Interaction and Domain Networks of Yeast

Stefan WUCHTY

European Media Laboratory

Corresponding author:

Stefan Wuchty

Schloß-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany

Voice: **49 6221 533 261

Fax: **49 6221 533 298

E-Mail: stefan.wuchty@eml.villa-bosch.de

Keywords: Yeast networks, scale-free and small-world topology, lethality and viability, transitivity of interactions, domain fusion events

Running title: Networks of Yeast

Summary

Data of currently available protein-protein interaction sets and protein domain sets of Yeast are used to set up protein and domain interaction and domain sequence networks. All of them are far from being random or regular networks. In fact, they turn out to be sparse and locally well clustered indicating so-called scale-free and partially small-world topology. This subtle topologies display considerable indirect properties which are measured with a newly introduced transitivity coefficient. Fairly small sets of highly connected proteins and domains shape the topologies of the underlying networks emphasizing a kind of backbone the nets are based on. The biological nature of these particular nodes is further investigated. Since highly connected proteins and domains accumulated a significant higher number of links by their important involvement in certain cellular aspects, their mutational effect on the cell is considered by a perturbation analysis. In comparison to domains of Yeast, it is investigated what factors force domains to accumulate links to other domains in protein sequences of higher eukaryotes.

INTRODUCTION

Tremendous amount of biological data currently available emphasize the necessity to investigate the mutual relationships of genes, proteins and metabolites. The latter were the starting point of considering metaboloms of prokaryotes as complex networks [1–3]. Quite similarly, proteomes offer an opportunity to examine domain architectures of their protein sequences from this perspective [4, 5]. Furthermore, efforts were made to enlighten interactions between families of protein domains. Structural domain data were mapped to a network linking interacting domain structures [6]. Finally, protein-protein interaction networks emerged by employing sets of protein interactions of *H. pylori* [7] and *S. cerevisiae* [8–12].

The results contradict the initial assumption that the connection topology of these biological networks is either completely regular or random. In fact, it appears that they employ subtle topologies situated between these two extremes. Two network models recently introduced result in topologies which are able to describe biological networks more accurately. Primarily, small-world networks were introduced by Watts and Strogartz [13]. This network type turned out to be sparse but much more highly clustered than an equally sparse random graph. It was illustrated with metabolic networks as well as with sociological and technical networks [13–15]. Subsequently, Barabási and Albert introduced a theoretical network model which emphasizes a characteristic connectivity distribution. [16]. A variety of networks emerging from different fields like biology, sociology, technology, linguistics and others have adopted this topology [17].

In this paper, I report the use of protein interaction data to generate an interaction network of *Saccharomyces cerevisiae*. Using the known nonredundant complete Yeast proteome, domain information is used to set up domain sequence and domain interaction networks. Since a comparison of these three types of networks is currently undone, the topologies of these networks will be comparatively studied and biological consequences discussed.

MATERIALS AND METHODS

Definition of networks

A protein-protein interaction graph, G_{p-p}, is defined by a set of nodes which contains a set of interacting Yeast proteins. In order to complete the definition of the network, protein-protein interactions are denoted by a set of undirected edges.

In a coarse grained way, a protein sequence can be computed as a linear arrangement of the domains it contains. Thus, a domain sequence graph, G_D, is formally defined by a set of nodes consisting of all domains which occur in the protein sequences of the Yeast proteome. Two domains are regarded as being undirectedly linked if they co-occur in one of these protein sequences [4].

Since $\sim 95\%$ of all Yeast proteins carry only one type of domain, the construction of a domain interaction graph, G_{d-d}, focuses on interactions involving these particular proteins. Thus, the set of nodes consists of domains which appear in interactions of single sorted domain proteins. Obviously, ambiguity arising from multi-domain interactions is thus avoided. An undirected edge between these domains indicates this relationship.

Sources of protein-protein interaction data

Sets of Yeast protein-protein interactions were collected from several overlapping data compilations [8–11] which employed Yeast two-hybrid experiments extensively. Other relevant interaction data was retrieved from several other protein interaction databases. The database of interacting proteins (DIP,

http://dip.doe-mbi.ucla.edu) scans the literature in order to provide a collection of all functional linkages of proteins obtained by experimental methods [18]. The MIPS Yeast Genome Database (MYGD, http://www.mips.biochem.mpg.de/) [19] is a collection of genetic data from literature which relies on the results of micro array expression experiments. Additionally, MIPS also contains data from Yeast two-hybrid and co-immunoprecipitation experiments.

The protein nomenclature of these data is inconsistent, therefore, the terms were translated to Swiss-Prot/TrEMBL annotations.

Proteome specific data

Yeast specific proteome and protein domain information came from the InterPro database (http://www.ebi.ac.uk/interpro) [20] and the Proteome Analysis database (http://www.ebi.ac.uk/proteome) [21]. Since InterPro employs Swiss-Prot annotation, every protein sequence is itemized with each of its domains.

Network properties

From a theoretical point of view, network topologies set up very differently. Small world networks emerge from regular graphs. With a probability $p > 0$, edges are clipped and randomly rewired. Considering $p = 1$, the completely rewired graph would end up as a random graph [13]. In contrast, scale-free networks emerge in a more 'evolutive' way. In fact, the generation procedure features continuous addition of new nodes considering the quotient $p_i = k_i / \sum_j k_j$ as the probability of connecting node v_i to the newly introduced

one [16]. Note, that the number of nodes of scale-free networks is subject to change, while the number of nodes of small world networks remains constant. Although these network types emerge completely different, both feature a sparse but highly clustered topology.

In order to unravel the topology of an otherwise unknown network, different characteristic values have been defined. In the networks, the degree k of a node is the number of other nodes to which it is connected.

The mean path length of a node v, L_v, is defined as the average of all shortest paths from node v to all other nodes. Accordingly, the mean path length L of the whole network is represented as the average of L_v over all v.

The clustering coefficient of a node v, C_v, measures the fraction of nodes connected to v which are also connected to each other. By extension, the clustering coefficient C of the graph is defined as the average of C_v over all v.

Provided that there exists a sequence of edges $a - b - c$, one might ask to which extent edge $a - c$ are undirectedly linked in the graph. The transitivity coefficient, T_v, represents the mean fraction of neighboring nodes of v which obey this relation. Accordingly, the mean transitivity coefficient, T, is defined as the average of all T_v over all v. This definition resembles the concept of 'transitive triples' which is used in sociology [22].

Some of these characteristic values enables the identification of several network types. Basically, these types are classified by the connectivity distribution $P(k)$ of the nodes. Exponential networks like the small-world [13] and random graph

model [23] are characterized by $P(k)$ which peaks at an average $\langle k \rangle$ and decays exponentially [24]. Both lead to fairly homogenous networks with nodes having approximately the same number of links $k \sim \langle k \rangle$. Furthermore, a small-world graph is defined as sparse, $L \geq L_{random}$, but much more clustered than an equally sparse random graph, $C \gg C_{random}$. By contrast, inhomogeneous networks known as scale-free networks show power-law decay $P(k) \sim k^{-\gamma}$ in their connectivity distribution. Compared to exponential networks, the probability that a node is highly connected ($k \gg \langle k \rangle$) is statistically significant in scale-free networks [16]. This result indicates a network free of a characteristic scale since the network properties are actually independent of the network size. Figure 1 schematically gives an idea of the introduced topologies.

Lethal and viable proteins

Information about lethality and viability of proteins was retrieved from the YPD database (http://www.proteome.com) [25]. Obviously, if one protein proves to be lethal all links in the protein-protein interaction network have to be considered as lethally affected. For the domain-related networks, only fractions of connections prove to be lethal or viable depending on the protein under consideration. Since these networks map protein specific information to a domain dependent space, every link between protein domains does not have to be inevitably proven either lethal or viable. If there exists a domain link which occurs both in one lethal and one viable protein fraction of lethal connections turns out to be 0.5. Hence, these nodes and edges are interesting objects to investigate in regards to their influence on network properties.

Domain fusion events

Since the domain interaction graph focuses on interactions of proteins which carry only one type of domain, the superposition of the protein domain sequence network onto the domain interaction network enables detection of all interaction processes which are accompanied by a domain fusion event on the sequence level. The extent to which domain interactions in Yeast coincide with domain fusions in higher eukaryotes is of particular interest. Species dependent proteome information was retrieved from Proteome Analysis database.

Graph tools

Graph analysis tools were written in C++ using the LEDA library of data types [26].

RESULTS

Network topologies

It is the intention of this work to provide a comparison of these three network types. Thus, some already known results are addressed partially in order to provide a thorough view.

All networks emerge as sparse networks providing mean numbers of edges per vertex $\langle k_v \rangle$ which are far smaller than the maximal possible degree per vertex as shown in Table 1.
Frequency distributions of links immediately reveal the presence of scale-free topology. Thus, frequency distributions follow a power-law $P(k) \sim k^{-\gamma}$ [4, 6, 12]. Figure 2 compares the frequency plots of the networks considered. As a result of this analysis, the curves of frequency distributions of InterPro domain interactions and protein-protein interactions almost coincide. Thus, the assumption that $\sim 95\%$ of the interacting proteins carry only one domain is well reflected.

Additionally, InterPro domain sequence networks have been found to exhibit small-world properties [4]. Addressing the relevant parameters of small-worldedness, the clustering coefficient C of InterPro domain sequence networks far exceeds the respective one of an equally sized random graph. However, the definition of small-world networks additionally demands $L \geq L_{random}$. It turns out that the major component of the domain sequence network which covers the majority of domains fulfills this demand. As a result,

protein and domain interaction networks both feature clustering coefficients which fulfill the definition of small-world networks (Table 2). However, respective numbers of L fail. Although protein interaction and domain interaction networks both feature huge 'major' components, neither of them satisfies this structural demand of small-world networks.

Biological hubs

Scale-free topology suggests that only a minority of the highly connected nodes shape the topology of the underlying network. Since highly connected hubs play a crucial role for information processing and integrity of networks, it is interesting to see which role these nodes play in biological networks. Table 3 shows the 15 most highly connected nodes of each network. Highest linked InterPro domains in domain sequence networks of Yeast were already found to be involved in signal transduction pathways. Other high linked domains appear in transcriptional/translational activities and energy maintenance [4].

In this analysis, the significance of signaling pathways is strongly emphasized in the domain interaction network by WD40 and zinc-finger motifs which are among the highest interacting domains.

Strongest interacting proteins are involved in nucleus related transportation processes. These include subunits of Importin and nucleoporins. Furthermore, cell-cycle regulating (MEC3, TEM1) and transcription processing proteins appear highly interacting.

Transitivity

Since scale-free and small-world networks were found to be sparse but highly clustered, the degree of transitivity can be questioned. The mean transitivity value, T, measures the extent to which indirect links are accompanied by direct ones. Such a 'back-up' of links reinforces the clustered nature of biological (sub)networks. Table 2 shows statistics of the networks under this consideration. Similarly to the behavior of C, T exceeds the respective number of random graphs of equal size, T_r, by far. However, it should be noted that the values are reasonably low.

It might be tempting to assume that T is closely related to C since an edge $a - c$ implies an increase of $C(b)$. In order to investigate the mutual relation of T and C, Figure 3 shows a scatterplot of T against C concerning all three types of networks. Considering Figure 3, symbols indeed arrange around the median axis. However, they are far from indicating a strong correlation.

From a biological point of view, it is interesting to discover the role of proteins which are involved in such a transitive organization. Table 4 shows a compilation of proteins and domains exhibiting highest T_v. Strikingly, the list of interacting proteins is headed by proteins which form enzymatic protein clusters. Among them are PMT and OST proteins setting up Dolichyl-Diphosphooligosaccharide-protein glycosyltransferase protein complex. Similarly, the interacting domains with the highest T-values are protagonists of functional clusters involved in transcription (TFIID-proteins and RNA pol β subunit) and signal transduction. However, the T-values of interacting domains are lower than those of interacting proteins. The picture

changes drastically if T-values of the domain sequence network are considered since these tend to be shifted to higher values.

Lethality and Viability

The separation of viable and lethal proteins allows one to observe protein interaction networks and domain related networks from a different perspective. The frequency distributions of both lethal and viable proteins in the protein-protein interaction network are shown in Figure 4. Initially, it was suggested that strongly interacting proteins can be assigned a lethal role [12]. In fact, this analysis shows significantly that this assumption is misleading since the latter plot indicates merely a slight trend of lethal proteins to accumulate higher numbers of interactions than viable ones. Since the transitivity coefficient takes the existence of alternative paths into account, it might be interesting to check if T is more suited to explain the latter correlation. Figure 5 shows a frequency plot of T regarding lethal and viable proteins. Confirming the latter assumption, lethal proteins indicate a slight trend to higher T. Regarding higher values of T, it clearly appears that lethal proteins tend to accumulate more alternative interaction paths. However, it should be noted that frequencies are considerably low. A similar view holds for lethal and viable fractions of domains in the respective networks. Figure 6 displays frequency distributions of fractions of lethal and viable domain interactions and domain connections in the respective graphs. Analogously, the plots suggest a slight shift to lower fractions of lethal connections in Figure 6.

Observing the mutational effects from a different perspective, Figure 7 displays

frequency distributions of the mean path lengths L and mean clustering coefficients C of the protein-protein interaction network. Results were obtained by deleting separately lethally and viably mutated proteins and subsequent calculation of these network properties. Both types of distributions are normally distributed. The distributions of lethally perturbed networks generally show slightly increased standard deviations. These observations also hold for domain related networks which were considered analogously by clipping fractions of links affected by lethal or viable mutations of the respective proteins.

Domain interactions and fusion events

Functional links between proteins have also been detected by analyzing fusion patterns of protein domains. Separate proteins A and B in one organism are found to be expressed as a fusion protein in other species. A protein sequence containing both A and B is termed a Rosetta Stone sequence [27]. In this analysis, however, this framework does not work in general but only in particular cases. The comparison of pairwise domain interactions and pairwise domain fusions in higher organisms enables an estimation of the extent to which domain interactions are indeed accompanied by a domain fusion event. Pairs of domain interactions and domain links correspond to edges in the respective networks. Considering every domain separately, edges in the Yeast domain interaction network are counted which co-occur in the domain sequence networks of *A. thaliana*, *C. elegans*, *Drosophila*, *H.sapiens* and Yeast, respectively. Subsequently, fractions of domain fusion per domain interaction of the mentioned organisms are calculated. As a result, Figure 8 summarizes an

increasing extent of domain fusions in the latter row of eukaryotes.

DISCUSSION

Completeness and quality of data

The protein-protein interaction data used for the set up of the interaction network are widely based on yeast two-hybrid analyses. However, yeast two-hybrid data are significantly flawed by high rates of false positive signals [28]. Moreover, many of the interactions identified merely rely on positive signals from one single technique and result from indirect observations. The observation that Importin α subunit protein (SRP1) (Table 3) interacts with that number of proteins is merely a result of the two-hybrid screen employed since a very small fraction of those interactions were shown by other methods.

The discovery of scale-freedom in protein and domain related networks alleviates the insurmountable problems arising from the current extent of incompleteness. Even though the current interaction data are far from complete and are somewhat noisy, these findings reinforce the argument that the topology of interaction networks will not change significantly as the amount of interaction data grows. Strictly speaking, the set up of the domain interaction network is an indirect one since interactions are inferred from protein interactions and domain sequence information. In contrast to other approaches, no structural information of domains was taken into account. Thus, it should be noted that domain interaction networks mediate a certain degree of simplification. However, even though the domain interaction network is simplified to a certain degree scale-freeness of the network confirms the assumption that the topology will not change with increasing amount and quality of domain and protein interaction data. Analogously, this assumption also holds for domain sequence networks since the

proteome data have been far better compiled and studied with the release of the complete genomic sequence of *Saccharomyces cerevisiae*. The assumption of scale-free characteristics leads to interaction and domain sequence graphs independent of the actual size of the underlying networks. Although the generation and compilation of interaction data is still at a basic level, these interaction graphs give tentative insights of the underlying network topology.

What do these network architectures tell?

The observation of scale-freedom in all three networks confirms the appearance of sparse but highly clustered nets. As a consequence, highly connected nodes emerge which predominantly shape the topology of the underlying network. Considering the shortest ways through the network, it will become immediately clear that these routes always pass highly connected nodes. Thus, these hubs illustrate crossways helping to transport information quickly to even remote parts of the network.

So, sparsity and strong local clustering of the scale-free nets offer a different view on the organization of the networks considered. Pathways defined by protein and domain interactions might be treated as highly clustered subnets which are sparsely interlinked to other ones. Accordingly, highly interacting proteins and domains can be considered as the ’backbone’ of the networks which interconnect pathways in the respective networks. Otherwise, these nodes might be central proteins and domains which shape a particular pathway. Thus, it is possible to get a good flavor of the general characteristics of the underlying networks without the knowledge of all interactions. This idea intuitively

becomes important since the current interaction data are from being complete as already mentioned in the latter section.

In order to get a flavor how frequent sequences of edges $a - b - c$ are accompanied by co-occuring edges $a - c$, a new measure, mean transitivity coefficient T, was introduced. Similarly to mean clustering coefficient C, T-values of scale-free and small-world networks exceed the respective numbers of equally sized random graphs strongly emphasizing a tendency to reinforce clustering. However, the T-values of all three networks of Yeast indicate the assumption that indirect or alternative linkage might be rather an exceptional than common feature. However, the latter point crucially depends on the set up of networks. Since domain networks were generated by considering domain nodes linked if they co-occur with other ones in proteins, T is subject to a shift to higher values. Nevertheless, this value reflects the extent to which particular sequences of nodes are 'backed up' by an inserted direct link. As already mentioned, proteins which are mainly involved in enzymatic clusters display a high degree of transitivity. Obviously, this result is based on the observation that these proteins nearly interact with each other in the respective protein clusters. Otherwise, two nodes - although already linked - might be connected indirectly by adding an intermediate node. Considering protein and domain interactions, intermediate proteins and domains might be considered as the entry to alternative pathways. Analogously, intermediate domains in domain sequence networks might display access to different domain architectures. Thus, high values of T imply domains which frequently co-occur with the same domains. Since a crucial role for the networks topology coincides with connectedness, highly transitive

nodes might be among the sets of highly connected nodes of the networks considered. However, no evidence to support this assumption was found. In fact, it turns out that rather the opposite is the case since frequent interacting proteins like JSN1 or YHR4 show transitivity coefficients around 0.03. The same holds for highest connected domains in the domain sequence network emphasizing pkinase and WD40 representing transitivity coefficients of 0.49 and 0.09, respectively. In contrast, domains of the domain interaction network apparently contradict this particular trend since highly interacting domains show reasonable high transitivity more frequently. WD40 and RRM which lead the list of highly interacting domains in Table 3 emerge as fairly transitive with values of 0.39 and 0.29, respectively. However, this apparent contradictory trend seems to be more the result of the small sample than a characteristic of interacting domains.

Scale-freedom vs. Small-worldedness

It is significant that the interaction networks differ from domain sequence networks in their lack of distinct small-worldedness. Small-world networks employ rewiring of an otherwise regular graph as the crucial point in their set up. In contrast, preferential attachment of newly introduced nodes results in the occurrence of highly connected hubs in scale-free networks. Obviously, the emergence of scale-free networks is a rough representation of evolution. However, recent works which considered loss of interactions by 'incomplete' implementation of interactions upon duplication of particular proteins [29] lead to networks which exhibit scale-free topology and fit the experimentally observed results very well. Thus, continuous addition and subsequent preferential

attachment are considered to be indirect consequences of this particular process. In contrast, protein domain sequence networks also employ small-world features which emphasise 'rewiring character' within the domain space. Studies of proteomes reveal a variety of domain architectures in higher eukaryotes which allow cellular operations to be maintained without tremendously expanding the size of the genome [4]. Extensive domain shuffling and domain accretion increases the combinatorial diversity of protein sets and is therefore procedurally more similar to the 'rewiring' of already existing domain nodes.

Evolutionary aspects

Compared to the Yeast proteome, domain fusions in the proteomes of higher organism are more frequent [27]. On the one hand, proteome complexity is particularly assumed to be the consequence of protein innovations. On the other hand, proteomes are generated by expansion of protein families and subsequent combinatorial arrangements of domains. Combinatorial diversity provides protein sets which are sufficient to preserve cellular procedures without dramatically expanding the absolute size of the protein complement. The list of highly connected domains in domain sequence networks immediately reveals a substantial lack of overlap in the compilation of single interacting domains. Although the ratios of fusions grow constantly towards organisms of increasing complexity, they remain considerably low. Subsequent fusion of interacting domains seems to be rather an exceptional than a common feature. Accordingly, single domain interactions seem to be no driving force for fusing domains in one sequence. Naturally, one might argue that the number of fusions will be subject

to tremendous change when the Yeast interactome will be further explored. Since the knowledge about the Yeast interactome is far from being complete, the overall trend of increasing numbers of fusions per domain interaction will still be reflected by improved numbers.

Considering highly connected domain nodes in Table 3, the abundance of domains involved in signal transduction pathways like kinases and zinc-finger motifs is conspicuous. Proteins which emerged by fusion of domains or combinatorial diversification of domain architectures are important parts of signal transduction and cell-cell communication pathways of Yeast emphasizing its role as a single cellular organism leading the way to multicellularity. Domains which proved to be fit in different cellular aspects of Yeast are rewarded with an increasing degree of connections in higher eukaryotes emphasizing a sort of 'fit-get-rich' regime [4]. Thus, it might be expectable that these partially highly connected proteins and domains identify as very crucial for the survivability of the cell. However, it turned out that this is not the case. Although perturbation analysis of all three types of networks indicates a tendency of lethal proteins and domains to slightly assemble more crucial effects on the networks, the results are far from offering a clear distinction between lethal and viable sets of proteins and domains. However, it should be kept in mind that this results might be based on the low complexity of Yeast and absence of highly comprehensive data sets. With protein specific data of higher organism, this question will be revisited.

ACKNOWLEDGMENTS

I especially wish to acknowledge useful discussions with Ursula Kummer. Fruitful discussions with Albert-László Barabási, Zsoltan Oltvai, Carel van Gend, Razif Gabdoulline, Isabel Rojas, Carola Busse and Rebecca Wade are also gratefully acknowledged. Ioannis Xenarios gave many information regarding the quality of protein interaction data. Kelly Elkins is gratefully acknowledged for carefully reading the manuscript. Thanks to the Klaus-Tschira-Foundation (KTF) for funding this project.

References

[1] Fell, D., Wagner, A., *Nature Biotech.* 2000, 189, 1121–1122.

[2] Wagner, A., Fell, D., *Proc. R. Soc. Lon. B* 2001, 268, 1803–1810.

[3] Jeong, H., Tombor, B., Albert, R., Oltvai, Z., Barabási, A.-L., *Nature* 2000, 407, 651–654.

[4] Wuchty, S., *Mol. Biol. Evol.* 2001, 18, 1694–1702.

[5] Apic, G., Gough, J., Teichmann, S., *J. Mol. Biol.* 2001, 310, 311–325.

[6] Park, J., Lappe, M., Teichmann, S., *J. Mol. Biol.* 2001, 307, 919–938.

[7] Rain, J.-C., Selig, L., DeReuse, H., Battaglia, V., *et al.*, *Nature* 2001, 409, 211–215.

[8] Ito, T., Tashiro, K., Muta, S., Ozawa, R., *et al.*, *Proc. Nat. Acad. Sci.* 2000, 97, 1143–1147.

[9] Uetz, P., Giot, L., Cagney, G., Mansfield, T., *et al.*, *Nature* 2000, 403, 623–627.

[10] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., *et al.*, *Proc. Nat. Acad. Sci.* 2001, 98, 4569–4574.

[11] Schwikowski, B., Uetz, P., Fields, S., *Nature Biotechn.* 2000, 18, 1257–1261.

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

[13] Watts, D., Strogatz, S., *Nature* 1998, 393, 440–442.

[14] Milgram, S., *Psychology Today* 1967, 2, 60–67.

[15] Newman, M., *Proc. Nat. Acad. Sci.* 2001, 98, 404–409.

[16] Barabási, A., Albert, R., *Science* 1999, 286, 509–512.

[17] Albert, R., Barabási, A., *Rev. Mod. Phys.* 2002, 74, 47.

[18] Xenarios, I., Fernandez, E., Salwinski, L., Duan, X., *et al.*, *Nucl. Acids Res.* 2001, 29, 239–241.

[19] Mewes, H., Frishman, D., Gruber, C., Geier, B., *et al.*, *Nucl. Acids Res.* 2000, 28, 37–40.

[20] Apweiler, R., Attwood, T., Bairoch, A., Bateman, A., *et al.*, *Nucl. Acids Res.* 2001, 29, 37–40.

[21] Apweiler, R., Biswas, M., Fleischmann, W., Kanapin, A., *et al.*, *Nucl. Acids Res.* 2001, 29, 44–48.

[22] Wasserman, S., Faust, K., *Social Network Analysis*, Cambridge University Press, New York, USA, 1994.

[23] Erdös, P., Rényi, A., *Publ. Math. Inst. Hung. Acad. Sci.* 1960, 5, 17–61.

[24] Barabási, A., Albert, R., Jeong, H., *Physica A* 1999, 272, 173–187.

[25] Costanzo, M., Crawford, M., Hirschmann, J., Kranz, J., *et al.*, *Nucl. Acids Res.* 2001, 29, 75–79.

[26] Mehlhorn, K., Naeher, S., *The LEDA Platform of Combinatorial Computing*, Cambridge University Press, Cambridge, 1999.

[27] Marcotte, E., Pellegrini, M., Ng, H.-L., Rice, D., *et al.*, *Science* 1999, 285, 751–753.

[28] Hazbun, T., Fields, S., *Proc. Natl. Acad. Sci.* 2001, 98, 4277–4278.

[29] Wagner, A., *Mol. Biol. Evol* 2001, 18, 1283–1292.

Table 1 Some basic data for the domain sequence, G_D, domain interaction, G_{d-d}, and protein-protein interaction graphs, G_{p-p}.

	G_D	G_{d-d}	G_{p-p}
n_v	1196	394	3212
$\langle k_v \rangle$	1,49	3,06	3,79
$n_{conn.\,comp.}$	653	19	89

Table 2 Mean path lengths, L, clustering coefficient, C, and transitivity coefficient, T of the domain sequence, G_D, domain interaction, G_{d-d} , protein-protein interaction graphs, G_{p-p} , and respective random graphs ($_r$).

	L/L_r	C/C_r	T/T_r
G_D	5,25/13,33	0,1989/0,0012	0,0401/4×10^{-4}
G_{d-d}	4,01/5,18	0,0816/0,0031	0,0296/0.0031
G_{p-p}	4,85/6,18	0,0806/4×10^{-4}	0.1288/9×10^{-4}

Table 3 The 15 most highly connected nodes of the protein-protein interaction graph, G_{p-p}, the domain interaction graph, G_{d-d} , and the domain sequence graph, G_D.

$v^{G_{p-p}}$	k_v	$v^{G_{d-d}}$	k_v	v^{G_D}	k_v
JSN1	230	WD40	53	zf-C3HC4	21
Importin α subunit	123	RRM	46	pkinase	19
ATP14	108	Zn2-CY6-fungal	39	Ser-Thr-kin-actsite	19
TIR1 precursor	107	snRNP-Sm	28	AAA	19
NUP116/NSP116	107	vATP-synt AC39	24	PH	18
SRB4	92	zf-C2H2	22	EF-hand	16
TFIIB	81	cyclin	20	C2	15
YHR4	72	Ser/Thr-phosphatase	18	WD40	14
VMA6	71	TPR	16	DEAD	14
PGDH	69	SH3	15	helicase C	14
MEC3	65	bZIP	12	ATP-GTP-A	14
TEM1 protein	61	TFIID	11	AA-tRNA-ligase-II	14
SOH1 protein	50	Myb-DNA-bind	11	CPSase	14

LYS14 protein	50	zf-CCHC	11	GATase-1	13
Importin β-1 subunit	40	Histone-core	11	FMN-binding enzyme	12

Table 4 The 15 most transitive nodes of the protein-protein interaction network, G_{p-p}, domain interaction network, G_{d-d} , and domain sequence network, G_D.

$v^{G_{p-p}}$	T_v	$v^{G_{d-d}}$	T_v	v^{G_D}	T_v
WBP1	0,92	STT3	0,50	DNAtopI-DNA-bind	1,0
PMT3	0,91	DAD	0,50	DNAtopI-ATP-bind	1,0
PMT4	0,91	RNA pol β s.u.	0.44	DNA pol β-like	1,0
PMT2	0,91	MCM	0,40	Interleukin-1	1,0
OST5	0,90	WD40	0,39	RNA-polII-repeat	1,0
OST3	0,90	T-SNARE	0,38	ATPase-α-β	1,0
OST2	0,90	Ribosomal-S12	0.36	Tubulin	1,0
STT3	0,90	BK-channel-α	0,33	CytC-heme-bind	1,0
SWP1	0,89	TFIID-31	0,31	6-P-fructo-2-kinase	1,0
UBC5	0,80	Synaptobrevin	0,31	RNA-pol-A	1,0
UBC4	0,80	TFIID-18	0,28	Gluc-transporter	1,0
OST4	0,79	Znf-CCHC	0,28	Dynamin	1,0
ALG5	0,76	Histone-core	0,28	Helix-hairpin-helix motif cl. 2	1,0

VPS16	0,75	Znf-C2H2	0,27	Middle domain of eIF4G	1,0
PEP3	0,75	DNA-RNApol-7kD	0,25	GHMP-kinase	1,0

Figure 1 Models of exponential and scale-free networks.

Diameters of circles indicate the number of connections respective nodes have. $P(k)$ is the frequency that nodes have k connections. Top: Exponential networks consist of nodes which show similar numbers of links to other nodes. Thus, the frequency distribution peaks at an average and decays exponentially. Bottom: In fact, biological networks adopt scale-free topology. A fairly small amount of highly connected nodes which show much higher numbers of connection than the average shapes a straight line in the log-log plot of the connectivity distribution.

Figure 2 The frequency distribution of the protein-protein interaction graph, G_{p-p} , domain interaction graph, G_{d-d} , and domain sequence graph, G_D.

The numbers of links to other vertices were logarithmically binned and frequencies thus obtained.

Figure 3 Scatterplot of mean clustering coefficient C_v vs. mean transitivity coefficient C_v concerning the protein-protein interaction graph, G_{p-p} , domain interaction graph, G_{d-d} , and domain sequence graph, G_D.

Figure 4 Frequency distributions of links which are set up by interactions of lethal and viable proteins.

The compilation of lethal and viable proteins was retrieved from the YPD database.

Figure 5 Frequency distributions of mean transitivity coefficient T_v which are set up by interactions of lethal and viable proteins.

The compilation of lethal and viable proteins was retrieved from the YPD database.

Figure 6 Frequency distributions of fraction of lethal and viable links per domain in domain interaction and domain sequence networks.

The compilation of lethal and viable proteins was retrieved from the YPD database.

Figure 7 Distributions of the mean path length L_v and mean clustering coefficient C_v of the protein-protein interaction network.

Lethally and viably mutated proteins were clipped and network parameters thus obtained. Protein information was retrieved from the YPD database.

Figure 8 Histogram of domain fusion events per domain interaction.

The co-occurrence of domain interactions found in *S.cerevisiae* and domain fusion events were detected in *S.cerevisiae*, *A. thaliana*, *C.elegans*, *Drosophila* and *H.sapiens*. Domain information was obtained from InterPro domain database.

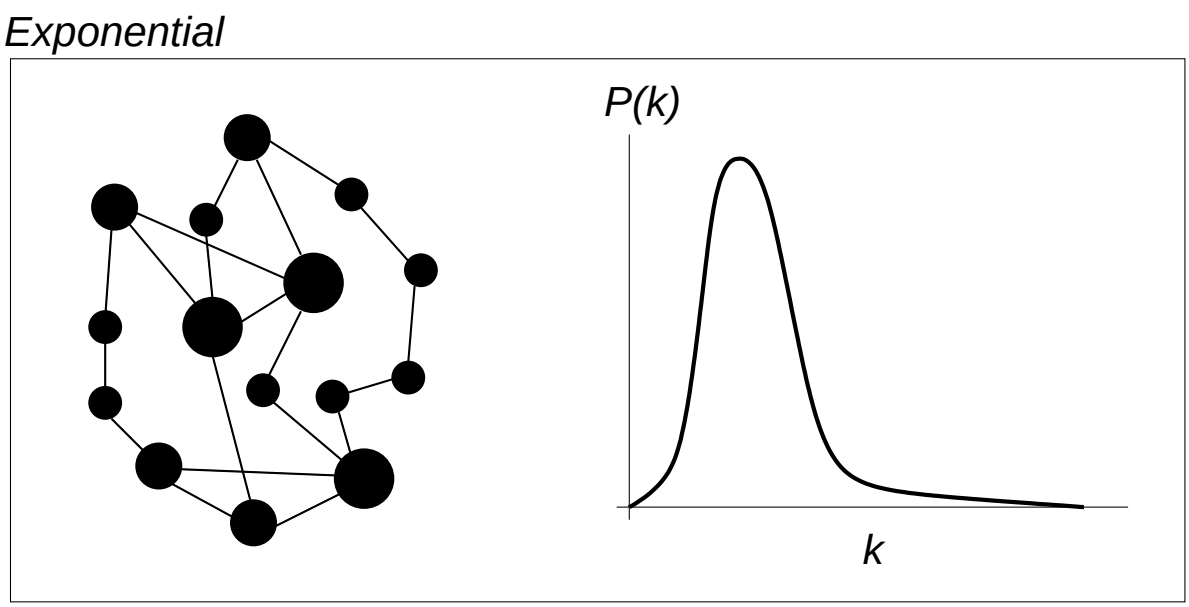

Figure 1:

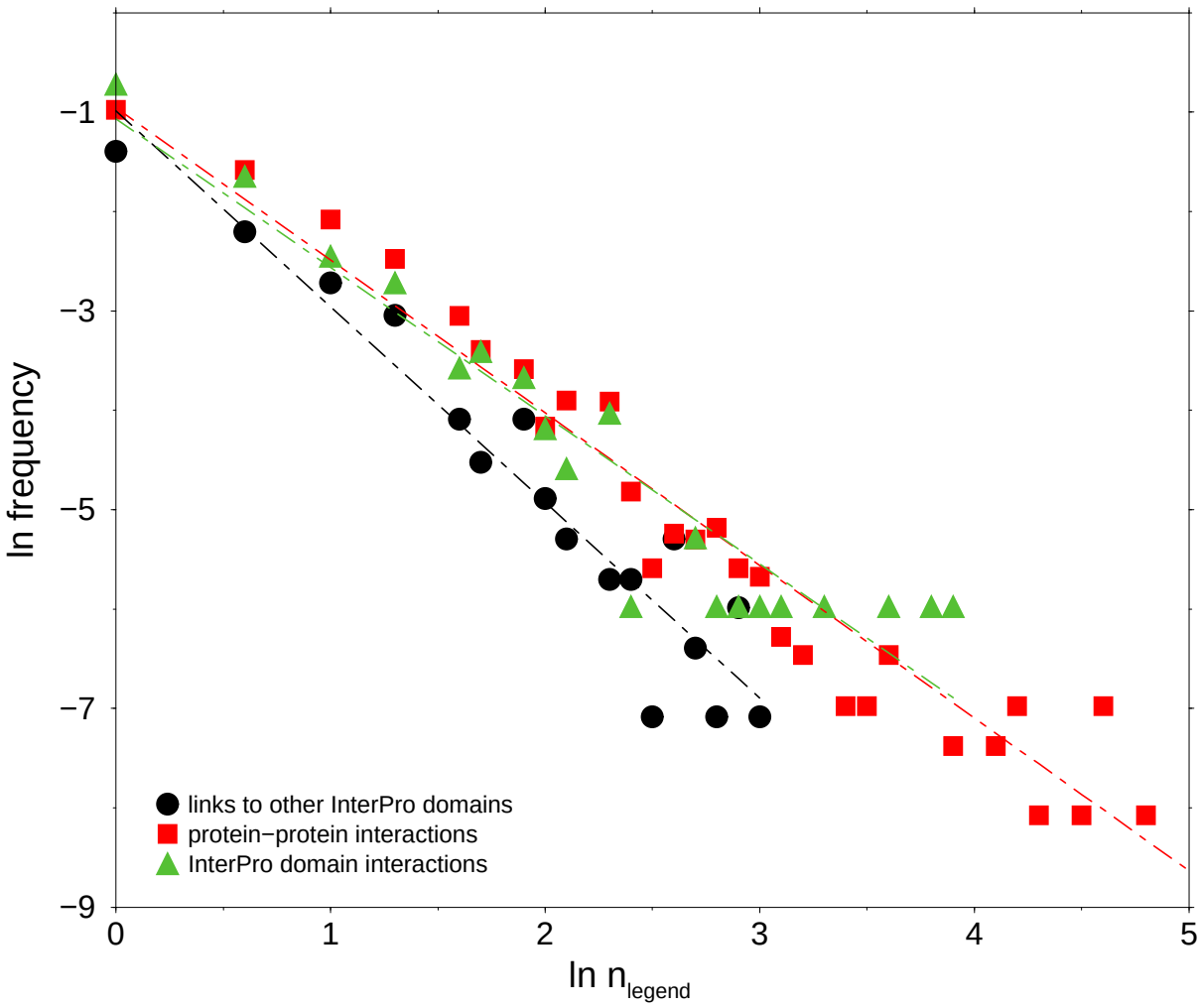
ln frequency
−1
−3
−5
−7
−9
0
1
2
3
4
5
ln n legend
links to other InterPro domains
protein−protein interactions
InterPro domain interactions

Figure 2:

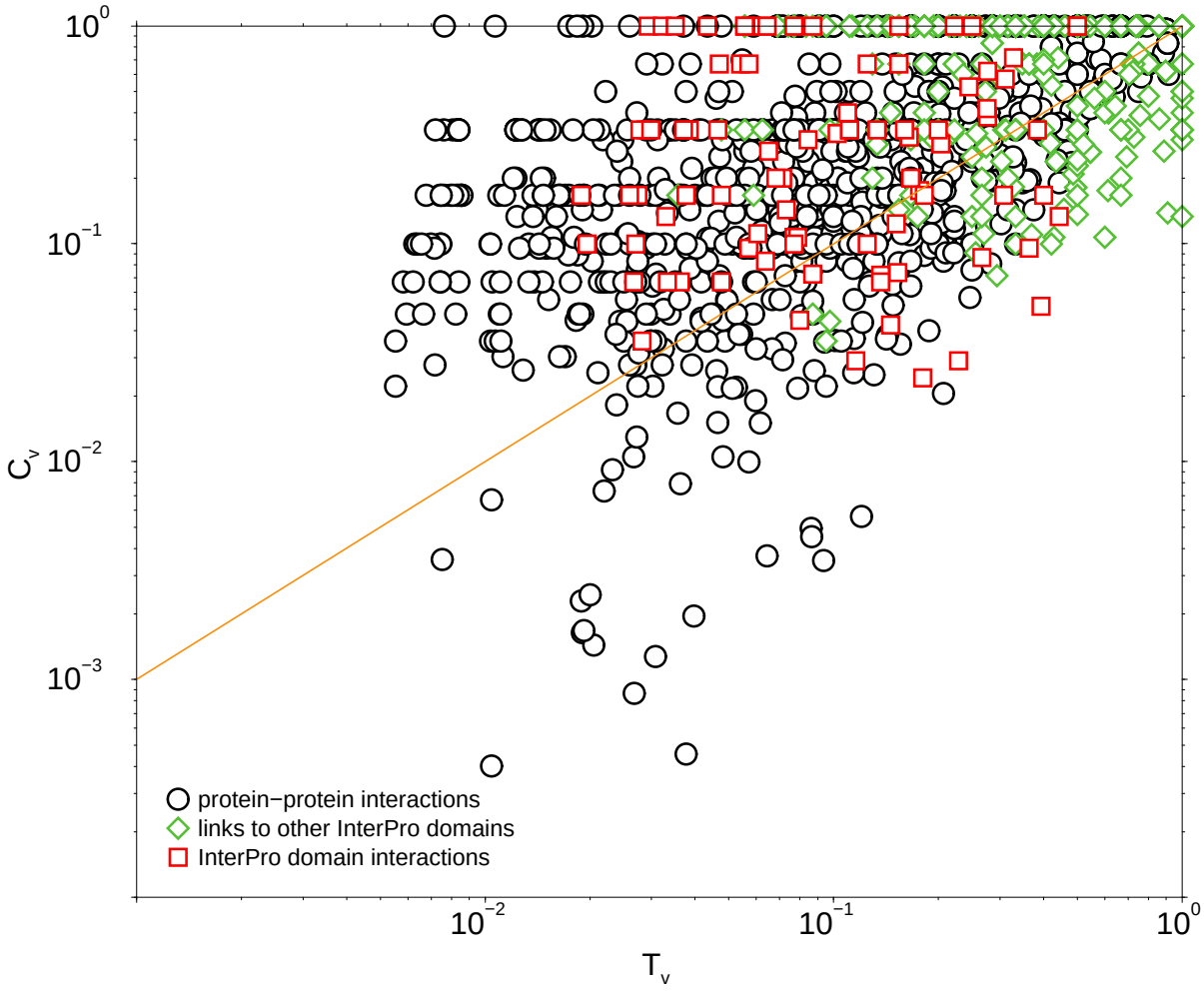
10^0
10^{-1}
10^{-2}
10^{-3}
C_v
10^{-2}
10^{-1}
10^0
T_v
protein−protein interactions
links to other InterPro domains
InterPro domain interactions

Figure 3:

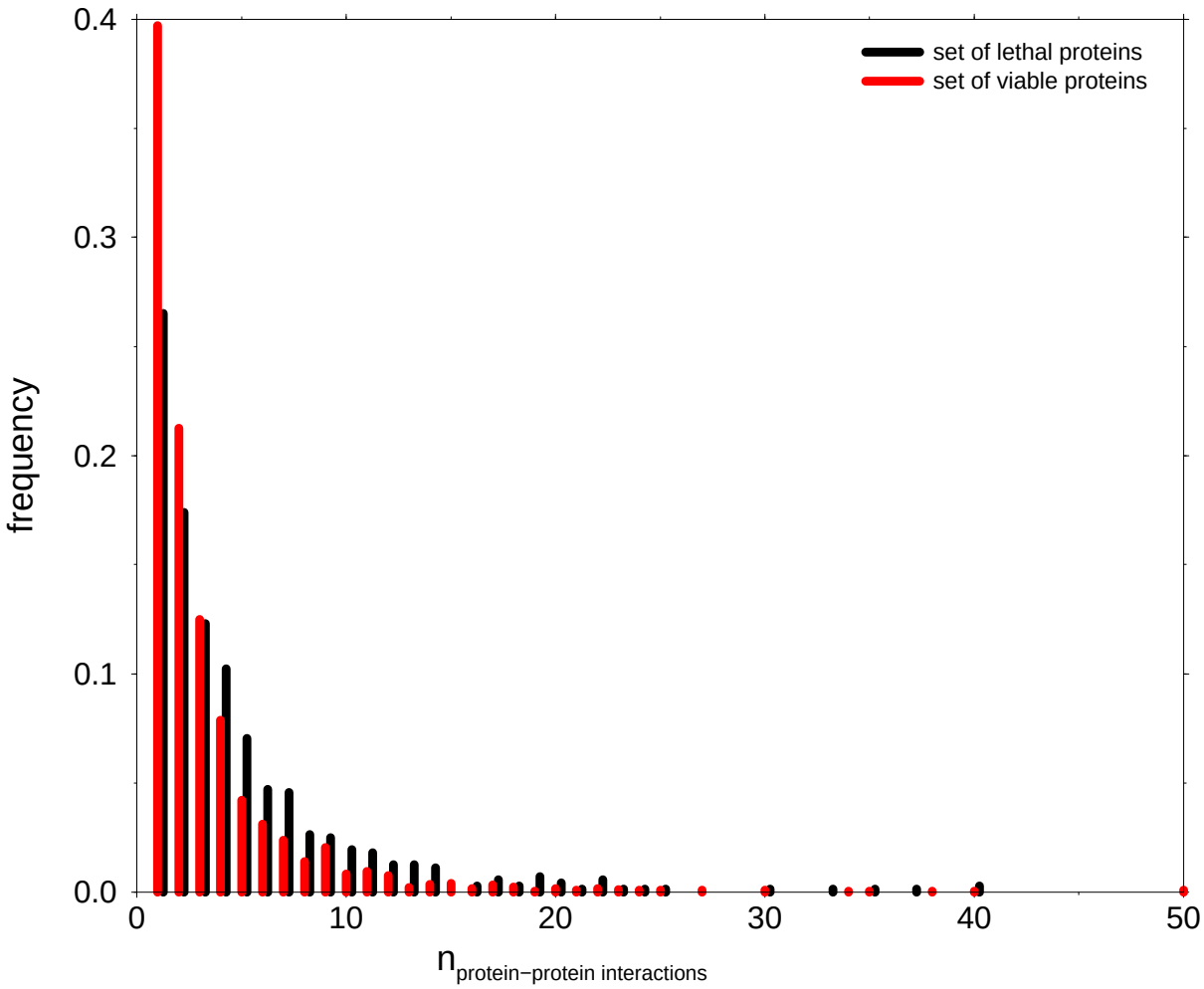
0.4
0.3
0.2
0.1
0.0
frequency
0
10
20
30
40
50
n protein−protein interactions
set of lethal proteins
set of viable proteins

Figure 4:

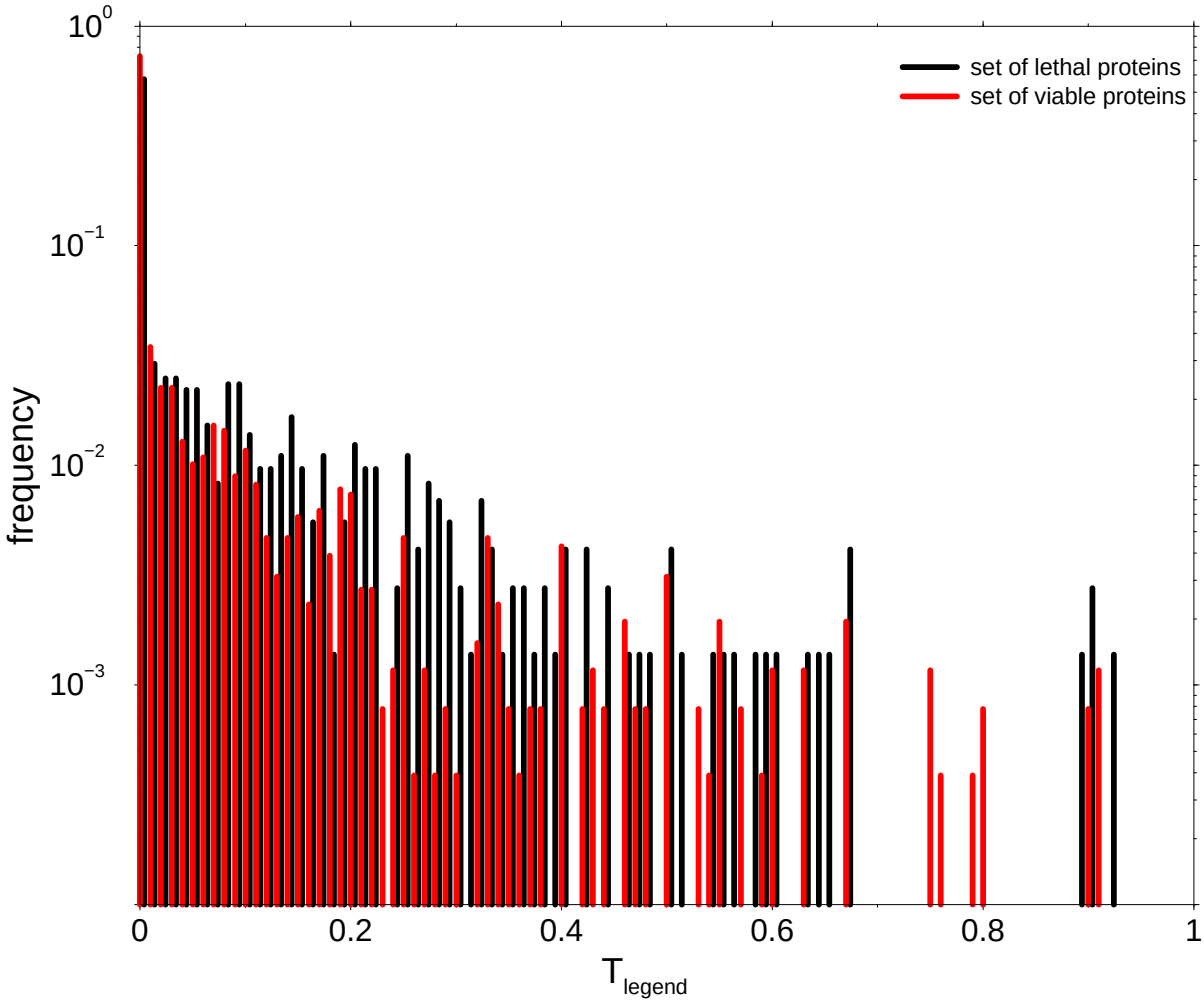
set of lethal proteins
set of viable proteins
frequency
10^{0}
10^{-1}
10^{-2}
10^{-3}
0
0.2
0.4
0.6
0.8
1
T_{legend}

Figure 5:

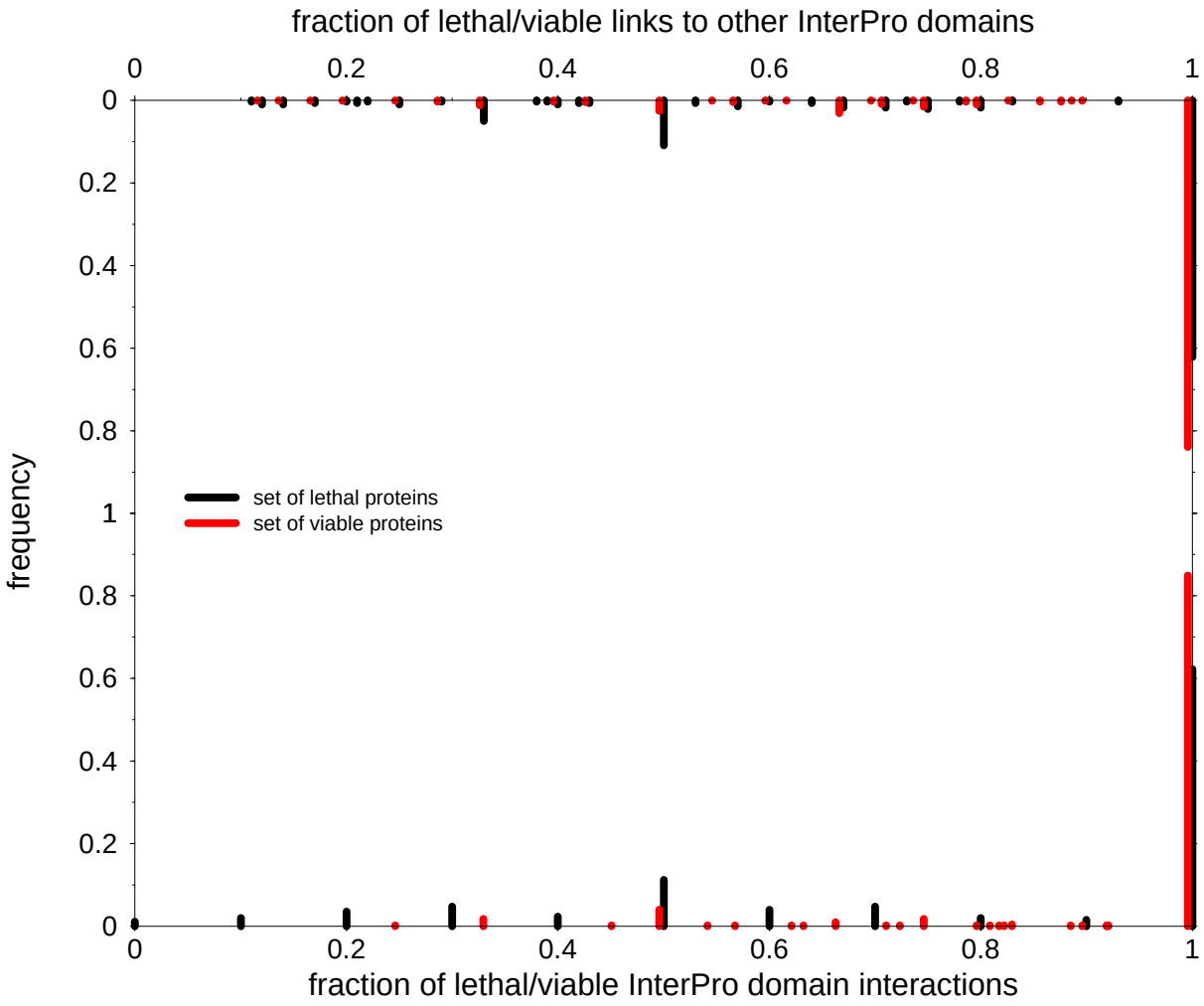
fraction of lethal/viable links to other InterPro domains
0
0.2
0.4
0.6
0.8
1
frequency
set of lethal proteins
set of viable proteins
fraction of lethal/viable InterPro domain interactions

Figure 6:

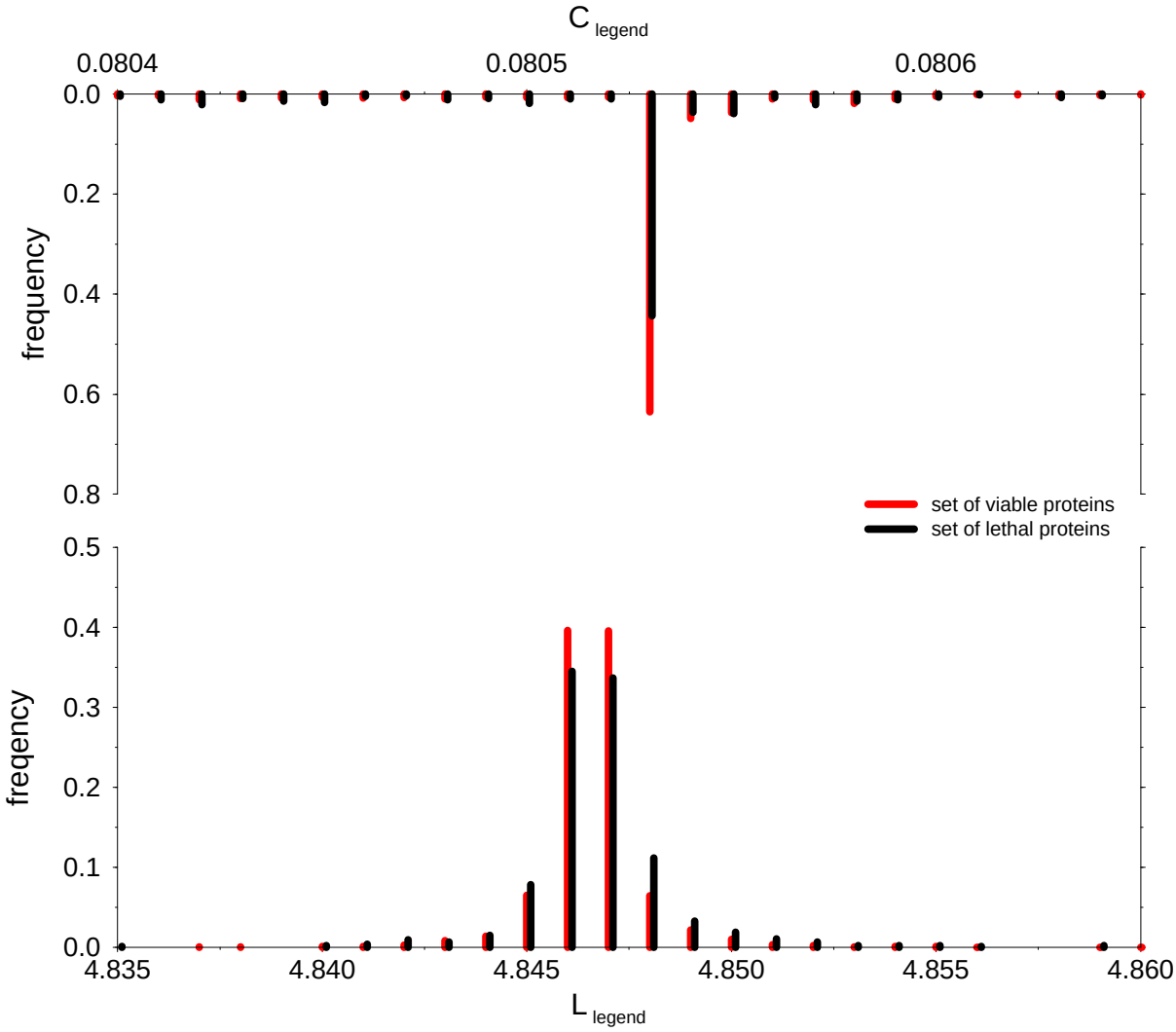
C legend
0.0804
0.0805
0.0806
0.0
0.2
0.4
0.6
0.8
frequency
set of viable proteins
set of lethal proteins
0.5
0.4
0.3
0.2
0.1
0.0
freqency
4.835
4.840
4.845
4.850
4.855
4.860
L legend

Figure 7:

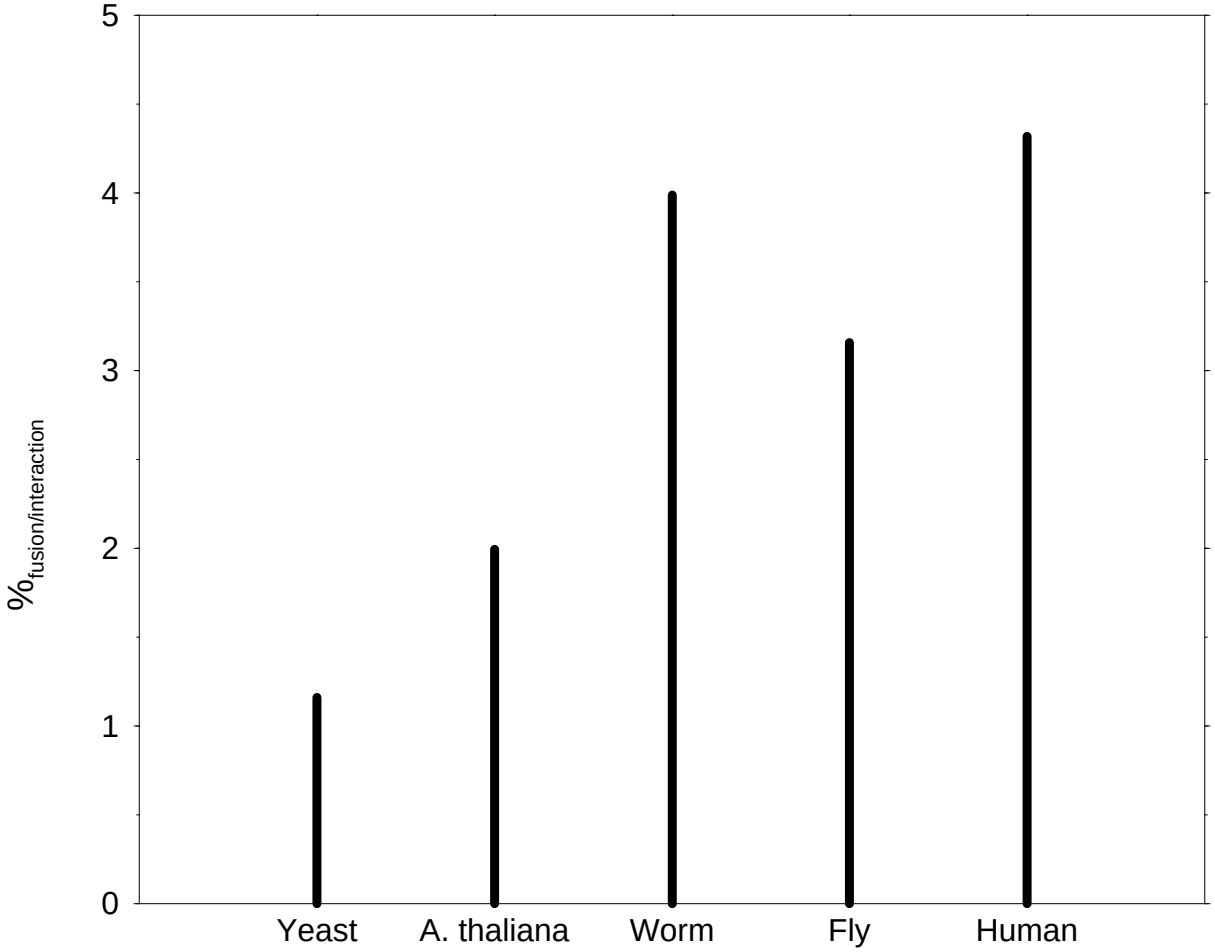

Figure 8: