Protein Family Expansions and Biological Complexity

Christine Vogel[1,2*], Cyrus Chothia[1]

[1] Medical Research Council Laboratory of Molecular Biology, Cambridge, United Kingdom, [2] Institute for Cellular and Molecular Biology, University of Texas at Austin, Austin, Texas, United States of America

During the course of evolution, new proteins are produced very largely as the result of gene duplication, divergence and, in many cases, combination. This means that proteins or protein domains belong to families or, in cases where their relationships can only be recognised on the basis of structure, superfamilies whose members descended from a common ancestor. The size of superfamilies can vary greatly. Also, during the course of evolution organisms of increasing complexity have arisen. In this paper we determine the identity of those superfamilies whose relative sizes in different organisms are highly correlated to the complexity of the organisms. As a measure of the complexity of 38 uni- and multicellular eukaryotes we took the number of different cell types of which they are composed. Of 1,219 superfamilies, there are 194 whose sizes in the 38 organisms are strongly correlated with the number of cell types in the organisms. We give outline descriptions of these superfamilies. Half are involved in extracellular processes or regulation and smaller proportions in other types of activity. Half of all superfamilies have no significant correlation with complexity. We also determined whether the expansions of large superfamilies correlate with each other. We found three large clusters of correlated expansions: one involves expansions in both vertebrates and plants, one just in vertebrates, and one just in plants. Our work identifies important protein families and provides one explanation of the discrepancy between the total number of genes and the apparent physiological complexity of eukaryotic organisms.

Citation: Vogel C, Chothia C (2006) Protein family expansions and biological complexity. PLoS Comput Biol 2(5): e48. DOI: 10.1371/journal.pcbi.0020048

Introduction

During the course of evolution, the complexity of organisms as measured by the total number of their cells and the number of different cell types has increased greatly. The different processes that have produced these increases in biological complexity are of fundamental interest, and the data available from complete genome sequences should allow us to eventually determine their general nature and relative contributions. Prior to the information available from the genome projects, it was believed that one central process is the formation of new genes by gene duplication, divergence, and combination [1–6]. Particular emphasis was placed on extensions in the repertoire of proteins involved in the regulation of expression and in signal transduction; for a review see Kirschner and Gerhart [7].

From analyses of prokaryote genome sequences, van Nimwegen [8] and Ranea et al. [9] have shown that the number of genes in different functional categories scales as a power-law of the total number of genes. For different functional categories, the exponent of the power-law has different values. High values, ~2, are indeed found for proteins involved in transcription and its regulation and for those involved in signal transduction. Low values, <0.5, are found for those involved in protein biosynthesis, the cell cycle, and DNA replication [8]. Other functional groups have intermediate values. Van Nimwegen also obtained somewhat similar results from an analysis of the eukaryote genome sequences available at the time he carried out that work [8].

In eukaryotes, a comparison of the predicted protein sequences of the unicellular yeast, *Saccharomyces cerevisiae,* and the multicellular nematode, *Caenorhabditis elegans,* also showed that the nematode has many more proteins, and types of proteins involved in transcription and its regulation, and in signal transduction, than yeast [10]. Subsequently, an analysis of the proteins with these functions in the plant [11], the fly *Drosophila melanogaster* [12], and in humans [13] showed that repertoire of these proteins becomes larger and more diverse as complexity increases.

However, in eukaryotes there is a complication not found in prokaryotes: the biological complexity of an organism is not correlated with its total number of genes (Figure 1A, $R = 0.54$). The current known number of genes are 26,750 in the plant *Arabidopsis,* 20,050 in the nematode *C. elegans,* 13,770 in *Drosophila,* and 22,220 in the vertebrate *Homo sapiens* [14] (Figure 1A). This implies that, whilst the expansion of some protein families can lead to the evolution of organisms of higher complexity, other protein families expanded to improve an organism's adaptation to its environment but without a substantial change in complexity. These two types of expansion have been called "progressive" and "conservative" protein family expansions, respectively [15].

In this paper, we determine the extent to which the expansion of individual protein families and combinations of

Editor: Philip Bourne, University of California San Diego, United States of America

Received June 10, 2005; **Accepted** March 27, 2006; **Published** May 26, 2006

DOI: 10.1371/journal.pcbi.0020048

Abbreviations: SCOP, Structural Classification of Proteins; TNF, tumour necrosis factor

* To whom correspondence should be addressed. E-mail: cvogel@mail.utexas.edu

Synopsis

One of the main goals in biology is to understand how complex organisms have evolved. Much of an organism's physiology, and hence complexity, is determined by its protein repertoire. The repertoire has been largely formed by the duplication, divergence, and combination of genes. This means that proteins can be grouped into families whose members are descended from a common ancestor. The authors have examined the sizes of 1,219 protein families in 38 eukaryotes of different complexity. Only a small fraction of protein families have expansions that are correlated with the number of cell types in the organisms. Half of these families are involved in regulation or extracellular processes. Other families do have expansions but in a lineage-specific manner. Thus, certain protein family expansions are "progressive" in that they lead to increases in biological complexity; other expansions are "conservative" in that they help an organism to adapt better to its environment, but do not increase its complexity. This means that there is no simple correlation between an organism's complexity and the number of its genes.

families correlates with increases in macroscopic complexity of organisms such as fungi, protozoa, plants, nematodes, insects, a urochordate, and vertebrates. We measure protein family expansions in terms of the number of proteins that contain domains of defined superfamilies [16] (Figure 1B and Dataset S1). Domains are the structural, functional, and evolutionary units that form proteins; and domains of common ancestry are grouped into superfamilies based on evidence from structure, sequence, and function [16]. Two proteins that contain a domain of the same superfamily are grouped into one protein family.

We carry out two sets of calculations. First, we determine whether families undergo different expansions in different genomes and whether or not the expansions are related to the complexity of the organisms from which they come. As a measure of biological complexity of an organism, we use its number of different cell types. Second, we determine the extent to which the abundance of different superfamilies within different sets of the genomes correlate with each other. We identify three major trends that encompass about half of the largest superfamilies.

Results/Discussion

The 38 eukaryote genomes used in this work comprise those from 11 vertebrates, among which five are mammals from one urochordate, five protostomia, including two nematodes and three insects, from ten fungi, eight protozoa, and from three plants, of which two are flowering (magnoliophyta). The phylogenetic relationships of these organisms are indicated in Figure 1A. We focus on the relationships between organisms and their protein repertoires at the level of major phylogenetic groups, such as uni- and multicellular plants, protozoa, fungi, protostomia, the urochordate, vertebrates, and, within the latter ones, mammals, but do not resolve relationships within these groups. Further, our analysis is limited to those 60% of proteins for which there is good information on the evolutionary relationships of the domains of which they are composed, as well as information on domain functions. Before turning to our results, we describe briefly how the repertoire of domain superfamilies is defined and predicted in genome sequences, and which types of functions they usually have.

Domain Superfamilies as Units of Protein Evolution

Our ability to detect the evolutionary relationships of proteins, or protein domains, by sequence comparisons is limited because they frequently diverge beyond the point where their true relationship can be recognised by such comparisons. Also, large proteins are formed by combinations of domains that often come from different superfamilies. The presence or absence of superfamily relationships and of different domains can be determined if the three-dimensional structure of proteins is known, and these relationships are described in the Structural Classification of Proteins database (SCOP) [16]. It is these proteins and domains of known structure and their clear homologues for which we searched in the genome sequences of the 38 organisms.

The SUPERFAMILY database [17] contains hidden Markov models of the one-domain proteins and of the individual domains in multidomain proteins that are in the SCOP database. The SUPERFAMILY database also contains a description of the significant matches that the hidden Markov models make to the protein sequences predicted to the known genomes. Matches are made to all or part of about 60% of the predicted proteins in each genome. We extracted from SUPERFAMILY the matches made by the hidden Markov models to the sequences of 38 eukaryotes and placed them in their respective superfamilies. This procedure resulted in 1,219 domain superfamilies that occur in at least one protein in at least one of the 38 genomes. In human, for example, we find 950 of these superfamilies, and they map to a total of 19,225 domains [18]. In our analysis, we sometimes refer to a subset of largest superfamilies; these are the 299 superfamilies that occur in at least 25 proteins in at least one of the genomes.

The Functions of Superfamily Members

In an extension of domain annotations described previously [19], we manually assigned each superfamily to one of 50 types of function from a scheme similar to that used in COGs (clusters of orthologous groups of proteins) [20]. The annotation is based on information taken from SCOP [21], Pfam [22], SwissProt [23], and literature. Each of these functions map to one of seven general categories (see Protocol S1 and http://polaris.icmb.utexas.edu/people/cvogel/HV): (1) Information: storage and maintenance of the genetic code, DNA replication/repair, general transcription/translation; (2) Regulation: regulation of gene expression and protein activity, information processing in response to environmental input, signal transduction, general regulatory or receptor activity; (3) Metabolism: anabolic and catabolic processes, cell maintenance/homeostasis, secondary metabolism; (4) Intracellular processes; cell motility/division, cell death, intracellular transport, secretion; (5) Extracellular processes: inter- and extracellular processes (e.g., cell adhesion), organismal processes (e.g., blood clotting), immune system; (6) General: general and multiple functions, interactions with proteins/ions/lipids/small molecules; and (7) Other/Unknown: unknown function, viral proteins/toxins.

We are aware that the members of some superfamilies, particularly the large ones, may have a variety of functions.

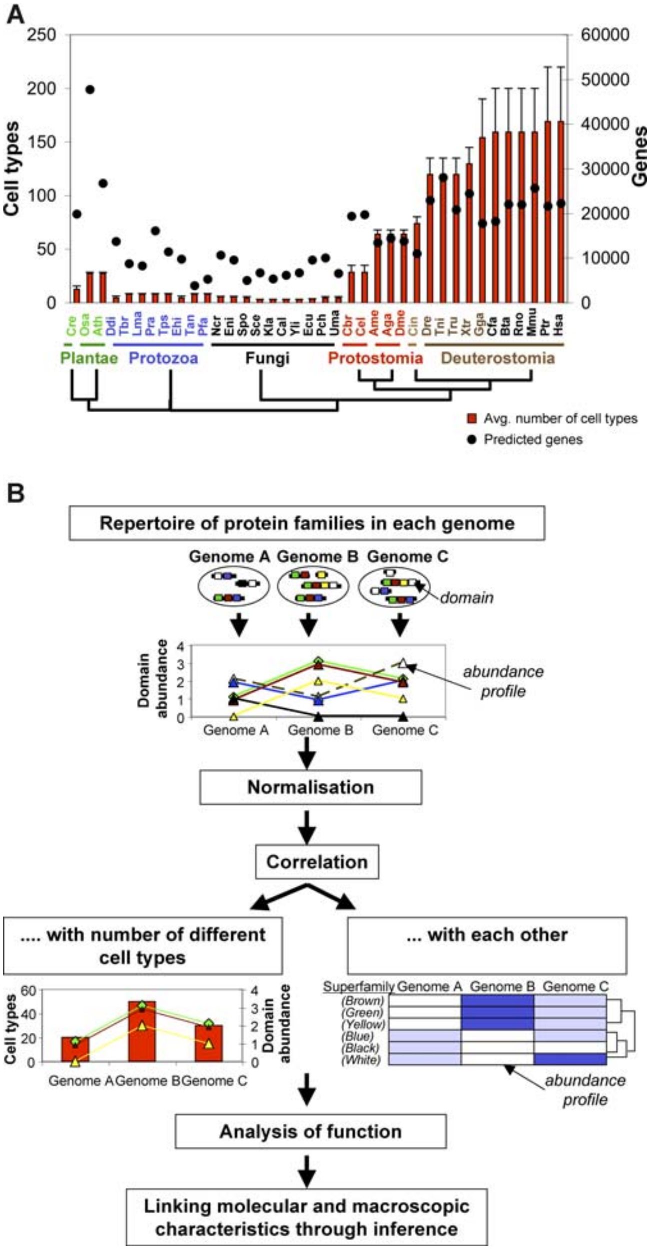

Figure 1. Motivation and Outline of the Analysis

(A) The number of genes and eukaryotic complexity are uncorrelated. The figure displays for 38 eukaryotic genomes the estimated number of different cell types [28,29] in relation to the predicted total number of genes. The tree indicates, in a simplified form, the phylogenetic relationships between the organisms as taken from the National Center of Biotechnology Information (NCBI) taxonomy server (http://www.ncbi.nlm.nih.gov/Taxonomy). The order of the organisms is the same in all figures and tables; their major groups are: plants (green), protozoa (blue), fungi (black), and animals (red and brown). The correlation between the number of different cell types and the number of genes is poor ($R^2 = 0.29$, $R = 0.54$).

Within the plants, we distinguish green algae *(Cre, Chlamydomonas reinhardtii),* and flowering plants *(Osa, O. sativa; Ath, Arabidopsis thaliana).* We include eight protozoa *(Ddi, Dictyostelium discoideum; Tbr, Trypanosoma brucei; Lma, Leishmania major; Pra, Phytophthora ramorum; Tps, Thalassiosira pseudonana; Ehi, Entamoeba histolytica; Tan, Theileria annulata; Pfa, Plasmodium falciparum),* and ten fungi *(Ncr, Neurospora crassa; Eni, Emericella nidulans; Spo, Schizosaccharomyces pombe; Sce, S. cerevisiae; Kla, Kluyveromyces lactis; Cal, Candida albicans; Yli, Yarrowia lipolytica; Ecu, Encephalitozoon cuniculi; Pch, Phanerochaete chrysosporium; Uma, Ustilago maydis).* Protostomia include two nematodes *(Cbr, Caenorhabditis briggsae; Cel, C. elegans),* and three insects *(Ame, Apis mellifera; Aga, Anopheles gambiae; Dme, D. melanogaster).* Deuterostomia include one urochordate *(Cin, Ciona intestinalis),* and 11 vertebrates, among which six are mammals *(Dre, Danio rerio; Tni, Tetraodon nigroviridis; Tru, Takifugu rubripes; Xtr, Xenopus tropicalis; Gga, Gallus gallus;* and *Cfa, Canis familiaris; Bta, Bos taurus; Rno, Rattus norvegicus; Mmu, Mus musculus; Ptr, Pan troglodytes;* and *Hsa, H. sapiens,* respectively).

(B) Outline of our analysis. For each of the 38 genomes (three, symbolised by circles), we collected information on the number of proteins (lines with boxes) that contain domains of particular superfamilies (boxes of particular colour). The resulting abundance profiles were normalised and compared both to the estimated number of different cell types in each organism, and to each other. Analysis of function of particular groups of domain superfamilies gives information on how their expansion in some organisms may have supported an increase in organismal complexity.

DOI: 10.1371/journal.pcbi.0020048.g001

For example, immunoglobulin domains are involved in cell adhesion, muscle structure, the extracellular matrix, and the immune system. The function categories here aim to describe the dominant and most widespread function for each superfamily, as far as it is known today. We annotated all 1,219 domain superfamilies of seven SCOP classes *a* to *g* [21] that occur in the 38 genomes. Close to half of all superfamilies (448) have metabolism-related functions, while each of the other categories comprises less than 15% of the domain superfamilies (Table 1). In humans, one-third of the superfamilies are metabolic (339/950), mapping to one-sixth of all domains (3,212/19,225). Some 10% of the superfamilies (122) have unknown functions (also see Figure S1).

Family Expansions and the Number of Different Cell Types

We aim to identify superfamilies whose expansions may have supported an increase in biological complexity in some eukaryotes as compared to others, thereby linking molecular characteristics to a macroscopic phenotype. As a measure of the biological complexity of an organism, we would ideally use information on both the number of cell types and the total number of cells. While some previous work is available for closely related organisms on correlates of their body mass [24–27], information on the total number of cells is not readily available for a wide range of organisms such as those used in our analysis. Therefore, we use here as a measure of complexity the estimates made for the number of different cell types found in different organisms [28,29]. Fungi and protozoa have the lowest complexity with five or fewer cell types; vascular plants have a similar number of different cell types as nematodes (i.e., ~30 or fewer), and vertebrates are of highest complexity with some 170 different cell types (Figure 1 and Dataset S1).

For each individual domain superfamily, we calculated the correlation between its abundance profile and the estimated number of different cell types per genome. The abundance is the number of proteins in a genome that contain at least one domain of a particular superfamily. The abundance profile is the collection of abundances of a domain superfamily across several genomes. In normalised form, the profile expresses relative domain abundances.

The Pearson correlation coefficient R is a measure of linear relationship between to sets of variables; R equals 1 or –1 if there is a perfect positive or negative linear correlation, respectively. $|R| \leq 0.20$ implies a very weak or nonexistent linear correlation. Squared R (R^2) is a measure for the proportion of variance in the data that are explained by a

Table 1. Few Domain Superfamilies Correlate Well with the Number of Different Cell Types

Process	R $\leq$ 0.20		R $\geq$ 0.80		Total	
Metabolism	248	0.45	28	0.14	448	0.37
Information	87	0.16	12	0.06	175	0.14
Intracellular processes	61	0.11	28	0.14	169	0.14
Regulation	57	0.10	56	0.29	163	0.13
Other	66	0.12	3	0.02	122	0.10
General	25	0.05	28	0.14	87	0.07
Extracellular processes	11	0.02	39	0.20	55	0.05
Total	555	1	194	1	1,219	1

A fraction of superfamilies (194/1,219) exhibit a good correlation between their abundance and the number of different cell types in 38 genomes (R $\geq$ 0.80). Close to half of all superfamilies (555/1,219) exhibit no or negative correlation with the number of cell types ($R \leq 0.20$). Each domain superfamily was mapped to one of seven function categories. The table shows the distribution of superfamilies across these functions. While domains of metabolism largely contribute to superfamilies with poor correlation with the number of different cell types, domains of regulation and extracellular processes (e.g., cell adhesion, the immune system, or blood clotting) have an unusually large contribution to superfamilies that are well correlated with the number of different cell types, indicated in red. The function distributions are significantly biased ($p < 0.001$).

DOI: 10.1371/journal.pcbi.0020048.t001

linear relationship between two variables, e.g. about two thirds of the variance are explained at $R = 0.80$ ($R^2 = 0.64$).

The distribution of correlation coefficients R between abundance profiles and the number of different cells types is shown in Figure 2. Only 15% of the superfamilies (194/1,219) show a strong correlation between their abundance and the number of different cell types per organism, with $R \geq 0.80$. These superfamilies expand mostly in vertebrates, have intermediate abundance in other animals and plants, and are of low frequency or absent in fungi and protozoa. Some 40% of the superfamilies (470/1,219) have a correlation coefficient R between 0.20 and 0.80. Close to 45% of the superfamilies (555/1,219) have correlation coefficients between 0.20 and negative values.

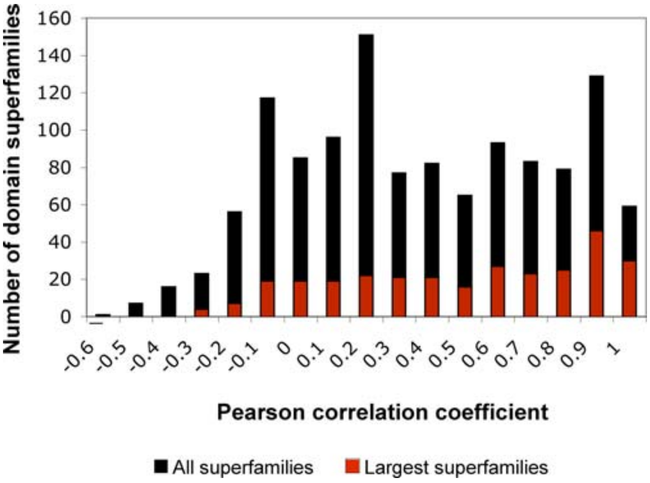

Figure 2. Some Family Expansions Correlate Well with the Number of Different Cell Types in Each Organism

For each of the 1,219 domain superfamilies and their profile of abundance in the 38 genomes, we calculated the correlation coefficient R of the profile with the number of different cell types per organism. The distribution of R values is plotted in black. For the subset of largest superfamilies (i.e., those with at least 25 proteins in one of the genomes) the distribution of R values is shown in red. There are few superfamilies with high correlation ($R \geq 0.80$), and many with poor correlation or slight anticorrelation ($R \leq 0.20$); this distribution is similar for both sets of superfamilies.

DOI: 10.1371/journal.pcbi.0020048.g002

Family Expansions with Good Correlation with the Number of Cell Types

We examined in detail the properties of those superfamilies that have strong correlations with the number of different cell types in 38 organisms ($R \geq 0.80$). These proteins are described in Tables 1–3, and Figure 3A. This group of superfamilies represents only 15% of all superfamilies (194/1,219), but they are found in more than 40% of human domains (7,825/19,225; Table 3). In contrast, they form only a tenth of the domains in *Arabidopsis* (1,884/19,323; Table 3).

We examined the functions of the domain superfamilies to understand how their duplications may have supported the emergence of novel cell types. For the 194 superfamilies with good correlation with the number of different cell types, all functional categories make some contribution, but two make disproportionally large ones ($p < 0.001$; Tables 1 and 2). These categories include superfamilies of extracellular processes (20%) and superfamilies of regulation (29%), encompassing close to one-half (49%) of the 194 superfamilies. In particular, these families include many domains of signal transduction (e.g., PH-, SH3, and SH2 domains), DNA-binding domains (e.g., C2H2 and C2HC zinc fingers, and winged helix domains), kinases and phosphatases and cell adhesion molecules (e.g., immunoglobulin, fibronectin type III, or EGF/laminin domains) (Table 2). We also observe several large families (e.g., dsRNA-binding, Tudor/PWWP/MBT, SAM/Pointed, or KH domains) that are known to have RNA-binding activity [30], some in addition to DNA-binding activity. The other function contributions usually include smaller superfamilies, and consist of domains involved in metabolism, intracellular processes, and information (Tables 1 and 2).

Generally, we can distinguish three types of expansion patterns among the 194 superfamilies with high correlations (Table 2 and Dataset S1). Close to one-third of the superfamilies are found in all 38 genomes, one-third exclusively occurs in animals, and one-third occurs in animals and has a spasmodic distribution in the other kingdoms. Of those found in all genomes, the abundance is usually highest in vertebrates, particularly in mammals, and moderate in the other animals and low in plants, protozoa, and fungi. Examples are the PH domains and GTPase activation domains, which both function in signal transduction (Table 2). Another example is voltage-gated potassium channels,

Table 2. Domain Families with Good Correlation with the Number of Different Cell Types

General Function	More Detailed Function	*R*	Name	*Ath*	*Ddi*	*Sce*	*Dme*	*Hsa*
Regulation	Signal transduction	0.91	PH domain-like	53	111	26	119	377
	Signal transduction	0.92	SH3-domain	5	29	23	59	204
	Signal transduction	0.85	C2 domain (Calcium/lipid-binding domain, CaLB)	104	58	10	46	147
	Signal transduction	0.89	PDZ domain-like	17	1	2	67	145
	Signal transduction	0.93	SH2 domain	2	13	0	32	110
	Signal transduction	0.81	GTPase activation domain, GAP	10	59	15	28	97
	Signal transduction	0.84	Growth factor receptor domain	1	5	0	29	82
	Signal transduction	0.80	DBL homology domain (DH-domain)	0	47	5	23	67
	Signal transduction	0.83	Ypt/Rab-GAP domain of gyp1p	23	29	10	24	53
	Signal transduction	0.90	Regulator of G-protein signalling, RGS	1	10	3	9	37
	Signal transduction	0.81	PYP-like sensor domain (PAS domain)	18	11	1	15	34
	Signal transduction	0.93	Frizzled cysteine-rich domain	0	4	0	8	19
	Signal transduction	0.92	Transducin (heterotrimeric G protein), gamma chain	0	1	0	5	17
	Signal transduction	0.87	Doublecortin (DC)	0	2	0	4	10
	Signal transduction	0.82	SPOC domain-like	2	2	2	4	6
	Signal transduction	0.92	Stathmin	0	0	0	1	5
	Signal transduction	0.87	Nuclear receptor coactivator interlocking domain	0	0	0	0	5
	Signal transduction	0.88	Phenylalanine zipper	0	0	0	1	3
	Signal transduction	0.83	N-terminal domain of cbl (N-cbl)	0	0	0	1	3
	Signal transduction	0.87	TRADD, N-terminal domain	0	0	0	0	1
	Signal transduction	0.83	Ran-GTPase activating protein 1 (RanGAP1), C-terminal domain	0	0	0	0	1
	RNA binding, m/tr	0.88	dsRNA-binding domain-like	22	11	6	14	38
	RNA binding, m/tr	0.92	R3H domain	6	3	1	4	11
	Receptor activity	0.90	Plexin repeat	0	0	0	5	30
	Receptor activity	0.85	SRCR-like	0	0	0	4	25
	Receptor activity	0.93	Hormone receptor domain (HRM, Pfam 02793)	0	0	0	6	19
	Receptor activity	0.92	GDNF receptor-like (Pfam 02351)	0	0	0	2	5
	Receptor activity	0.86	Notch domain	0	1	0	2	4
	Receptor activity	0.86	alpha-2-Macroglobulin receptor associated protein (RAP) domain 1	0	0	0	1	1
	Other regulatory function	0.82	Sec7 domain	8	7	5	5	15
	Other regulatory function	0.87	Carboxypeptidase regulatory domain	1	3	0	3	13
	Other regulatory function	0.85	Mob1/phocein	4	4	2	4	7
	Other regulatory function	0.89	GCM domain	0	0	0	2	2
	Other regulatory function	0.86	Midkine	0	0	0	0	2
	Other regulatory function	0.81	Agouti-related protein	0	0	0	0	2
	Other regulatory function	0.91	Suppressor of Fused, N-terminal domain	0	0	0	1	1
	Kinases/phosphatases	0.80	(Phosphotyrosine protein) phosphatases II	24	46	14	37	109
	Kinases/phosphatases	0.87	Myosin phosphatase inhibitor 17kDa protein, CPI-17	0	0	0	1	5
	Kinases/phosphatases	0.83	FAT domain of focal adhesion kinase	0	0	0	1	2
	Kinases/phosphatases	0.82	Protein serine/threonine phosphatase 2C, C-terminal domain	0	0	0	2	2
	DNA-binding	0.93	C2H2 and C2HC zinc fingers	22	6	34	197	659
	DNA-binding	0.80	Winged helix	194	47	34	80	186
	DNA-binding	0.93	SAM/Pointed domain	11	23	5	39	99
	DNA-binding	0.81	HMG-box	21	4	7	26	63
	DNA-binding	0.89	p53-like transcription factors	0	8	1	20	46
	DNA-binding	0.86	SMAD/FHA domain	21	25	14	25	45
	DNA-binding	0.88	Bromodomain	27	15	10	19	41
	DNA-binding	0.91	A DNA-binding domain in eukaryotic transcription factors	0	0	0	4	22
	DNA-binding	0.88	lambda repressor-like DNA-binding domains	4	2	1	6	21
	DNA-binding	0.85	ARID-like	10	2	2	6	14
	DNA-binding	0.94	SAND domain-like	0	0	0	3	11
	DNA-binding	0.80	Putative DNA-binding domain	1	1	2	6	9
	DNA-binding	0.85	SMAD MH1 domain	0	0	0	4	8
	DNA-binding	0.93	Transcription factor STAT-4 N-domain	0	0	0	0	7
	DNA-binding	0.91	STAT	0	4	0	1	7
	DNA-binding	0.91	CCHHC domain	0	0	0	2	7
Intracellular processes	Transport	0.89	LDL receptor-like module	0	0	0	33	41
	Transport	0.90	Lipocalins	3	2	0	7	36
	Transport	0.82	Clathrin adaptor appendage domain	7	5	6	6	12
	Transport	0.87	Phoshotransferase/anion transport protein	0	1	0	2	9
	Protein modification	0.82	Hect, E3 ligase catalytic domain	7	5	5	14	28
	Protein modification	0.86	Proteasome activator reg(alpha)	0	1	0	1	5
	Proteases	0.86	Metalloproteases ("zincins"), catalytic domain	17	13	8	85	104
	Proteases	0.96	Kazal-type serine protease inhibitors	0	0	0	19	38
	Proteases	0.86	Serpins	12	0	0	30	34
	Proteases	0.83	Cystatin/monellin	15	3	0	5	22
	Proteases	0.90	TIMP-like	0	0	0	3	21
	Proteases	0.93	Thyroglobulin type-1 domain	0	0	0	3	17
	Proteases	0.93	Elafin-like	0	0	0	1	14

Table 2. Continued

General Function	More Detailed Function	*R*	Name	*Ath*	*Ddi*	*Sce*	*Dme*	*Hsa*
	Proteases	0.81	Macro domain-like	6	5	2	11	11
	Proteases	0.80	Calpain large subunit, middle domain (domain III)	1	2	0	3	10
	Proteases	0.85	Trefoil	0	0	0	0	9
	Phospholipid m/tr	0.87	Annexin	8	3	0	7	15
	Phospholipid m/tr	0.90	Phospholipase A2, PLA2	5	0	0	7	13
	Ion m/tr	0.87	Voltage-gated potassium channels	36	6	2	42	121
	Ion m/tr	0.80	Small-conductance potassium channel	0	0	0	1	4
	Cell motility	0.89	Moesin tail domain	0	0	0	2	5
	Cell cycle, Apoptosis	0.96	DEATH domain	0	0	0	9	79
	Cell cycle, Apoptosis	0.93	Cystine-knot cytokines	0	0	0	13	65
	Cell cycle, Apoptosis	0.96	Bcl-2 inhibitors of programmed cell death	0	0	0	2	14
	Cell cycle, Apoptosis	0.89	Caspase-like	8	2	1	7	13
	Cell cycle, Apoptosis	0.82	Inhibitor of apoptosis (IAP) repeat	0	0	1	4	10
	Cell cycle, Apoptosis	0.84	Telomeric repeat binding factor (TRF) dimerisation domain	0	0	0	0	3
	Cell cycle, Apoptosis	0.91	C-terminal domain of DFF45/ICAD (DFF-C domain)	0	0	0	0	1
Extracellular processes	Toxins/defense	0.93	Snake toxin-like	1	0	0	7	30
	Toxins/defense	0.92	Anaphylotoxins (complement system)	0	0	0	0	5
	Toxins/defense	0.88	Neurophysin II	0	0	0	0	2
	Immune response	0.94	Complement control module/SCR domain	0	0	0	14	51
	Immune response	0.92	TNF-like	0	1	0	1	49
	Immune response	0.87	MHC antigen-recognition domain	0	0	1	0	49
	Immune response	0.87	Interleukin 8-like chemokines	0	0	0	0	47
	Immune response	0.84	4-helical cytokines	0	0	0	0	47
	Immune response	0.94	Cytokine (FGF and IL-1)	2	0	0	3	35
	Immune response	0.88	Tetraspanin	0	0	0	33	30
	Immune response	0.92	TNF receptor-like	0	0	0	0	21
	Immune response	0.84	Interferon-induced guanylate-binding protein 1 (GBP1), C-terminal domain	3	1	0	1	9
	Immune response	0.91	Class II MHC-associated invariant chain ectoplasmic trimerization domain	0	0	0	0	1
	Cell adhesion	0.97	Immunoglobulin	0	1	0	123	551
	Cell adhesion	0.91	Fibronectin type III	3	0	1	56	184
	Cell adhesion	0.94	EGF/Laminin	8	10	1	46	167
	Cell adhesion	0.85	vWA-like	36	39	10	14	87
	Cell adhesion	0.88	Cadherin-like	0	0	1	19	84
	Cell adhesion	0.91	TSP-1 type 1 repeat	0	1	0	14	66
	Cell adhesion	0.91	Spectrin repeat	0	1	0	9	31
	Cell adhesion	0.86	Sema domain	0	0	0	7	31
	Cell adhesion	0.96	Integrin domains	0	0	0	7	26
	Cell adhesion	0.91	Integrin alpha N-terminal domain	1	8	0	7	22
	Cell adhesion	0.94	PGBD-like	6	2	0	2	20
	Cell adhesion	0.95	SEA domain	0	0	0	2	17
	Cell adhesion	0.94	TB module/8-cys domain	0	0	0	0	7
	Cell adhesion	0.93	Integrin beta tail domain	0	0	0	1	7
	Cell adhesion	0.90	Somatomedin B domain	0	0	0	2	7
	Cell adhesion	0.86	alpha-catenin/vinculin	0	2	0	3	6
	Cell adhesion	0.87	TSP type-3 repeat	0	0	0	2	5
	Cell adhesion	0.89	Hedgehog/DD-peptidase	0	1	0	1	3
	Cell adhesion	0.89	CAPPD, an extracellular domain of amyloid beta A4 protein	0	0	0	1	3
	Cell adhesion	0.90	Fibronectin type I module	0	0	0	0	2
	Cell adhesion	0.85	Invasin/intimin cell-adhesion fragments	1	1	0	1	2
	Cell adhesion	0.83	A middle domain of Talin 1	0	0	0	1	2
	Cell adhesion	0.81	beta-catenin-interacting protein ICAT	0	1	0	0	1
	Blood clotting	0.83	Fibrinogen C-terminal domain-like	0	0	0	15	31
	Blood clotting	0.95	Kringle-like	0	0	0	2	30
	Blood clotting	0.94	Blood coagulation inhibitor (disintegrin)	0	0	0	5	23
Metabolism	Transferases	0.81	MIR domain (Pfam 02815)	1	1	7	5	10
	Transferases	0.82	Methylated DNA-protein cysteine methyltransferase domain	0	0	0	1	1
	Secondary metabolism	0.83	ADP-ribosylation	10	17	0	2	22
	Redox	0.83	Multiheme cytochromes	1	0	3	14	307
	Redox	0.83	Aromatic amino acid monoxygenases, catalytic and oligomerization domains	0	1	0	3	4
	Redox	0.89	Nitric oxide (NO) synthase oxygenase domain	0	0	0	1	3
	Polysaccharide m/tr	0.92	Ricin B-like lectins	3	12	0	14	28
	Photosynthesis	0.84	Non-globular alpha+beta subunits of globular proteins	0	0	0	0	3
	Photosynthesis	0.82	PRC-barrel domain	0	0	0	1	3
	Other enzymes	0.82	Alkaline phosphatase-like	13	7	6	42	33
	Other enzymes	0.89	Carbonic anhydrase	8	0	0	15	22
	Other enzymes	0.80	NHL repeat	8	3	0	11	16
	Other enzymes	0.86	Transglutaminase, two C-terminal domains	0	0	0	1	9
	Other enzymes	0.88	Peptidylarginine deaminase Pad4, middle domain	0	0	0	0	5
	Other enzymes	0.83	FMT C-terminal domain-like	2	0	0	2	4

Table 2. Continued

General Function	More Detailed Function	*R*	Name	*Ath*	*Ddi*	*Sce*	*Dme*	*Hsa*
	Other enzymes	0.88	Proguanylin	0	0	0	0	2
	Other enzymes	0.84	ADP ribosyl cyclase-like	0	0	0	0	2
	Other enzymes	0.88	CNF1/YfiH-like putative cysteine hydrolases	0	0	0	0	1
	Other enzymes	0.81	RNA 3′-terminal phosphate cyclase, RPTC, insert domain	0	1	0	1	1
	Nucleotide m/tr	0.91	Nucleoside diphosphate kinases	5	4	1	4	11
	Lipid m/tr	0.86	Lipase/lipooxygenase domain (PLAT/LH2 domain)	14	0	0	3	23
	Lipid m/tr	0.85	YWTD domain	0	3	0	11	15
	Lipid m/tr	0.85	Apolipoprotein	0	0	0	0	6
	Lipid m/tr	0.90	Colipase-like	0	0	0	0	3
	Lipid m/tr	0.87	Rab geranylgeranyltransferase alpha-subunit, insert domain	0	0	0	0	1
	Energy	0.90	Mitochondrial cytochrome c oxidase subunit VIIa	0	0	0	2	4
	Energy	0.93	GFP-like	0	0	0	1	3
	Carbohydrate m/tr	0.89	Galactose-binding domain-like	45	25	4	21	73
Information	Translation	0.93	Second domain of FERM	1	5	0	19	43
	Translation	0.83	Elongation factor TFIIS domain 2	9	2	2	3	9
	Translation	0.86	L27 domain	0	0	0	1	5
	Translation	0.86	HBS1-like domain	0	0	0	1	1
	RNA processing	0.80	Eukaryotic type KH-domain (KH-domain type I)	32	5	7	28	37
	DNA replication/repair	0.93	Tudor/PWWP/MBT	30	5	3	22	53
	DNA replication/repair	0.86	Barrier-to-autointegration factor, BAF	0	0	0	1	3
	DNA replication/repair	0.88	Eukaryotic DNA topoisomerase I, dispensable insert domain	0	0	0	1	2
	DNA replication/repair	0.81	N-terminal domain of MutM-like DNA repair proteins	1	0	0	0	2
	DNA replication/repair	0.83	XRCC4, N-terminal domain	0	0	0	0	1
	Chromatin structure	0.86	Nucleoplasmin-like core domain	0	0	0	2	13
	Chromatin structure	0.81	Smc hinge domain	5	4	4	4	6
General	Structural protein	0.86	AFP III-like domain (Pfam 01354)	0	0	0	1	1
	Small molecule binding	0.92	Mannose-6-phosphate receptor binding protein 1 (Tip47), C-terminal domain	0	0	0	0	3
	Small molecule binding	0.88	A heparin-binding domain	0	0	0	1	3
	Protein interaction	0.80	Ankyrin repeat	119	88	19	85	260
	Protein interaction	0.80	POZ domain	98	35	6	82	179
	Protein interaction	0.91	WW domain	9	5	6	21	41
	Protein interaction	0.84	UBA-like	31	9	15	19	38
	Protein interaction	0.92	Hemopexin-like domain	0	0	0	2	23
	Protein interaction	0.91	Dimerization-anchoring domain of cAMP-dependent type II PK regulatory subunit	0	0	0	3	8
	Protein interaction	0.91	Polo-box domain	0	1	1	2	6
	Protein interaction	0.83	NSFL1 (p97 ATPase) cofactor p47, SEP domain	4	1	1	2	5
	Protein interaction	0.91	p53 tetramerization domain	0	0	0	0	3
	Protein interaction	0.86	Dimerization cofactor of HNF-1 alpha	0	0	0	0	2
	Protein interaction	0.82	TAF(II)230 TBP-binding fragment	1	0	0	1	2
	Protein interaction	0.80	IP3 receptor type 1 binding core, domain 2	0	0	0	1	2
	Lipid/membrane binding	0.86	LEM domain	0	0	0	2	6
	Ligand binding	0.85	Supernatant protein factor (SPF), C-terminal domain	5	2	4	9	12
	Ion binding	0.80	Pyk2-associated protein beta ARF-GAP domain	17	12	6	8	32
	Ion binding	0.88	Amyloid beta a4 protein copper binding domain (domain 2)	0	0	0	1	3
	General	0.80	WD40 repeat-like	261	191	111	188	278
	General	0.86	EF-hand	179	71	18	105	239
	General	0.81	Ubiquitin-like	129	83	21	61	140
	General	0.88	Calponin-homology domain, CH-domain	16	34	4	31	71
	General	0.84	Spermadhesin, CUB domain	0	0	0	24	51
	General	0.89	GLA-domain	0	0	0	0	14
	General	0.85	Pentein	3	3	1	2	9
	General	0.92	PKD domain	0	0	0	0	7
	General	0.87	Serum albumin-like	0	0	0	0	5
Unknown function	Unknown function	0.83	E set domains	26	77	9	58	81
	Unknown function	0.87	Domain from hypothetical 2610208m17rik protein	0	0	0	0	3
	Unknown function	0.85	Prion-like	0	0	0	0	2

The table lists 194 domain superfamilies with good correlation with the number of different cell types (i.e., $R \geq 0.80$), their general and detailed functions, and abundance in five representative genomes. The complete data are available in Dataset S1. The abundance profiles of some of the largest superfamilies are displayed in Figure 3A. The superfamilies are sorted according to their function and their abundance in human. Their names are taken directly from SCOP [21].

m/tr, metabolism/transport.

DOI: 10.1371/journal.pcbi.0020048.t002

whose vertebrate-specific expansion is possibly linked to their function in neural signalling. Two-thirds of the 194 superfamilies are only commonly found in animals, but are absent or occur in very low frequencies in the other kingdoms (Table 2). Examples are tumour necrosis factor (TNF)–like, TNF receptor–like, and also DEATH domains, all of which are known to function in apoptosis. Some 12% of these superfamilies are vertebrate specific, and examples of these include proteins of the immune system, such as major histocompatibility complex antigen–recognition domains, or four-helical cytokines.

Table 3. Contribution of Different Groups of Domain Superfamilies to the Overall Composition of Genomes

Organism	Estimated Number of Cell Types	$R \leq 0.20$	Fraction of Total Assigned Domains	$R \geq 0.80$	Fraction of Total Assigned Domains	Total Number of Assigned Domains
A. thaliana	27	7,620	0.39	1,884	0.10	19,323
D. discoideum	5	2,469	0.29	1,418	0.17	8,376
S. cerevisiae	3	1,358	0.32	533	0.13	4,243
D. melanogaster	64	2,048	0.21	2,669	0.27	9,722
H. sapiens	169	2,206	0.11	7,825	0.41	19,225

The number of different cell types is an average of different estimates (see Dataset S1). Many domains in plants, protozoa, and fungi are of superfamilies with poor correlation with the number of different cell types ($R \leq 0.20$). Vice versa, human has the largest fraction of domains from superfamilies with good correlation with the number of different cell types ($R \geq 0.80$), *Arabidopsis* has the smallest fraction.

DOI: 10.1371/journal.pcbi.0020048.t003

Family Expansions with No or Inverse Correlation with the Number of Cell Types

There are 555 of the 1,219 superfamilies whose abundances have correlation coefficients of less than 0.20; examples of these families are shown in Figure 3B. Most of the expansions do not correlate with the number of different cell types ($|R| \leq 0.20$). Only 95 superfamilies show weak inverse correlation ($R \leq -0.20$), and these superfamilies are usually small (see Dataset S1). No superfamily expansion displays strong negative correlation ($R \leq -0.80$). For this reason we regard all superfamilies with $R \leq 0.20$ as one group.

Domains from these superfamilies are found in 39% of the domains in *Arabidopsis* (7,620/19,323), but only 11% of the domains in humans (2,206/19,225; Table 3). Similar to the well-correlated superfamily expansions described above, the superfamily expansions with little correlation to the number of different cell types are significantly biased in their functions ($p < 0.001$). The major contributions to this set of superfamilies come from metabolism (45%) and information storage and replication (16%); smaller contributions are made by the other function categories (Table 1).

In many cases, the distribution of these superfamilies follows one of three patterns. They (1) generally occur in very small numbers with a spasmodic distribution (167 superfamilies seen in a total of ten or fewer proteins); (2) are widely spread with low frequencies (about half of the superfamilies occur in ≤5 proteins/organism on average); or (3) have expansions in one or a few genomes and low frequencies elsewhere. Many of these expansions have their highest abundance in plants, and intermediate or low numbers in animals.

Some examples of these superfamilies are shown in Figure 3B, sorted with respect to their abundance in *Arabidopsis.* Large superfamilies include Tetratricopeptide-repeat–like or F-box protein interaction domains that are typical for *Arabidopsis* and often involved in ubiquitination [30], and many enzymes. Another large superfamily in *Arabidopsis* is the ribonuclease inhibitor–like domains that are leucine rich and known to inhibit ribonucleases but also to bind other proteins and function in nucleic acid processing [30]. Thus they may also have a role in RNA interference, a process that is common in plants [31].

Correlated Expansions Show Three Major Trends

One implication of our work is an evaluation of the correlation between domain superfamily expansion profiles (i.e., an identification and description of the different duplication patterns of duplications that formed eukaryotic protein repertoires). To do so, we calculated the correlation coefficient (R value) for each pair of relative abundance profiles for the 299 largest superfamilies and then grouped the families sharing a high R value. The result of such clustering is shown in Figure 4. In the figure, each row denotes a domain superfamily; each column denotes a particular genome. The relative abundance of the domain superfamily in each genome is colour-coded, and the abundance profiles are hierarchically clustered.

When applying an R-value cutoff of 0.90, we obtained 26 clusters with correlated abundance profiles, and all clusters with four or more members are described in Table 4. We observed three major trends of domain superfamily expansions and several trends with fewer superfamilies involved. One-sixth of the domain superfamilies (52/299) expand specifically in vertebrates, have intermediate abundance in the other animals and plants, and very low abundance in protozoa and fungi. Unsurprisingly, all of these superfamily expansions, except for one, belong to the 194 expansions described above as strongly correlated with the number of different cell types. Further, we observed a group of 26 superfamilies that expand in both plants and vertebrates as compared to other organisms, and have intermediate numbers in the other animals. A third group of 33 superfamilies expand in plants, but have very low abundance in all other organisms.

In contrast to these three major trends, most superfamilies do not belong to the groups of vertebrate- and/or plant-specific expansions, but are members of 23 smaller clusters that have expansions specific to one or few genomes other than plants or vertebrates. These families are often of low abundance. The contribution of these superfamilies to the protein repertoire confirms previous findings on the prevalence of "lineage-specific gene family expansions" that is the emergence of different domain architectures, through domain accretion and domain shuffling, in different phylogenetic lineages [10,11,32–34].

We conducted similar clustering with all 1,219 superfamilies and a range of R-value cutoffs (see Protocol S1 and Figure S2). The results for these procedures are qualitatively the same as those described above. A more detailed description of the different expansion patterns (Figure 4 and Table 4) is very interesting, but goes beyond the scope of this paper and will be left to future analysis.

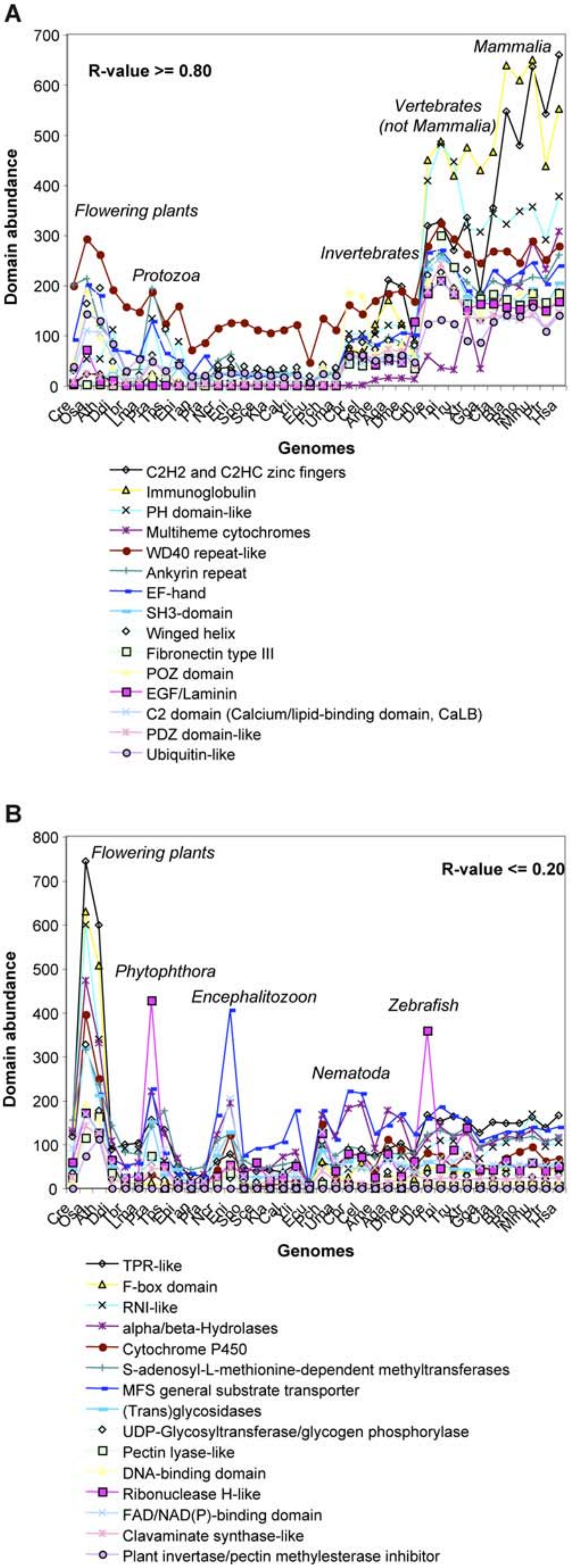

Figure 3. Examples of Family Expansions with Good or Poor Correlation with the Number of Different Cell Types

There are 194 superfamilies with good ($R \geq 0.80$; [A]) and 555 superfamilies with poor or negative ($R \leq 0.20$; [B]) correlation with the number of different cell types, and the diagrams shows 15 examples of each. Some of the peaks are annotated in italics. The genomes are in the same order as in Figure 1A. The lines between counts of domain abundance are for better visualisation only. Abbreviations are as in Figure 1.

DOI: 10.1371/journal.pcbi.0020048.g003

Future Refinements of Our Work

In the work described here we have often given precise numbers and descriptions. For a variety of reasons we expect that these statements will be refined or modified in future research. First, gene predictions, especially of recently sequenced genomes, often contain some errors. For example, we noticed a large number of *Xenopus*-specific expansions (59 superfamilies, $R \geq 0.80$), many proteins of which may in fact come from erroneously included bacterial sequences. Second, the prediction of protein domains is obviously biased by our current knowledge of domain structure, and only the completion of the systematic structural genomics projects will provide an accurate survey of domain distributions across genomes. Currently we are able to predict these domains in ~60% of the sequences [18]; increases in coverage will improve our knowledge, especially about smaller protein families. Third, we have used here rough general annotations of the functions carried out by members of different superfamilies. This annotation needs to be refined, and new experimental characterisation will reveal further functional variety within domain superfamilies. Last but not least, while the number of different cell types is a valid measure of organismal complexity, it crucially depends on how these cell types are defined [28,29]. These definitions still need to be improved, although first, most valuable steps have already been taken [35]. However, given these qualifications, we would expect that future work will not upset the broad conclusions derived by our analysis.

Summary and Conclusions

We present here one of the first studies that directly links protein family expansions to increases in eukaryotic complexity. We go beyond what has been known before in several ways. First, we use the number of different cell types as a measure of biological complexity. Second, we address a larger number of eukaryote genomes than previous analyses: we particularly focus on 17 completely sequenced animals, including two nematodes, three insects, one urochordate, and 11 vertebrates, and compare them to the genomes of three plants, eight protozoa, and ten fungal genomes (Figure 1). Third, as the proteins of these distantly related organisms are highly diverged, we include information on protein structure to accurately determine the family relationships [16,36], using domains as the smallest structural, functional, and evolutionary unit.

We demonstrate that, independent of the total number of genes, particular protein families expand in concert with increases in biological complexity and functions of these superfamilies can be linked to the evolution of more intricate physiological features. These family expansions are largely caused by gene duplications rather than by domain accretion [33] or by invention of new superfamilies: the average protein

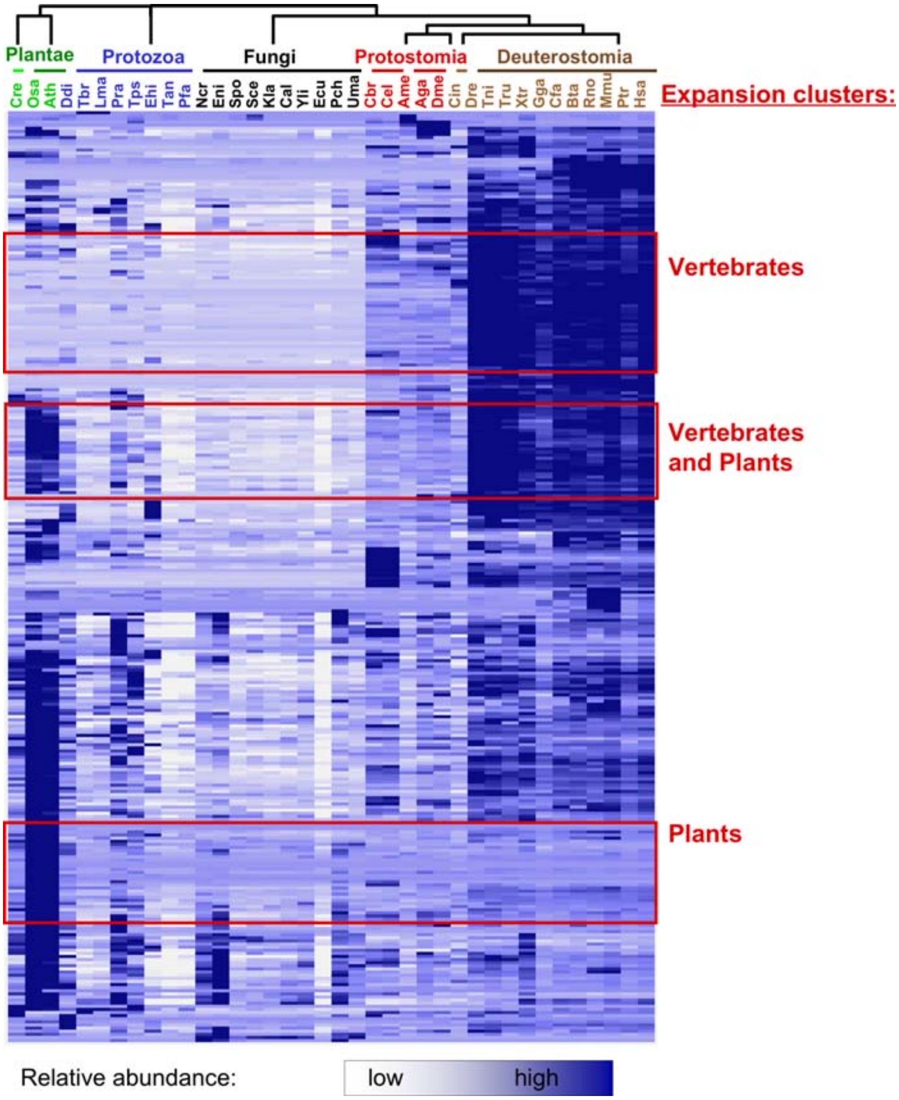

Figure 4. Domain Superfamilies Show Different Expansion Patterns

The matrix shows the 299 largest domain superfamilies that occur in $\geq$25 proteins in at least one of the genomes, hierarchically clustered. Each row represents one superfamily. Colour-coded profiles show the normalised abundance of each domain superfamily across the different eukaryotic genomes: white, low relative abundance; blue, high relative abundance. Each column represents one genome. All genomes are abbreviated and organised as in Figure 1A. A grouping of superfamily pairs with $R \geq 0.90$ results in 26 clusters, and the three largest clusters are indicated in red boxes: expansions in vertebrates (52 superfamilies) and expansions in plants (33 superfamilies), and expansions in vertebrates and plants (26 superfamilies). Further descriptions can be found in Table 4 and at http://polaris.icmb.utexas.edu/people/cvogel/HV.

DOI: 10.1371/journal.pcbi.0020048.g004

lengths, which indicate the average number of domains per protein, are similar for all 38 eukaryotes (Figure S3). As domain reshuffling is known to be linked to domain duplication [37,38], domain superfamily expansions also provide the basis for an increase in the number of domain combinations and multidomain proteins in vertebrates [13,39], which in turn increases proteome complexity.

Our work suggests that the two basic types of duplication have different relative contributions to proteomes. "Conservative expansions" do not correlate with an increase in the number of different cell types, but simply enlarge the genome size. Most protein families belong to this group. The functions of the domains involved define organism-specific properties [10,11,32,34]: they help the organism to adapt to environmental challenges. Examples are the expansions of chemoreceptors in nematodes [40] and metabolic domains in plants whose expansions correlate with an increase in the number of secondary metabolites.

In contrast, "progressive expansions" correlate with an increase in physiological complexity, but they represent only a small fraction of all domain superfamilies. We identify ~200 domain superfamilies that are both correlated with each other in their abundance pattern and correlate with the number of different cell types in eukaryote organisms; these are candidates for progressive expansions. Domains of these expansions are likely to have enabled the emergence of novel cell types and the communication between these cells [41], for example, by their functions in extracellular, or regulatory processes. An increased number of cell adhesion molecules, but also apoptotic domains, supports the intricate embryonic development found in animals. Large families of transcription factors or proteins with protein-binding domains

Table 4. Patterns of Domain Superfamily Expansions

Superfamily Number[a]	Description of Expansion Profile	Examples of the Largest Domain Superfamilies	Main Domain Functions	Comment	Node[b]
52	Specific to vertebrates; moderate expansions in the other animals; absent or in very low numbers in fungi/ protozoa/plants	C2H2 and C2HC zinc fingers (659 in humans); PH domain (377 in humans); SH3 domain (204 in humans); voltage-gated potassium channels (121 in humans)	Extracellular processes (18) (e.g., cell adhesion); regulation (15) (e.g., signal transduction)	Best correlation with the number of different cell types	158
33	Specific to multicellular plants	F-box domain (629 in *Arabidopsis*); TPR-like domain (598 in *Arabidopsis*); alpha/beta-hydrolases (472 in *Arabidopsis*); RNI-like domain (338 in *Arabidopsis*)	Metabolism (16)	No correlation with the number of different cell types	160
26	Major expansions in vertebrates and plants; minor expansions in the other animals; absent or infrequent in fungi/ protozoa	Homeo-domain (351 in *Arabidopsis*, 275 in humans); ARM repeat (301 in *Arabidopsis*, 258 in humans); RNA-binding domain (268 in *Arabidopsis*, 258 in humans); WD40-repeat (261 in *Arabidopsis*, 278 in humans)	General or multiple function (8); regulation (5) (e.g., DNA-binding)	Both the expansion pattern and composition of domain function make this group intermediate to vertebrate-specific (Node 158) and plant-specific expansions (Node 160)	156
13	Rice-specific expansions; moderate numbers in cress and vertebrates	Protein-kinase (1,744 in rice, 523 in humans); L domain (938 in rice, 195 in humans); RING/U-box domain (609 in rice, 285 in humans)	Intracellular processes (6)	Similar to vertebrate-/ plant-specific expansions (Node 156)	154
7	Nematode-specific expansions; moderate in vertebrates, less in the other animals	Glucocorticoid receptor DNA binding domain (315 in *C .elegans*); nuclear receptor ligand-binding domain (278 in *C. elegans*); C-type lectin-like domain (275 in *C. elegans*)	Regulation (3) (e.g., receptors); intracellular processes (3)	Nematode-specific expansions of nuclear receptors have been described before, e.g. [40], but the reasons for these are unknown.	119
6	Specific to *E. histolytica* and vertebrates; minor expansions in the other animals	DBL homology domain (108 in *Entamoeba*); GTPase activation domain, GAP (120 in *Entamoeba*)	Regulation (6) (e.g., signal transduction)	Similar to vertebrate-specific expansions (Node 158)	155
5	Rodent-specific expansions	Glyceraldehyde-3-phosphate dehydrogenase (152 in mice); L30e-like domain (99 in mice)	Other (3) (e.g.. viral proteins)	Possibly due to contamination of the rodent genomes with bacterial/viral sequences	94
5	Mammalia-specific expansions, moderate in other vertebrates, absent in plants/fungi/ protozoa	Multiheme cytochromes (307 in humans); KRAB domain (307 in humans)	Extracellular processes (3) (e.g. immune response)	Similar to vertebrate-specific expansions (Node 158)	140
4	Major expansions in rodents (mouse and rat), moderate in vertebrates/ plant, less in the other animals	HMG-box (125 in mice, 63 in humans)	Regulation (2) (e.g., DNA- or RNA-binding); intracellular processes (2) (e.g., proteases)	Similar to vertebrate-/ plant-specific expansions (Node 156) but with additional expansions in rodents	152

The table lists, for the set of the 299 largest superfamilies, the clusters of similar abundance profiles (see also Figure 4). The *R*-value cutoff is 0.90, and the node number refers to the respective node in the hierarchical clustering (see http://polaris.icmb.utexas.edu/people/cvogel/HV). The table lists the number of superfamily members in the cluster, examples of superfamilies, main functions, other comments on the particular cluster, and the node number as taken from the hierarchical clustering.

[a] Number of domain superfamilies in the cluster.

[b] Node number/name of the cluster; see http://polaris.icmb.utexas.edu/people/cvogel/HV.

DOI: 10.1371/journal.pcbi.0020048.t004

result in complex intra- and intercellular signalling and regulatory networks. Further, the expansion of some protein families correlates with the emergence of animal- or vertebrate-specific traits, such as the immune system. Finally, some families (e.g., RNA-binding domains) may support regulation of alternative splicing that plays a significant role in humans [42], and, together with other post-transcriptional and -translational modifications, further increase the complexity of vertebrate proteomes.

In general, plants and animals went separate paths with large, kingdom-specific expansions. We observed, however, some plant-specific expansions, which also occurred in vertebrates, but were not as pronounced in the other animals. These patterns may correlate with macroscopic features yet to be identified. For example, future work may include measures such as the number of cells, the body mass, or even population size [43] in a description of organismal complexity.

Finally, we hope that our analysis provides a framework for more detailed studies of family expansions. One example for such studies is domains of the immunoglobulin superfamily that expanded in number in *Drosophila* as compared to *C. elegans.* Most of the fly-specific proteins have been shown to be cell-surface receptors and cell-adhesion molecules that are involved in axon pathfinding during the embryonal development of the nervous system [15,44,45]—this illustrates one of the factors that allows the fly to have a more complex nervous system than the worm.

Materials and Methods

Datasets. The 38 eukaryotic genomes used in our analysis are listed in Figure 1A, and our analysis is outlined in Figure 1B. The gene predictions and domain assignments to the gene predictions were taken from the SUPERFAMILY database version 1.69, updated in September 2005 [18]; information on genome versions and source can be found at http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY. The domain superfamilies are defined in the SCOP database [21], and our analysis focuses on the seven well-defined classes *a* to *g*, respectively. All domains within a SCOP superfamily are related and can be regarded as descendants from one common ancestral domain. The *Arabidopsis* and animal genomes were made nonredundant with respect to predicted splice variants: for each gene only the longest transcript was included. Information on predicted splice variants was unavailable for the fungal genomes (except for *S. cerevisiae*), *Oryza sativa,* and the protozoa. Information on the estimated number of different cell types was taken from literature [28,29] and is detailed in Dataset S1.

Correlations between superfamilies and the number of different cell types. The abundance of a domain superfamily in each genome was measured as the number of proteins with at least one predicted hit of the respective superfamily (Figure 1 and Dataset S1). Many domain superfamilies occur in only one or two genomes and in only a few proteins. For each superfamily, changes in abundance across different eukaryotes can be described in an abundance pattern or profile. The abundance counts for one superfamily across different genomes were normalised according to *A__n = (A__i – A__avg) / A__sdv,* where *A__i* and *A__n* are the absolute and normalised abundance count in a particular genome, respectively, and *A__avg* and *A__sdv* are the average abundance and standard deviation across all genomes for that superfamily, respectively. This means the abundance of a superfamily in one genome is described relative to its abundance in other genomes.

Similar to what has been done for gene expression data in other studies, each expansion pattern was colour-coded, using the matrix2png [46] and treeview programs (http://rana.lbl.gov/EisenSoftware.htm) for visualisation. In Figure 4, each row depicts the profile for one superfamily: blue denotes high, and white denotes low relative abundance. The genomes are arranged in the same order as in Figure 1A, and the rows (superfamilies) are hierarchically clustered using the XCluster software (http://genetics.stanford.edu/~sherlock/cluster.html). A cutoff of $R \geq 0.90$ results in clusters of highly similar expansion patterns with largest clusters indicated in red (Figure 4). The clusters are also described in Table 4 and at http://polaris.icmb.utexas.edu/people/cvogel/HV.

Similar to correlating the abundance profiles of superfamilies with each other, we also correlated them with the number of different cell types per organism. In an extension of what we published previously [38], we assigned each domain superfamily to one of 50 small functional categories (see Protocol S1). Each of the small categories maps to one of seven larger functional categories.

Supporting Information

Dataset S1. Characteristics of the 38 Genomes and the 1,219 Superfamilies

The spreadsheet genome__characteristics lists the 38 genomes used in our analysis, as taken from SUPERFAMILY version 1.69 [18], http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY. Please refer to the SUPERFAMILY database for further information on the origin of the genome sequences. The two-letter abbreviation for each genome is the one used in the SUPERFAMLY MySQL database. The three-letter abbreviation for each genome is used in our paper. The information on the total number of genes (non-redundant in terms of splice variants), average gene length, and the total number of domains predicted by SUPERFAMILY (domains of all classes) were taken directly from the SUPERFAMILY database [18].

The estimated number of different cell types are taken from the publication by Valentine et al. [28] and Hedges et al. [29]. The average of these values represents the estimated number of different cell types used in this analysis.

The spreadsheet superfamily__data contains information on the abundance of the 1,219 superfamilies in 38 genomes. The superfamilies are annotated in terms of their general and more detailed type of function, their identifier used in the SCOP [21] and in the SUPERFAMILY [18] database, and their correlation with the estimated number of different cell types.

Found at DOI: 10.1371/journal.pcbi.0020048.sd001 (982 KB XLS).

Figure S1. Distributions of Domain Functions

(A) Distribution of functions in terms of domain superfamilies defined in SCOP [21]. Domain superfamilies of metabolism (e.g., enzymes) are the most abundant category. (B) shows the distribution of superfamilies across the function categories; this distribution is similar for all genomes, five of which are shown. This means that invention of domain superfamilies specific to some genomes did not significantly change the overall composition in terms of function. This is different when taking gene duplication into account (C): the composition in terms of domain functions varies within the five genomes shown. While the largest category in plant is metabolism, in human it is regulation.

Previous work reported a linear relationship between genome size and the number of metabolic proteins for bacteria and eukaryotes [8,9]. Such a linear relationship would result in a constant fraction of metabolic domains across genomes, but this is not what we observe when comparing five different eukaryotes (D): the fraction of domains in metabolism is lower in invertebrates and vertebrates (fly and human) than in the other organisms. These differences observed may be due to different datasets (domains used instead of whole proteins) and different function annotation procedures.

Abbreviations are as in Figure 1.

Found at DOI: 10.1371/journal.pcbi.0020048.sg001 (46 KB PDF).

Figure S2. Expansion Profiles of all 1,219 Superfamilies

Similar to Figure 4, the matrix displays the relative abundance profiles for each of the 1219 superfamilies (rows) in the 38 genomes (columns) in a colour-coded format. Blue denotes high, and white denotes low relative domain abundance in some organisms as compared to others. As for the subset of 299 largest superfamilies (Figure 4), three major trends become apparent: expansions specific to vertebrates, expansions specific to plants, and expansions that occur in plants and vertebrates.

Abbreviations are as in Figure 1.

Found at DOI: 10.1371/journal.pcbi.0020048.sg002 (678 KB TIF).

Figure S3. Relationship between the Number of Different Cell Types, Total Number of Domain Superfamilies, Total Number of Domains per Genome, and Sequence Length

The number of different cell types is only weakly correlated with the number of different domain superfamilies found ($R^2 = 0.52$, [A]), the total number of genes predicted for an organism ($R^2 = 0.54$, Figure 1A), and with the total number of domains ($R^2 = 0.59$, [B]). Part of the latter correlation can be explained by the fact that more domains are known and assigned to vertebrates than to protists and plants. There are no large differences in the average sequence length of fungi, protists, plants, or vertebrates ($R^2 = 0.02$, [C]). Thus, the higher number of domains in some organisms as compared to others must largely arise from duplication of whole genes rather than the addition of domains to existing proteins.

The number of different domain superfamilies can be taken as a measure of invention of novel families in an organism, while the total number of domains is a measure of duplication. Thus, duplication correlates better than invention with increases in biological complexity as measured in the number of different cell types, and may have been one of the driving forces behind the emergence of novel cell types.

Abbreviations are as in Figure 1.

Found at DOI: 10.1371/journal.pcbi.0020048.sg003 (64 KB PDF).

Protocol S1. Notes on Domain Function Annotation and Clustering Procedure

Found at DOI: 10.1371/journal.pcbi.0020048.sd002 (144 KB PDF).

Table S1. Summary of Key Terms Used in the Paper

Found at DOI: 10.1371/journal.pcbi.0020048.st001 (109 KB DOC).

Table S2. Groups of Domain Function and the Number of Different Cell Types

Domains in the function categories are non-overlapping subsets of all domains in each organism. Only two function categories (i.e., domains of extracellular processes and regulation) show very good correlation of domain abundance and the number of different cell types (i.e. $R \geq 0.80$).

Found at DOI: 10.1371/journal.pcbi.0020048.st002 (110 KB DOC).

Additional supporting material can be found at http://polaris.icmb.utexas.edu/people/cvogel/HV. The Web site contains several files: (1) mapping of the 50 more detailed function categories to the seven main function categories; and (2) names, SCOP identifiers, and SUPERFAMILY identifiers of all SCOP superfamilies, v. 1.69 [18]. The Web site also has links to additional Web pages, which display clusterings of superfamily expansion profiles using different cutoffs. Each of the Web pages describes clusters of similar expansion profiles, and the number and size of the clusters depends on the cutoffs used and the distribution of domain functions. Each cluster is labelled with a unique node number, and this number is taken directly from output of the XCluster program at http://genetics.stanford.edu/~sherlock/cluster.html.

Acknowledgments

We thank Siarhei Maslau, Jose Pereira-Leal, Matthew Bashton, Kevin Hannay, Emmanuel Levy, Madan Babu Mohan, Sarah Teichmann, Mark Tsechansky, and Claudia Voelckel for commenting on the work and manuscript.

Author contributions. CV and CC conceived and designed the experiments. CV performed the experiments. CV and CC analyzed the data. CV and CC wrote the paper.

Funding. CV acknowledges funding by the Boehringer Ingelheim Fonds, the Medical Research Council, and the Human Frontier of Science Program.

Competing interests. The authors have declared that no competing interests exist. ■

References

1. Rossmann MG, Moras D, Olsen KW (1974) Chemical and biological evolution of nucleotide-binding protein. Nature 250: 194–199.
2. Perutz MF, Kendrew JC, Watson HC (1965) Structure and function of haemoglobin: II. Some relations between polypeptide chain configuration and amino acid sequence. J Mol Biol 13: 669–678.
3. Patthy L (1985) Evolution of the proteases of blood coagulation and fibrinolysis by assembly from modules. Cell 41: 657–663.
4. Williams AF, Barclay AN (1988) The immunoglobulin superfamily—Domains for cell-surface recognition. Annu Rev Immunol 6: 381–405.
5. Bork P (1991) Shuffled domains in extracellular proteins. FEBS Lett 286: 47–54.
6. Chothia C (1992) Proteins—1,000 families for the molecular biologist. Nature 357: 543–544.
7. Kirschner M, Gerhart J (1998) Evolvability. Proc Natl Acad Sci U S A 95: 8420–8427.
8. van Nimwegen E (2003) Scaling laws in the functional content of genomes. Trends Genet 19: 479–484.
9. Ranea JA, Buchan DW, Thornton JM, Orengo CA (2004) Evolution of protein superfamilies and bacterial genome size. J Mol Biol 336: 871–887.
10. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, et al. (1998) Comparison of the complete protein sets of worm and yeast: Orthology and divergence. Science 282: 2022–2028.
11. Aravind L, Subramanian G (1999) Origin of multicellular eukaryotes—Insights from proteome comparisons. Curr Opin Genet Dev 9: 688–694.
12. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, et al. (2000) Comparative genomics of the eukaryotes. Science 287: 2204–2215.
13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860–921.
14. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, et al. (2006) Ensembl 2006. Nucleic Acids Res 34: D556–D561.
15. Vogel C, Teichmann SA, Chothia C (2003) The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity. Development 130: 6317–6328.
16. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540.
17. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903–919.
18. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J (2004) The SUPERFAMILY database in 2004: Additions and improvements. Nucleic Acids Res 32: D235–D239.
19. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA (2004) Supra-domains—Evolutionary units larger than single protein domains. J Mol Biol 336: 809–823.
20. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4: 41.
21. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, et al. (2004) SCOP database in 2004: Refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226–D229.
22. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, et al. (2006) Pfam: Clans, web tools and services. Nucleic Acids Res 34: D247–D251.
23. Boeckmann B, Blatter MC, Famiglietti L, Hinz U, Lane L, et al. (2005) Protein variety and functional diversity: Swiss-Prot annotation in its biological context. C R Biol 328: 882–899.
24. Schmid PE, Tokeshi M, Schmid-Araya JM (2000) Relation between population density and body size in stream communities. Science 289: 1557–1560.
25. Enquist BJ, Niklas KJ (2001) Invariant scaling relations across tree-dominated communities. Nature 410: 655–660.
26. Carbone C, Gittleman JL (2002) A common rule for the scaling of carnivore density. Science 295: 2273–2276.
27. Finlay BJ (2002) Global dispersal of free-living microbial eukaryote species. Science 296: 1061–1063.
28. Valentine JW, Collins AG, Meyer CP (1994) Morphological complexity increase in metazoans. Paleobiology 20: 131–142.
29. Hedges SB, Blair JE, Venturi ML, Shoe JL (2004) A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 4: 2.
30. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2005) InterPro, progress and status in 2005. Nucleic Acids Res 33: D201–D205.
31. Matzke MA, Matzke AJ (2004) Planting the seeds of a new paradigm. PLoS Biol 2: E133.
32. Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 12: 1048–1059.
33. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. Cell 101: 573–576.
34. Copley RR, Schultz J, Ponting CP, Bork P (1999) Protein families in multicellular organisms. Curr Opin Struct Biol 9: 408–415.
35. Bard J, Rhee SY, Ashburner M (2005) An ontology for cell types. Genome Biol 6: R21.
36. Lesk AM, Chothia CH (1986) The response of protein structures to amino-acid-sequence changes. Philos Transact A Math Phys Eng Sci 317: 345–356.
37. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310: 311–325.
38. Vogel C, Teichmann SA, Pereira-Leal JB (2005) The relationship between domain duplication and recombination. J Mol Biol 346: 355–365.
39. Li WH, Gu Z, Wang H, Nekrutenko A (2001) Evolutionary analyses of the human genome. Nature 409: 847–849.
40. Sluder AE, Maina CV (2001) Nuclear receptors in nematodes: themes and variations. Trends Genet 17: 206–213.
41. Szathmary E, Jordan F, Pal C (2001) Molecular biology and evolution. Can genes explain biological complexity? Science 292: 1315–1316.
42. Brett D, Pospisil H, Valcarcel J, Reich J, Bork P (2002) Alternative splicing and genome complexity. Nat Genet 30: 29–30.
43. Lynch M, Conery J (2003) The origins of genome complexity. Science 302: 1401–1404.
44. Hutter H, Vogel BE, Plenefisch JD, Norris CR, Proenca RB, et al. (2000) Cell biology: Conservation and novelty in the evolution of cell adhesion and extracellular matrix genes. Science 287: 989–994.
45. Hynes RO, Zhao Q (2000) The evolution of cell adhesion. J Cell Biol 150: F89–F96.
46. Pavlidis P, Noble WS (2003) Matrix2png: A utility for visualizing matrix data. Bioinformatics 19: 295–296.