

**ANALYSIS OF THE EVOLVING PROTEOMES:  
PREDICTIONS OF THE NUMBER OF PROTEIN DOMAINS  
IN NATURE AND THE NUMBER OF GENES IN  
EUKARYOTIC ORGANISMS**

VLADIMIR A. KUZNETSOV\*

*The Laboratory of Integrative and Medical Biophysics,  
National Institute of Child Health and Human Development,  
13 South Drive, Bethesda, MD 20892, USA  
vk28u@nih.gov*

VALERY V. PICKALOV

*Institute of Theoretical and Applied Mechanics SB RAS, Novosibirsk, 630090, Russia  
pickalov@itam.nsc.ru  
<http://www.itam.nsc.ru/lab17>*

OLEG V. SENKO

*Computer Center of Russian Academy of Sciences,  
Vavilov str. 40, 117967 Moscow, Russia  
senko@ccas.ru*

GARY D. KNOTT

*Civilized Software, Inc., 12109 Heritage Park Circle,  
Silver Spring, MD 20906, USA  
knott@civilized.com*

Received 16 March 2002

Accepted 5 May 2002

Motivation: Obtaining accurate estimates of the numbers of protein-coding genes and protein domains in a proteome, and the number of protein domains in nature is a daunting challenge. Computational analysis of the protein domain sets in the proteomes of many species allows us to estimate these numbers and to find their evolution relationships.

Results: We have analyzed the distributions of the number of occurrences of protein domains in sample proteomes of the 70 fully sequenced genome organisms of three major kingdoms of life: Archaea, Bacteria and Eukaryota. We found that a large fraction of the identified distinct protein domains (i.e., unique domains and homologous domain families) in these 70 proteomes (1051 (23%) out of 4493) are found in at least one organism in each of these kingdoms of life and that 43 (1%) of these domains are common to all the 70 organisms. All the observed domain occurrence frequency distributions for these 70 proteomes are well fitted by a family of Pareto-like functions, associated with the steady state distributions of a linear Markov random process. We present explicit

\*Corresponding author.

formulas that accurately predict the number of distinct protein domains and the number of protein-coding genes for a given organism as functions of the number of non-redundant domain-to-protein links in the proteomes. These functions allow us to predict that there are 42,740, 27,900, and 21,200 protein-coding genes/open reading frames in the human, *A. thaliana*, and mouse genomes, respectively. We also estimate that there are 5271, 2955, and 4915 distinct protein domains in the human, *A. thaliana*, and mouse proteomes, respectively, and about 5500 distinct protein domains in the entire “proteome world”.

*Keywords:* Proteome complexity; evolution; number of genes; number of protein domains; Pareto-like distribution; stochastic process.

1991 Mathematics Subject Classification: 22E46, 53C35, 57S20

## 1. Introduction

A protein domain is an amino acid sequence that is essential to the biological function(s) of the protein in which it occurs [1, 4, 19]. A protein domain often includes the “active” binding site(s) of a protein. Alternatively, a protein domain may serve as a necessary shape-determinant “building block”. Protein domains are considered to be sufficiently-long homologous amino acid sequences encoded by evolutionarily-conserved DNA sequences. Many common protein domains are found in phylogenetically different organisms. The evolutionary history of such domains can accommodate modest differences; thus, strictly, a class of homologous DNA sequences defines a domain.

In this work, we consider protein domains as the basic structural and functional parts of proteins, which allows us to simplify analysis of complex proteome information. From this point of view, a protein in a proteome may be considered as compositions of domains, specifically linked within the protein chains in the course of evolution. Of course, the protein domains themselves cannot completely determine all protein functions. However, knowledge of the specific distinct domains and their *ordering* in each protein of a set of proteomes can be used for comparative analysis of the proteome complexity. Analysis of such proteome characteristics for many species may provide understanding of the mechanisms of proteome complexity and changes due to evolution history.

Statistical analysis of DNA sequences that code protein domains within the fully-sequenced genomes of Archaea, Bacteria and Eukaryotic organisms allow us to estimate some fundamental numbers in biology: (a) the number of protein-coding genes in genomes that are only partially known, (b) the number of distinct protein domains in a proteome, and (c) the number of distinct protein domains in nature.

A proteome can be defined as a complete set of distinct proteins encoded by protein-coding genes in the entire genome. In this work, the proteome size is defined by the number of distinct protein-coding sequences in the organism. Previous estimates of the size of the human proteome have used experimental approaches such as measuring the complexity of cellular RNA, reassociation kinetics, CpG island determination, or, more recently, genome sequence analysis. The International Human Genome Sequencing Consortium (IHGSC) has estimated that there are 29,700

putative genes/ORFs in the human genome [17]; Celera Genomics has estimated 26,000–39,000 putative genes/ORFs [21]; Ewing and Green, using expressed sequence tag (EST) clustering incorporating quality scores estimated 35,000 transcripts [7], and Kuznetsov, using a large human SAGE transcriptome data sets estimated a lower limit of 31,200 transcripts [14]. However, even after publication of human genome sequence drafts by Celera Genomics and by the IHGSC in February 2001, a reliable estimate of the total number of human protein-coding genes has not yet been obtained due to experimental errors and ambiguities in the data [5, 9, 14]. For humans, the number of protein-coding genes is currently so uncertain that geneticists keep arguing about its exact value (<http://www.ensembl.org/Genesweep/>).

The actual number of existing protein domains in proteomes and the total number of protein domains in nature (or in the “proteome world”) are both uncertain. Estimates of the number of domains in organisms and in nature range between a few thousand to twenty-three thousand or more [3, 14, 20, 22, 23].

In this paper, we have analyzed the functional relationships between the number of identified protein-coding genes, the number of identified protein domains in Archaea, Bacteria and Eucaryotic proteomes having fully-sequenced genomes. For each organism, we analyzed the statistical distribution of the number of proteins containing a random domain within a proteome. We estimate there are 42,740, 27,900, and 21,200 protein-coding genes in the human, plant *Arabidopsis thaliana*, and mouse genome, respectively. We predict that there are about 5500 distinct proteome domains in all of nature. This is about 1000 more than the 4500 distinct domains in the InterPro database (March 12, 2002; [www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)).

## 2. Databases, Definitions and Methods

Practically, protein domains have been determined based on expert knowledge and probabilistic models of amino acid sequence similarity. Sequences similar to a given protein domain sequence can be often identified computationally and are taken to be the same domain functionally. We used the InterPro database (Integrated Resource of Protein Domains and Functional Sites; [www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)) which integrates data from several domain/protein family/functional site databases (Pfam [3], SMART [18] etc.) and determines domains based on a signature recognition approach [12].

The InterPro set of domains is constructed by building groups of homologous polypeptide patterns (domains/protein families/motifs), including patterns, if any such exist at least in one of the integrated databases. Each such group has a unique identification number and is recorded as a unique domain in the InterPro database by the InterPro criteria [12]. All the sequences in a domain group must be sufficiently long and must be homologous enough to warrant this grouping. The InterPro database identifies the location of all InterPro domains boundaries found within sequences in the comprehensive SWISS-PROT/TrEMBL databases which,

in total, include amino acid sequence entries found in large numbers of completely and incompletely sequenced genomes. Note that in each organism, a unique InterPro domain can be represented by a single polypeptide sequence for the protein domain or by two or more homologous polypeptide patterns. Thus, equivalence classes define the *distinct domains*. On March 12, 2002, 4493 distinct InterPro domains had been collected within the proteins of 70 organisms (see Appendix 1).

We developed a local Protein Domain Database Analyzer (PDDA) program which accesses the Interpro database and optionally stores data in a local MySQL relational database. Our basic data consists of a table in which a row is for each unique Interpro domain and a column is for each organism. Let  $N_j$  denote the number of identified distinct domains of organism  $j$ . The same domain may be repeated in a given protein many times, as well as occurring in many different proteins of the proteome. The  $(i, j)$ -th entry of the table,  $m_{ij}$ , is the number of proteins containing one or more instances of distinct domain  $i$  in the sample proteome of organism  $j$ . Thus, we count the number of non-redundant occurrences of the domain  $i$  in all proteins of organism  $j$ . By non-redundant, we mean that a distinct domain is counted only once even if it occurs several times in the same protein. We also refer to the number  $m_{ij}$  as *the number of non-redundant domain-to-protein links* of a distinct domain in the proteome of the organism. Let  $M_j$  denote the sum of the numbers of domain-to-protein links in the sampled proteome of organism  $j$ . Thus  $M_j = \sum_i m_{ij}$ . We call  $M_j$  *the connectivity number of a proteome*. Let  $G_j$  denote the number of protein-coding genes/ORFs (open reading frames) for organism  $j$ .

Our PDDA also possesses data mining features which allow us: (1) to select rows or columns of the table which meet certain criteria (logical functions), (2) to construct empirical histograms of occurrence counts of proteins containing given domains for any given organism or for a group of organisms, (3) to count the numbers of elements in subsets of 2-set and 3-set Venn diagrams. The PDDA has been developed using the MySQL database and the Apache Web Server, and was written using the HTML, PHP and C languages. The PDDA also links to other electronic resources on protein domains/domain families (Pfam) and active sites (SMART) which allows us to obtain comprehensive information on all studied protein patterns, proteins, and their functions.

The model fitting and statistical analysis were performed using the MLAB mathematical modeling software (Civilized Software, Inc., [www.civilized.com](http://www.civilized.com)). The programs are available from V. A. Kuznetsov ([vk28u.nih.gov](mailto:vk28u.nih.gov)).

In this work, we analyzed the protein domain data for the 12 Archaeal, 50 Bacterial and 6 Eukaryotic organisms of the InterPro database (see Appendix 1), which provides information on 68 fully-sequenced genomes, as well as for the nearly completely-sequenced and partially annotated mouse and human genomes (March 12, 2002; [www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)).

### 3. Results

#### 3.1. Increasing complexity trends in evolving proteomes

One of the most surprising discoveries of recent years has been the degree to which genes are conserved across species. Different organisms share many distinct protein coding genes, even over widely divergent evolutionary paths. Mechanisms for horizontal exchange and lineage-specific gene loss are thought to be major evolutionary forces [17, 21]. However, while gene loss is conveniently justified in the literature, the impact of horizontal gene transfer in evolution is still under discussion. A quantitative comparison of protein domain sets in fully-sequenced genomes representing major kingdoms of life could help us to identify protein domains shared by biological functions and biological processes in phylogenetically distant species.

Table 1 shows the 7 distinct subsets of the 4493 InterPro Protein Domains in three domains of life, represented by 12 Archaea, 50 Bacteria and 8 Eukaryotic organisms. Table 1 shows that a very large fraction of domains (1051 of all 4493 domains; 23.4%) occurs at least once in all three major kingdoms of life. These 1051 domains found in each of the Archaea, Bacteria and Eukaryotic kingdoms make up the largest fraction of domains found in Archaea (67%) and in Bacteria (40%). In Eukaryotes, this fraction is 30% (1051 of 3509 domains). Based on this finding, we can suggest that these 1051 domains are associated with functionally-important genes that are significant for evolving organisms.

The number of Archaeal domains of the Archaea domain set A, which also appear in the Bacterial domain set B, but not in the Eukaryotic domain set E, is more than 2 times greater than the number of Archaea domains that appear in the Eukaryotic domain set E, but not in the Bacterial domain set B. Only 75 (4.8%) of the 1550 Archaea domains found in the InterPro database are unique to Archaea proteomes. Thus, even with environmental specialization, isolation, differences in complexity of biological organization and divergent evolutionarily paths, a vast majority of Archaea protein domains are also used in Bacteria or in Eukaryotic proteins. However, Archaea are more similar to Bacteria than to higher Eukaryota, measured in terms of number of domains in common.

Tables 2a and 2b show the average number of non-redundant domain-to-protein links of 43 common protein domains observed one or more times in the proteomes of all 12 Archaea, 50 Bacteria and 8 Eukaryotic organisms. We call these domains the evolutionarily *super-conserved* domains. The column "Sum" contains the total number of non-redundant domain-to-protein links in the proteomes of all 70 organisms. The other columns show the mean number per organism for the three kingdoms of life. Table 2 shows that most of the evolutionarily super-conserved domains are associated with the translation machinery of a cell. The average number of non-redundant domain-to-protein links per proteome are strongly correlated in all column pairs: Archaea-Bacteria ( $r = 0.79$ ), Archaea-Eukaryota ( $r = 0.8$ ), Bacteria-Eukaryota ( $r = 0.73$ ) (using Spearman correlation coefficient). These high correlations indicate that the distributions of the number of occurrences of the

Table 1. Distribution of the distinct domains in three major kingdoms of life represented by sets of InterPro protein domains for Archaea (A), Bacteria (B), and Eukaryota (E).  $N_i = N_A, N_B, N_E$ .  $n$  is the number of the distinct domains in a domain subset.  $N = 4493$ : the total number of distinct InterPro domains.

Subset of domains	$n$	$n/N$	$n/N_i$
$E-(A \cup B)$	1687	0.38	0.48
$(A \cap E)-B$	136	0.03	0.04
$(B \cap E)-A$	635	0.14	0.18
$A \cap B \cap E$	1051	0.23	0.30
$N_i = N_E$	3509		1.0
$B-(A \cup E)$	621	0.14	0.24
$(A \cap B)-E$	288	0.06	0.11
$(B \cap E)-A$	635		0.25
$A \cap B \cap E$	1051		0.41
$N_i = N_B$	2595		1.0
$A-(B \cup E)$	75	0.02	0.05
$(A \cap B)-E$	288		0.19
$(A \cap E)-B$	136		0.09
$A \cap B \cap E$	1051		0.68
$N_i = N_A$	1550	1.0	1.0

Table 2a. Twenty most common protein domains shared by all 70 organisms. Mean occurrence values of the given domain for individual Archaea, Bacteria and Eukaryota organism are represented. The column “Sum” contains the sums of occurrence values for the 70 organisms.

	Domain description	Sum	Archaea	Bacteria	Eukaryotes
1	ABC transporter	3715	33.25	56.26	62.88
2	ATP-binding protein; ATPase-like	1327	4.92	23.08	14.25
3	DEAD/DEAH box helicase	1320	10.42	10.58	83.25
4	Helicase; C-terminal	1303	9.92	10.40	83.00
5	AAA ATPase; central region	929	7.75	9.44	45.50
6	Elongation factor Tu; domain 2	508	5.08	6.46	15.50
7	Elongation factor; GTP-binding	517	4.67	6.42	17.50
8	S1 RNA binding domain	490	4.42	7.20	9.63
9	KH domain	479	5.33	4.24	25.38
10	S4 domain	415	1.92	6.86	6.13
11	Helix-hairpin-helix DNA-binding motif class 1	409	6.08	5.98	4.63
12	tRNA synthetases; class-II (G; H; P and S)	343	5.00	4.54	7.00
13	Toprim domain	328	4.00	5.20	2.50
14	GTP1/OBG family	313	3.33	3.84	10.13
15	Ribosomalprotein L24/bacterial NUSG	217	3.33	2.16	8.63
16	Replication factor C conserved domain	208	2.50	2.34	7.63
17	Elongation factor G; C-terminal	141	1.00	1.88	4.38
18	Type 2 KH domain	118	1.00	1.88	1.50
19	DNA-directed RNA polymerase; beta subunit	100	1.42	1.00	4.13
20	RNA polymerase; alpha subunit	100	1.17	1.04	4.25

Table 2b. Ribosomal protein domains shared by all 70 proteomes. Mean occurrence values of a given domain for the Archaea, Bacteria and Eukaryota organisms are represented.

	Domain description	Sum	Archaea	Bacteria	Eukaryotes
21	Ribosomal protein L15	91	1.75	1.00	2.50
22	Ribosomal protein L1	89	1.00	1.00	3.38
23	Ribosomal protein S5	99	1.00	1.00	4.63
24	Ribosomal protein L11	95	1.00	1.04	3.88
25	Ribosomal protein S14	96	1.00	1.14	3.38
26	Ribosomal protein L13	87	1.00	1.00	3.13
27	Ribosomal protein L10	83	1.00	1.00	2.63
28	Ribosomal protein L4/L1e	86	1.00	1.00	3.00
29	Ribosomal protein S2	97	1.00	1.00	4.38
30	Ribosomal protein L2	83	1.00	1.00	2.63
31	Ribosomal protein S12	82	1.00	1.00	2.50
32	Ribosomal protein S7	85	1.00	1.02	2.75
33	Ribosomal protein S17	83	1.00	1.00	2.63
34	Ribosomal protein L3	87	1.00	1.00	3.13
35	Ribosomal L23 protein	96	1.00	1.00	4.25
36	Ribosomal protein S11	81	1.00	1.00	2.38
37	Ribosomal protein S15	81	1.00	1.00	2.38
38	Ribosomal protein S9	83	1.00	1.00	2.63
39	Ribosomal protein L22/ L17	89	1.00	1.00	3.38
40	Ribosomal protein L14b/L23e family	82	1.00	1.00	2.50
41	Ribosomal protein L6	82	1.00	1.00	2.50
42	Ribosomal protein L5	78	1.00	1.00	2.00
43	Ribosomal protein S3	75	1.00	1.00	1.63

evolutionarily conserved domains in proteins are functionally related in all three major kingdoms of life. The average number of non-redundant domain-to-protein links increases in the order: Archaea  $\rightarrow$  Bacteria  $\rightarrow$  Eukaryota (Table 2). Differences between pairs of these groups are significant:  $p = 0.017$  for Archaea-Bacteria,  $p = 0.002$  for Archaea-Eukaryota, and  $p = 0.0035$  Bacteria-Eukaryota (using Wilcoxon matched pairs test on column pairs in Table 2b).

Interestingly, for the vast majority of ribosomal protein domains, the mean number of the non-redundant domain-to-protein links of super-conserved domains per proteome is increased only in Eukaryotic organisms (1.6–4.6 times; Table 2b). We observed such a trend for many other sets of conserved protein domains. Thus, the number of occurrences of evolutionarily-conserved protein domains in the proteins tends to increase with progressive evolution.

These results suggest that (1) many protein domains appearing in nature are widely distributed in phylogenetically distant organisms; (2) the more frequent domains found in the proteome of relatively simple and evolutionarily older organisms, tend to be used in the proteomes of many other organisms; (3) more frequent domains of evolutionarily-older and relatively simple organisms tend to be involved in more domain-to-protein links (and, probably, in more biochemical interactions) in more complex and evolutionarily-younger organisms.

To better understand proteome complexity, we also studied all combinations of a given protein domain with other domains co-occurring in different proteomes. We have used the Simple Modular Architecture Research Tool (SMART; <http://Smart.embl-heidelberg.de/>) to look up these combinations. For example, Table 3 shows the frequencies of occurrences of proteins in Archaea, Bacteria and Eukaryotic organisms which contain the S4-domain. The S4-domain is a relatively small protein domain consisting of 60–65 amino acid residues that was detected in ribosomal proteins and, perhaps, mediates binding to RNA. This domain matches 34, 97 and 68 proteins in 16 Archaea, 12 Bacteria and 8 Eukaryota organisms, respectively, found in the SMART database. The average number of protein encoded DNA sequences matched by the S4-domain in Archaea is significantly smaller than in Bacteria or in Eukaryotic organisms (2.1, 8.1, 7.9, respectively). This data is consistent with the data for the InterPro S4-domain presented in Table 2a.

Table 3 shows that the fraction of single-domain proteins containing the S4-domain decreases in the order Archaea→Bacteria→respectively). The number of multi-domain proteins containing the S4-domain increases in the same order: Archaea → Bacteria → Eukaryota. There are co-occurrences of the S4 domain with other specific domains in multi-domain proteins. Some combinations of domains may appear, preferentially, in one of the major kingdoms of life. For example, the combinations of the S4-domain with the ribosomal S4, or combination of the S4-domain with the PseudoU-synth-2-domain and the tRNA\_synth\_1b-domain

Table 3. Frequency of occurrences of the proteins containing S4 domain in 16 Archaea, 12 Bacteria and 9 Eukaryotic proteomes. 1: S4: S4 RNA binding domain (IPR002942); 2: KOW: KFDVGNVVMVTGGRNRGRVGVKRNREKH (SM0739); 3: Ribosomal S4: the Pfam domain with the length 103 amino acids (aa); 4: the PseudoU-synth-2: the Pfam domain with the length 147 aa; 5: tRNA\_synth\_1b: the Pfam domain with 302 aa; 6: the signal peptide with the length (22–44 aa); 7: the transmembrane segment LVPLFALKALFYLFVFFFWMV; 8: S\_TKcc: Serine/Threonine protein kinases, catalytic domain (SM0220); 9: EFh: the EF-hand, calcium binding motif (SM0054); 10: dCMP\_cyt\_deam: the Pfam domain with the length 112 aa; 11: the coiled-coil region with the length 70 aa. See more information in <http://smart.embl-heidelberg.de/smart/> and <http://www.sanger.ac.uk/cgi-bin/Pfam/>

Domain combinations and order	Archaea,%	Bacteria,%	Eukaryota,%
Number of proteins	34	97	68
1	68	38	24.7
1+2	32	0	28.0
1+3	0	11	26.5
1+4	0	41	8.8
5+1	0	10	3.0
6+5+1	0	0	1.5
1+6	0	0	3.0
7+1	0	0	1.5
8+9+9+9+1+4	0	0	1.5
1+4+10	0	0	1.5
6+11+1	0	0	1.5
Number of proteomes	16	12	9

often appears in Eukaryotic and Bacterial proteins, but these combinations of domains are not observed in Archaea proteins. The S4-domains and the KOW-domains often co-occur in Archaea and Eukaryotic proteins, but that combination is not observed in all studied Bacteria proteomes.

For many evolutionarily super-conserved protein domains, more complex organisms exhibit more complex combinations of domains in protein sequences. For example, in all studied Archaea proteomes, the S4-domain occurs only with the KOW-domain. However, in Eukaryotic proteomes, we observed 10 different combinations of the S4-domain with other domains and some of these combinations provide very diverse protein functions (for example, the *A. thaliana* protein Q9LQR4 (which belongs to the serine/thirosine family of protein kinases) contains the S\_TKc-domain, the EF-hand motif (3 times), the S4-domain and the PseudoU-synth-2-domain (Table 3). Similar domain combination complexity occurs in complex organisms, primarily Eukaryotes, for many orthologous sets of proteins [1, 2].

Interestingly, the orders of the specific domains in the different multi-domain protein sequences containing the S4-domain are the *same* in all proteins of the studied proteomes. Table 3 shows these orders. For example, “5+1” means domain 5 (the tRNA\_synth..1b-domain) is always before domain 1 (the S4-domain) in each protein sequence where 5 and 1 both occur. We observed also that many other domain pairs occur within proteins of different organisms in the same sequential order; for example, the NGN-domain (transcription termination-antitermination factor domain) and the KOW-domain occurs in Eukaryotic, Bacteria and Archaea proteins in the order: NGN-domain, then KOW-domain wherever they co-occur. Similarly, conserved orders of other domain pairs in proteins have been reported [1]. These results indicate that many domain pairs arose in evolution from single recombination events and such events form the super-family of domains. Apparently, larger combinations of domains in proteins occurs more frequently as organism complexity increases. This trend suggests an appearance of more complex and more specific biochemical networks for the super-conserved domains in the course of progressive evolution.

### 3.2. Evolutionary propagation of the *G. theta* protein domains

Apparently, increased domain usage in new proteomes coincides with the increase of proteome size and organism’s complexity. In each studied proteome, we counted the occurrence of protein domains of the cryptomonad *Guliardia theta*, which is the smallest eukaryote with a compact genome. The *Guliardia theta* genome contains 451 genes and 213 distinct InterPro domains. This chimerical cryptomonad underwent evolutionary compaction hundred of millions of years ago, and eliminated nearly all genes for metabolic functions, but left hundreds of genetic-housekeeping genes. Most known functions of *G. theta* genes have important roles in all Eukaryotes [6]. We found 199 of the 213 *G. theta* domains in the other 7 studied Eukaryotic proteomes.

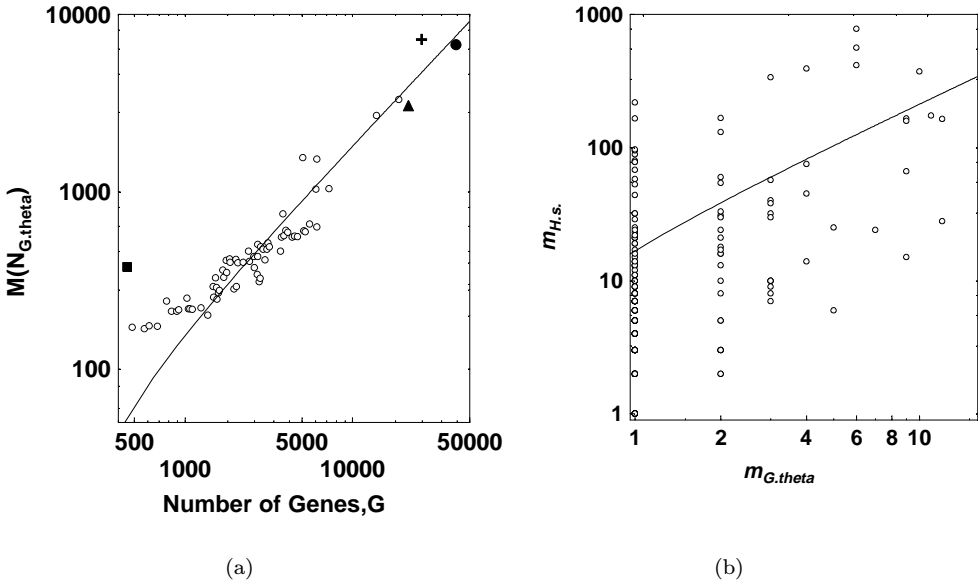


Fig. 1. The occurrence the protein domains of *Guliardia theta* in proteins of each of the other 69 organisms. (a) Correlation between the total numbers of proteins containing the *G. theta*'s domains,  $M(N_{G.theta})$ , counted in each of the other 69 sample proteomes, and estimates of the total number of protein-coding genes/ORFs,  $G$  in that proteomes. ■ : *G. theta*, ○: other 66 organisms; ▲: mouse; +: *A. thaliana* and ●: human. The number of protein-coding genes for mouse, *A. thaliana* and human proteomes equals to 21,140, 27,913, and 42,682, respectively (see Sec. 3.4: Estimation of the number of protein-coding genes in a proteome). Solid line: regression line  $M(N_{G.theta}) = -29.5 + 0.183 * G$  fitted to data excluding *G. theta*, mouse, *A. thaliana* and human data points; (b) Scatter plot: Correlation between the numbers of proteins containing the same domain in the human and *G. theta* proteomes. Solid line: regression line  $m_{H.s.} = -4.9 + 21.7 * m_{G.theta}$  fitted to data excluding *G. theta*, mouse, *A. thaliana* and human data points. ( $r = 0.47$  ( $p < 0.01$ )).

Figure 1(a) shows that the number of non-redundant *G. theta* domain-to-protein links, counted in proteomes of each of the other 69 organisms, tends to be larger when the number of protein-coding genes is larger. These results demonstrate that many domains of one organism (i.e., *G. theta*) are included in many other Archaea, Bacteria and Eukaryotic proteomes and the number of domain-to-protein links tend to increase in larger proteomes (Fig. 1(a)). It is interesting that the ratio of the number of non-redundant *G. theta* domain-to-protein links in other organisms versus the total number of non-redundant domain-to-protein links for the proteome does not depend on the total number of genes; this ratio is around  $0.19 \pm 0.046$ .

A vast majority of domains observed in the *G. theta* proteome are also found in the other 69 organisms. For example, only 20 of the 213 *G. theta* domains were lost in the human proteome sample. Moreover, a domain which occurs in many proteins of an organism has a higher than average chance to appear in many proteins in many other organisms, even over widely divergent evolutionary paths. Figure 1(b) shows that domains represented by a larger number of links in the *G. theta* proteome

tend to also appear more often in the human proteome. We have observed a similar trend in the yeast proteome [14].

Because about 90% of *G. theta* domains represented in the human proteome are much older than many other domains contained in the human proteome, one can expect that the *G. theta* domains have had more time to appear in the human genome and to establish more links to human proteins, than have other, “younger” domains. We found that, on average, one *G. theta* protein domain has 35 non-redundant domain-to-protein links to proteins in the human proteome. On the other hand, all other domains in the human proteome has, on average, only 13.6 non-redundant links to proteins of the human proteome. These results, together with data presented in Fig. 1(b), indicate that, for a given evolutionarily “younger” protein domain in the human proteome, the probability of acquisition of new non-redundant links to proteins is roughly proportional to the number of occurrences of this domain in other evolutionarily “older” non-human proteomes. Such correlations between the evolutionary ages of organisms and the number of non-redundant occurrences of “older” domains in the “younger” proteins might shed a light on the direction and intensity of horizontal gene transfer.

### 3.3. Numbers of single-domain proteins and the most common domains in the proteomes

All proteins have one or more domains, with the exception of some disordered proteins. The single-domain proteins are encoded by *de novo* origin genes appearing due to mutations, gene duplication and horizontal gene transfer. What is the relationship between the fraction of single-domain proteins in a proteome and the proteome size? In our previous analysis of the frequency distributions of protein domain occurrences in proteomes for 20 fully-sequenced genome organisms [14], we observed that for each organism, a fraction of the number of proteins containing a single protein domain is larger than the fractions of proteins containing 2-, 3-, etc. distinct protein domains. We can estimate the fraction  $p_1$  of distinct domains which non-redundantly link exactly one protein in the proteome as follows:  $p_1 \approx n_1/N$ , where  $n_1$  is the number of distinct proteins in the sample proteome containing a single domain and  $N$  is the number of distinct domains in the sample proteome. Here we observe that a larger number of protein-coding genes in a proteome leads to a decrease of the fraction  $p_1$  (Fig. 2(a)).

On the other hand, the number of non-redundant domain-protein links of the most common domain in the proteome of the organism (denoted by  $J$ ) positively correlates with the number of protein-coding genes/ORFs in the genome (Fig. 2(a)). The regression line  $J = 0.031G$  ( $r = 0.96$ ) was fitted to data points for 67 organisms (excluding mouse, *A. thaliana* and human) (Fig. 2(b)).

Figure 2 shows that as the proteome size increases, the probability of the number of occurrences of single-domain proteins decreases, but the number of the most common domains in the proteome increases. Figure 2 illustrates a general trend of

the evolution of proteome complexity: to solve adaptation problems, more complex organisms decrease using “new” single-domain proteins, and, preferentially, increase using already-existing “old” protein domains and motifs sequences in constructing new and more diverse multi-domain proteins. Note that the Eukaryotic organisms exhibit lower  $p_1$ -values than the Bacteria or Archaea organisms. These observations imply that there is a limited or slowly-growing repertoire of domains in nature, and, therefore, in creating new biological functions in more complex organisms. Apparently, in the course of proteome evolution, nature more frequently combines already-existing domains rather than use new ones.

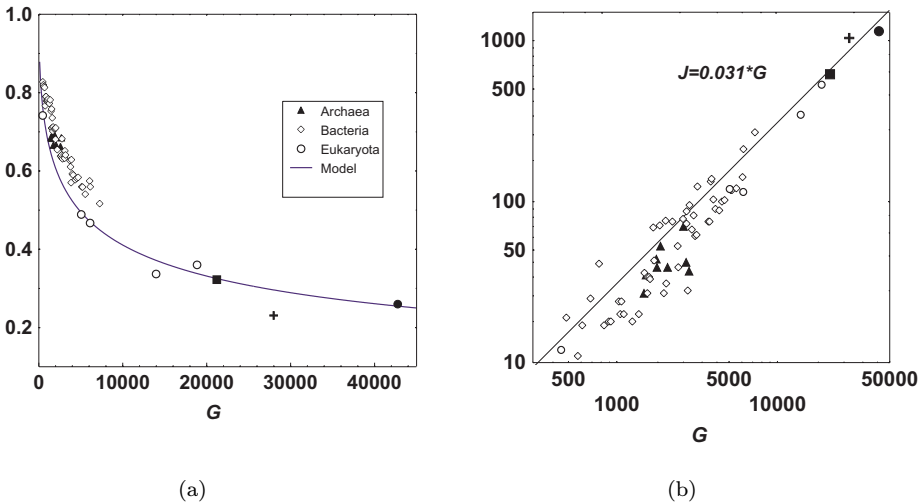


Fig. 2. Trends of proteome complexity evolution. (a)  $\blacktriangle$ ,  $\diamond$ ,  $\circ$ ,  $\blacksquare$ ,  $+$ ,  $\bullet$ : fractions of distinct domains which non-redundant link exactly one protein in the proteome versus the number of protein-coding genes/ORFs in the organism; solid line: best fit curve by Eq. (8), where  $M = G/1.03$  (see also Figs. 3 and 6); (b) number of non-redundant domain-to-protein links of the most common domain in the proteome,  $J$ , versus the number of protein-coding genes/ORFs in the organism.  $\circ$ : data point of the Eukaryotic organisms, excluding  $\blacksquare$ : mouse;  $+$ : *A. thaliana* and  $\bullet$ : human. The number of protein-coding genes for mouse, *A. thaliana* and human proteomes equals 21,140, 27,913, and 42,682, respectively (see Sec. 3.4: Estimation of the number of protein-coding genes in a proteome).

### 3.4. Estimation of the number of protein-coding genes in a proteome

The total number  $M$  of the non-redundant domain-to-protein links in a sample proteome characterizes the proteome complexity. Figure 3 shows a remarkable linear relationship between the number of protein-coding genes,  $G$ , and the connectivity number,  $M$ , presented in a sample proteome. This data is fitted by the regression line,  $G = 1.03M$  (Fig. 3). This line fits the data for 67 of the 70 organisms (excluding data points for human, *A. thaliana*, and mouse). By extrapolation of the line

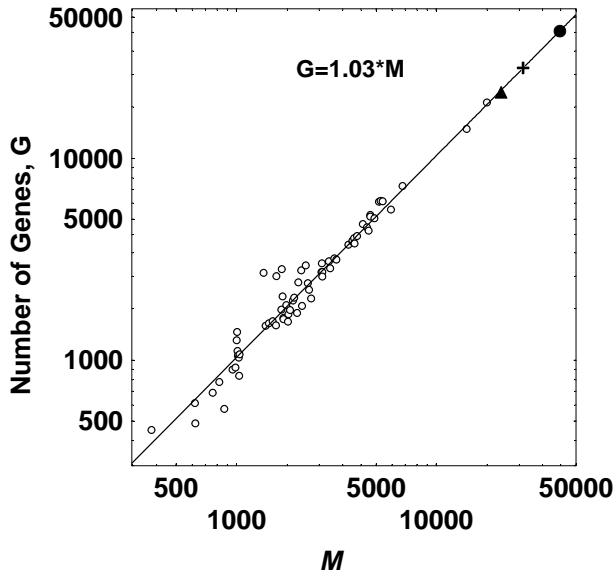


Fig. 3. Relationships between the number of protein-coding genes/ORFs,  $G$ , and the observed connectivity number in the proteomes,  $M$ , of 70 organisms ordered by ascending of the connectivity number (see Appendix 1). Data points for mouse ( $\blacktriangle$ ), *A. thaliana*, (+) and human ( $\bullet$ ) were estimated using the regression line  $G = 1.03M$ ; this regression model was fitted to other 67 data points ( $r = 0.99$ ).

$G = 1.03M$ , we can predict 42,743 [41,489, 43,997], 27,975 [27,184, 28,766] and 21,202 [20,622, 21,781] protein-coding genes in human, *A. thaliana* and mouse proteomes, respectively.

Note that a similar linear relationship ( $G = 1,015M$ ) was obtained for smaller domain sets of 18 Archaea, Bacteria, and Eukaryotic fully-sequenced genome organisms, retrieved from the InterPro database released on December 22, 2001 [14].

Our predictions are based on the InterPro database, released on March, 12, 2002, which had incomplete mouse genome sequences. The latest (May 4, 2002) mouse draft sequence based on whole genome shotgun analysis covering 96% of the mouse euchromatic DNA predicts 22,444 genes (Mouse Genome Assembly v.3; [http://www.ensembl.org/Mus\\_musculus/](http://www.ensembl.org/Mus_musculus/)). Our prediction of the number of protein-coding genes for human and *A. thaliana* are higher than estimates based on genome sequence analysis. In humans,  $\sim 30,000$  genes was predicted by the IHGSC [17], and 26,000–39,000 genes was predicted by Celera Genomics [21]; in *A. thaliana*, a current estimate is 25,773 genes ([www.ebi.ac.uk/interpro/](http://www.ebi.ac.uk/interpro/)).

### 3.5. Estimation of the number of protein domains and the average number of domain-to-protein links in the proteomes

For a given proteome,  $P$ , the total number of functionally distinct protein sequences is determined by the number of protein-coding genes,  $G$ . There are two

fundamental numbers:  $N_P$ , the total number of distinct domains occurring in  $P$ , and  $M_P$ , the connectivity number in  $P$ ; both  $N_P$  and  $M_P$  are characteristics of proteome complexity. However, currently only some of the distinct proteins, and some of the protein domains existing in nature are known, and, correspondingly, the list of non-redundant domain-to-protein links seen in any proteome is significantly incomplete. Therefore, for a given organism, we do not generally know the true numbers  $N_P$  and  $M_P$ . However, for a sample proteome, we can obtain  $N$ , the number of distinct protein domains, and  $M$ , the observed connectivity number in proteome, which are lower boundaries for  $N_P$  and  $M_P$ , respectively. We can also count  $g$ , the number of distinct proteins of the proteome  $P$  that matched at least one InterPro domains. Let  $\gamma = g/G$  denote the fraction of identified protein-coding genes matched by InterPro domains in the genome for  $P$ . There is no correlation between  $\gamma$  and  $N$  in the 70 sample proteomes. We also observed that  $\gamma$  is relatively large (typically, 0.55–0.75) for each of 70 studied organisms. These observations imply that  $g/G$  is a specific characteristics of incompleteness of domain set of a proteome sample. Let us assume that each gene contains at least one protein domain coding sequence. As more domains are discovered,  $g$  must approach  $G$  and  $N$  must approach  $N_P$ . Taking the equation  $g/G \approx N/N_P$  and using the regression function  $G = \lambda M$ , we can obtain the estimator  $N_P \approx \lambda N M / g$ .

This formula allows us to estimate the number of protein domains in a proteome. For example, it predicts 5271, 2955, and 4915 distinct domains in the human, *A. thaliana*, and mouse proteomes, respectively. The estimate of the mean value of  $N_P$  in Archaea, Bacteria and Eukaryotic organisms is  $1147 \pm 178$ ,  $1617 \pm 519$ , and  $3120 \pm 1181$  (excluding *G. theta*), respectively.

The linearity of the graph in Fig. 3 is interesting. This may imply that the mean value of the number of non-redundant domain-to-protein links per protein is an invariant relative to the proteome size for all studied organisms. We can estimate the ratio  $\alpha = M/g$  for each studied organism. The mean values of  $\alpha$  equal 1.34, 1.49, and 1.51 for Archaea, Bacteria and Eukaryotic organisms, respectively. These mean values show that, most frequently, the proteins in a given proteome contain only one- or two-domains. This also indicates that the mean values of the connectivity numbers are similar in different kingdoms of life. However, we also observed a positive correlation between the  $\alpha_p$  values and the number of protein-coding genes,  $G$ , for the Eukaryotic organisms ( $r = 0.8$ ;  $p < 0.05$ ; Spearman correlation coefficient) and significant differences in the distributions of  $\alpha_p$ -values for the three kingdoms of life ( $p < 0.01$ ; Kruskal–Wallis ANOVA by rank test). We need further analysis to explain these observations. In the next section, we will present our analysis of the statistical distribution of the number of non-redundant domain-to-protein links in the 70 proteomes.

### 3.6. Statistics of the number of protein domains arising in Bacteria, Archaea and Eukaryotic proteomes

We define the protein domain profile as the list of all distinct domains in a sample proteome, together with the number of non-redundant domain-to-protein links for each domain for the sample proteome. Let  $X$  denote the number of non-redundant domain-to-protein links of a random domain within proteins in a given proteome. Then we can define the domain occurrence probability function (DOPF)  $f(m) = P(X = m)$  for occurrence values  $m = 1, 2, \dots, J$ , where  $J$  denotes the maximum non-redundant domain-to-protein link number for the given proteome. The function  $f(m)$  denotes the probability that a random distinct domain occurs non-redundantly exactly  $m$  times within the proteome.

We found that the observed DOPFs in all 70 proteomes are similar; there are few frequent, and many rare distinct domains. Table 3 shows that all empirical DOPFs are fitted well by the Generalized Discrete Pareto (GDP) probability function [13, 14]:

$$\hat{f}(m) = \frac{1}{\zeta_1(m+b)^{k+1}}. \quad (1)$$

The function  $\hat{f}(m)$  involves two unknown parameters,  $k$ , and  $b$ , where  $k > 0$ , and  $b > -1$ . The normalization factor  $\zeta_1$  is the generalized Riemann Zeta-function value [11]:  $\zeta_1 = \sum_{j=1}^J \frac{1}{j+b}^{k+1}$ .  $J$  is the largest non-redundant occurrence number of protein domains in proteins of the proteome.

The parameter  $J$  positively correlates with the total number of distinct protein-coding genes/ORFs in the genome,  $G$  ( $J = 0.031G$ ; Fig. 2(b)) and with the observed connectivity number  $M$  in the sample proteome ( $J = 0.03M$ ). The observed DOPFs do not show the scale-invariant property associated with a simple power law distribution: rather the distribution explicitly depends on  $M$  (Fig. 4). As the size of the proteome increases, the shape of the empirical DOPF changes systematically. Recall  $p_1 = \hat{f}(1)$  is the probability that a random domain occurs non-redundantly exactly once within proteins in the proteome. As  $M$  becomes larger,  $p_1$  decreases. Also the parameter  $J$  increases in proportion to  $M$ . For Archaea organisms, the parameters  $b$  and  $k$  do not show strong trends (Table 4a), as  $M$  increases. However,  $b$  becomes larger for Eukaryotic organisms (Table 4c),  $b$  increases and then slowly decreases for Bacterial organisms (Table 4b), as  $M$  increases. For Bacterial and Eukaryotic proteomes, the parameter  $k$  shows a quite similar size-dependent trend:  $k$  increases and then decreases, as  $M$  increases (Tables 4b and 4c).

### 3.7. Analysis of stochastic birth-death processes of protein domains in a proteome

Although the GDP model appears to fit empirical DOPFs for our studied organisms and it provides a justifiable way to compare the DOPFs for different organisms, this model does not capture causal mechanisms and is only a descriptive model,

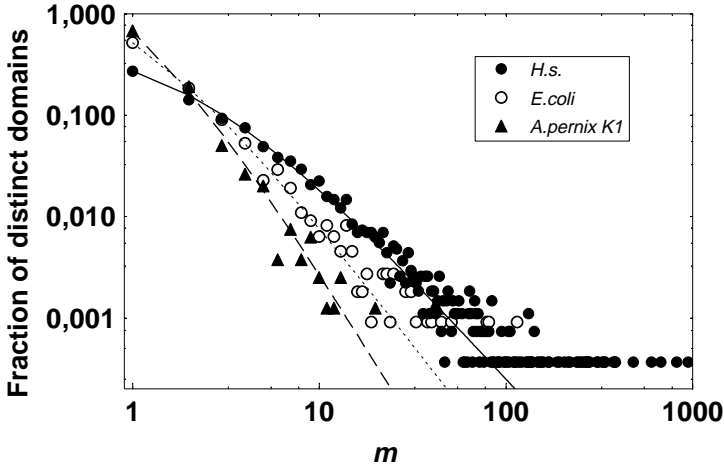


Fig. 4. Empirical DOPFs for the Archaea, Bacteria and Eukaryotic proteomes, represented by *A. Pernix K1*, *E. Coli* and human, respectively. Log-log plot: Fitting of the empirical frequency distributions by the GDP model for human (●: at  $k = 1.02$ ;  $b = 2.17$ ), *E. coli* (○: at  $k = 1.41$ ;  $b = 0.88$ ), and *A. pernix K1* (▲: at  $k = 2.07$ ;  $b = 0.77$ ), respectively. (Points of the GDP probability distributions are joined by the smooth curves for display purposes.)

not an explanatory model.

A simple model of the DOPF can be described in terms of a Markov stochastic process. We will characterize a proteome of an organism by its  $N$  distinct domains  $d_1, \dots, d_i, \dots, d_N$ . Let  $m_i$  denote the number of proteins containing  $d_i$ . Note  $\sum_i^N m_i = M$ . Given an organism, we may imagine the evolutionary path starting from its unknown ancestor species at time zero that leads to the evolution of this organism. Evolutionary progress along this path, in general, results in the appearance of new domains, and the increase/decrease in the number of uses of domains, and even the dropping of domains (i.e., due to lineage-specific gene loss). Many distinct species may occur on this path as precursors of the given organism. Let the random variable  $D_t(d, P)$  be the number of proteins containing the distinct domain  $d$  in the organism occurring at time  $t$  in the evolutionary path  $P$  of some end-point organism.  $D_t(d, P)$  is a realization of a continuous-time stochastic process  $\{D_t, t > 0\}$ . This process can be considered as a birth-death process where protein domains are “born” and “die” during evolution for some evolutionary path. A birth indicates an increase in the number of proteins that contain the random fixed domain  $d$ , whereas a death indicates a decrease in this number. We assume that the proteome size associated with any evolutionary path is finitely-bounded above and becomes nearly constant as evolution progresses along that path.

Let us assume that  $D_t$  is a Markov birth-death random process. Let functions  $\lambda_m(t)$  and  $\mu_m(t)$  be the intensities of the birth and death processes for a given domain occurring in  $m$  proteins in the proteome, where  $m = 0, 1, 2, \dots$ . The intensities  $\lambda_m$  and  $\mu_m$  are related to the transition probability  $P_{i,j}(t, s)$  which is the

Table 4a. Fitting of the domain occurrence probability distributions for the Archaea organisms.  $M$ : the connectivity number of the proteome;  $N$ : number of distinct domains in the proteome;  $k, b$ : parameters of the GDP model; EAE: Average absolute error; MSC: Model Selection Criterion [13]. The values of the MSC are ranged between “excellent” [11,8], “very good” [8,6], and “satisfactory” [6,4].

Organism	$M$	$N$	$k$	$b$	EAE	MSC
Aeropyrum pernix K1	1372	765	2.213	0.87	$1.4 \times 10^{-3}$	7.0
Thermoplasma acidophilum	1403	751	1.61	0.322	$1.5 \times 10^{-3}$	7.8
Thermoplasma volcanium	1460	778	1.74	0.44	$1.5 \times 10^{-3}$	7.2
Pyrobaculum aerophilum	1586	771	1.68	0.47	$6.5 \times 10^{-4}$	8.0
Methanobacterium thermoautotrophicum	1683	909	1.724	0.38	$1.2 \times 10^{-3}$	7.5
Sulfolobus tokodaii	1693	770	1.393	0.293	$1.5 \times 10^{-3}$	7.5
Pyrococcus horikoshi	1706	843	1.56	0.356	$1.9 \times 10^{-3}$	6.5
Methanococcus jannaschii	1780	933	1.66	0.94	$1.5 \times 10^{-3}$	6.3
Pyrococcus abyssi	1859	900	1.47	0.26	$1.1 \times 10^{-3}$	7.8
Halobacterium sp. NRC-1	2048	910	1.49	0.47	$1.5 \times 10^{-3}$	6.9
Sulfolobus solfataricus	2231	867	1.302	0.308	$1.1 \times 10^{-3}$	7.3
Archaeoglobus fulgidus	2286	1009	1.45	0.361	$1.1 \times 10^{-3}$	7.4

Table 4b. Fitting of the domain occurrence probability distributions for the Bacteria organisms.

Organism	$M$	$N$	$k$	$b$	EAE	MSC
Mycoplasma genitalium	624	445	1.4	-0.31	$1.6 \times 10^{-3}$	8.6
Mycobacterium leprae	1725	961	1.62	0.218	$1.2 \times 10^{-3}$	7.8
Aquifex aeolicus	1818	1068	1.948	0.487	$1.2 \times 10^{-3}$	8.2
Listeria monocytogenes	2953	1193	1.36	0.28	$6.8 \times 10^{-4}$	7.4
Escherichia coli O157_H7 substrain RIMD 0509952	4683	1618	1.48	0.76	$4.5 \times 10^{-4}$	7.9
Escherichia coli O157_H7 strain EDL933	4688	1619	1.48	0.74	$5.1 \times 10^{-4}$	7.5
Rhizobium meliloti	5275	1357	1.07	0.234	$3.9 \times 10^{-4}$	7.0
Rhizobium loti	6780	1461	1.06	0.402	$3.2 \times 10^{-4}$	6.8

Table 4c. Fitting of the domain occurrence probability distribution for the Eukaryotic organisms.

Organism	$M$	$N$	$k$	$b$	EAE	MSC
Guillardia theta algal nucleomorph	377	213	0.547	-0.584	$5.88 \times 10^{-3}$	6.3
Schizosaccharomyces pombe	4911	1484	1.62	1.38	$4.46 \times 10^{-4}$	7.56
Saccharomyces cerevisiae	5401	1484	1.514	1.366	$7.2 \times 10^{-4}$	6.38
Drosophila melanogaster	14241	1949	1.197	1.87	$3.35 \times 10^{-4}$	5.95
Caenorhabditis elegans	17967	1838	0.947	1.086	$1.83 \times 10^{-4}$	6.15
Mus musculus	20575	2485	1.13	1.74	$1.397 \times 10^{-4}$	6.64
Arabidopsis thaliana	27167	1948	1.08	2.85	$1.56 \times 10^{-4}$	5.11
Homo sapiens	41540	2757	0.982	2.23	$1.397 \times 10^{-4}$	4.81

probability that  $D_{t+s} = j$  given that  $D_s = i$ . For a birth-death process  $P_{i,j}(t, s)$  depends only on  $i, j$ , and  $t$ . Thus we may write  $P_{i,j}(t, s) = P_{i,j}(t)$ . Also,  $P_{i,j}(t) = 0$  for  $|i - j| > 1$ . An additional fact is that the time  $T_m$  that the process  $D_t = m$

before making a transition to a different value is exponentially-distributed so that  $P(T_m \leq \alpha) = 1 - \exp(-\alpha/v_m)$ , where  $v_m = E(T_m)$ . Now we can specify the meaning of  $\lambda_m$  and  $\mu_m$ :  $\lambda_m(t) = v_m P_{m,m+1}(t)$  and  $\mu_m = v_m P_{m,m-1}(t)$ . We will consider  $D_t$  to be a Markov random process such that the intensities are the linear functions of  $m$ :

$$\begin{aligned} \lambda_m &= \lambda_1^* + \lambda_2^* m, \\ \mu_m &= \mu_1^* + \mu_2^* m, \quad (m = 0, 1, 2, \dots), \end{aligned} \tag{2}$$

where the constants  $\lambda_1^* > 0$ ,  $\lambda_2^* > 0$ ,  $\mu_1^* > 0$ ,  $\mu_2^* > 0$ ;  $\mu_0 = 0$ . Hence, during an interval  $(t, t + h)$  where  $h$  is small, we assume there are four independent possible events: the spontaneous “birth” or “death” of a protein containing the random fixed domain  $d$ , with constant intensities  $\lambda_1^*$ , and  $\mu_1^*$ , respectively, and the “flows” of the domains with the intensities proportional to the number of proteins already containing that domain  $(\lambda_2^* m, \mu_2^* m)$ .

Note that, probabilistically, the intensities  $\lambda_1^*$ , and  $\mu_1^*$  are the intensities of Poisson processes. During the interval  $(t, t + h)$  where  $h > 0$  is small, the intensity  $\lambda_1^*$  is proportional to a transitional (birth) probability of a spontaneous increase in the number of proteins containing the fixed random domain  $d$  due to gene duplication events, mutations or horizontal gene transfer from another species. During the same interval  $(t, t + h)$ , the intensity  $\mu_1^*$  is proportional to a transitional (death) probability of spontaneous mutation, deletion or modification of a domain or gene loss for a protein containing  $d$ .

Suppose that in the most evolving near end-point organisms, the random birth and death processes of domains are keeping near equilibrium. Let  $p_m(t) = P(D_t = m)$ .  $p_m$  denote the probability function associated with the random process  $\{D_t, t > 0\}$ . Using the forward Kolmogorov equations (see Appendix 2), we can obtain the non-zero limiting probability function for the random process  $D_\infty$  ( $D_\infty = \lim_{t \rightarrow \infty} D_t$ ):

$$p_m^* = \lim_{t \rightarrow \infty} p_m(t). \tag{3}$$

The distribution  $p_m^*$  depends on three parameters. Let

$$a = \frac{\lambda_1^*}{\lambda_2^*}; \quad \theta = \frac{\lambda_2^*}{\mu_2^*}; \quad b = \frac{\mu_1^*}{\mu_2^*}. \tag{4}$$

Appendix 2 shows that

$$p_0 = \left( 1 + \sum_{m=1}^{\infty} \prod_{i=1}^m \theta \frac{(a + i - 1)}{(b + i)} \right)^{-1}, \tag{5}$$

$$p^* m = b p_0 \frac{a^{[m]}}{b^{[m+1]}} \theta^m, \tag{6}$$

where  $z^{[m]} = z(z + 1) \dots (z + m - 1)$ ,  $m \geq 0$ . We call  $p_m^*$  the Kolmogorov–Waring (KW) probability function.

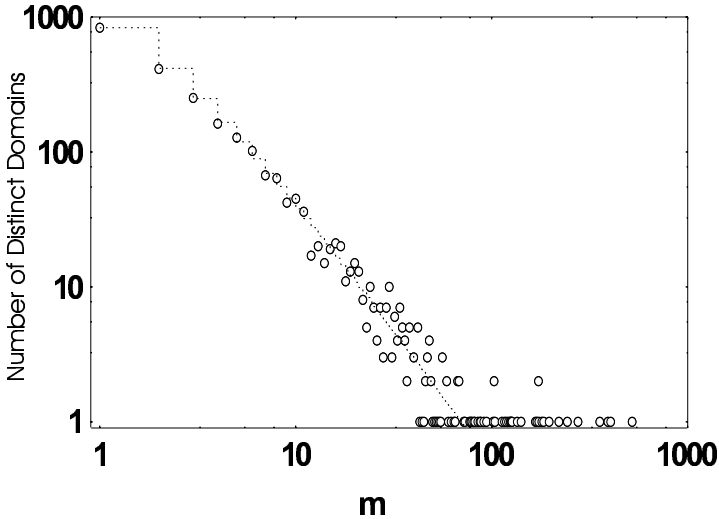


Fig. 5. Fitting of the zero-truncated Kolmogorov–Waring probability function to the empirical DOPF for Eukaryotic organism (represented by mouse).  $\circ$ : number of distinct domains at  $m = 1, 2, \dots, 520$ . Step-function: the best-fit function  $n(m) = N * (b/a)p_m^*$  at  $a = 0.993$ ;  $b = 1.988$ ,  $\theta = 1.0$ ,  $N$  is the number of distinct domains in the sample proteome.

Figure 5 shows that the KW function  $p_m^*$  fits the DOPF for the mouse protein domain data set. This model fits the data points well for the entire dynamical range of the number of proteins containing a random fixed domain in the proteome. Note that differences between the best-fit KW model and the best-fit GDP model (data not presented) are very small. The best-fit parameters of the KW probability function suggest that (1) the corresponding random process  $D_\infty$  has approximately similar birth and death rates ( $\theta = \lambda_2^*/\mu_2^* \approx 1$ ) for domains which already exist in the proteome and (2) the rate of random appearance of new domains,  $\lambda_1^*$ , is smaller than the rate of random deletion/modification of domains,  $\mu_1^*$ . Note that similar approximation by the KW model was obtained for different Eukaryotic, Bacterial and Archaea organisms (data not presented). These results imply that in observed organisms, the birth and death intensities with transition probabilities proportional to the number of proteins containing a random fixed domain in the proteome appears to approach an equilibrium. Thus, an organism appears to hold the number of domains it used near steady state at  $\lambda_2^*/\mu_2^* \approx 1$  and  $\mu_1^* > \lambda_1^*$ .

### 3.8. Probabilistic model of a domain's evolution

Recently, we have developed a probabilistic model of the distribution of protein domains in proteins of a proteome in the course of evolution [14]. Briefly, this model assumes: (1) the number of distinct protein domains in nature is a finite number; (2) each protein domain has a positive probability of occurring in time

in any given proteome and (3) the number of occurrences of a given protein domain is determined by its intrinsic properties (thermodynamic stability of the domain sequence, hydrophobic groups, etc.) and the evolutionary history of the domain (i.e., evolution age), but is statistically independent of the number of occurrences of other domains in the proteome. Of course, co-occurrence of distinct domains in the proteome certainly occurs (for instance in multidomain proteins). However, the fraction of such multidomain proteins in the proteome is small (see Fig. 4 and previous sections) and correlation between the numbers of occurrences for a given domain and the number of occurrences for other hundreds or thousands domains in the evolved proteome would likely be statistically insignificant. The mathematical description of our model of evolution of the DOPF is based on the multinomial distribution [13, 14]. Let  $X$  be the number of proteins containing the domain  $d$ . Our probabilistic model relates  $N$ , the number of distinct domains in the proteome,  $M$ , the connectivity number of these domains in the proteome and the probability  $P(X = m)$ .

In our model,  $M$  is considered as the independent variable,  $M = 1, 2, \dots$ . We will take  $N$  to be a function of  $M$ ,  $N(M)$ . When  $M$  is large enough, we can obtain the probability function  $p_m := P(X = m)$ , in terms of the value  $M$  and the value  $N(M)$  as follows:

$$p_m \approx (-1)^{m+1} \frac{1}{N} \cdot \frac{M!}{m!(M-m)!} \frac{d^m N}{dM^m}, \quad \text{where } m = 1, 2, \dots \quad (7)$$

(see [13,14]). The function  $N = N(M)$  is a differentiable function of  $M$ . The function  $p_m$  is called the Binomial Differential (BD) probability function. The value  $p_1$  is the probability that the domain  $d$  occurs non-redundantly exactly once within proteins in the proteome.

Taking the derivative of  $p_1$  with respect to  $M$ , one can show that  $p_1$  is a decreasing function of  $M$ . Using this property, we found the empirical approximation

$$p_1 = \frac{1 + (1/d)^c}{1 + (M/d)^c}, \quad (8)$$

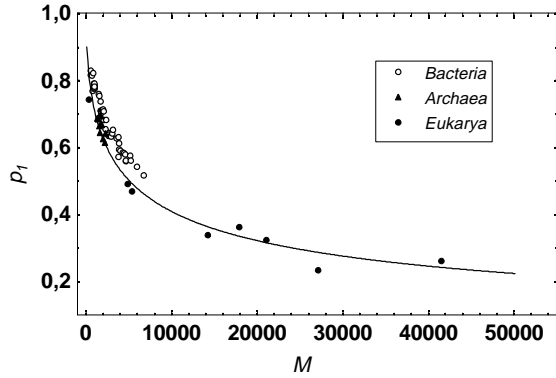
where the  $c$  and  $d$  are positive constants [15, 16]. Then taking Eq. (8) at  $m = 1$ , we have the differential equation:

$$\frac{dN}{dM} = p_1 \frac{N}{M}, \quad (9)$$

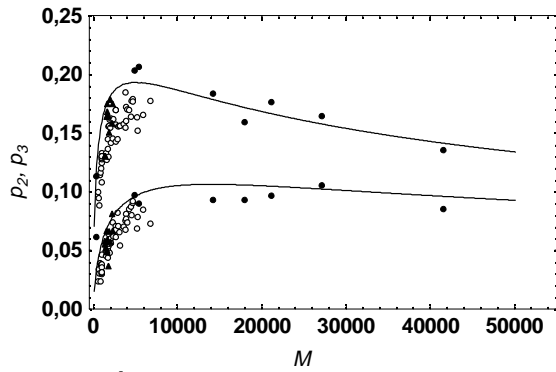
with  $N(1) = 1$ . The function  $N$  increases when  $M$  becomes larger, but this function has a limit when  $M$  approaches infinity (see below). Equation (9) has an exact solution:

$$N(M) = \left( M^c \frac{1 + 1/d^c}{1 + (M/d)^c} \right)^{\frac{1+1/d^c}{c}}, \quad (10)$$

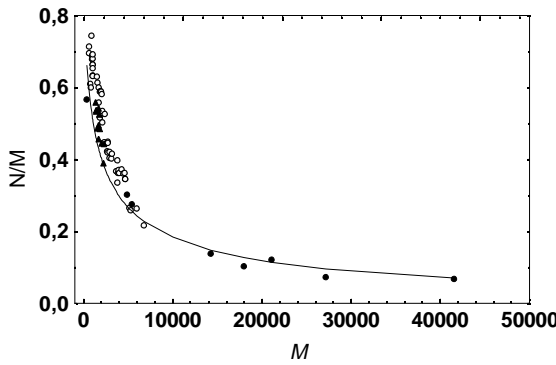
with  $N(1) = 1$ . The parameters  $c$  and  $d$  are positive constants. Equation (10) defines the “growth” function of the number of distinct protein domains  $N(M)$  in the course of evolution (see Fig. 6).



(a)



(b)



(c)

Fig. 6. Analysis and prediction of the evolution trends in the proteomes. (a)  $\circ$ : Relationships  $p_1$  vs  $M$  for 70 sample proteomes. Solid line: Best-fit curve Eq.(8) at the parameters  $c = 0.49 \pm 0.02$ ;  $d = 4,500 \pm 300$  (data points for *A. thaliana* and *C. elegans* have been excluded from curve-fitting analysis); (b) Prediction of the function  $N$  in terms of variable  $M$  based on Eq.(7); (c) Prediction of the functions  $p_2, p_3$  in terms of variable  $M$  based on Eq. (10).  $\blacktriangle$ : Archaea,  $\circ$ : Bacteria and  $\bullet$ : Eukaryota.

We assume that  $M \rightarrow \infty$  as the proteome complexity increases. Let  $N_t$  be the total number of protein domains in the entire “proteome world”. Then Eq. (10) provides an asymptotic estimate:

$$N_t = \lim_{M \rightarrow \infty} N(M) = (1 + d^c)^{\frac{1+1/d^c}{c}}. \quad (11)$$

We can fit the function  $p_1(M)$  given by Eq. (8) to data for all Archaea, Bacteria and Eukaryotic organisms, excluding *A. thaliana* and *C. elegans* (because many genes in the genomes of these organisms are massively duplicated (65% and 46%, duplicated genes, respectively)), (Fig. 6(a)), and then we may predict the values of the functions  $N, p_2, p_3, \dots$  in terms of the variable  $M$  and the value  $p_1$  (Figs. 6(b) and 6(c)). The best-fit estimate of the parameters  $c$  and  $d$  are  $c = 0.49 \pm 0.02$  and  $d = 4,500 \pm 300$ . To evaluate the robustness of our results, we also fitted Eqs. (7), (8) and (10) to data presented in Figs. 6(a), 6(b) and 6(c) for 68 studied organisms (excluding *A. thaliana* and *C. elegans* data points) and we obtained quite similar values of the parameters  $c$  and  $d$  (data not presented).

Figure 6(b) shows that as the connectivity number  $M$  in the proteome increases, the number of distinct protein domains,  $N$ , tends to a finite limit. By Eq. (11), this limit is estimated to be  $5360 \pm 400$  distinct domains.

#### 4. Discussion

In this work, we defined the proteome complexity in terms of the *numbers* of proteins, numbers of protein domains and the number of domain-to-protein links in the proteomes of an organism.

Our results suggest that many protein domains appearing in nature are widely distributed in phylogenetically distant species. We found that 23% of 4493 identified InterPro protein domains are common in three major domains of life, Archaea, Bacteria, and Eukaryota, and that 43 (1%) protein domains are common to all the 70 organisms. The major fraction of the latter, “super-conserved” domains is exhibited in the protein biosynthesis machinery of cells and the number of occurrence of those domains tends to increase in the order: Archaea, Bacteria, and Eukaryota. Surprisingly, only 75 (4.8%) of all 1550 Archaea domains found in InterPro database are unique to Archaea proteomes.

Our statistical analysis of the distributions of the number of domain occurrences in different species implies certain rules of progressive evolution of the proteomes:

- (1) more frequent domains found in the proteome of relatively simple and evolutionarily older organisms tend to be used in the proteomes of many other organisms;
- (2) for a given protein domain found in an evolutionarily “younger” proteome, the probability of acquisition of new non-redundant links to new proteins is roughly proportional to the number of occurrences of this domain in the other evolutionarily “older” proteomes;

- (3) for a given super-conserved protein domain, the proteins in more complex organisms that exhibit that domain show more diverse combinations of domains within the proteins of their proteomes;
- (4) more complex organisms decrease using “new” single-domain proteins, and, preferentially, increase using already-existing “old” “building blocks” (protein domain and motifs sequences) in constructing new and more diverse multi-domain proteins. This implies that there may be a limited repertoire of domains in nature, and, therefore, in creating new biological functions in more complex organisms.
- (5) for studied Archaea, Bacteria and Eukaryotic organisms, the protein domain occurrence frequency distributions belong to a family of skewed Pareto-like functions whose shape depends in a predictable manner on the total number of non-redundant domain-to-protein links in the sample proteome.

Our new probabilistic BD model allows us to predict a general trend of the evolution of the probability distribution function of the number of proteins containing 2, 3 and more distinct protein domains in a proteome (Fig. 4(b)). Moreover, we can estimate the total number of protein domains in nature (Fig. 6(c)).

Thus, our results imply that in the course of progressive evolution, the number of distinct domains in the proteome increases occasionally over time, but ever more slowly (Fig. 6(c)). Nature more frequently combines the already-existing domains rather than using new ones (Figs. 2, 4, 6(a) and 6(b)). Our model predicts that life is currently based on about 5500 protein domains as defined by our analysis of the InterPro data sets.

Interestingly, the *A. thaliana* and *C. elegans* data points on Fig. 6(a) significantly deviate from the best-fit prediction function (Fig. 6(b)): in both cases the number of observed distinct domains are much less than the model predicts. These differences correlate with massive gene duplication in these organisms (65% and 45%, respectively). Due to positive correlation between proteome size (represented as the number of protein-coding genes) and the number of duplicated genes in larger Bacteria and Eukaryotic proteomes, our estimate  $N_t = 5360 \pm 400$  distinct protein domains is, probably, a conservative estimate of the true number of distinct protein domains in nature. Note also that the true number of protein domains may be higher than our estimate, due to errors determining domains including redundant identification of domains (i.e., the same domain can be identified as several distinct domains). Current classifications of domains are partially “non-overlapping” (i.e., not all sequence-based domains are known, and not all existing proteins are used for identification of domains) and this affects the estimate of the number of domains. A population of pseudogenes, i.e., disabled copies of genes that do not produce a functional, full-length copy of protein, is an additional source of uncertainty in identification of protein domains. However, our data analysis suggests the total number of structurally and functionally distinct protein domains in nature is very much limited.

Analysis of the relationships between the numbers of distinct protein-coding genes/ORFs, the numbers of observed domains and the numbers of domain-protein links in 70 Archaea, Bacteria, and Eukaryotic proteome samples allow us to estimate the total numbers of distinct domains in these proteomes, even for incomplete sequenced genome organisms. Surprisingly enough, the numbers of distinct protein domains in the mouse and human proteomes (4915 and 5271 InterPro domains) approach our estimate of the total number of protein domains in nature.

Could the number of protein domains grow in the future, and, if so, could that number grow without bound as evolution creates more and more complex and varied species? Clearly nature can evolve new protein domains, both from non-domain sequences and from modification and combination of existing domains. Indeed, homologous sequences that are classified as the same domain may rather be the beginnings of new domains.

The answer to this question also depends on the definition of a domain, i.e., on the functional-structural “radius” of the cluster of polypeptide sequences that will be taken to be the same domain, and on the “birth” and “death” rates of domains used in “newer” species imposed by evolution.

We found that the total numbers of non-redundant domain-protein links in sample proteomes strongly correlate with the number of protein-coding genes/ORFs in different organisms. This finding allows us to estimate the number of protein-coding genes/ORFs in the mouse, *A. thaliana* and human proteomes. However, the number of genes, as well as the number of domains used in our analysis, have been obtained *in silico* and significant fractions of “genes” and “protein domains” have not yet been discovered. Further experimental verification of the protein-coding sequences and computational analysis of that information will allow us to correct the estimates of the total number of the protein-coding genes in complex eukaryotic organisms and the number of protein domains in the entire “proteome world”.

## Acknowledgment

Thanks to Robert Bonner, Ralph Nossal for critical comments on this work. Thanks to Paul Kersey for his comments on InterPro data.

## Appendix 1

List of the organisms ordered by the total number of the non-redundant domain-to-protein links,  $M$ , in the sample proteome. *Guillardia theta* (algal nucleomorph), *Ureaplasma parvum*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Mycoplasma pulmonis*, *Buchnera aphidicola* (subsp. *Acyrtosiphon pisum*), *Chlamydia trachomatis*, *Chlamydia muridarum*, *Borrelia burgdorferi*, *Rickettsia conorii*, *Chlamydia pneumoniae* strain AR39, *Chlamydia pneumoniae* strain CWL029, *Treponema pallidum*, *Rickettsia prowazekii*, *Chlamydia pneumoniae* strain J138, *Aeropyrum pernix* K1, *Thermoplasma acidophilum*, *Thermoplasma volcanium*, *Helicobacter*

*pylori strain 26695, Helicobacter pylori strain J99, Pyrobaculum aerophilum, Methanobacterium thermoautotrophicum, Sulfolobus tokodaii, Pyrococcus horikoshi, Campylobacter jejuni, Mycobacterium leprae, Methanococcus jannaschii, Aquifex aeolicus, Streptococcus pyogenes strain SF370, Pyrococcus abyssi, Neisseria meningitidis strain Z2491 (serogroup A), Neisseria meningitidis strain MC58 (serogroup B), Haemophilus influenzae, Halobacterium sp. NRC-1, Xylella fastidiosa, Thermotoga maritima, Sulfolobus solfataricus, Archaeoglobus fulgidus, Lactococcus lactis (subsp. lactis) strain IL1403, Pasteurella multocida, Clostridium perfringens, Listeria innocua, Staphylococcus aureus strain Mu50, Staphylococcus aureus strain N315, Deinococcus radiodurans, Listeria monocytogenes, Brucella melitensis, Synechocystis sp. PCC 6803, Caulobacter crescentus, Clostridium acetobutylicum, Yersinia pestis, Bacillus halodurans, Vibrio cholerae, Bacillus subtilis, Salmonella typhi, Salmonella typhimurium, Escherichia coli K-12, Escherichia coli O157:H7 sub-strain RIMD 0509952, Escherichia coli O157:H7 strain EDL933, Schizosaccharomyces pombe, Anabaena sp. strain PCC 7120, Rhizobium loti, Saccharomyces cerevisiae, Pseudomonas aeruginosa, Rhizobium meliloti, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, Arabidopsis thaliana, Homo sapiens.*

**Appendix 2. A Stochastic Model of Macromolecular Evolution in a Proteome**

Let  $p_m(t) = P\{D_t = m\}$  be the probability function associated with the random process  $\{D_t, t > 0\}$ ; ( $m = 0, 1, 2, \dots$ ). We used the forward Kolmogorov equations [8]:

$$\begin{aligned} \frac{dp_0(t)}{dt} &= -\lambda_0(t)p_0(t) + \mu_1(t)p_1(t), \\ \frac{dp_m(t)}{dt} &= -(\lambda_m(t) + \mu_m(t))p_m(t) \\ &\quad + \lambda_{m-1}(t)p_{m-1}(t) \\ &\quad + \mu_{m+1}(t)p_{m+1}(t), \end{aligned} \tag{12}$$

where the functions  $\lambda_m(t)$  and  $\mu_m(t)$  are the intensities of the birth and death processes for a given value  $m$ , respectively. The initial probability distribution  $p_m(0) \geq 0$  ( $m = 0, 1, 2, \dots$ ) satisfies to the condition:  $\sum_{i \geq 0} p_i(0) = 1$ .

Here, we consider the limiting random process such that the intensities are linear functions of  $m$  defined by Eq. (3).

Let  $a = \lambda_1^*/\lambda_2^*$ ;  $\theta = \lambda_2^*/\mu_2^*$  and  $b = \mu_1^*/\mu_2^*$ . Let us denote the factorial power  $z^{[m]} = z(z + 1) \dots (z + m - 1)$ . A non-zero limiting (or steady state) solution of Eq. (12) with the intensities defined by Eq. (2) exists and can be defined from the stationary conditions  $dp_m(t)/dt = 0$ ;  $m = 0, 1, \dots$ . Define

$$p_m^* = \lim_{t \rightarrow \infty} p_m(t), \quad (m = 0, 1, 2, \dots), \tag{13}$$

then we can obtain

$$p_0 = \left( 1 + \sum_{m=1}^{\infty} \prod_{i=1}^m \theta \frac{(a+i-1)}{(b+i)} \right)^{-1}. \quad (14)$$

and

$$p_m^* = bp_0 \frac{a^{[m]}}{b^{[m+1]}} \theta^m, \quad (15)$$

(Kuznetsov, submitted). We call Eqs. (15) and (14) the Kolmogorov–Waring (KW) probability function. The non-zero limiting solution of Eq. (12) exists at Eq. (2) if  $\frac{\lambda_2^*}{\mu_2^*} < 1$  or if ( $\frac{\lambda_2^*}{\mu_2^*} = 1$  and  $\mu_1^* > \lambda_1^*$ ).

Note, the limiting KW probability distribution includes a family of skewed Paretian distributions when

$$\frac{p_{m+1}^*}{p_m^*} \leq 1, \quad m = 1, 2, \dots,$$

i.e.,

$$\theta \leq \frac{(b+1+m)}{(a+m)}; \quad m = 1, 2, \dots$$

Note it is always true that  $\theta \leq 1$  (or  $\lambda_2^*/\mu_2^* \leq 1$ ).

Several well known distributions can be derived from Eqs. (14) and (15). For example, the Waring distribution [10, 11] arises when  $p_0 = (1 - a/b)$  and  $\theta = 1, b > a > 0$ . When  $\theta = 1, a = 1$  and  $b > 1$ , the limiting KW distribution is the Yule distribution. When  $\theta = 1$ , the left tails of the both Waring and Yule distributions converge to the power law distribution  $p_k = c * m^k$  with power parameter  $k$  equals to  $-2$  (the Lotka–Zipf distribution) [11] (see also Figs. 4 and 5).

## References

- [1] Apic G., Gough J., and Teichmann S. A., An insight into domain combinations, *Bioinformatics* **17** (suppl.1) (2001) pp. S83–S89.
- [2] Aravind L., Dixit, V. M. and Koonin E. V., Apoptotic molecular machinery: Vastly increased complexity in vertebrates revealed by genome comparisons, *Science* **291** (2001) pp. 1279–1284.
- [3] Bateman A., Birney E., Cerruti L., Durbin R., Ewlinger L., Eddy S. R., Griffiths-Jones S., Howe K. L., Marshall M. and Sonnhammer E. L. L., The Pfam protein families database, *Nucleic Acids Res.* **30**(1) (2002) pp. 276–280.
- [4] Copley R. R., Doerks T., Letunic I. and Bork P., Protein domain analysis in the era of complete genomes, *FEBS Letters* **513** (2002) pp. 129–134.
- [5] Daly M., Estimating the human gene count, *Cell* **109** (2002) pp. 283–284.
- [6] Douglas S., Zauner S., Fraunholz M., Beaton M., Penny S., Deng L.-T., Wu X., Reith M., Caveller-Smith T. and Maler U.-G., The highly reduced genome of an enslaved algal Nucleus, *Nature* **410** (2001) pp. 1091–1096.
- [7] Ewing B. and Green P., Analysis of expressed sequence tags indicates 35,000 human genes, *Nat. Genet.* **25** (2000) pp. 232–234.
- [8] Feller W., *An Introduction to Probability Theory and Its Applications* (Vol. 1, 3rd Edition) (Wiley, New York, 1968).

- [9] Hogenesch J. B., Ching K. A., Batalov S., Su A. I., Walker J. R., Zhou Y., Kay S. A., Schultz P. G. and Cooke M. P., A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes, *Cell* **106**(4) (2001) pp. 413–415.
- [10] Irwin J. O., The place of mathematics in the medical and biological sciences, *J. Royal Stat. Soc.* **A126** (1963) pp. 1–45.
- [11] Johnson N. L., Kotz S. and Kemp A. W., *Univariate Discrete Distributions* (Wiley and Sons, New York, 1993).
- [12] Kanapin A., Apweiler R., Biswas M., Fleischmann W., Karavidopoulou Y., Kersey P., Kriventseva E. V., Mittard V., Mulder N., Oinn T., Phan L., Servant F. and Zdobnov E., Interactive InterPro-based comparisons of proteins in whole genomes, *Bioinformatics* **18**(2) (2002) pp. 374–375.
- [13] Kuznetsov V. A., Distribution associated with stochastic processes of gene expression in a single eukaryotic cell, *EURASIP J. on Applied Signal Processing* **4** (2001) pp. 285–296.
- [14] Kuznetsov V. A., Statistics of the numbers of transcripts and protein sequences encoded in the genome, *Computational and Statistical Methods to Genomics*, ed. by Zhang W. and Shmulevish I. (Kluwer Academic Press, Boston, 2002) pp. 125–171.
- [15] Kuznetsov V. A. and Pickalov V. V., The numbers of protein domain sequences and protein coding genes in the evolved proteomes, in *Proc. Third Int. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002)* (July 14–20, 2002, IC&G: Novosibirsk, Russia) **3** pp. 160–163.
- [16] Kuznetsov V. A., Knott G. D. and Bonner R. F., General statistics of stochastic process of gene expression in eukaryotic cells, *Genetics* **161**(3) (2002) pp. 1321–1332.
- [17] Lander E. S., Linton L. M., Birren B., Nusbaum C., Zody M. C., Baldwin J., Devon K., Dewar K., Doyle M. and FitzHugh W. *et al.*, Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium, *Nature* **409** (2001) pp. 860–921.
- [18] Letunic I., Goodstadt G., Dickens G. N. J., Doerks T., Schultz J., Mott M., Ciccarelli F., Copley R. R., Ponting C. P. and Bork P., Recent improvements to the SMART domain-based sequence annotation resource, *Nucl. Acids Res.* **30** (2002) pp. 242–244.
- [19] Marchler-Bauer A., Panchenko A. R., Shoemaker B. A., Thiessen P. A., Geer L. V. and Bryant S. H., CDD: A database of conserved domain alignments with links to domain three-dimensional structure, *Nucl. Acids Res.* **30**(1) (2002) pp. 281–283.
- [20] Rzhetsky A. and Gomez S. M., Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome, *Bioinformatics* **17** (2001) pp. 988–996.
- [21] Venter J. C., Adams M. D., Myers E. W., Li P. W., Mural R. J., Sutton G. G., Smith H. O., Yandell M., Evans C. A. and Holt R. A. *et al.*, The sequence of the human genome, *Science* **291** (2001) pp. 1304–1351.
- [22] Wolf Y. I., Grishin N. V. and Koonin E. V., Estimating the number of protein folds and families from complete genome data, *J. Molec. Evolution* **299** (2000) pp. 897–905.
- [23] Wuchty S., Scale-free behavior in protein domain networks, *Molec. Biol. Evol.* **18**(9) (2001) pp. 1694–1702.
- [24] Yule G. U., A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis, F.R.S., *Philosophical Transactions of the Royal Society of London. Ser.* **B213** (1924) pp. 21–87.