Scale-Free Behavior in Protein Domain Networks

Stefan Wuchty

European Media Laboratory, Heidelberg, Germany

Several technical, social, and biological networks were recently found to demonstrate scale-free and small-world behavior instead of random graph characteristics. In this work, the topology of protein domain networks generated with data from the ProDom, Pfam, and Prosite domain databases was studied. It was found that these networks exhibited small-world and scale-free topologies with a high degree of local clustering accompanied by a few long-distance connections. Moreover, these observations apply not only to the complete databases, but also to the domain distributions in proteomes of different organisms. The extent of connectivity among domains reflects the evolutionary complexity of the organisms considered.

Introduction

Many diverse systems may best be described as networks with complex topologies. Often, the connection topology is assumed to be either completely regular or completely random (Erdös and Rényi 1960). Watts and Strogatz (1998) revealed a new class of network topologies that lies somewhere between these two extremes. Originally, these small-world networks were generated by randomly rewiring nodes in a regular network. Small-world networks combine the local clustering of connections characteristic of regular networks with occasional long-range connections between clusters, as can be expected to occur in random networks. By defining measures that distinguish these three types of networks, Watts and Strogatz (1998) showed that several biological, technological, and social networks are of the small-world type. A small-world graph is formally defined as a sparse graph which is much more highly clustered than an equally sparse random graph. Barthélémy and Amaral (1999) provided evidence that the appearance of small-world behavior is not a phase transition, but a crossover phenomenon which depends on both the network size and the degree of disorder. Small-world graphs were first illustrated with friendship networks (Milgram 1967) in sociology, often referred to as ‘‘six degrees of separation’’ (Guare 1990). The architecture of the power grid of the western United States, the structures of some sociological networks dealing with mathematical collaborations on publications, and the casting of actors in movies were found to be small-world graphs (Watts and Strogatz 1998).

Barabási and Albert (1999) introduced a theoretical model that generates graphs demonstrating a connectivity distribution which decays as a power-law. This feature was found to be a direct consequence of the following two generic mechanisms: (1) networks are allowed to expand continuously by the addition of new vertices, and (2) these newly added nodes attach preferentially to sites that are already well connected (Barabási and Albert 1999). Since this feature is independent of the actual size of the network, this class of inhomogeneous networks was called scale-free networks. The topology of the World Wide Web was investigated by considering HTML documents as vertices connected by links pointing from one page to another (Albert, Jeong, and Barabási 1999; Barabási and Albert 1999; Barabási, Albert, and Jeong 2000). The latter net, as well as the Internet which emerges from connecting different servers, demonstrates scale-free properties. Both nets display a high degree of robustness against errors (Albert, Jeong, and Barabási 2000). However, these networks are highly vulnerable to perturbations of the highly connected nodes.

Biological Networks

Recently, scale-free and small-world behaviors have also been found in biological networks. Watts and Strogatz (1998) reported the architecture of the *Caenorhabditis elegans* nervous system to show significant small-world behavior. Fell and Wagner (2000) assembled a list of stoichiometric equations representing the central routes of energy metabolism and small-molecule building block synthesis in *Escherichia coli*. A substrate graph, defined by a vertex set consisting of all metabolites that occur in the network, was constructed. Two metabolites were considered to be linked if they occurred in the same reaction. Fell and Wagner (2000) found the substrate graph to be sparse, with glutamate, coenzyme A, 2-oxoglutarate, pyruvate, and glutamine having the highest degree of connectivity. This sample of metabolites might be viewed as a core of *E. coli* metabolism which was found without any subjective criteria.

Most recently, Jeong et al. (2000) comparatively analyzed metabolic networks of organisms representing all three domains of life. The metabolic network is represented by nodes, the substrates, connected by directed edges symbolizing the actual reaction. The topologies of these networks are best described by a scale-free model. Furthermore, the diameters of the nets remain the same for all of these networks regardless of the number of substrates found in the given species. Interestingly, the ranking of the most connected substrates is largely identical for all organisms, thus indicating hubs which dominate the topology of the nets. Like the technical networks, the *E. coli* network theoretically has high tolerance to random errors but severe sensitivity to the removal of the highly connected nodes.

Key words: protein domains, scale-free and small-world topology, evolution of protein architectures.

Address for correspondence and reprints: Stefan Wuchty, European Media Laboratory, Schloß-Wolfsbrunnenweg 33, D-69118 Heidelberg, Germany. E-mail: stefan.wuchty@eml.villa-bosch.de.

Mol. Biol. Evol. 18(9):1694–1702. 2001

Another biochemical network is formed by sets of domains which are linearly arranged in protein sequences. This might generate graphs comprising interesting features. Since the topology of graphs thus generated is still unknown, it is worth considering this way of treating domain architectures.

Domain Organization

Protein crystallography reveals that the fundamental unit of protein structure is the domain. Independent of neighboring sequences, this region of a polypeptide chain folds into a distinct structure and mediates biological functionality (Janin and Chothia 1985). Most proteins contain only one single domain (Doolittle 1995). Some sequences appear as multidomain proteins adopting different linear arrangements of their domain sets. On average, such domain architectures comprise two to three domains; however, some human proteins contain up to 130 domains (Li et al. 2001).

Similar to the discussion about the role of certain metabolites in the emergence of metabolism, there has been a debate about the actual number of existing domains and their origin. One view treats all past and present proteins as the result of shuffling of a large set of primordial polypeptides (Dorit and Gilbert 1991). These are assumed to result from splicing events involving exons separated by introns (Gilbert and Glynias 1993). The other view deals with the existence of a few small polypeptides in the early stages of life; these are the predecessors of most contemporary proteins (Doolittle 1995). Gene duplication and subsequent modification were employed to form the latter molecules from this small set of polypeptides. Independent of the timing for the introduction of introns, recombination in introns provides a mechanism for the exchange of exons between genes. This mechanism for the acquisition of new functions by eukaryotic genes is commonly known as ‘‘exon shuffling.’’ It was assumed that primitive proteins were encoded by exons that were spliced together (Seidel, Pompliano, and Knowles 1992). However, such shuffling events take on biological significance only if the exons involved carry a functional or structural domain. Although many examples of exon shuffling have been found, no significant correspondence between exons and units of protein structure has been detected (Stoltzfus et al. 1994).

It is common to find that newly sequenced proteins are homologous to some other known proteins over parts of their lengths. Thus, most proteins may have descended from relatively few ancestral types. The sequences of large proteins often show signs of having evolved by the joining of preexisting domains in new combinations. Such a mechanism is commonly known as ‘‘domain shuffling’’ and appears as two types: domain duplication and domain insertion (Doolittle 1995). Domain duplication refers to the internal duplication of at least one domain in a gene. Domain insertion denotes the process by which structural or functional domains are exchanged between proteins or inserted into a protein. Shuffling of domains has more biological significance than exon shuffling because domains are real structural and functional units in proteins, while exons often are not.

Functional links between proteins have also been detected by analyzing the fusion patterns of protein domains. Two separate proteins A and B in one organism may be expressed as a fusion protein in other species. A protein sequence containing both A and B is termed a Rosetta Stone sequence. However, this framework applies only in a minority of cases (Marcotte et al. 1999).

Protein Domain Databases

Currently, there are a large variety of databases, each collecting protein domain information in completely different ways. The Prosite database (http://expasy.proteome.org.au/prosite/) consists of biologically significant motifs and profiles determined and formulated with appropriate computational tools. Uncharacterized proteins are assigned to certain protein families with the aid of weight matrices and profiles (Hofmann et al. 1999). The majority of Prosite documentation refers to motifs thus providing combined motif and domain information. Release 16.0 of Prosite contains 1,374 different patterns, rules, and profiles.

Another database is Pfam (http://www.sanger.ac.uk/Software/Pfam/index.shtml), which is a large collection of multiple-sequence alignments of protein families and profile hidden Markov models (Bateman et al. 2000). Moreover, Pfam contains curated documentation for all 2,478 families in version 5.5, covering nearly 65% of SwissProt release 38 and SP-TrEMBL release 11.

Many more protein families are found, however, in ProDom (http://www.toulouse.inra.fr/prodom.html) (Corpet et al. 2000), which contains all protein domain families that can be generated automatically from the SwissProt and TrEMBL sequence databases (Bairoch and Apweiler 2000). Expert-validated families are extended by using Pfam seed alignments to build new ProDom families with the Psi-Blast database searching algorithm (Altschul et al. 1997). Other families are generated by recursive use of Psi-Blast. ProDom, version 99.2, has 157,648 domain families, covering almost 95% of SwissProt release 37 and TrEMBL release 10. ProDom offers higher coverage than Pfam. However, ProDom tends to overpredict the number of protein families which can be discovered as subsets of larger families.

Finally, InterPro (http://www.ebi.ac.uk/interpro) (Apweiler et al. 2001*a*) is an integrated documentation resource of protein families, domains, and functional sites rationalizing the complementary efforts of the Prosite, Pfam, ProDom, and Prints (Attwood et al. 2000) database projects. InterPro contains manually curated documentation and diagnostic signatures from these databases and uses these to create a unique, nonredundant characterization of protein families, domains, and functional sites.

Proteome Databases

The advent of fully sequenced genomes of various organisms has facilitated the investigation of proteomes.

The Proteome Analysis database (http://www.ebi.ac.uk/proteome) (Apweiler et al. 2001*b*) has been set up to provide comprehensive statistical and comparative analyses of the predicted proteomes of fully sequenced organisms. The analysis is compiled using mainly InterPro and CluSTr (Kriventseva et al. 2001) and is performed on the nonredundant complete proteome sets of SwissProt and TrEMBL entries. The latest release provides 41 nonredundant proteomes of genomes of archaea, bacteria, and eukaryotes.

Most recently, SwissProt and Ensembl have prepared a complete nonredundant human proteome set consisting of 30,585 sequences. The set consists of the combination of the SwissProt/TrEMBL nonredundant human proteome set (15,691 sequences) and additional nonredundant peptides predicted by Ensembl (14,894 sequences). Ensembl (http://www.ensembl.org) provides complete and consistent annotation across the human genome.

In this paper, domain networks generated with data from ProDom, Pfam, and Prosite domain databases will be presented. Furthermore, InterPro domain networks of different species that are generated with complete proteome sets provided by the Proteome Analysis database will be considered. Subsequently, the topology of these networks will be investigated, and biological and evolutionary consequences will be discussed.

Materials and Methods

A domain graph $G_D = (V_D, E_D)$ is formally defined by a vertex set V_D consisting of all domains found within proteins. Two domains are regarded as being adjacent if they occur together in one protein at least once. An undirected edge connecting these two vertices indicates this relationship. Such connections define the edges set E_D. In this graph, the degree k of a vertex is the number of other vertices to which it is linked. The mean path length L from a vertex to any other vertex of the graph is defined as the average of the path lengths to all other vertices. Another important quantity is the clustering coefficient $C(v)$ of a vertex v. It measures the fraction of the vertices connected to v which are also connected to each other. In extension, the clustering coefficient C of the graph is defined as the average of $C(v)$ over all v.

Growing amounts of empirical and theoretical data about the topologies of large complex networks indicate the emergence of several network types. Basically, these types are classified by the connectivity distribution $P(k)$ of nodes. Exponential networks are characterized by $P(k)$, which peaks at an average $\langle k \rangle$ and decays exponentially. Prominent protagonists of this type are the random graph model (Erdös and Rényi 1960) and the small-world model (Watts and Strogatz 1998). Both lead to fairly homogenous networks with nodes comprising approximately the same number of links $k \sim \langle k \rangle$ (Barabási, Albert, and Jeong 1999). Furthermore, a small-world graph adopts a sparse topology, $L \geq L_{\mathrm{random}}$, but remains more highly clustered than an equally sparse random graph, $C \gg C_{\mathrm{random}}$ (Watts and Strogatz 1998). By contrast, in the class of inhomogeneous networks called scale-free networks, the connectivity distribution decays as a power-law $P(k) \sim k^{-\gamma}$. The latter result indicates a network free of a characteristic scale. Compared with exponential networks, the probability that a node is highly connected ($k \gg \langle k \rangle$) is statistically significant in scale-free networks (Barabási and Albert 1999).

Table 1
Some Basic Data for the ProDom, Prosite, and Pfam Graph

	ProDom	Pfam	Prosite
n_v	5,995	2,478	1,360
$\langle k_v \rangle$	2.33	1.12	0.77
$n_{\mathrm{conn.comp.}}$[a]	1,394	1,396	809
$n_{\mathrm{unconn.dom.}}$[a]	975	1,316	577

[a] $n_{\mathrm{conn.comp.}}$ denotes the number of connected components of the underlying graph. $n_{\mathrm{unconn.dom.}}$ denotes the number of domains which appear unlinked in the respective graph.

In this study, protein domain information was retrieved from the ProDom, Prosite, and Pfam databases. Sixty-five percent of all ProDom sequences correspond to families containing 10 or more members. In order to restrict the size of the network, the sample of ProDom domains focuses on these families. Thus, 5,995 ProDom domains were obtained. The Prosite database declares false-negative entries which were filtered out of the sample used for the network construction. Sequence entries of each database provide SwissProt annotation. Thus, every protein sequence was itemized with each domain that it contained. This was done for each database separately. Domains which were listed due to their occurrence in one protein sequence represent vertices which are connected to each other in the domain graphs.

Complete proteome data sets of different species were retrieved from the Proteome Analysis database, which uses InterPro annotation of protein domains. Such proteome data sets adopt SwissProt, TrEMBL, TrEMBLnew, and Ensembl annotation of proteins. Analogously, InterPro domains which appear along with other domains in a protein sequence represent vertices which are connected to each other in the domain graphs. The numbers of links to other domains in such graphs were logarithmically binned, and frequencies were thus obtained. Such pairs of values were subjected to a linear regression procedure.

PAJEK (the Slovene word for spider), a program for large-network analysis and visualization, was used for the calculation of the latter values (Batagelj and Mrvar 1998). This program is available at http://vlado.fmf.uni-lj.si/pub/networks/pajek/.

Results

The domain graphs are sparse, with small average degrees (table 1) compared with the maximal possible degree $k = n - 1$, where n is the number of vertices. In this respect, the results of figure 1 are interesting. The vertices which denote Prosite domains were ranked by the frequency of their connectivity. The curve is similar to a generalized Zipf's law curve, in which it is ob-

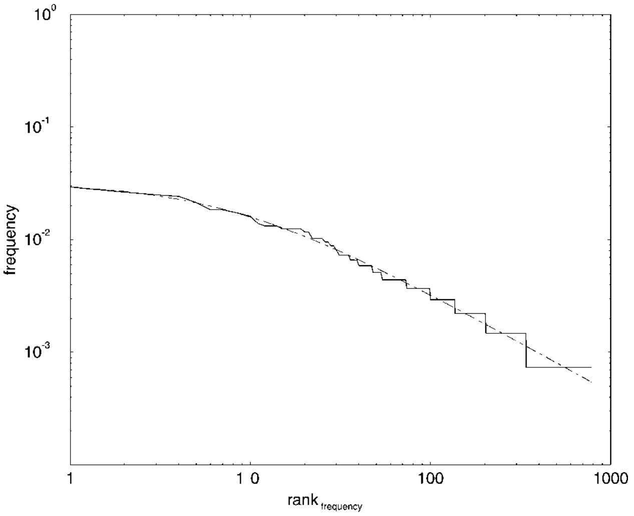

FIG. 1.—The frequency distribution of Prosite domain connectivity. The number of links to other domains are ranked by their frequencies, which follow a generalized Zipf's law: $f(x) = a(b + x)^{-c}$, with x being the rank and $f(x)$ being its frequency. Parameter values of the best fit (dot-dashed curve) are $a = 0.21$; $b = 7.93$; and $c = 0.89$.

served that the frequency of occurrence of some event $f(x)$ as a function of the rank x is a power-law function $f(x) = a(b + x)^{-c}$, with the exponent c close to unity. The plot of Prosite domains in figure 1 satisfies the latter condition, with $c = 0.89$. We are thus dealing with relatively few highly connected domains and many rarely connected ones. Essentially, the frequency distributions of ProDom and Pfam domains are similar. However, they fit the generalized Zipf's law less well. Distributions following Zipf's law have also been observed in the context of literary vocabulary (Miller and Newman 1958), frequency of secondary structures of RNA (Schuster et al. 1994), lattice proteins (Bornberg-Bauer 1997), and hits per web page on the World Wide Web (Huberman et al. 1998). This observation is in accordance with the picture of scale-free networks which are topologically dominated by a few highly connected hubs.

As illustrated in figure 2, frequency distributions of vertices with degree k follow a distribution comparable to a power-law distribution. Although the shapes of the distribution curves are different, they share an area of linearity. Regarding these latter areas, the frequency distribution of links from ProDom domains follows $P(k) \approx k^{-\gamma}$ with $\gamma = 2.5$. By contrast, the distributions of degrees of Pfam and Prosite domains follow the same law with $\gamma = 1.7$. Although the curves do not follow exactly the proposed curvature of the frequency of degrees in the original scale-free model, one can observe a type of scale-free dependence even if the scale-free model is a raw approximation of the real situation. Obviously, the topology of such domain graphs is better described by a highly heterogenous scale-free or small-world model than by an exponential model.

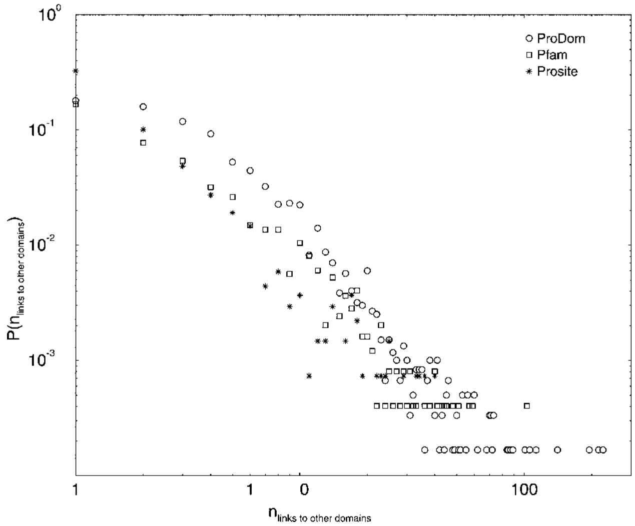

FIG. 2.—The frequency distribution of domain connections within protein sequences. Domain data were obtained from the ProDom, Pfam, and Prosite protein databases.

In table 2 it can be observed that the domain graphs partially satisfy the structural properties of small-world graphs. While clustering coefficients C_υ of the domain graphs by far exceed the respective coefficients of corresponding random graphs, the characteristic path lengths L_υ do not accomplish the demanded qualifications of a small-world graph. Emphasizing the observation that the vast majority of proteins contains only one domain (Marcotte et al. 1999), the domain networks contain a huge amount of unconnected vertices (see table 1). This feature of domain distribution among protein sequences illustrates in particular the large number of connected components in domain graphs. Although domain graphs are thus highly scattered, every graph contains a major subnet among its connected components which gathers the majority of domains. These major components feature L_υ and C_υ values that satisfy the demands of small-world graphs by exceeding the respective values of random graphs of equal size. Thus, this study focuses on the analysis of the major components exhibiting small-world and scale-free behavior. In order to clarify the graph topology, figure 3 displays the major component of the network which was generated by proteome data of *Saccharomyces cerevisiae.*

The investigations carried out so far consider all domains without taking into account their origin. Presumably, the degree of connectivity is different if one focuses on different species. All domain connections of six species which developed differently in the course of evolution were extracted from the complete proteome sets provided by the Proteome Analysis database. As illustrated in figures 4 and 5, the frequency distributions of links regarding humans, *C. elegans*, *Drosophila*,

Table 2
Characteristic Path Length *L* and Clustering Coefficient *C* of ProDom, Pfam, and Prosite Domain Nets

	L_{actual}	L_{random}	C_{actual}	C_{random}
ProDom.	4.96	5.81	0.51	0.0008
Pfam	4.54	9.05	0.15	0.0003
Prosite	5.44	6.46	0.33	0.0044

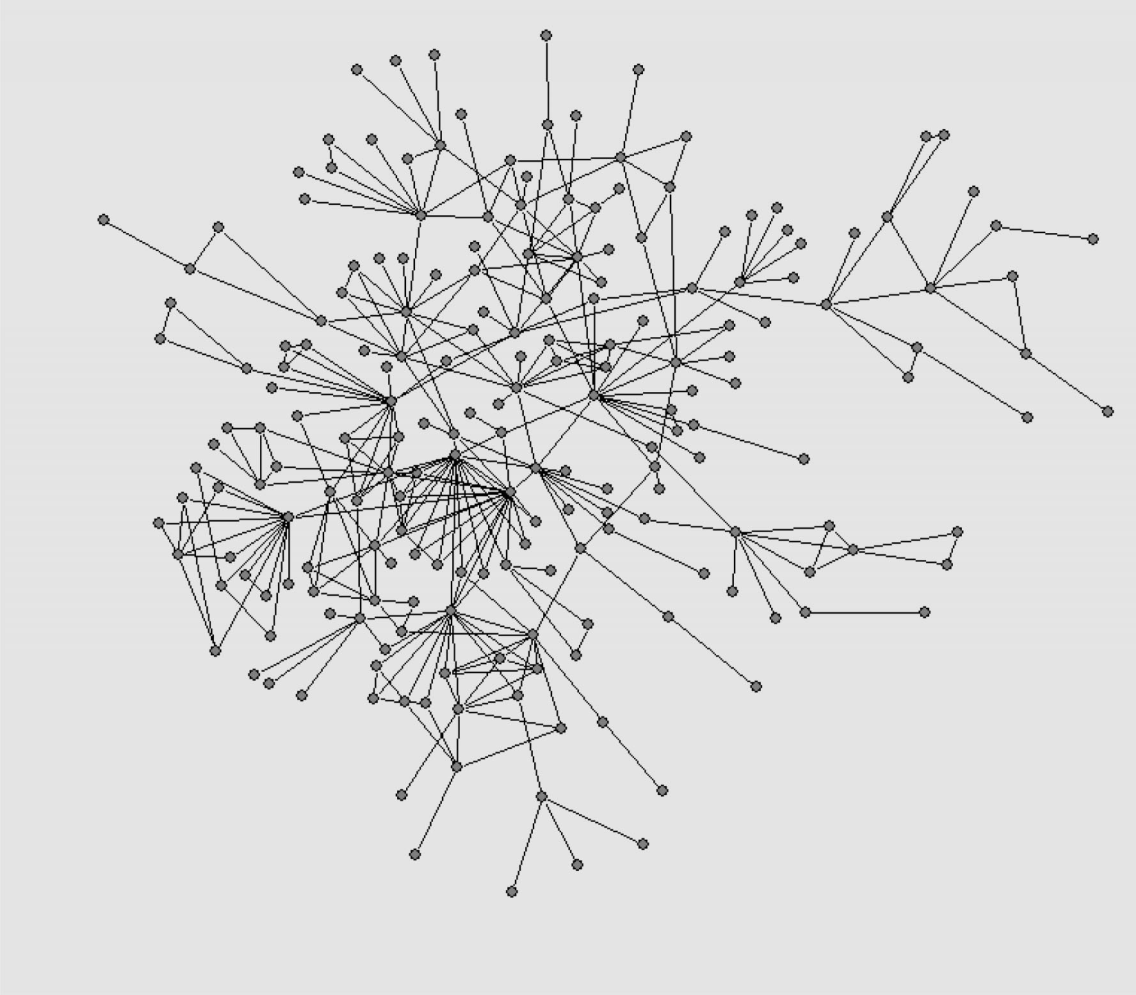

FIG. 3.—Major component of the domain network of *Saccharomyces cerevisiae,* comprising 204 vertices and 347 edges.

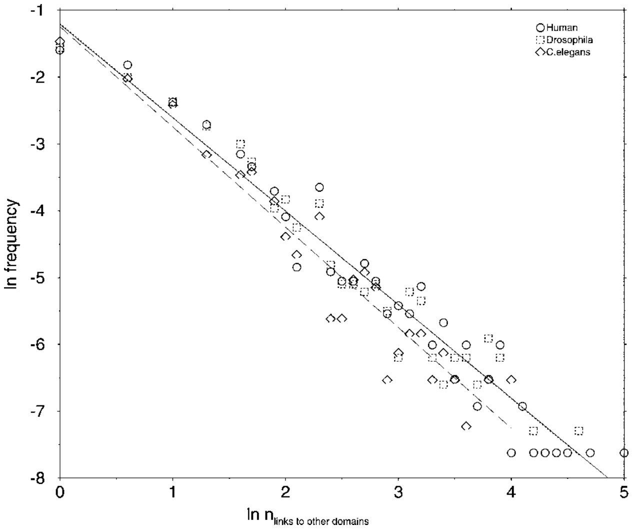

FIG. 4.—The frequency distribution of domain connections within protein sequences of *Caenorhabditis elegans, Drosophila,* and humans. The domain data were obtained from the Proteome Analysis database. The numbers of links to other domains were logarithmically binned, and frequencies were thus obtained. These pairs of values were subject to a linear regression procedure. Regression lines of *Drosophila* and human coincide.

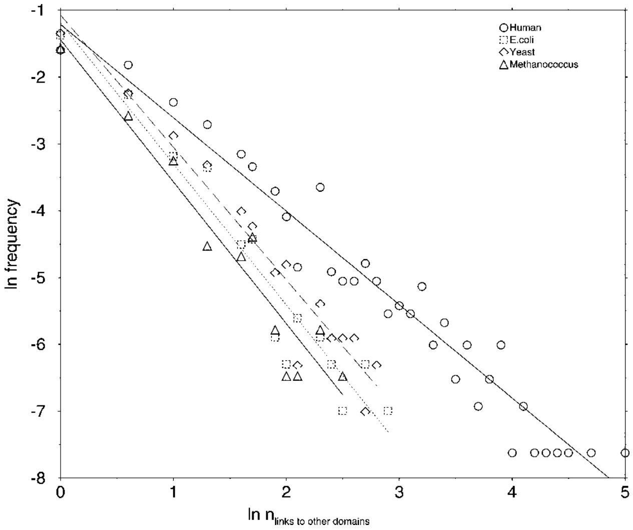

FIG. 5.—The frequency distribution of domain connections within protein sequences of *Methanococcus, Escherichia coli,* yeast, and humans. The domain data were obtained from the Proteome Analysis database. The numbers of links to other domains were logarithmically binned, and frequencies were thus obtained. These pairs of values were subject to a linear regression procedure.

yeast, *E. coli*, and *Methanococcus* still follow the expected power-law. However, the slopes of the lines are slightly different. Interestingly, the slopes of humans and *Drosophila* nearly coincide in figure 4. Moreover, the regression lines show almost the same interception in comparison with *C. elegans*. In figure 5, the situation changes slightly. While the slopes in comparison with humans are significantly steeper, the regression lines of yeast, *E. coli*, and *Methanococcus* run nearly parallel. Thus, it is tempting to assume a trend which guides multicellular organisms to higher domain connectivity.

Interestingly, the majority of highly connected InterPro domains appear in signaling pathways, as the list of the 10 best linked domains of different species in table 3 reveals. Obviously, the evolutionary trend toward compartmentalization of the cell and multicellularity demands a higher degree of organization. Therefore, more emphasis is put on the maintenance of inter- and intracellular signaling channels, cell-cell contacts, and integrity. Hence, proteomes have to provide protein sets which cover such cellular demands. The growing number of highly linked domains of signaling and extracellular proteins seen in comparisons of archaea, prokaryotes, and eukaryotes confirms this assumption.

Discussion

What might be the functional, phylogenetic, or bioinformatic implications of the power-law distribution of the connectivity of domains and the small-world behavior of the domain networks studied?

Completeness and Quality of Data

Regardless of whether Pfam, Prosite, or ProDom domain information is used, the qualitative topology of domain networks remains unchanged. Since these databases differ significantly in size and methodology, the argument is tempting that even though the current domain data are far from complete, the topology of domain networks will not change significantly with the growing amount of domain data. This assumption is supported by the characteristics of scale-free networks leading to domain graphs which are independent of the actual size of the underlying networks. Hence, the major observation that the topology of domain graphs is mainly dominated by few highly linked domains will not be changed entirely with the incorporation of new protein domain data. InterPro gathers and streamlines mostly distinct domain information from the above-mentioned domain databases, providing a centralized annotation resource to reduce the amount of duplication between the database resources. Hence, scale-free characteristics of InterPro domain networks which were generated with the aid of complete proteomes of different species do not change significantly in comparison to networks generated with domain information from a single database. However, it should be noted that the acquisition of protein domain information is biased to a certain extent, since eukaryotic and mammalian proteins are far better studied and documented in databases on average than archeae or prokaryotic proteins.

Table 3
The Ten Most Highly Connected InterPro Domains of *Methanococcus*, *Escherichia coli*, Yeast, *Caenorhabditis elegans*, *Drosophila*, and Humans

METHANOCOCCUS		E. COLI		YEAST		C. ELEGANS		DROSOPHILA		HUMANS	
Domain	k_v	Domain	k_v	Domain	k_v	Domain	k_v	Domain	k_v	Domain	k_v
SAM	13	NAD-BINDING	20	pkinase	18	pkinase	57	PRICHEXTENSN	101	ATP-GTP-A	169
fer4	11	ESTERASE	16	P-KINASE-ST	18	EGF	57	pkinase	70	GPCRRHODOPSN	162
FMN-ENZYMES	10	SAM	15	PH	16	PH	46	zf-C2H2	53	PRICHEXTENSN	110
NAD-BINDING	9	fer4	13	zf-C3HC4	14	efhand	45	ank	52	EGF	98
AA-TRNA-LIGASE-I	8	AA-TRNA-LIGASE-II	12	AA-TRNA-LIGASE-II	14	ank	37	EGF	50	pkinase	89
intein	7	FMN	12	efhand	14	P-KINASE-ST	35	SH3	48	ig	79
pyr-redox	7	HIS-KIN	11	C2	13	EGF-CA	34	ANTIFREEZEI	46	PH	72
ATP-GTP-A	6	AA-TRNA-LIGASE-I	11	CPSase-L-chain	13	zf-C3HC4	33	efhand	45	efhand	64
CBS	6	HIS REC	10	GATase	13	ig	30	PH	45	SH3	61
N6-MTASE	6	PAS	9	WD40	12	SH3	30	P-KINASE-ST	44	zf-C2H2	58

Another important consideration regards aspects of acquisition of proteome information. Proteome data which were entirely extracted by genome translation might not sufficiently explain the setup of all cellular processes. Domain networks were generated with the aid of translated genome databases which did not cover effects that include alternative splicing and domain usage. Alternative pre-mRNA splicing is an important mechanism for regulating gene expression in higher eukaryotes (Smith, Patton, and Nadal-Ginard 1989). By recent estimates, the primary transcripts of ~30% of human genes are subject to alternative splicing. Thus, the connectivity of domains found in higher eukaryotes might be significantly higher than it is ''in silico.''

In addition, the differences in frequency distributions between higher eukaryotes, bacteria, and archaea in figures 4 and 5 might also be related to the numbers of domain architectures that were found in the different organisms. Since eukaryotes and mammals developed much more distinct domain architectures (International Human Genome Sequencing Consortium 2001), the respective distributions of domain connections are statistically more reliable than those of prokaryotes and archaea. Therefore, future studies should clarify whether the small number of domain architectures leads to slight artifacts in the slope of prokaryotic and archeal organisms.

Evolutionary Aspects

Are the observed topologies a direct consequence of domain evolution? The model of Barabási and Albert (1999) generates scale-free networks by preferential attachment of newly added vertices to already well connected ones. Consequently, Fell and Wagner (2000) argued that vertices with many connections in a metabolic network were metabolites originating very early in the course of evolution and shaping a core metabolism. Analogously, highly connected domains could also have originated very early. If one compares the lists of the most highly linked domains in table 3, this assumption does not hold. The majority of more highly linked domains in *Methanococcus* and *E. coli* are mainly concerned with the maintenance of metabolism. Given that in *Methanococcus* and *E. coli* nearly none of the highly linked domains in the higher organisms can be found, and vice versa, the focus of domain connection shifts to domain hubs involved in signal transduction, transcription, and cell-cell interactions. In addition, helicase C has roughly similar degrees of connection in all organisms. However, the ankyrin repeat motif (ank) is one of the few domains which can be found to be unlinked in *E. coli,* whereas it possesses a growing degree of connectivity in higher eukaryotes.

Apparently, the majority of highly connected domains seem to have arisen late in eukaryotes of larger proteome size. The evolutionary trend toward multicellularity requires proteomes which feature new and additional complex cellular processes like signal transduction or cell-cell contacts. One way of accomplishing growing demands is the expansion of already-existing protein sets. Indeed, many protein families are expanded in humans relative to *Drosophila* and *C. elegans.* These are mainly involved in inter- and intracellular signaling pathways, apoptosis (Aravind, Dixit, and Koonin 2001), development, and immune and neural functions (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). Although many protein families of these organisms exhibit great disparities in abundance, C2H2-type zinc finger motifs and eukaryotic protein kinase (pkinase) are among the top 10 most frequent domain families (Rubin et al. 2000; Tupler, Perini, and Green 2001) and the best-connected domains in table 3. At least in higher eukaryotes, both domains tend to increase their connections to other domains in a way similar to that of the already-mentioned ankyrin repeat motif (ank).

Although the human phenotypic complexity exceeds the respective ones of *Drosophila* and *C. elegans* by far, proteome dimensions remain considerably low. Thus, combinatorial aspects of domain arrangements might have a major impact on the preservation of cellular processes. Among chromatin-associated proteins, transcription factors, and especially apoptosis proteins, a significant portion of protein architecture is shared between humans and *Drosophila.* However, substantial innovation in the creation of new protein architectures was significantly detectable (International Human Genome Sequencing Consortium 2001). Apparently, expansion of particular domain families and accompanying evolution of complex domain architectures from presumably preexisting domains coincides with the increase of the organism's complexity. In this regard, the different slopes in figures 4 and 5 indicate this evolutionary trend to higher connectivity of domains (e.g., pkinase, SH3, and EGF in table 3), as well as a growing complexity in the arrangement of domains within proteins. In comparison to noneukaryotes, *Drosophila* developed more complex domain architectures. Thus, the frequency distributions of the latter organisms can be clearly separated in figure 5, where lower complexity in domain architecture is indicated by steeper slopes. The first point is well reflected by the slightly different slopes of humans, *Drosophila,* and *C. elegans* in figure 4.

In conclusion, a variety of arguments point to an increase in the complexity of the proteome from the single-celled yeast to multicellular vertebrates such as humans. Essentially, the expansion of protein families coincides with the increase of connectivity of the respective domains. Extensive shuffling of domains to increase combinatorial diversity might provide protein sets which are sufficient to preserve cellular procedures without dramatically expanding the absolute size of the protein complement. Hence, the relatively greater proteome complexity of higher eukaryotes, and especially humans, cannot be simply a consequence of genome size but, to a certain extent, must also be a consequence of innovations in domain arrangements. Thus, highly linked domains represent functional centers in various different cellular aspects. They could be treated as evolutionary hubs which help to organize the domain space by occasionally linking them to numerous other functionally related domains.

Quality of the Basic Models

The view that new protein architectures can be created by shuffling, adding, and deleting domains, resulting in new proteins from old parts, is well reflected by the emergence of such domain hubs. However, there exist a variety of domain arrangements which contradict the ideal image of continuous addition of new domain links to already-existing hubs in the sense of scale-free networks. The S1 RNA-binding domain is linked to helicase C in *E. coli*, while it is found to be connected to RNB, KH domain, and RNAse PH in humans. Neither the procedure of generating a small-world graph in the original sense nor the scale-free model provides the deletion of vertices. However, the assumption that domains emerge and disappear occasionally is a basic demand of protein evolution. Thus, scale-free and small-world models can obviously only be a rough approximation to the real situation.

Acknowledgments

The inspiration for this work came from Ursula Kummer, who was always ready to discuss this work, as well as Erich Bornberg-Bauer, Isabel Rojas, Carel van Gend, Martin Vingron, and Rebecca Wade. Ioannis Xenarios gave me numerous valuable hints. I was supported by Andrej Mrvar and Vladimir Batagelj during the work with their PAJEK program. The Klaus Tschira-Foundation (KTF) is gratefully acknowledged for funding this project.

LITERATURE CITED

ALBERT, R., H. JEONG, and A. BARABÁSI. 1999. Diameter of the World Wide Web. Nature **401**:130–131.

———. 2000. Error and attack tolerance of complex networks. Nature **406**:378–382.

ALTSCHUL, S., T. MADDEN, A. SCHAEFFER, J. ZHANG, Z. ZHANG, W. MILLER, and D. LIPMAN. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

APWEILER, R., T. ATTWOOD, A. BAIROCH et al. (26 co-authors). 2001*a.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res. **29**:37–40.

APWEILER, R., M. BISWAS, W. FLEISCHMANN et al. (11 co-authors). 2001*b.* Proteome Analysis Database: online application of InterPro and CluSTr for the functional classification of proteins in whole genomes. Nucleic Acids Res. **29**: 44–48.

ARAVIND, L., V. DIXIT, and E. KOONIN. 2001. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. Science **291**:1279–1284.

ATTWOOD, T., M. CRONING, D. FLOWER, A. LEWIS, J. MABEY, P. SCORDIS, J. SELLEY, and W. WRIGHT. 2000. PRINT-S: the database formerly known as PRINTS. Nucleic Acids Res. **28**:225–227.

BAIROCH, A., and R. APWEILER. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28**:45–48.

BARABÁSI, A., and R. ALBERT. 1999. Emergence of scaling in random networks. Science **286**:509–512.

BARABÁSI, A., R. ALBERT, and H. JEONG. 1999. Mean-field theory for scale-free random networks. Physica A **272**:173–187.

———. 2000. Scale-free characteristics of random networks: the topology of the World-Wide Web. Physica A **281**:69–77.

BARTHÉLÉMY, M., and L. AMARAL. 1999. Small-world networks: evidence for a crossover picture. Phys. Rev. Lett. **82**:3180–3183.

BATAGELJ, V., and A. MRVAR. 1998. PAJEK—program for large network analysis. Connections **21**:47–57.

BATEMAN, A., E. BIRNEY, R. DURBIN, S. EDDY, K. HOWE, and E. SONNHAMMER. 2000. The Pfam protein families database. Nucleic Acids Res. **28**:263–266.

BORNBERG-BAUER, E. 1997. How are model protein structures distributed in sequence space? Biophys. J. **5**:2393–2403.

CORPET, F., F. SERVANT, J. GOUZY, and D. KAHN. 2000. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. Nucleic Acids Res. **28**: 267–269.

DOOLITTLE, R. 1995. The multiplicity of domains in proteins. Annu. Rev. Biochem. **64**:287–314.

DORIT, R., and W. GILBERT. 1991. The limited universe of exons. Curr. Opin. Genet. Dev. **1**:464–469.

ERDÖS, P., and A. RÉNYI. 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. **5**:17–61.

FELL, D., and A. WAGNER. 2000. The small world of metabolism. Nat. Biotech. **189**:1121–1122.

GILBERT, W., and M. GLYNIAS. 1993. On the ancient nature of introns. Gene **135**:137–144.

GUARE, J. 1990. Six degrees of separation: a play. Vintage Books, New York.

HOFMANN, K., P. BUCHER, L. FALQUET, and A. BAIROCH. 1999. The PROSITE database, its status in 1999. Nucleic Acids Res. **27**:215–219.

HUBERMAN, B., P. PIROLLI, J. PITKOW, and R. LUKOSE. 1998. Strong regularities in World Wide Web surfing. Science **280**:95–97.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

JANIN, J., and C. CHOTHIA. 1985. Domains in proteins: definitions, location, and structural principles. Methods Enzymol. **115**:420–430.

JEONG, H., B. TOMBOR, R. ALBERT, Z. OLTVAI, and A.-L. BARABÁSI. 2000. The large-scale organization of metabolic networks. Nature **407**:651–654.

KRIVENTSEVA, E., W. FLEISCHMANN, E. ZDOBNOY, and R. APWEILER. 2001. CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. Nucleic Acids Res. **29**:33–36.

LI, W.-H., Z. GU, H. WANG, and A. NEKRUTENKO. 2001. Evolutionary analyses of the human genome. Nature **409**:847–849.

MARCOTTE, E., M. PELLEGRINI, H.-L. NG, D. RICE, T. YEATES, and D. EISENBERG. 1999. Detecting protein function and protein-protein interactions from genome sequences. Science **285**:751–753.

MILGRAM, S. 1967. The small-world problem. Psychol. Today **2**:60–67.

Miller, G., and E. Newman. 1958. Tests of a statistical explanation of the rank-frequency relation for words in written English. Am. J. Psychol. **71**:209–218.

Rubin, G., M. Yandell, J. Wortmann et al. (52 co-authors). 2000. Comparative genomics of the eukaryotes. Science **287**:2204–2215.

Schuster, P., W. Fontana, P. Stadler, and I. Hofacker. 1994. From sequences to shapes and back: a case study in RNA secondary structures. Proc. R. Soc. Lond. B Biol. Sci. **255**:279–284.

Seidel, H., D. Pompliano, and J. Knowles. 1992. Exons as microgenes. Science **257**:1489–1490.

Smith, C., J. Patton, and B. Nadal-Ginard. 1989. Alternative splicing in the control of gene expression. Annu. Rev. Genet. **23**:527–577.

Stoltzfus, A., D. Spencer, M. Zuker, J. Logsdon Jr., and W. Doolittle. 1994. Testing the exon theory of genes: the evidence from protein structure. Science **265**:202–207.

Tupler, R., G. Perini, and M. Green. 2001. Expressing the human genome. Nature **409**:832–833.

Venter, J., M. Adams, E. Myers et al. (271 co-authors). 2001. The sequence of the human genome. Science **291**: 1304–1351.

Watts, D., and S. Strogatz. 1998. Collective dynamics of ‘small-world’ networks. Nature **393**:440–442.

William Taylor, reviewing editor

Accepted May 14, 2001