Evolution of protein domain promiscuity in eukaryotes

Malay Kumar Basu, Liran Carmel, Igor B. Rogozin and Eugene V. Koonin

Genome Res. 2008 18: 449-461; originally published online Jan 29, 2008;
Access the most recent version at doi:10.1101/gr.6943508

Letter

Evolution of protein domain promiscuity in eukaryotes

Malay Kumar Basu, Liran Carmel, Igor B. Rogozin, and Eugene V. Koonin[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

Numerous eukaryotic proteins contain multiple domains. Certain domains show a tendency to occur in diverse domain architectures and can be considered "promiscuous." These promiscuous domains are, typically, involved in protein–protein interactions and play crucial roles in interaction networks, particularly those that contribute to signal transduction. A systematic comparative-genomic analysis of promiscuous domains in eukaryotes is described. Two quantitative measures of domain promiscuity are introduced and applied to the analysis of 28 genomes of diverse eukaryotes. Altogether, 215 domains are identified as strongly promiscuous. The fraction of promiscuous domains in animals is shown to be significantly greater than that in fungi or plants. Evolutionary reconstructions indicate that domain promiscuity is a volatile, relatively fast-changing feature of eukaryotic proteins, with few domains remaining promiscuous throughout the evolution of eukaryotes. Some domains appear to have attained promiscuity independently in different lineages, for example, animals and plants. It is proposed that promiscuous domains persist within a relatively small pool of evolutionarily stable domain combinations from which numerous rare architectures emerge during evolution. Domain promiscuity positively correlates with the number of experimentally detected domain interactions and with the strength of purifying selection affecting a domain. Thus, evolution of promiscuous domains seems to be constrained by the diversity of their interaction partners. The set of promiscuous domains is enriched for domains mediating protein–protein interactions that are involved in various forms of signal transduction, especially in the ubiquitin system and in chromatin. Thus, a limited repertoire of promiscuous domains makes a major contribution to the diversity and evolvability of eukaryotic proteomes and signaling networks.

[Supplemental material is available online at www.genome.org.]

Large proteins typically contain multiple domains, either with the same or with different structural folds (Doolittle 1995; Vogel et al. 2004; Orengo and Thornton 2005; Fong et al. 2007; Han et al. 2007). Some of these domain combinations are stable during evolution whereas others are more labile. Accordingly, domains differ in their tendency to appear in variable multidomain contexts, with some being "promiscuous," i.e., combining with many other domains (Marcotte et al. 1999). Combination of domains with different structures and functions within multidomain proteins is a major mode of creation and modulation of molecular functionality, especially for signal transduction. The common modes of action of promiscuous domains involve connecting components of signal transduction networks through specific protein–protein interactions and delivering effectors to the sites of their action, in particular, the chromatin (Chervitz et al. 1998; Hofmann 1999; Aravind et al. 2001; Patthy 2003; Templeton et al. 2004). In addition, mobile small-molecule-binding domains provide for the allosteric regulation of the activities of diverse enzymes by the same ligands and feeding intracellular and environmental cues into signal transduction pathways (Anantharaman et al. 2001). On some occasions, the functions of individual domains combine in a multidomain protein to yield a novel function (Bashton and Chothia 2007).

It appears likely that the increase in the complexity of domain organization of proteins would substantially contribute to the evolution of organismal complexity owing to the increased potential for interactions and formation of signal transduction pathways (Koonin et al. 2000, 2002; Patthy 2003; Tordai et al. 2005; Itoh et al. 2007). Analyses of the links between multidomain organization of proteins and organismal complexity yielded somewhat ambiguous results. It has been shown that the frequency of occurrence of proteins with an increasing number of distinct domains (single-domain, two-domain, three-domain, etc., proteins) follows an exponential decay law, which is compatible with a model of random, nonselective domain recombination (Wolf et al. 1999). The substantial randomness of domain recombination during evolution has been independently supported by the demonstration of a positive correlation between the abundance of a domain and the number of multidomain combinations in which it is involved (Vogel et al. 2005). However, the slope of the domain number distributions decreased in the three superkingdoms of life, in the order archaea > bacteria > eukaryotes, indicating that the likelihood of the formation of a multidomain protein was greater in eukaryotes than in prokaryotes and suggesting a link between the abundance of multidomain proteins and biological complexity (Koonin et al. 2002). Concordantly, comparative analyses of multidomain proteins in archaea, bacteria, and eukaryotes have revealed a substantially greater fraction of multidomain proteins in the more complex eukaryotic organisms (Apic et al. 2001; Wang and Caetano-Anolles 2006). Similar conclusions have been reached through the analysis of domain co-occurrence networks, namely, that more complex organisms displayed greater connectivity of

[1]Corresponding author.
E-mail koonin@ncbi.nlm.nih.gov; fax (301) 435-7793.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6943508. Freely available online through the *Genome Research* Open Access option.

the co-occurrence graph (Wuchty 2001; Ye and Godzik 2004). Furthermore, it has been noticed, first anecdotally and then systematically, that a phenomenon dubbed "domain accretion" occurs during evolution of some orthologous sets of eukaryotic genes: proteins of complex life forms, in particular, animals, accrete additional domains compared to orthologs from simpler eukaryotes (Koonin et al. 2000, 2004). A detailed survey by Pathy and coworkers clearly supported the notion that the genomes of more complex organisms, in particular, animals, encoded a greater fraction of multidomain proteins than the genomes of simpler eukaryotes and, especially, prokaryotes (Tordai et al. 2005).

We were interested in the evolutionary aspects of domain promiscuity, that is, the propensity of protein domains to combine with different other domains in multidomain proteins (Marcotte et al. 1999). We sought to develop an objective, quantitative measure of domain promiscuity and to apply this measure to compare the sets of promiscuous domains from different eukaryotes and to reconstruct the evolution of domain promiscuity. We further attempted to identify functional and evolutionary correlates of domain promiscuity.

Results and Discussion

Domains and domain combinations in eukaryotic genomes

Domains were identified in the protein sequences from 28 genomes of diverse eukaryotes (Supplemental Table S1), representing most of the major eukaryotic lineages, using the RPS-BLAST program and a collection of position-specific scoring matrices (PSSMs) from the Conserved Domain Database (CDD), which combines domain data from the SMART and Pfam databases (see Methods for details; Marchler-Bauer et al. 2005). There is a monotonic increase in the number of detectable domains with the increase in the apparent organismal complexity, from ~800–1700 detectable domain types in unicellular eukaryotes to almost 3000 domain types in vertebrates (Fig. 1A). We then identified all pairs of domains that are neighbors on a protein sequence ("bigrams," a standard term for pairs of adjacent words in computational linguistics) (Manning and Schütze 1999) in each of the genomes (see Supplemental Fig. S1). The number of bigrams shows a considerably steeper growth with organismal complexity than the total number of domains (Fig. 1B). There are substantially fewer bigrams than domains in unicellular eukaryotes, roughly the same number of domains and bigrams in plants, and more bigrams than domains in animals (Fig. 1A), emphasizing the previously noticed trend toward domain accretion in multicellular organisms, especially, animals (Chervitz et al. 1998; Koonin et al. 2000, 2004; Apic et al. 2001; Tordai et al. 2005; Itoh et al. 2007).

When the number of bigrams was plotted against the corresponding domain count (the number of domains that are involved in 0, 1, 2, etc., bigram types) separately for each analyzed species, each distribution closely followed a power-law (Fig. 2; Supplemental Table S2). The degrees of the distributions were statistically indistinguishable not only within each lineage but even between pairs of species that drastically differ in organismal complexity, in particular, animals versus unicellular eukaryotes. However, when combined distributions were compared, the degree of the bigram frequency distribution for animals was significantly smaller than that for protists, fungi, and even plants (Supplemental Table S2). The degree of the distribution decreased from ~2.5 in protists and fungi to ~1.9 in animals. Thus, in animals, a greater number of domains gather higher bigram counts, producing a fat tail and resulting in a decrease in the slope of the distribution (Fig. 2). This appears to reflect the significant tendency toward domain accretion in animals (Chervitz et al. 1998; Koonin et al. 2000, 2004; Tordai et al. 2005).

Promiscuous domains

As a quantitative measure of domain promiscuity, we used the weighted bigram frequency (derived from the Kullback-Leibler information gain formula):

$$\pi_i = \beta_i \times \log\left(\frac{\beta_i}{f_i}\right)$$

Here, β_i is the bigram frequency:

$$\beta_i = \frac{T_i}{\frac{1}{2}\sum_{j=1}^{t} T_j}$$

where t is the number of distinct domain types, T_i is the number of unique domain neighbors of domain i, and f_i is the frequency of domain i in the genome, calculated as n_i/N, where n_i is the total count of domain i, and N is the total number of domains detected in the given genome:

$$N = \sum_{i=1}^{t} n_i.$$

This formula was chosen to normalize the number of bigrams over the abundance of a given domain, in order to weight against the more abundant domains that would otherwise produce a substantial fraction of bigrams. This weighting scheme is based on the assumption that the formation of multidomain proteins is a random process. Although this hardly can be true of each particular domain combination, previous analyses have shown that the distribution of the number of domains in proteins does not dramatically deviate from the predictions of the stochastic null model (Wolf et al. 1999; Koonin et al. 2002). Furthermore, it has been shown that the number of domain combinations in which a given domain is involved is proportional to the domain's abundance (Vogel et al. 2005). Accordingly, normalization over abundance is a logically straightforward approach to detect domains that are more prone to form diverse domain combinations than expected by chance, that is, can be appropriately classified as promiscuous.

If the promiscuity value of a singleton, a domain present only once in the genome and having only one bigram type, is taken as the cutoff—that is, all domains with π values greater than that of a singleton were considered promiscuous—then there were 1089 promiscuous domains in the analyzed eukaryotic genomes taken together. This definition is quite liberal because many domains with π values greater than that of a singleton have only a few bigram types and, intuitively, do not appear to be particularly promiscuous. Therefore, we also developed a stringent criterion of domain promiscuity that is based on the assumption that the expected frequency of domain combinations in a "random" genome follows the Poisson distribution. A significant deviation from a single Poisson distribution can be represented as a mixture of two or more Poisson distributions. Separation of mixtures of standard distributions is a common problem in computational biology, and several algorithms have

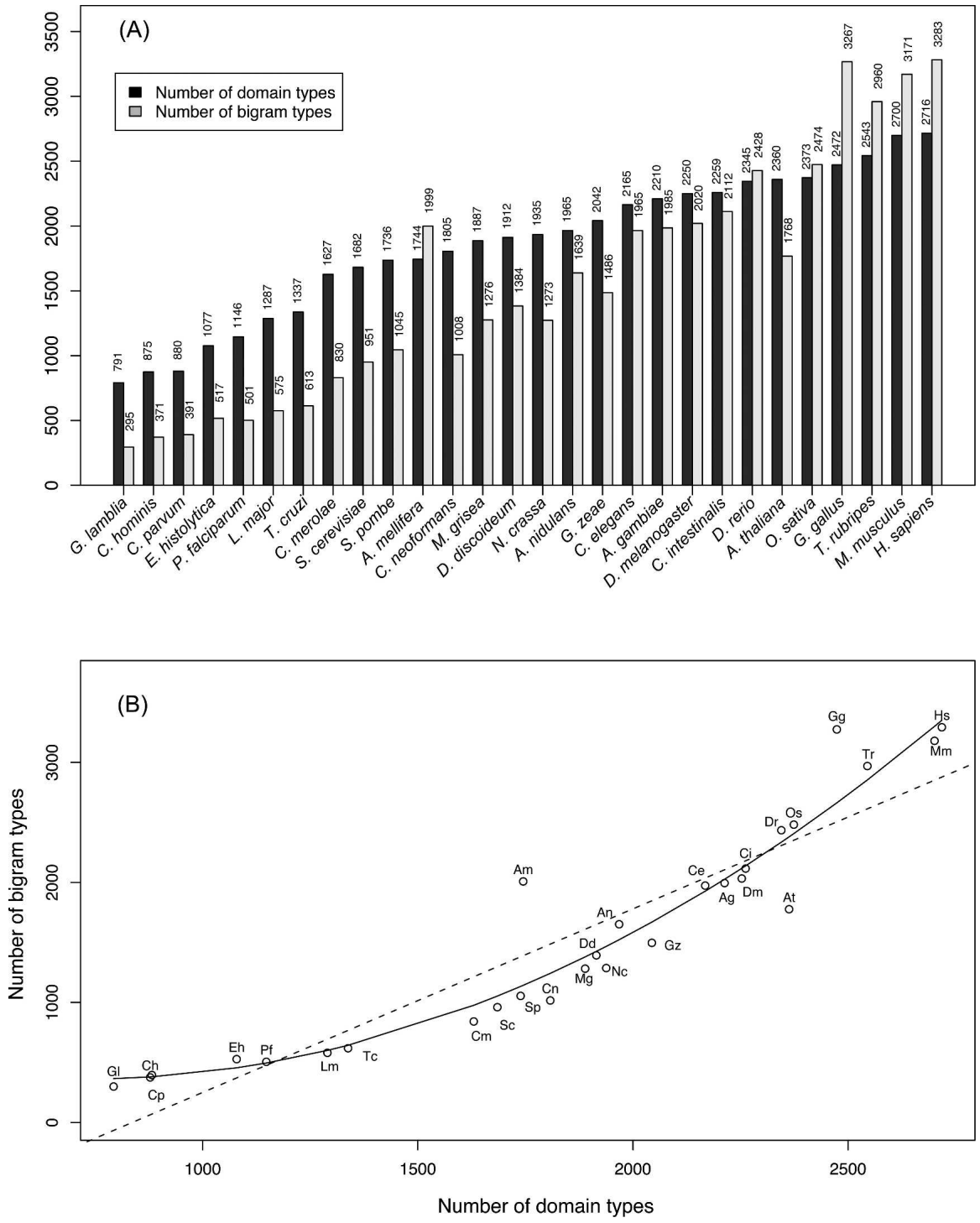

Figure 1. (*A*) The counts of distinct domain types and distinct bigram types in the analyzed species. (*B*) The dependence of the number of bigrams types on the number of domain types encoded in a genome. The linear (dotted) and quadratic (solid) regression lines are shown. The quadratic function is a better fit than the linear function (Pearson's product-moment correlation: 0.92; *P*-value ~ 0.005). Each point is labeled with the species abbreviations as described in Methods.

been developed to solve it (Fickett and Guigo 1993; Bohning et al. 1998; Glazko et al. 1998).

We used the C.A.MAN program (Bohning et al. 1998) to analyze the frequency distributions of domain combinations. For each genome, the distribution of the raw numbers of unique bigrams for all domains identified as promiscuous by the liberal criterion was decomposed into at least two Poisson distributions. The class of domains with the largest mean of the Poisson distribution was considered promiscuous (Supplemental Table S3). This class included 215 highly promiscuous domains (see the

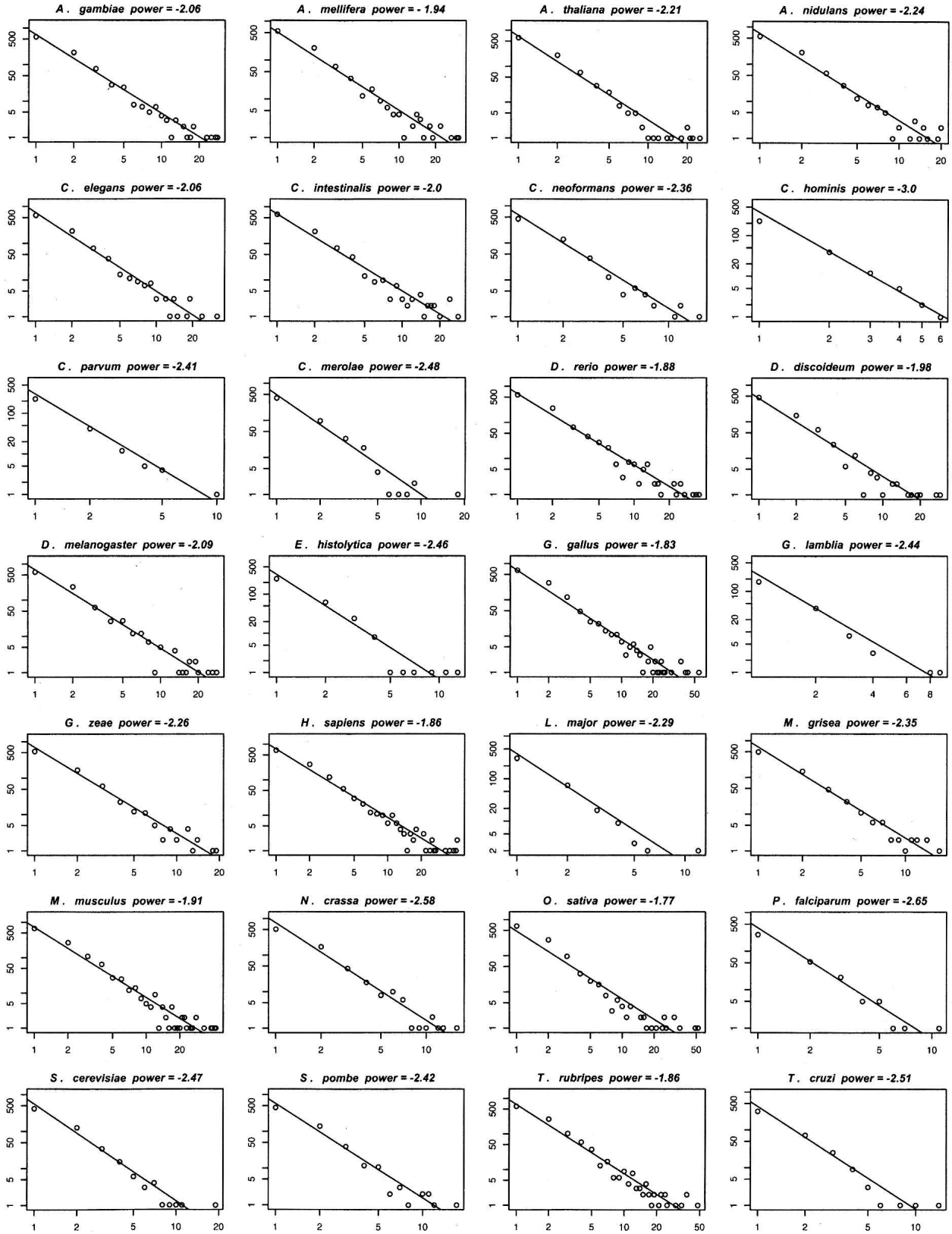

Figure 2. Power law distributions of bigram frequencies in 28 eukaryotes. The linear regression is shown on each plot. Each panel shows log-log plots of the count of bigram types on the *X*-axis and the domain count (number of domains participating in that many bigram types as *X* coordinate) on the *Y*-axis. The species name and the power of the regression line are shown at the *top* of each plot.

complete list of these domains in Supplemental Table S4) that, obviously, comprise a subset of the 1089 domains identified with the liberal criterion.

The normalization procedure described above defines promiscuity of protein domains not as the sheer number or frequency of unique domain combinations (bigrams) in which a domain is involved, but as a function of both this number and the overall abundance of the respective domain, such that high-abundance domains are down-weighted. As a result, it is possible for a domain that forms a substantial number of distinct combinations to be considered non-promiscuous owing to its overall high abundance, a result that potentially could be construed as counterintuitive. To assess the magnitude of this effect, we compared the lists of domains in each species ranked by the weighted bigram frequency (π) values with two alternative rankings, one that employed π calculated for the occurrences of domains in multidomain proteins only (i.e., after removing all single-domain proteins) and another that ranked domains by the raw bigram frequency (β) values that reflect the share of the bigrams containing the given domain among all unique bigrams in a particular species. For all species, the two alternative lists of promiscuous domains strongly, positively correlated with the original list obtained after normalization (ranking correlation coefficients >0.75 for all 28 species; Supplemental Tables S5 and S6), indicating that most of the domains that form numerous bigrams were, indeed, labeled promiscuous, the normalization over abundance notwithstanding.

The taxonomic distribution of promiscuous domains and the excess of promiscuous domains in animals

There was a steep increase in the number of promiscuous domains with increasing organismal complexity (Fig. 3A) and a strong linear dependence between the number of promiscuous domains and the number of domain types (Fig. 3B). Among the 215 promiscuous domains identified with the strict criterion, 147 domains were identified in animals, 81 in plants, and 58 in fungi, with 25 domains present in all three kingdoms (Fig. 4; Table 1). The number of these "universal" promiscuous domains significantly exceeded the random expectation (P-value = 4×10^{-4}, calculated using χ^2 with the background frequency determined by Monte Carlo simulation with 10,000 replicates). The fraction of promiscuous domains in animals (~4.9%) was significantly greater (P-value < 0.01 by the Fisher's exact test) than in fungi (~3.3%) or plants (~3.5%), whereas the latter two values were statistically indistinguishable. The present estimate of promiscuous domains in vertebrates is conservative in that we did not take into account alternative splicing. An analysis of all splice isoforms could reveal an even greater excess of promiscuous domains.

A tree of eukaryotes inferred from a comparison of domain promiscuities

Various features of genomes beyond sequence per se, such as gene composition and gene order, have been used to construct "genome trees" that, generally, combine the phylogenetic signal

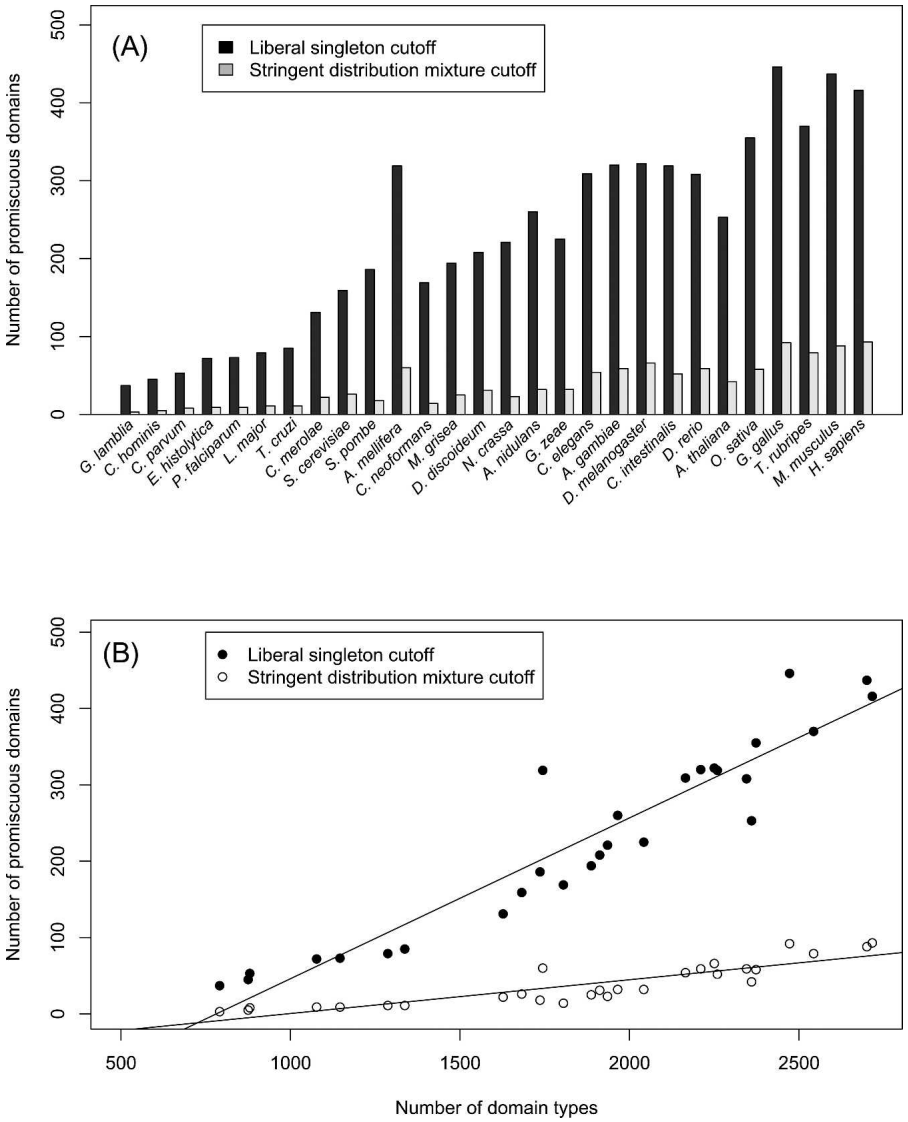

Figure 3. Distribution of promiscuous domains in eukaryotes. (*A*) Promiscuous domains in the analyzed eukaryotic species. (Black bars) Promiscuous domains defined using weighted bigram frequency with the cutoff determined by the liberal singleton method; (gray bars) promiscuous domains defined using the strict distribution mixture criterion (see text for details). (*B*) The number of promiscuous domains (on the *Y*-axis) increases with the number of unique domain types (on the *X*-axis). (Black circles) Promiscuous domains determined by the liberal singleton cutoff method (Pearson's correlation 0.94, P-value 4.4×10^{-14}); (empty circles) promiscuous domains determined with the strict distribution mixture criterion (Pearson's correlation 0.88, P-value 4.6×10^{-10}).

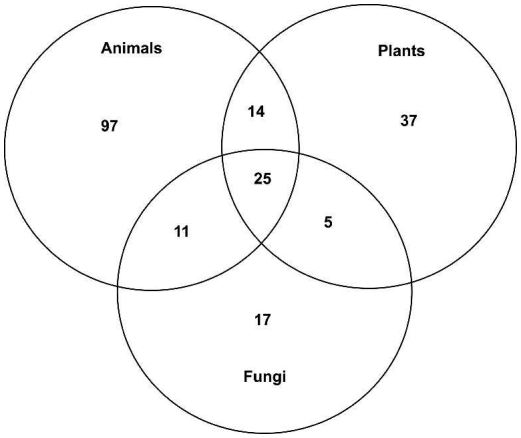

Figure 4. Distribution of promiscuous domains in animals, plants, and fungi. The overlap exceeds the random expectation with a *P*-value of 9.9×10^{-5} (χ^2 with the background probability calculated using the Monte-Carlo method).

with signals reflecting the lifestyles of the compared organisms, for example, parallel gene loss in parasites (Snel et al. 1999, 2005; Wolf et al. 2002, 2004; Wang and Caetano-Anolles 2006). In particular, protein domain combinations have been used as a phylogenetic character to address hard phylogenetic problems, such as the Coelomata–Ecdysozoa conundrum in the evolution of animals (Wolf et al. 2004; Wang and Caetano-Anolles 2006). Here, we used the domain promiscuity profiles of the analyzed eukaryotic species to generate a new variant of a genome-tree. If species A has domains $(A_1,A_2,A_3\cdots A_n)$ with π values $(P_1,P_2,P_3\cdots P_n)$ and species B has domains $(B_1,B_2,B_3\cdots B_n)$ with π values $(Q_1,Q_2,Q_3\cdots Q_n)$, then the similarity value of these two species is calculated using the angular separation method:

$$Sim(A,B) = \frac{\sum_{i=1}^{n} (P_i \times Q_i)}{\sqrt{\sum_{i=1}^{n} (P_i)^2 \times \sum_{i=1}^{n} (Q_i)^2}}$$

When a domain was absent from a given genome, a π value of 0 was assigned (Webb 2002). The distance between these two species, then, was defined as

$$Dis(A,B) = 1 - Sim(A,B)$$

The calculated distances were then used to construct a neighbor-joining tree (Fig. 5). The resulting tree retained most of the major eukaryotic clades, including the animal–fungal clade. Depending on the root position, the tree topology could be viewed as compatible with either the "crown-group" topology, under which several lineages of unicellular eukaryotes are basal to the "crown group" that includes animals, fungi, and plants, along with some unicellular forms (Hedges 2002; Templeton et al. 2004); or the unikont–bikont tree, where the root is between the animal–fungal and plant lineages (Stechmann and Cavalier-

Table 1. The 25 promiscuous domains shared by animals, plants, and fungi

Domain	SMART/Pfam ID	No. of genomes where domain is promiscuous	Function/comments
RING	smart00184	19	Ubiquitin signaling: E3 component of ubiquitin ligases
AAA	smart00382	19	ATPase involved in various functions, including chaperone roles and various forms of signal transduction
UCH	pfam00443	18	Ubiquitin signaling: Ubiquitin C-terminal hydrolase
PH	smart00233	18	Protein–protein interactions; various signaling processes, in particular, inositol phosphate signaling
PHD	smart00249	17	Protein–protein interactions, primarily, in chromatin
SET	smart00317	17	Methyltransferase methylating histones and other chromatin-associated proteins
ANK	smart00248	15	Diverse protein–protein interactions, signaling
UBQ	smart00213	15	Ubiquitin signaling: Ubiquitin and homologous domains
C2	smart00239	15	Phospholipid and inositol phosphate binding, protein–protein interactions; lipid-related signaling
BROMO	smart00297	15	Acetyl-lysine-binding, binds to acetylated histone tails, modulator of chromatin structure
Biotin_lipoyl	pfam00364	14	Coenzyme-binding domain of various metabolic enzymes
MYSc	smart00242	13	ATPase domain of myosins, combines with a variety of tail domains
S_TKc	smart00220	13	Serine-threonine protein kinase
DEXDc	smart00487	13	C-terminal domain of superfamily 2 helicases: Extremely diverse functions in regulation of translation, transcription, repair
DnaJ	smart00271	12	Protein–protein interactions, various chaperone functions
BRCT	smart00292	12	Phosphoserine-binding domain, protein–protein interactions: Repair, cell cycle regulation
CHROMO	smart00298	11	Protein–protein interactions, modulation of chromatin structure
UBA	smart00165	9	Ubiquitin signaling: Ubiquitin-binding domain, present, in particular, in chromatin-associated proteins
Cyt-b5	pfam00173	9	Heme/steroid binding domain, steroid signaling.
GTP_EFTU	pfam00009	7	GTPase P-loop domain involved in translation and a variety of regulatory processes; combines with a variety of domains, typically, at the C terminus
Pyr_redox	pfam00070	7	NADH-binding domain combining with other domains in a variety of oxidoreductases
Thioredoxin	pfam00085	6	Widespread disulfide redox domain
adh_short	pfam00106	6	Domain present in a wide variety of dehydrogenases
RRM	smart00360	5	The most common RNA-binding domain found, mostly in proteins involved in splicing, nucleocytoplasmic RNA transport, and chromatin remodeling
RVT	pfam00078	4	Reverse transcriptase domain combining with other domains in a broad variety of mobile elements

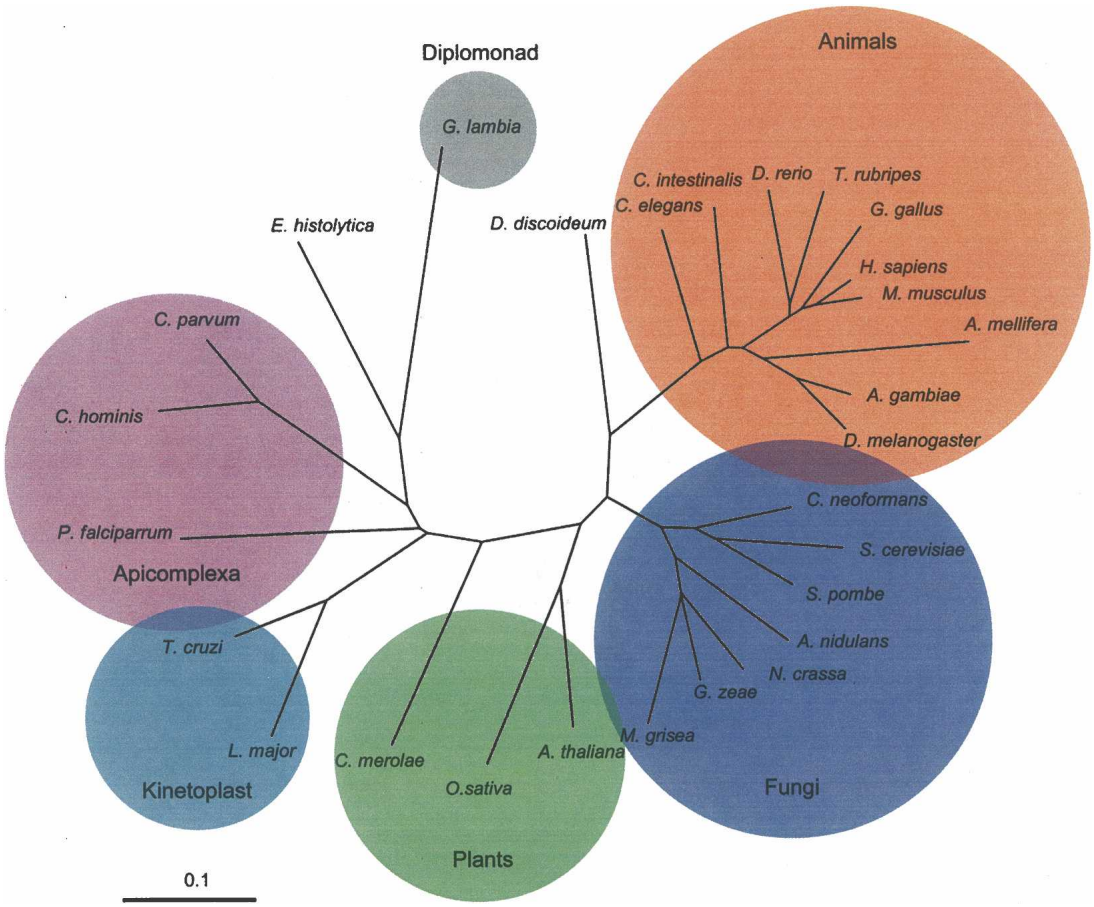

Figure 5. A tree of eukaryotes derived using the correlation values from the ordered list of promiscuity for domains in each of the analyzed species. The tree is color-coded according to the major groups of eukaryotes as follows: (orange) Animals; (green) Plantae; (dark blue) Fungi; (light blue) Kinetoplastida; (magenta) Apicomplexa; (gray) Diplomonada.

Smith 2003). There are several deviations from all versions of the currently accepted eukaryotic phylogenies. In particular, the slime mold *Dictyostelium discoideum* branches with animals, which reflects the previously noticed high diversity of domain architectures of *Dictyostelium* proteins (Eichinger et al. 2005). The second amoebozoan, *Entamoeba histolytica*, grouped within the unicellular part of the tree, emphasizing the distinction between these two organisms formally included within *Amoebozoa* (Song et al. 2005). The urochordate *Ciona intestinalis* fails to cluster with the chordates, and similarly, the unicellular rhodophyte *Cyanidioschyzon merolae* fell outside the plant clade; apparently, this reflects the paucity of domain architectures in these species. These anomalies notwithstanding, examination of the tree topology shows that the profile of domain promiscuity carries a strong phylogenetic signal. Notably, trees constructed using π values calculated after excluding single-domain proteins or raw β values failed to show phylogenetically sensible topology (data not shown), suggesting that the normalized bigram frequency is, indeed, the more robust measure of domain promiscuity.

Evolution of domain promiscuity in eukaryotes

To investigate how promiscuity evolved during eukaryotic evolution, we performed a parsimonious reconstruction of the ancestral sets of promiscuous domains (Fig. 6; Supplemental Figs. S2, S3). Here, the evolutionary tree topology is given, and the parsimony principle is applied to reconstruct the most parsimonious evolutionary scenario, that is, the scenario with the minimum number of events. We used two characters for reconstruction: first, domain presence–absence, and, second, domain promiscuity. For the domain presence–absence reconstruction, we applied Dollo parsimony (Farris 1977; Rogozin et al. 2005) to the set of the 215 stringently defined promiscuous domains. The crucial assumption of the Dollo parsimony method is that a character can be gained only once in a given tree, whereas multiple, independent losses are allowed. This assumption is reasonable for the reconstruction of the gain and loss of individual domains but not for the analysis of domain promiscuity. Therefore, for the latter reconstruction, we used general parsimony. Both reconstructions were performed using two alternative topologies of the eukaryotic evolutionary tree, namely, the "crown-group" tree (Hedges 2002) and the unikont–opisthokont tree (Stechmann and Cavalier-Smith 2003).

Under the crown-group topology, 84 of the 215 promiscuous domains were inferred to have been present in the last eukaryotic common ancestor (LECA) but only one domain, the AAA+ ATPase, was inferred to be promiscuous in LECA, with the status of five domains remaining uncertain (Fig. 6; Supplemental Fig. S2). Under the unikont–opisthokont topology, 180 of the 215 promiscuous domains were inferred to have been present in LECA, with two domains, AAA+ ATPase and BROMO, inferred to

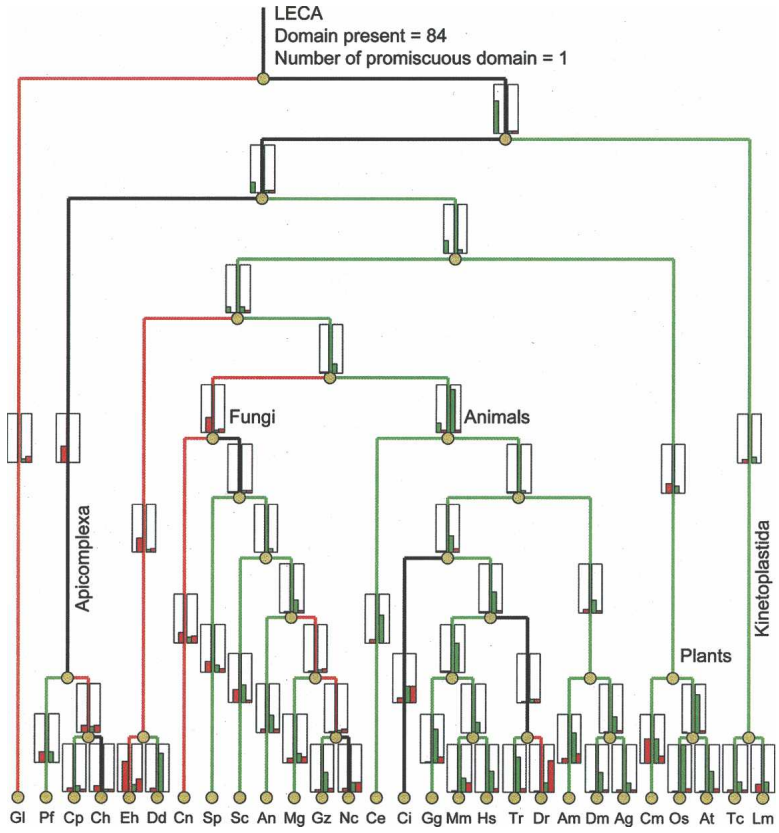

Figure 6. Gain and loss of domains and domain promiscuity during the evolution of eukaryotes. The number of domains gained and lost in each branch, inferred using Dollo parsimony, is shown on the bar plot to the *left* of the branch, and the number of gained and lost promiscuous domains, inferred using DNAPARS, is shown on the bar plot to the *right* of the branch. (Green bars) gain; (red bars) loss. The bars are normalized to the highest gain (green bars) or highest loss (red bars) of all the nodes. Additionally, each edge is colored to indicate (green) the greater number of gained promiscuous domains, (red) the greater number of lost promiscuous domains, and (black) equal contributions of gain and loss of promiscuity. The root node represents the Last Eukaryotic Common Ancestor (LECA). As gain and loss cannot be inferred for LECA, the presence of domains and the number of domains ascertained to be promiscuous are given by numbers. The major branches of eukaryotes are labeled. The tree has the "crown group" topology (Hedges 2002). For additional information, see Supplemental Figure S2. The species abbreviations are as described in Methods.

be promiscuous, and 13 domains assigned an uncertain state (Supplemental Fig. S3). Thus, domain promiscuity appears to be a feature that was poorly conserved in eukaryotic evolution; at least, domains did not retain their promiscuity through major evolutionary transitions. There seems to be a trend of gain of both domains and, particularly, promiscuity during eukaryotic evolution. The highest gain is inferred to have occurred at the base of the animal clade, where 16 domains were gained and 25 domains became promiscuous (Fig. 6; Supplemental Fig. S2). As noted above, there was a significant overlap between the sets of promiscuous domains in animals, fungi, and plants. Combining this observation with the reconstruction results, one has to conclude that parallel, independent gain of promiscuity by the same domain in different major branches was fairly common in eukaryotic evolution.

We further analyzed the evolution of promiscuity and domain combinations by examining bigram frequencies. For each promiscuous domain, a matrix D(*i*, AB) was constructed, where each element is the number of proteins from species *i* that contain the bigram AB. First, we determined the frequency of columns where a bigram was found in all genomes. A surprisingly large number of domains, 54, were found to form at least one ubiquitous bigram, a significant excess over the random expectation ($P < 0.001$, calculated using Monte-Carlo simulations). Next, we calculated the distribution of bigram frequency in the genomes for all promiscuous domains (Fig. 7). The vast majority of bigrams were found in a small fraction of genomes (<10%, the first bin); however, the distribution had a fat right tail (Fig. 7). An analysis using the C.A.MAN program suggested that this distribution is a mixture of two Poisson distributions separated at 0.4–0.5 (fraction of species containing the given bigram). The right tail (the second Poisson distribution) represents bigrams that were found in many species. Matrices for 90 domains contained at least one bigram that was found in >90% genomes, and matrices for almost all highly promiscuous domains (201) contained at least one bigram that was present in >50% genomes. These findings suggest that promiscuous domains persist within a "reservoir" of evolutionarily stable domain combinations (the right tail of the distribution in Fig. 7) from which numerous rare combinations (unique for a few species represented by the left half of the distribution in Fig. 7) emerge during evolution.

We then used general parsimony to reconstruct the scenario of gain and loss of domain combinations (bigrams) during eukaryotic evolution. In agreement with the reconstructions of the evolution of domain promiscuity described in the preceding section, we found that only a small fraction of bigrams, ~1% for the crown-group topology (Fig. 8; Supplemental Fig. S4) and ~2% for the unikont–opisthokont tree (Supplemental Fig. S5) mapped to LECA (Supplemental Table S7). This reconstruction supported the notion of remarkable volatility of domain promiscuity and, accordingly, domain combinations during eukaryotic evolution.

Domain promiscuity is correlated with the number of structural interactions

To investigate the relationship between domain promiscuity and physical interactions between domains, we determined the correlation between π and the number of unique interactions of the corresponding domain in the iPfam database, which collects structural information from the PDB database (Finn et al. 2005). The iPfam database includes interaction data for 740 of the 2715 domains detected in human proteins. Of these 740 domains, 173 were promiscuous according to our liberal definition. A relatively weak but statistically significant, positive correlation was detected between the promiscuity values of these 173 domains and the number of interactions reported in the iPfam database

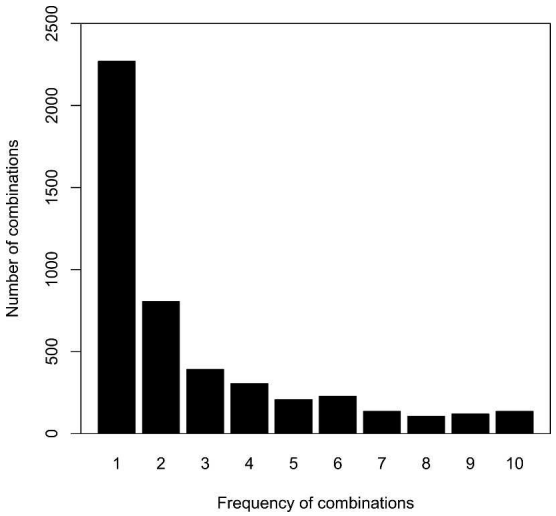

Figure 7. Distribution of bigram frequency in the analyzed genomes for all promiscuous domains. The bigram occurrence data were separated into 10 bins (bin 1, bigrams found in 0%–10% genomes; bin 2, bigrams found in 11%–20% genomes; . . .; bin 10, bigrams found in 91%–100% genomes).

(Supplemental Fig. S6). This is a low, conservative estimate because, first, the data in the iPfam database are extremely sparse, and second, the interactions are extracted from the PDB, which is a collection of structures from all organisms and is not specific to humans such that the number of interactions between domains of human proteins might be underestimated. Nevertheless, the correlation suggests that promiscuity measured by domain adjacency on a protein sequence reflects the ability of domains to participate in physical interactions.

Promiscuous domains are subject to strong purifying selection

We further investigated potential connections between promiscuity and the evolution of domain sequences. The ratio of nonsynonymous (K_a) and synonymous (K_s) substitution rates of domains (between the orthologous sequences from human and mouse), which reflects the strength of purifying selection (Li 1997; Hurst 2002), showed moderate but statistically highly significant negative correlation with domain promiscuity (Supplemental Fig. S7). Conceivably, and in agreement with the observation in the preceding section, the multiple interactions of promiscuous domains, which require multiple binding surfaces, constrain sequence evolution to a greater extent than it is constrained in domains with a smaller number of interaction partners.

Functional implications of domain promiscuity

To gain insight into the biological functions that were affected by the promiscuity increase during eukaryotic evolution, the 215 promiscuous domains were classified into functional categories (Tatusov et al. 2003). The excess of promiscuous domains in proteins involved in various forms of signaling is immediately apparent (Fig. 9). Examination of the list of the 25 domains that come across as promiscuous in animal, fungi, and plants further emphasizes and sharpens this conclusion (Table 1). In particular, two themes are prominent among these widespread promiscuous domains, namely, chromatin remodeling, which is a major contributor to the regulation of gene expression in eukaryotes (PHD, SET, BROMO, CHROMO, BRCT, and, in part, the AAA ATPase domains); and ubiquitin signaling (RING, UBQ, UCH, and UBA domains). Indeed, these signaling systems are conserved in all eukaryotic lineages and must have been present in LECA, although few domain combinations seem to have survived since that time (see above).

An examination of the top 10 promiscuous domain lists in animals, fungi, and plants shows both considerable coherence and interesting differences (Table 2) (very similar top 10 lists were obtained when the domains were ranked by promiscuity values obtained after removal of single-domain proteins or by the raw bigram frequency values; see Supplemental Tables S8 and S9). Three universally promiscuous domains—PH, PHD, and RING—are present in each of the three lists, and several others are shared

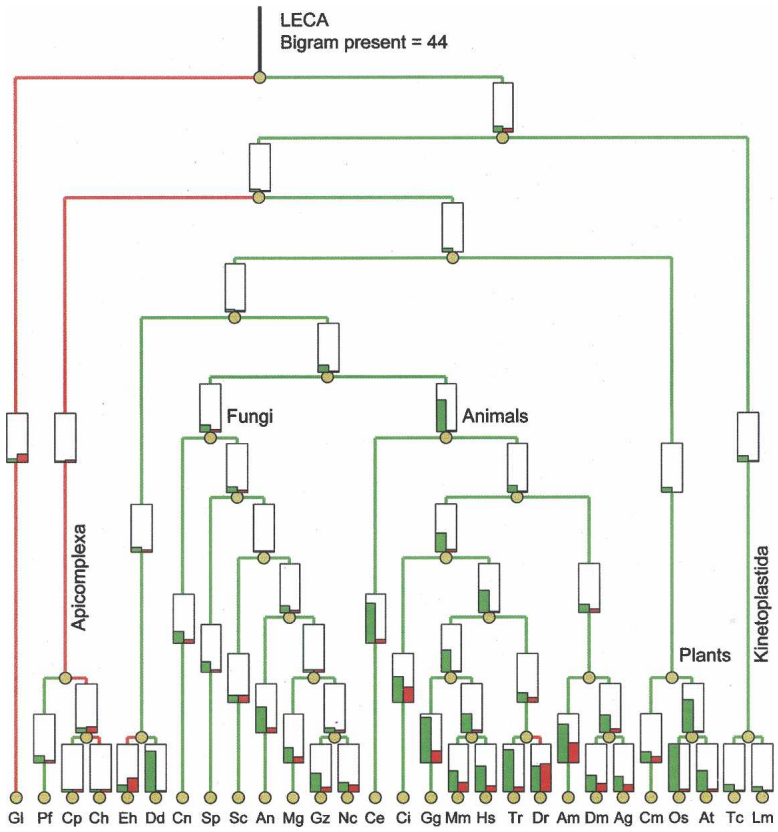

Figure 8. Gain and loss of domain bigrams during the evolution of eukaryotes. The parsimonious scenario of gains and losses was reconstructed using the DNAPARS program for the "crown group" topology of the eukaryotic phylogenetic tree. (Bar plots) The number of bigrams (green) gained and (red) lost in each branch. The other designations are as in Figure 6. For additional information, see Supplemental Figure S4.

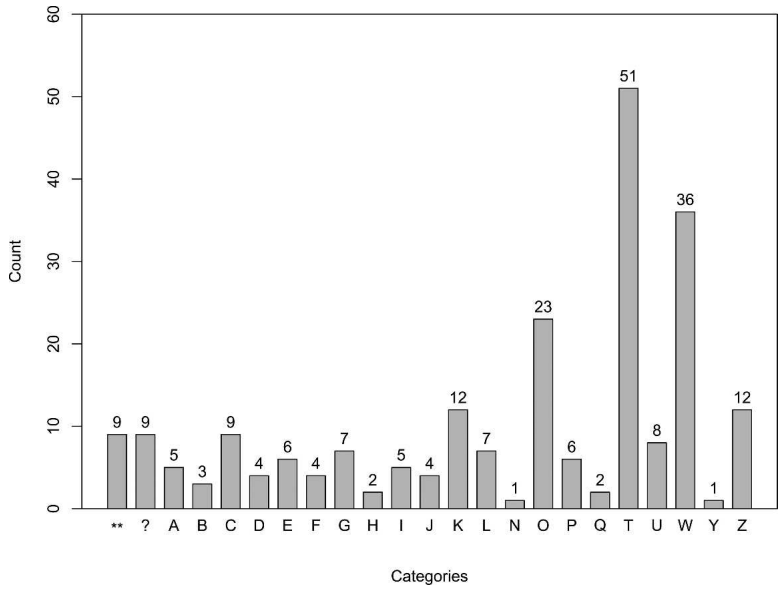

Figure 9. Distribution of promiscuous domains among functional categories of eukaryotic proteins. The categories are indicated with single-letter abbreviations on the *X*-axis, and the exact count of promiscuous domains in that category is shown on the *top* of each bar. If a domain is classified in more than one category, it is counted more than once. Abbreviations for functional categories (Tatusov et al. 2003) on the *X*-axis are: (A) RNA processing and modification; (B) chromatin structure and dynamics; (C) energy production and conversion; (D) cell cycle control, cell division, chromosome partitioning; (E) amino acid transport and metabolism; (F) nucleotide transport and metabolism; (G) carbohydrate transport and metabolism; (H) coenzyme transport and metabolism; (I) lipid transport and metabolism; (J) translation, ribosomal structure and biogenesis; (K) transcription; (L) replication, recombination, and repair; (N) cell motility; (O) post-translational modification, protein turnover, chaperones; (P) inorganic ion transport and metabolism; (Q) secondary metabolites biosynthesis, transport, and catabolism; (T) signal transduction; (U) intracellular trafficking, secretion, and vesicular transport; (W) extracellular structures and cell–cell signaling; (Y) nuclear structure; (Z) cytoskeleton; (**) various functions; (?) unknown function.

between two kingdoms. However, there is also a notable kingdom-specific component, for example, the two varieties of the EGF domain in animals and the cellulose-binding domain in fungi. It is of further note that even the shared top promiscuous domains never have the same most frequent bigram partner in any two kingdoms, an observation that emphasizes the volatility of domain combinations in eukaryotic evolution (see above).

Conclusions

The analysis of eukaryotic promiscuous domains reported here confirms and quantifies previously noticed and intuitively expected trends, such as the increase in domain promiscuity in phenotypically complex life forms (Koonin et al. 2000, 2004; Apic et al. 2001; Tordai et al. 2005; Wang and Caetano-Anolles 2006). A more unexpected series of observations reveals the low level of conservation of domain promiscuity and domain combinations in the course of eukaryotic evolution. The results suggest that very few, if any, domains have retained their promiscuous character throughout the history of eukaryotes, and very few domain combinations that involve promiscuous domains remained stable. Some caution is due in the interpretation of these findings because the evolutionary reconstructions were performed using parsimony approaches that have an inherent tendency to overestimate gain and underestimate loss of characters. In a similar setting, a series of recent studies on the gain and loss of introns in eukaryotic genes illustrates that maximum likelihood methods of evolutionary reconstruction yield substantially more intron-rich ancestors than parsimony methods (Csuros 2005; Nguyen et al. 2005; Rogozin et al. 2005; Carmel et al. 2007). Thus, the parsimony estimates give the low bound of domain promiscuity in ancestral eukaryotic forms. Because of the low counts of promiscuous domains, the use of maximum likelihood instead of parsimony is impractical. However, the demonstration of the low conservation of domain bigrams does not depend on reconstruction methods. Therefore, the conclusion that domain promiscuity is an evolutionarily volatile feature appears solid. There is little doubt that promiscuous domains comprise a major reservoir of eukaryotic evolvability, in particular, for the evolution of lineage-specific signaling networks.

Methods

Identification and analysis of protein domains

Proteins from each of the 28 analyzed eukaryotic species were extracted from the RefSeq database (National Center for Biotechnology Information, NIH) unless another database is specified (Supplemental Table S1). The species used in this study with abbreviations are as follows (also see Supplemental Table S1): *Giardia lamblia* (Gl), *Trypanosoma cruzi* (Tc), *Leishmania major* (Lm), *Cryptosporidium hominis* (Ch), *Cryptosporidium parvum* (Cp), *Plasmodium falciparum* (Pf), *Entamoeba histolytica* (Eh), *Dictyostelium discoideum* (Dd), *Saccharomyces cerevisiae* (Sc), *Schizosaccharomyces pombe* (Sp), *Aspergillus nidulans* (An), *Neurospora crassa* (Nc), *Cryptococcus neoformans* (Cn), *Gibberella zeae* (Gz), *Magnaporthe grisea* (Mg), *Arabidopsis thaliana* (At), *Oryza sativa* (Os), *Cyanidioschyzon merolae* (Cm), *Caenorhabditis elegans* (Ce), *Ciona intestinalis* (Ci), *Anopheles gambiae* (Ag), *Apis mellifera* (Am), *Drosophila melanogaster* (Dm), *Danio rerio* (Dr), *Takifugu rubripes* (Tr), *Gallus gallus* (Gg), *Mus musculus* (Mm), and *Homo sapiens* (Hs). For those species in which alternative splicing is common, the RefSeq database includes the major splice isoform. The sequences were searched for the presence of known domains using RPS-BLAST against the Conserved Domain Database (CDD) (Marchler-Bauer et al. 2005) with the *E*-value cutoff of 0.001 after masking low-complexity regions. Our experimentation with cutoffs showed that using the more liberal cutoff of 0.01 yielded an unacceptable level of false positives for many domains, even when the low-complexity filtering was applied (data not shown). The searches were done on the NCBI Linux cluster, and the data were stored in the SQLite database (http://www.sqlite.org/). The results from each of these searches were then filtered, and hits from the SMART (Letunic et al. 2006) and Pfam (Finn et al. 2006) databases were selected. The final list of domains from each species was then made using the following rules: (1) All overlapping hits from the same database were discarded (e.g., when two hits from SMART database overlapped, they both were discarded from the analysis). (2) Whenever hits from the SMART and Pfam databases overlapped, the SMART hits were retained and the Pfam hits were discarded.

Table 2. The 10 most promiscuous domains in animals, fungi, and plants

Domain	Average promiscuity (π)	Most frequent bigram partner	No. of occurrences	Functions/comments
Animals				
PH (smart00233)	972.18	SH3 (smart00326)	96	Protein–protein interactions; various signaling processes, in particular, inositol phosphate signaling
PDZ (smart00228)	675.6	SH3 (smart00326)	166	Protein–protein interactions; various forms of signaling
SH3 (smart00326)	556.45	GuKc (smart00072)	197	Protein–protein interactions; various forms of signaling
C1 (smart00109)	479.35	C2 (smart00239)	85	Small-molecule-binding and protein–protein interaction domains present, primarily in protein kinases; various forms of signaling
PHD (smart00249)	464.83	BROMO (smart00297)	123	Protein–protein interactions, primarily in chromatin
RING (smart00184)	441.26	BBOX (smart00336)	128	Ubiquitin signaling: E3 component of ubiquitin ligases
TyrKc (smart00219)	413.74	FN3 (smart00060)	223	Tyrosine kinase, various signaling process, primarily membrane receptors
EGF_CA (smart00179)	397.07	CUB (smart00042)	55	Ca-binding epidermal growth factor domain; various forms of extracellular signaling
SAM (smart00454)	371.45	TyrKc (smart00219)	138	Protein–protein interactions; various signaling processes, both extracellular and nuclear
EGF (smart00181)	353.07	LamG (smart00282)	155	Epidermal growth factor domain; various forms of extracellular signaling
Fungi				
SH3 (smart00326)	913.71	RasGEFN (smart00229)	13	Protein–protein interactions, various forms of signaling
AAA (smart00382)	839.63	Peptidase_M41 (pfam01434)	15	ATPase involved in various functions, including chaperone roles and various forms of signal transduction
GATase(pfam00117)	682.93	CPSase_sm_chain(pfam00988)	14	Glutamine amidotransferase domain found in a variety of metabolic enzymes
PH (smart00233)	654.23	Oxysterol_BP (pfam01237)	11	Protein–protein interactions; various signaling processes, in particular, inositol phosphate signaling
Cyt-b5 (pfam00173)	581.03	FMN_dh (pfam01070)	19	Heme/steroid binding domain, steroid signaling
Biotin_lipoyl (pfam00364)	568.33	Biotin_carb_C (pfam02785), E3_binding (pfam02817)	13	Coenzyme-binding domain of various metabolic enzymes
RING (smart00184)	444.48	DEXDc (smart00487)	33	Ubiquitin signaling: E3 component of ubiquitin ligases
PHD (smart00249)	432.18	JmjC (smart00558)	9	Protein–protein interactions, primarily in chromatin
fCBD (smart00236)	407.25	Glyco_hydro_61 (pfam03443)	9	Cellulose-binding domain involved in cell wall biogenesis
UCH (pfam00443)	371.29	ZnF_UBP (smart00290)	14	Ubiquitin signaling: ubiquitin C-terminal hydrolase
Plants				
AAA (smart00382)	828.41	Petpidase_M41 (pfam01434)	27	ATPase involved in various functions, including chaperone roles and various forms of signal transduction
PHD (smart00249)	666.67	BAH (smart00439)	9	Protein–protein interactions, primarily in chromatin
RING (smart00184)	510.53	DEXDc (smart00487)	18	Ubiquitin signaling: E3 component of ubiquitin ligases
CHROMO (smart00298)	407.19	DEXDc (smart00487)	9	Protein–protein interactions; modulation of chromatin structure
PH (smart00233)	356.27	DUF1336 (pfam07059)	7	Protein–protein interactions; various signaling processes, in particular, inositol phosphate signaling
UBA (smart00165)	341.61	UBQ (smart00213)	10	Ubiquitin signaling: ubiquitin-binding domain, present, in particular, in chromatin-associated proteins
RNA_pol_Rpb2_6 (pfam00562)	340.86	RNA_pol_Rpb2_7 (pfam04560)	15	One of the accessory domains of RNA polymerases; promiscuous because of the diversity of domain architectures
WD40 (smart00320)	321.71	Coatomer_WDAD (pfam04053)	12	Protein–protein interactions in diverse forms of signaling and RNA processing
UCH (pfam00443)	312.35	zf-MYND (pfam01753), ZnF_UBP (smart00290), DUSP (smart00695)	8	Ubiquitin signaling: ubiquitin C-terminal hydrolase
SET (smart00317)	311.58	PreSET (smart00468)	26	Methyltransferase methylating histones and other chromatin-associated proteins

These rules enabled us to create a map of non-overlapping, distinct SMART and Pfam domains for each genome. However, for a few domains, where the SMART-derived and Pfam-derived position-specific scoring matrices (PSSMs) were substantially different, one of these domains showed up in the hits, whereas the other one did not. For example, the Pfam PSSM for the Ank domain was compiled using more diverse families of proteins than a SMART PSSM matrix for the same domain. As a result, in many proteins, some Ank domains were detected by the Pfam PSSM but not by the SMART PSSM. To produce nonredundant and maximally complete lists of occurrences for such domains, the CDD domain neighbor facility was used. For each Pfam domain that had a unique identifiable domain neighbor in the SMART database, each occurrence of the Pfam domain was replaced with the corresponding SMART ID; for all further analyses, these two domains were treated as synonyms.

In the course of domain analysis, it was noticed that, despite the applied filtering for low complexity regions, the Pfam Myosin_tail_1 matrix (PF01576) and SMC_hinge (PF06470) that include extended coiled-coil structures produce numerous spurious hits in diverse, unrelated proteins. Therefore, these two domains were eliminated from all the analyses.

For the analysis of domain interactions, the iPfam interaction table was downloaded from the iPfam FTP site (ftp://ftp.sanger.ac.uk/pub/databases/Pfam/database_files; Finn et al. 2005). The number of interactions reported for each Pfam domain (except self-interaction) was counted. Because our list of promiscuous domains also contained SMART domains, we used the CDD database domain neighbor list to extract synonymous Pfam domain IDs.

For the separation of frequency distributions of domain bigrams, we used the C.A.MAN software package (Bohning et al. 1992). C.A.MAN takes into account the possibility that the observed data were generated from a complex distribution function that is a mixture of simple distributions. The software performs maximum-likelihood estimation of the parameters of the constituent simple distribution functions, as well as the mixing coefficients. C.A.MAN allows for a wide variety of distributions from the exponential family to be used in the mixture model, and it provides statistics to determine the optimal number of distributions in the mixture.

Synonymous and nonsynonymous substitutions

For the analysis of nonsynonymous and synonymous substitutions, 8023 orthologous mouse and human proteins were extracted from the HomoloGene database (Wheeler et al. 2006). The corresponding cDNA sequences were retrieved from the NCBI server using the NCBI E-utility facility (http://eutils.ncbi.nlm.nih.gov/entrez/query/static/eutils_help.html) and a custom Perl script. The proteins were aligned using MUSCLE (Edgar 2004). The regions corresponding to each domain were extracted from each alignment, and the nucleotide sequences for each of these regions were then aligned using the protein alignment as a guide. The K_a/K_s ratio for each of these domain alignments was calculated using the method of Nei and Gojobori (1986) implemented in the yn00 program of the PAML package (Yang 1997). All domains with $K_a/K_s > 1$ were discarded to eliminate probable misaligned regions. The K_a/K_s values were averaged over all instances of a particular domain. This analysis included all 92 domains in human and 86 domains in mouse that were found to be promiscuous.

Phylogenetic analysis

The genome tree of eukaryotes based on the distances between genome-specific domain promiscuity profiles was constructed using the NEIGHBOR program. The reconstruction of the gain and loss of domains was performed using the DOLLOP program, and the reconstruction of the gain and loss of domain promiscuity using the DNAPARS program. All programs were from the PHYLIP package (Felsenstein 2005).

Functional classification of proteins and domains

The functional categories of proteins were from the COG classification of proteins (ftp://ftp.ncbi.nih.gov/pub/COG/COG/fun.txt; Tatusov et al. 2003; Koonin et al. 2004), with minor modifications: RNA processing and modification (A); chromatin structure and dynamics (B); energy production and conversion (C); cell cycle control, cell division, chromosome partitioning (D); amino acid transport and metabolism (E); nucleotide transport and metabolism (F); carbohydrate transport and metabolism (G); coenzyme transport and metabolism (H); lipid transport and metabolism (I); translation, ribosomal structure and biogenesis (J); transcription (K); replication, recombination, and repair (L); cell motility (N); post-translational modification, protein turnover, chaperones (O); inorganic ion transport and metabolism (P); secondary metabolites biosynthesis, transport, and catabolism (Q); signal transduction (T); intracellular trafficking, secretion, and vesicular transport (U); extracellular structures and cell–cell signaling (W); nuclear structure (Y); cytoskeleton (Z). Two additional categories were introduced: "various functions (**)" for domains that contribute to more than one category (mostly, protein–protein interaction domains such as SH3) and domains with unknown functions ("?").

Acknowledgments

We thank Yuri Wolf, Kira Makarova, and Anna Panchenko for useful discussions. This work was supported in part by the Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS.

References

Anantharaman, V., Koonin, E.V., and Aravind, L. 2001. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.* **307:** 1271–1292.

Apic, G., Gough, J., and Teichmann, S.A. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310:** 311–325.

Aravind, L., Dixit, V.M., and Koonin, E.V. 2001. Apoptotic molecular machinery: Vastly increased complexity in vertebrates revealed by genome comparisons. *Science* **291:** 1279–1284.

Bashton, M. and Chothia, C. 2007. The generation of new protein functions by the combination of domains. *Structure* **15:** 85–99.

Bohning, D., Schlattmann, P., and Lindsay, B. 1992. Computer-assisted analysis of mixtures (C.A.MAM): Statistical algorithms. *Biometrics* **48:** 283–303.

Bohning, D., Dietz, E., and Schlattmann, P. 1998. Recent developments in computer-assisted analysis of mixtures. *Biometrics* **54:** 525–536.

Carmel, L., Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. 2007. Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.* **17:** 1034–1044.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282:** 2022–2028.

Csuros, M. 2005. Likely scenarios of intron evolution. In *Comparative genomics*. Proceedings of the RECOMB 2005 International Workshop, RCG 2005, Dublin, Ireland, September 18–20, 2005. Lecture Notes in Computer Science (eds. A. McLysaght and D.H. Huson), Vol. 3678, pp. 47–60. Springer, Berlin.

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

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high

accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.
Eichinger, L., Pachebat, J.A., Glockner, G., Rajandream, M.A., Sucgang, R., Berriman, M., Song, J., Olsen, R., Szafranski, K., Xu, Q., et al. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature* **435:** 43–57.
Farris, J.S. 1977. Phylogenetic analysis under Dollo's Law. *Syst. Zool.* **26:** 77–88.
Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author, Department of Genome Sciences, University of Washington, Seattle.
Fickett, J.W. and Guigo, R. 1993. Estimation of protein coding density in a corpus of DNA sequence data. *Nucleic Acids Res.* **21:** 2837–2844.
Finn, R.D., Marshall, M., and Bateman, A. 2005. iPfam: Visualization of protein–protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics* **21:** 410–412.
Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., et al. 2006. Pfam: Clans, web tools and services. *Nucleic Acids Res.* **34:** D247–D251.
Fong, J.H., Geer, L.Y., Panchenko, A.R., and Bryant, S.H. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. *J. Mol. Biol.* **366:** 307–315.
Glazko, G.V., Milanesi, L., and Rogozin, I.B. 1998. The subclass approach for mutational spectrum analysis: Application of the SEM algorithm. *J. Theor. Biol.* **192:** 475–487.
Han, J.H., Batey, S., Nickson, A.A., Teichmann, S.A., and Clarke, J. 2007. The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.* **8:** 319–330.
Hedges, S.B. 2002. The origin and evolution of model organisms. *Nat. Rev. Genet.* **3:** 838–849.
Hofmann, K. 1999. The modular nature of apoptotic signaling proteins. *Cell. Mol. Life Sci.* **55:** 1113–1128.
Hurst, L.D. 2002. The K_a/K_s ratio: Diagnosing the form of sequence evolution. *Trends Genet.* **18:** 486.
Itoh, M., Nacher, J.C., Kuma, K.I., Goto, S., and Kanehisa, M. 2007. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol.* **8:** R121. doi: 10.1186/gb-2007-8-6-r121.
Koonin, E.V., Aravind, L., and Kondrashov, A.S. 2000. The impact of comparative genomics on our understanding of evolution. *Cell* **101:** 573–576.
Koonin, E.V., Wolf, Y.I., and Karev, G.P. 2002. The structure of the protein universe and genome evolution. *Nature* **420:** 218–223.
Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Rao, B.S., et al. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol.* **5:** R7. http://genomebiology.com/2004/5/2/R7.
Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., and Bork, P. 2006. SMART 5: Domains in the context of genomes and networks. *Nucleic Acids Res.* **34:** D257–D260.
Li, W.H. 1997. *Molecular evolution*. Sinauer, Sunderland, MA.
Manning, C.D. and Schütze, H. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.
Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., et al. 2005. CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res.* **33:** D192–D196.
Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O., and Eisenberg, D. 1999. Detecting protein function and protein–protein interactions from genome sequences. *Science* **285:** 751–753.
Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3:** 418–426.
Nguyen, H.D., Yoshihama, M., and Kenmochi, N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.* **1:** e79. doi: 10.1371/journal.pcbi.0010079.
Orengo, C.A. and Thornton, J.M. 2005. Protein families and their evolution—A structural perspective. *Annu. Rev. Biochem.* **74:** 867–900.
Patthy, L. 2003. Modular assembly of genes and the evolution of new functions. *Genetica* **118:** 217–231.
Rogozin, I.B., Wolf, Y.I., Babenko, V.N., and Koonin, E.V. 2005. Dollo parsimony and the reconstruction of genome evolution. In *Parsimony, phylogeny, and genomics* (ed. V.A. Albert), pp. 190–200. Oxford University Press, Oxford.
Snel, B., Bork, P., and Huynen, M.A. 1999. Genome phylogeny based on gene content. *Nat. Genet.* **21:** 108–110.
Snel, B., Huynen, M.A., and Dutilh, B.E. 2005. Genome trees and the nature of genome evolution. *Annu. Rev. Microbiol.* **59:** 191–209.
Song, J., Xu, Q., Olsen, R., Loomis, W.F., Shaulsky, G., Kuspa, A., and Sucgang, R. 2005. Comparing the *Dictyostelium* and *Entamoeba* genomes reveals an ancient split in the Conosa lineage. *PLoS Comput. Biol.* **1:** e71. doi: 10.1371/journal.pcbi.0010071.
Stechmann, A. and Cavalier-Smith, T. 2003. The root of the eukaryote tree pinpointed. *Curr. Biol.* **13:** R665–R666.
Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., et al. 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics* **4:** 41. doi: 10.1186/1471-2105-4-41.
Templeton, T.J., Iyer, L.M., Anantharaman, V., Enomoto, S., Abrahante, J.E., Subramanian, G.M., Hoffman, S.L., Abrahamsen, M.S., and Aravind, L. 2004. Comparative analysis of apicomplexa and genomic diversity in eukaryotes. *Genome Res.* **14:** 1686–1695.
Tordai, H., Nagy, A., Farkas, K., Banyai, L., and Patthy, L. 2005. Modules, multidomain proteins and organismic complexity. *FEBS J.* **272:** 5064–5078.
Vogel, C., Bashton, M., Kerrison, N.D., Chothia, C., and Teichmann, S.A. 2004. Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14:** 208–216.
Vogel, C., Teichmann, S.A., and Pereira-Leal, J. 2005. The relationship between domain duplication and recombination. *J. Mol. Biol.* **346:** 355–365.
Wang, M. and Caetano-Anolles, G. 2006. Global phylogeny determined by the combination of protein domains in proteomes. *Mol. Biol. Evol.* **23:** 2444–2454.
Webb, A.R. 2002. *Statistical pattern recognition*. Wiley, New York.
Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., DiCuccio, M., Edgar, R., Federhen, S., et al. 2006. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **34:** D173–D180.
Wolf, Y.I., Brenner, S.E., Bash, P.A., and Koonin, E.V. 1999. Distribution of protein folds in the three superkingdoms of life. *Genome Res.* **9:** 17–26.
Wolf, Y.I., Rogozin, I.B., Grishin, N.V., and Koonin, E.V. 2002. Genome trees and the tree of life. *Trends Genet.* **18:** 472–479.
Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. 2004. Coelomata and not Ecdysozoa: Evidence from genome-wide phylogenetic analysis. *Genome Res.* **14:** 29–36.
Wuchty, S. 2001. Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18:** 1694–1702.
Yang, Z. 1997. PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13:** 555–556.
Ye, Y. and Godzik, A. 2004. Comparative analysis of protein domain organization. *Genome Res.* **14:** 343–353.

Received July 22, 2007; accepted in revised form November 28, 2007.