considerations, the total proteome size (N_P^*)of the analysed species was assumed to be equal to the number of open reading frames, i.e., 1788 for H. pylori, 4285 for E. coli, 6307 for S. cerevisiae, 14218 for D. melanogaster and 18944 for C. elegans (Liu and Rost, 2001) or 28952 for family members of A. thaliana (Horan et al., 2005). Due to division by the scaling factor , where:

$$sc = \frac{n_0 + N_{k>0}}{N_P^*} \qquad (6)$$

describes the ratio of the extrapolated size of the analysed probe to the size of the total proteome, the DEL model amplitudes for accessed data, a1 and a2, were transformed into hypothetical values $a^*_1 = a_1/SC$ and $a^*_2 = a_2/SC$, for the total species proteome (see Appendix 1). In eq. 6 the unknown value n_0 was replaced by $a_1 + a_2$. Then, the expected amount of proteins in considered contributions 1 and 2 to the total proteome was estimated by the sum of infinite geometrical series

$$N_i^* = \sum_{k=0}^{\infty} a_i^* \exp(-d_i k) \qquad i = 1,\ 2 \qquad (7)$$

leading to:

$$N_1^* = \frac{a_1^*}{1 - \exp(-d_1)} \qquad (8)$$

$$N_2^* = \frac{a_2^*}{1 - \exp(-d_2)} \qquad (9)$$

In the estimation of the parameters of the model of protein network evolution (Appendix 2) eqs. A.2.8-11 were applied.

Results

It was observed that the distribution histogram of node degree of S. cerevisiae protein-protein interaction network exhibits a well-ordered pattern in the range $1 \le k \le 25$ (Fig. 1).

Above that range statistical fluctuations prevailed and quantization perturbed the continuity of analysed characteristics of the network. Attempts to describe the investigated distribution by a PL: $A = (1.65 \pm 0.02) \cdot 10^3$, $\gamma = (1.27 \pm 0.02)$ (upper line), or by a GPL-EC: $A = (2.4 \pm 0.7) . 10^3$, $k_0 = (0.3 \pm 0.7)$,

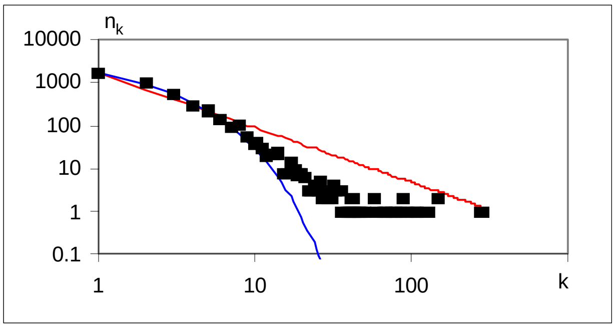


Figure 1. The distribution histogram (nk) of node degree (k) of S. cerevisiae protein-protein interaction network. Presented fits are: the upper line - a power law (PL): $n_k = 1.65 . 10^3 k^{-1.27}$; the bottom line - a generalized power law supplemented with an exponential cut-off (GPL-EC): $n_k = 2.4 \cdot 10^3 (k + 0.3)^{-0.5} \exp(-k / 3.0)$. Zero values are not shown.

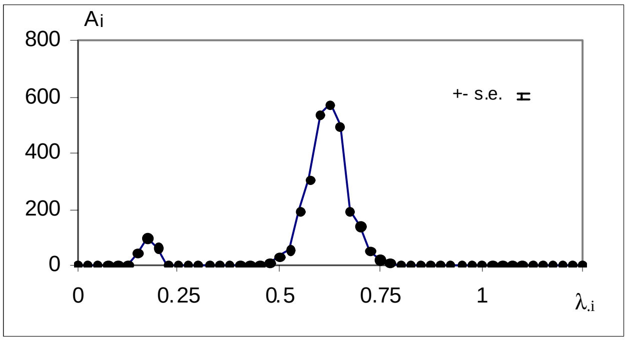


Figure 2. Linear combination of exponential decays method (LCED) applied to the data for S. cerevisiae (Fig.1). Two regions of decay constants (λ) spectrum with dominant amplitudes Ai at 7 =0.175 and 25 = 0.625 are clearly seen. Shown values are averages of adequate amplitudes of 20 multi-exponential fits mean standard error (s.e.) is also presented.

$\gamma=(0.5 \pm 0.3)$, $k_c=(3.0 \pm 0.4)$ (bottom line) gave good results only in the range $1 \leq k \leq 10$. The PL parameters obtained, $\gamma = 1.27$ and $A/N_p = 0.40$, are consistent with $\gamma = 1.32$ and $A/N_p = 0.42$ for the whole yeast interaction network (Yu et al., 2004). A different picture is seen in case of the GPL-EC model. One can notice a big discrepancy between our result and those for a small sample of 1870 nodes (Pastor-Satorras et al., 2003), which may indicate the narrow area of applicability of the cut-off formula.

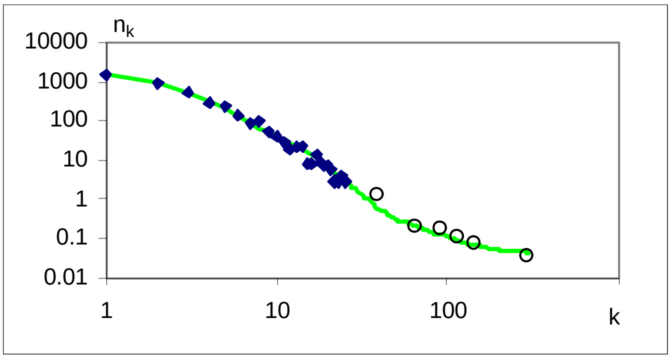
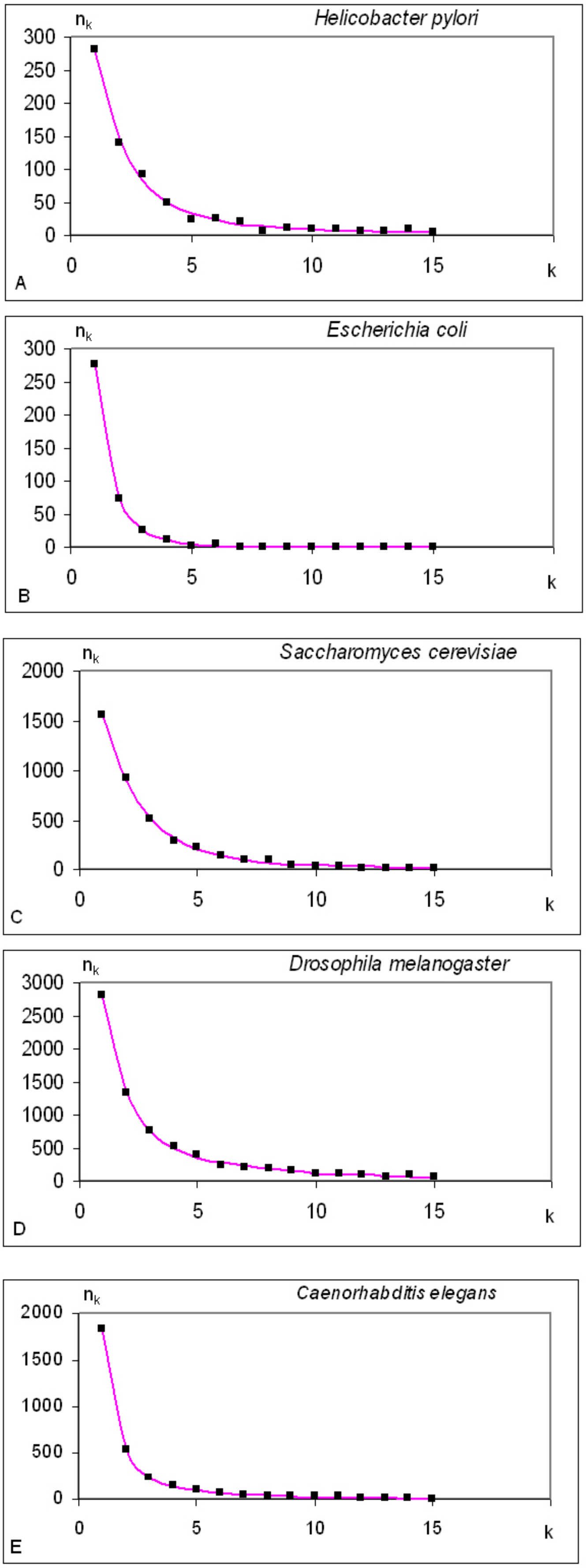


Figure 3. An example of one of 20 fits to the experimental data (Fig. 1) performed to obtain the decay constants spectrum (Fig.2) The n_K is the number of k-degree nodes. For clarity, the open circles denote group averages. Zero values are not shown.

The proposed LCED method (Fig. 2) revealed two narrow ranges of decay constants spectrum with dominant amplitudes at λ_7 = 0.175 and λ_{25} = 0.625 (characteristic values of node degree: $1/\lambda_7$ = 5.7, $1/\lambda_{25}$ = 1.6). Half-width of the observed peaks equals 0.025 and 0.050, respectively. An example of one in 20 fits performed to obtain the above spectrum is also presented (Fig.3). As it is seen here, and in the case of other fits (data not shown), their qualities, especially in the range of values $k > 10$, are better than the estimation with standard or modified power law.

As a result of the above, it was hypothesized that our combination, even reduced to a double exponential formula, could provide a better description of the node degree distribution than the con-



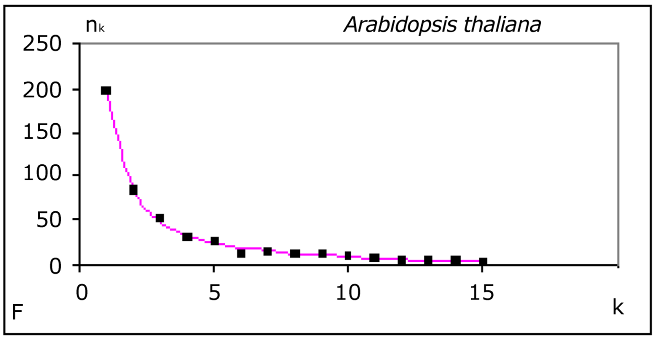


Figure 4. The distribution histogram (nk) of node degree (k) for different species. Continuous line is the fit of a double exponential law (DEL). Parameters of the DEL models are presented in Table 1.
A. Helicobacter pylori.
B. Escherichia coli.
C. Saccharomyces cerevisiae.
D. Drosophila melanogaster.
E. Caenorhabditis elegans.
F. Arabidopsis thaliana.

sidered power law type models. The examples of yeast and five other species were analysed for k<15. Corresponding fits of proposed DEL models are presented in Fig. 4a-f and Table 1. Their qualities are confirmed by AICc values, which favour bi-exponential approximation in 5/6 of the investigated cases (Table 2). Plots of alternative fits are not shown.

Some parameters of DEL models vary with proteome size. The size N_1^* and N_2^* of distinguished protein groups increases with the total number of proteins N_P^* (Fig. 5). There was no detected essential dependence of decay constant d_1 and d_2 on the proteome size.

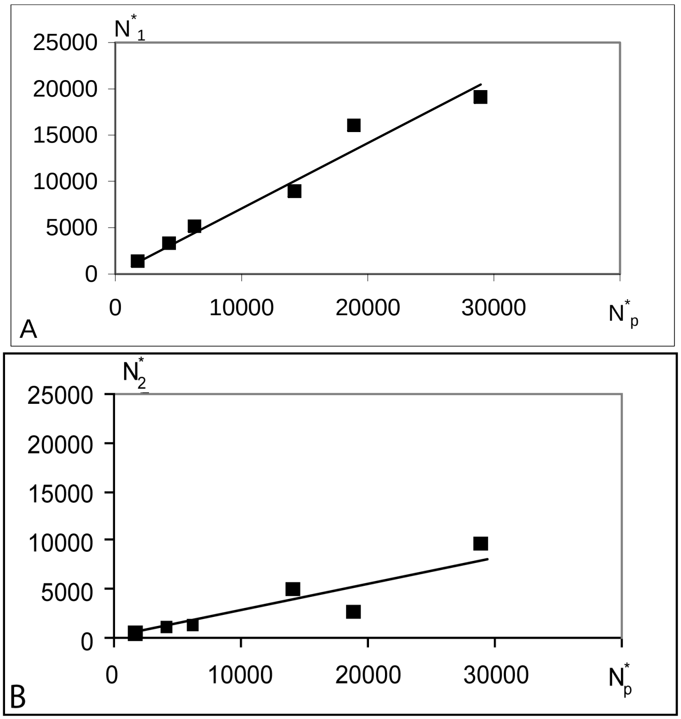


Figure 5. A-B

The variation in the estimated number of proteins N_F^* and N_S^* of a given protein class with proteome size N_P^*. The following data points represent: (from left) H. pylori, E. coli, S. cerevisiae, D. melanogaster, C. elegans and A. thaliana. A. Protein class F. B. Protein class S. Continuous line - linear trend.

	a_1	a_2	d_1	d_2
H. pylori	507.409	44.529	0.743	0.157
E. coli	1166.020	219.041	1.898	0.762
S. cerevisiae	2592.380	197.464	0.616	0.170
D. melanogaster	5783.780	837.777	1.005	0.187
C. elegans	7307.120	389.915	1.564	0.278
A. thaliana	486.548	68.659	1.234	0.220

Table 1. Parameters of the fitted DEL models.

	DEL	PL	GPL-EC
H. pylori	56.4 [1]	73.2 [3]	58.3 [2]
E. coli	6.0 [1]	43.0 [3]	6.4 [2]
S. cerevisiae	94.6 [1]	135.8 [3]	112.4 [2]
D. melanogaster	97.4 [1]	122.1 [3]	109.4 [2]
C. elegans	52.5 [1]	66.5 [2]	90.9 [3]
A. thaliana	38.9 [2]	37.2 [1]	130.5 [3]

Table 2. AICc ranking of the models[1].

Discussion

The results presented above confirm recent reports (Goldberg et al., 2005) suggesting the "break" of a power law in the global description of the protein interaction network. Actually, we can suggest that this "break" may be caused by the second exponential term in node degree distribution, which does not affect strongly the formula in the range of the node degree smaller than 10, but may be essential elsewhere.

Initial inspection of the data shown in Fig. 1 reveals that GPL-EC, the 4-parameter improvement of PL (bottom line), fits better than PL alone (upper line), but is still a very long way from perfect. Hence we decided to introduce a more general description.

In accordance with our idea, protein interaction network consists of subpopulations of vertexes described by a similar statistical formula, but with different parameters. As a universal formula we choose exponential decay, which is consistent with the suggested model of network evolution (see Appendix 2).

The proposed LCED method revealed the spectrum of decay constants and the magnitude of subpopulational contributions into the degree global distribution (Fig. 2). Two classes of nodes with the values of decay constant lying closely together were clearly distinguished. Good quality of fits (Fig.3) testifies to the utility of the method and the acceptance of the formula.

Reducing the huge number of parameters of a general model and taking into account the above observation, we propose to limit the number of decay components to only two items, indexed by 1 and 2. It did not weaken the fitting abilities for different species in the range $1 \le k \le 15$ (Fig. 4a-f, Table 1), which was confirmed by the AICc criterion. As seen in Table 2, the DEL models are the best in 5/6 of investigated cases and just a little worse (2nd place) than the winner in one case. Generally, they are more effective for networks with big proteomes (the PL model for a small probe of A. thaliana may be an exception) than for sets with a small protein number; PL or GPL-EC models may give similar results.

Documented changes in the dimensions of the indexed protein classes with proteome size (Fig. 5) indicate a similar tendency for linear increase for the first (a) and the second (b) component of proteome. This way the ratio $N_1^* / N_2^* \approx 2.5$ seems to be a universal constant for a wide group of organisms.

The contribution of each class of proteins to the summary distribution was shown in the example of a yeast probe (Fig. 6).

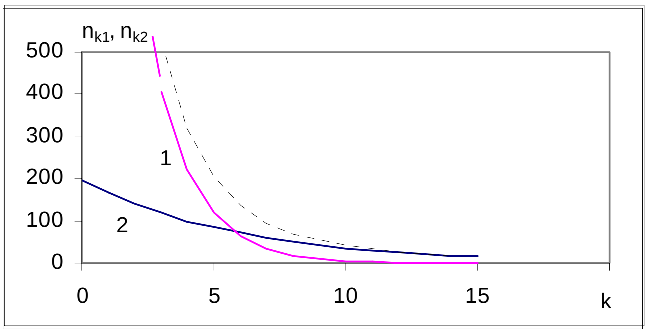


Figure 6. The contribution of "F" and "S" protein class to the overall distribution. The insets $n_{k1}=a_1\exp(-d_1 k)$ and $n_{k2}=a_2\exp(-d_2 k)$ were plotted (continuous lines 1 and 2) for the parameters of S. cerevisiae (Table 1). The broken line represents fitted summary distribution $n_k = n_{k1} + n_{k2}$.

As seen for small values of k, the two classes contribute to the global distribution. For $k > 10$, the first class vanishes and the second class clearly dominates. The latter class may be related to so called hubs. It is worth stressing that the second class of proteins may bear only a few links, too.

It seems that the proposed double-exponential model is a simplification of a hypothetical multi-componential model describing the full spectrum of contributions from different classes of proteins. The analysed data indicate that there probably exists the third, small amplitude class of yeast proteins (not visible in Fig. 2), which may be related to the "super" hubs connecting hundreds of nodes; however, a "false positive" error cannot be excluded.

Although the two protein classes clearly dominate, the analysed subpopulations do not form spikes along the decay constant axes, but have some definite width. We believe that more sophisticated analysis of discussed contributions, considering their continuous representation, should fully describe protein network statistics and reveal new properties of the proteome system.

As mentioned beforehand, to specify our hypothesis, we proposed a simple mathematical model of protein network evolution (Appendix 2). The applied assumptions permit duplication events to occur even more often than the appearance of "new" types of protein encoding genes. Such behaviour is suggested by the observation that gene-copy number within a family is often changed during the process of speciation (Cheng et al., 2005; Ma and Gustafson, 2005; Ting et al., 2004). However, to avoid an enormous expansion of the system, we assumed that the speciation processes are no more frequent than deletion episodes effectively leading to the elimination of proteins. On the other hand, one can detect evolutionary conservation of genes present even in different kingdoms. Therefore, the probability of multiplication of old "proteins" is similar to the probability of multiplication of "young" proteins in a given genome. The facts mentioned above were "silently" included in the model. It relates amplitudes and decay constants to the emergence rates, q_1 and q_2, effective elimination rates, γ_1 and γ_2, and interaction gaining rates, v_1 and v_2 of the two classes of proteins, with different dynamics of evolutional performance. This difference in dynamics of the evolution of proteins manifests in the observed difference between "fast" and "slow" tendency in the variation of the node degree distribution along k-axis. In general the above parameters may differ for different evolutional pathways.

According our model, the linear trend in Fig. 5 may be related to the stable dynamics of evolution of investigated classes of proteins during the inter space progress. Indeed, with equations A.2.12-14 it is easy to show that the observed dependence calls for stability of the ratio. $q_1\gamma_2/q_2\gamma_1$ This linear trend also suggests that for the total proteomes the corresponding amplitudes of calculated probability (frequency) of the occurrence of a node with a given degree may remain approximately constant. In a sense, we showed not a scale-free distribution but a scale-free evolution.

As the analysed decay constants d_1 and d_2 do not exhibit a clear tendency to change, we may simply imagine that during evolution γ_1, γ_2, v_1 and v_2 remain approximately constant (see eqs. A.2.10-11). According to this picture, q_1 and q_2 slowly evolve in a stable manner ($q_1 / q_2 = const$), governed, for example, by the varying amount of DNA, which accounts for the change in the global protein pool (see eq. A.2.12).

To make our considerations more quantitative we estimated values q_1, q_2, γ_1 and γ_2, assuming that $v_1 = v_2 = 0.1$ [1/mln years] (Berg et al., 2004). It is seen (Table 3) that first class of proteins may be characterized by a relatively rapid emerging rate q_1 and also relatively rapid elimination γ_1 rate (or short characteristic residence time) when to compare with the second class of proteins.

The proposed mathematical model of evolution suggests unexpected explanation of the observation of Barabasi and co-workers that more densely interconnected parts, "motives" of the interaction network, are more strictly evolutionary conserved (Wuchty et al., 2003). Intuitively, one can suppose that proteins belonging to such motives are evolutionary conserved because they are required for maintaining the connections in such motives. But the results of our simulations suggest an exactly opposite explanation: the old proteins (evolutionary conserved proteins) are more interconnected because they are simply old enough. This explanation although surprising for us, does in fact have sense. Since the majority of the proteins are not interacting (for example, protein-protein interaction network of yeast contains only approximately 30000 protein-protein interactions (according to the estimation of Kumar and Snyder, 2000) and more than 36000000 protein-protein pairs), and the protein interaction network is evolutionary conserved (see, for example, Matthews et al., 2001), it is likely that the majority of interactions have biological significance and that interactions appear gradually during the process of evolution. It is also likely that new "proteins" have no interactions or have a small number of interactions. During the process of evolution these proteins slowly gain new "useful" interactions. If they belong to the class 2, they may even gain many such interactions. This process leads to a well-ordered protein-protein interaction network in which proteins are not randomly connected and in which one can distinguish "modules" of interacting proteins.

As we have already referred to in the Introduction, our results support the hypothesis of Stuart Kaufman that natural selection, random mutations and the process of evolution are the source of order in biological systems. This paper shows a random process of evolution leading to complex and non-random systems. Although it remains an open question whether the random process is rapid enough to lead to creation of structures as complex as multi-enzymatic complexes or flagelles, we believe that a right step in the proper direction has been taken.

Acknowledgements

We would like to thank Doktor Christophe Pakleza for his help in the application acknowledgement of Mathematica to numerical calculations. Fianancial support from the ministry of Science and Higher Education geant PBZ-MNil_2/1/2005 and postdoctoral fellowship from the foundation for polish Science to S.K are gratefully acknowledged.

References

1. Barabási AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nat Rev Genet 5: 101-113. » CrossRef » Pubmed » Google Scholar
2. Berg J, Lässig M, Wagner A (2004) Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications. BMC Evolutionary Biology 4: 51-62. » CrossRef » Pubmed » Google Scholar
3. Burnham KP, Anderson DR (2004) Multimodel Inference: understanding AIC and BIC in Model Selection. Sociological Methods & Research 33: 261-304. » CrossRef » Google Scholar
4. Cheng Z, Ventura M, She X, Khaitovich P, Graves T, et al. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437: 88-93. » CrossRef » Pubmed » Google Scholar
5. Goldberg DS, Franklin G, Roth FP (2005) Breaking the Power Law: Improved Model Selection Reveals Increased Network Complexity. In: Poster Session of the Ninth Annual International Conference on Research in Computational Molecular Biology. RECOMB (2005), Cambridge, MA. » CrossRef » Google Scholar
6. Horan K, Lauricha J, Bailey-Serres J, Raikhel N, Girke T (2005) Genome cluster database. A sequence family analysis. Platform for Arabidopsis and rice. Plant Physiology 138: 47-54. » CrossRef » Pubmed » Google Scholar
7. Jeong H, Mason S, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42. » CrossRef » Pubmed » Google Scholar
8. Kaufman SA (1993) The origin of order self-organization and selection in evolution. Oxford Univeristy Press. » CrossRef » Google Scholar
9. Kumar A, Snyder M (2002) Protein complexes take the bait. Nature 415: 123-124. » CrossRef » Pubmed » Google Scholar
10. Liu J, Rost B (2001) Comparing function and structure between entire proteomes. Protein Science 10: 1970-1979. » CrossRef » Pubmed » Google Scholar
11. Ma XF, Gustafson JP (2005) Genome evolution of allopolyploids: a process of cytological and genetic diploidization. Cytogenet Genome Res 109: 236-249. » CrossRef » Pubmed » Google Scholar
12. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome Res 11: 2120-2126.» CrossRef » Pubmed » Google Scholar
13. Pastor-Satorras R, Smith E, Sole RV (2003) Evolving protein interaction networks through gene duplication. J Theor Biol 222: 199-210. » CrossRef » Pubmed » Google Scholar
14. Pereira-Leal JB, Audit B, Peregrin-Alvarez JM, Ouzounis CA (2005) An exponential core in the heart of the yeast protein interaction network. Mol Biol Evol 22: 421-425. » CrossRef » Pubmed » Google Scholar
15. Thomas A, Cannings R, Monk NA, Cannings C (2003) On the structure of protein-protein interaction networks. Biochem Soc Trans 31: 1491-1496. » CrossRef » Pubmed » Google Scholar
16. Ting CT, Tsaur SC, Sun S, Browne WE, Chen YC, et al. (2004) Gene duplication and speciation in Drosophila: evidence from the Odysseus locus. Proc Natl Acad Sci USA 101: 12232-12235. » CrossRef » Pubmed » Google Scholar
17. Yu H, Zhu X, Greenbaum D, Karro J, Gerstein M (2004) TopNet: a tool for comparing biological sub-networks, correlating protein properties with topological statistics. Nucleic Acids Res 32: 328-337. » CrossRef » Pubmed » Google Scholar
18. Wagner A (2001) The Yeast Protein Interaction Network Evolves Rapidly and Contains Few Redundant Duplicate Genes. Mol Biol Evol 18: 1283-1292. » CrossRef » Pubmed » Google Scholar
19. Wilhelm T, Nasheuer HP, Huang D (2003) Physical and functional modularity of the protein network in yeast. Mol Cell Prot 2: 292-298. » CrossRef » Pubmed » Google Scholar
20. Wuchty S, Oltvai ZN, Barabasi AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. Nat Genet 35: 176-179. » CrossRef » Pubmed » Google Scholar

Appendix 1

It was assumed that each data set analysed is only a homogenous part of the total proteome of a given species. Then the fitted DEL model formula and the hypothetical distribution of the total population of proteins of a given organism (see Appendix 2) are related in the proportion:

$$\frac{n_k}{n_k^*} \overset{def}{=} \frac{a_1 \exp(-d_1 k) + a_2 \exp(-d_2 k)}{a_1^* \exp(-d_1 k) + a_2^* \exp(-d_2 k)} = \frac{N_P}{N_P^*} \quad \text{(A.1.1)}$$

where a_1^* and a_2^* are the amplitudes of a hypothetical distribution for the total population, N_P is the extrapolated size of the analysed probe and N_P^* is the total size of proteome.

In the above ratio, N_P value includes interacting proteins ($N_{k>0}$) and also non-interacting ones (n_0) - not included in the investigated data sets, so that:

$$N_P = n_0 + N_{k>0} \qquad \text{(A.1.2)}$$

As eq. A.1.1 is fulfilled for each node degree and for different decay constants d_1 and d_2 , it should be:

$$a_1^* = a_1 / sc \qquad \text{(A.1.3)}$$

$$a_2^* = a_2 / sc \qquad \text{(A.1.4)}$$

where the scaling factor equals to:

$$sc = \frac{n_0 + N_{k>0}}{N_P^*} \qquad \text{(A.1.5)}$$

Appendix 2

Let us consider protein interaction network containing two classes of proteins (namely 1 and 2) characterized by different dynamics of evolutionary performance, i.e., emerging with the rates q_1 and q_1 (as non-interacting at the beginning), then gaining some interactions with the rates ν_1 and ν_2, and being eliminated with the rates γ_1 and - per protein. All mentioned rates are assumed as being distinct and constant.

A number of selected proteins of a given class δN_i^* (i=1,2), originated within small period of time , vanishes with age a according the equation

$$\frac{d\delta N_i^*}{da} = -\gamma_i \Delta N_i^* \qquad i = 1,\ 2 \qquad \text{(A.2.1)}$$

with an initial condition

$$N_i^*\big|_{a=0} = q_i \delta t \qquad i = 1,\ 2 \qquad \text{(A.2.2)}$$

The resolution of eqs. A.2.1 and A.2.2 represents the exponentially diminishing course

$$\delta N_i^* = q_i \delta t \exp(-\gamma_i a) \qquad i = 1,\ 2 \qquad \text{(A.2.3)}$$

The assumed continuous approximation and linear increase in protein connectivity

$$k = \nu_i a \qquad \text{(A.2.4)}$$

and also the relationship , let us to transform eq. A.2.3 into the formula

$$\delta N_i^* = \frac{q_i}{\nu_i} \delta k \exp(-\frac{\gamma_i}{\nu_i} k) \qquad i = 1,\ 2 \qquad \text{(A.2.5)}$$

which integrated within successive intervals [k, k+1] indicates the number of k-degree proteins of class "i" , , equal to

$$n_{ki}^* = \frac{q_i}{\gamma_i}\left(1 - \exp(-\frac{\gamma_i}{\nu_i})\right)\exp(-\frac{\gamma_i}{\nu_i} k) \qquad i = 1,\ 2 \qquad \text{(A.2.6)}$$

Now, the total distribution of node degree, n_k^* , where $n_k^* = n_{k1}^* + n_{k2}^*$, may be written in the double-exponential form:

$$n_k^* = a_1^* \exp(-d_1 k) + a_2^* \exp(-d_2 k) \qquad \text{(A.2.7)}$$

The symbols introduced above mean

$$a_1^* = \frac{q_1}{\gamma_1}\left(1 - \exp(-\frac{\gamma_1}{\nu_1})\right) \qquad \text{(A.2.8)}$$

$$a_2^* = \frac{q_2}{\gamma_2}\left(1 - \exp(-\frac{\gamma_2}{\nu_2})\right) \qquad \text{(A.2.9)}$$

$$d_1 = \frac{\gamma_1}{\nu_1} \qquad \text{(A.2.10)}$$

$$d_2 = \frac{\gamma_2}{\nu_2} \qquad \text{(A.2.11)}$$

A contribution of "i " class proteins in eqs. A.2.7 formally vanishes for $k > \tau_e \nu_i - 1$, where is the time of evolution of interactome. Thus the index k should not exceed $\max[\tau_e \nu_1 - 1,\ \tau_e \nu_2 - 1]$

Assuming a relatively high value τ_e ($>> 1/\nu_i$), by summation of a superposition of geometrical series n_k^* described by the eq. A.2.7 over $0 \le k \le \infty$, one can obtain the total size of proteome : N_P^*

$$N_P^* = \frac{q_1}{\gamma_1} + \frac{q_2}{\gamma_2} \qquad \text{(A.2.12)}$$

with a distinguished levels of class contribution

$$N_1^* = \frac{q_1}{\gamma_1} \qquad \text{(A.2.13)}$$

and

$$N_2^* = \frac{q_2}{\gamma_2} \qquad \text{(A.2.14)}$$

ISSN:0974-276X JPB, an open access journal