BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 3. 205–216
Advance Access publication January 16, 2009
doi:10.1093/bib/bbn057

Domain mobility in proteins: functional and evolutionary implications

Malay Kumar Basu, Eugenia Poliakov and Igor B. Rogozin

Submitted: 4th September 2008; Received (in revised form): 8th December 2008

Abstract

A substantial fraction of eukaryotic proteins contains multiple domains, some of which show a tendency to occur in diverse domain architectures and can be considered mobile (or 'promiscuous'). These promiscuous domains are typically involved in protein–protein interactions and play crucial roles in interaction networks, particularly those contributing to signal transduction. They also play a major role in creating diversity of protein domain architecture in the proteome. It is now apparent that promiscuity is a volatile and relatively fast-changing feature in evolution, and that only a few domains retain their promiscuity status throughout evolution. Many such domains attained their promiscuity status independently in different lineages. Only recently, we have begun to understand the diversity of protein domain architectures and the role the promiscuous domains play in evolution of this diversity. However, many of the biological mechanisms of protein domain mobility remain shrouded in mystery. In this review, we discuss our present understanding of protein domain promiscuity, its evolution and its role in cellular function.

Keywords: *mobile domain; promiscuous domain; domain network; domain architecture; domain evolution*

PROTEIN DOMAINS

Protein domains are the structural and functional units of proteins. It is now well established that proteins carry out their functions primarily through their constituent domains. They can be gained by proteins to acquire new function. Domains are, therefore, considered to be the units through which proteins evolve. In structural biology, domains are defined as independent folding units in a protein. However, domains are generally identified as highly conserved regions of the protein sequence. This apparent contradiction in definition of protein domain disappears upon scrutiny: domains identified by sequence conservation alone have been shown to have distinct structural identity [1, 2]. Numerous sequence- and structure-based domain databases enable protein domain detection with very high accuracy, such as Pfam [3], SMART [4], CDD [5], INTERPRO [6], SCOP [7], ProDom [8], DALI [9] and CATH [10]. These databases either use sequence- or structure-based methods to identify regions in protein sequences that belong to specific domain families.

Despite decades of study, the biological mechanisms shaping the domain architecture in proteins are largely unknown. However, it is now known that domains differ in their propensity to form multi-domain proteins. While some domains are present only in specific combinations, others participate in diverse domain architectures. Domains of the latter types are called 'promiscuous' or mobile domains, and are very important in creating the observed diversity in protein domain architectures. They play a major role in signaling network in the cell by bringing together domains with different functionalities into one protein sequence, and thus promoting crosstalk in signaling. Their central role in evolution cannot be overemphasized, but only

Corresponding author. Malay Kumar Basu, J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA. Tel: +1 301 795 7890; Fax: +1 301 795 7060; E-mail: malaykbasu@gmail.com

Malay Kumar Basu presently is a Senior Bioinformatics Engineer at the J. Craig Venter Institute (Rockville MD, USA).

Eugenia Poliakov is a staff scientist at the Laboratory of Retinal Cell and Molecular Biology, National Eye Institute, National Institutes of Health (Bethesda MD, USA).

Igor B. Rogozin is a staff scientist at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health (Bethesda MD, USA).

recently we have begun to understand the role of selection in shaping the domain promiscuity. In this review, we will discuss our current understanding of protein domain promiscuity, its evolution and its role in cellular function. After briefly discussing the multidomain architecture in proteins, we will discuss how promiscuous domains are identified, and how the domain promiscuity can be measured. Finally, we will discuss the functional and evolutionary significance of the promiscuous domains.

DOMAIN STRUCTURE OF PROTEINS

The number of unique domains in an organism is roughly proportional to its genome size. In unicellular eukaryotes, such as apicomplexans, diplomonads and protozoans, the unique number of domains is ∼1000, whereas in plants, fungi and animals, the numbers can be as high as ∼3000. The average size of domain is ∼100 amino acids [11]. The number of domains per gene (modularity) follows the power-law (see below) distribution [12], and it has been shown that tissue-specific genes have higher modularity [12, 13].

The estimation of the frequency of multidomain proteins in the three superkingdoms of life (bacteria, archaea and eukaryotes) varies with the methodologies and database used [14–18], but the emerging consensus is that prokaryotes have fewer multidomain proteins than eukaryotes. The tendency of formation of multidomain proteins increases from archaea to bacteria to eukaryotes [1, 19]. Although within eukaryotes, particularly in animals, there is a distinct tendency towards formation of multidomain proteins (39% of metazoan proteins contain more than one Pfam domain, whereas the corresponding number for unicellular eukaryotes is smaller, 32% [20]), a large fraction of the proteins in all three super-kingdoms of life contain 0–1 domain [2, 18, 20, 21]. However, we have to keep in mind that poor description of domains in some lineages may create problems for this analysis. Proteins with zero domains may actually lack domains, or such proteins may contain domains that are yet unknown. However, it was suggested that the differences between different evolutionary lineages are unlikely to be due to differences in annotation coverage [20]. As shown by Ekman and co-workers [17], the Pfam domain coverage is similar for archaea, bacteria and eukaryota: in each group about 70% of the proteins have at least one Pfam domain. In agreement with this conclusion, analyses by Tordai and co-workers [20] have also shown that Pfam coverage is similar for bacteria, archaea, protozoa, plants, fungi and metazoa. It is, therefore, reasonable to infer that the differences in the number of multidomain proteins in archaea, bacteria and eukaryotes are indeed true.

The propensity of protein domains to form multidomain architecture increases with organismal complexity. Though complexity is a contentious issue in evolution, here we define it as the number of cell types in an organism. The phenomenon that organisms with higher complexity tend to acquire more multidomain proteins is called 'domain accretion' [22], which could translate into increasing interaction amongst the domains. This may be one of the explanations of the apparent lack of correlation between the complexity and number of genes in a genome (G-value paradox): flies have fewer genes than nematodes; humans have fewer genes than rice [23]. Increasing modularity through domain accretion, at least in theory, can overcome the shortcoming posed by fewer genes in the genome. The biological mechanisms dictating domain accretion is not known. But, there is evidence that domains involved in the same functional pathway tend to come together in one protein sequence [24]. This phenomenon has been used to determine the functions of unknown domains in proteins, in what is called the 'Rosetta Stone' approach [24].

Given the large number of domains present in an organism, the possible combinatorial arrangements are enormous. However, in eukaryotic genomes domains are present only in a limited set of arrangements in multidomain proteins. This suggests that evolutionary constraints play an important role in the selection of domain architectures observed in multidomain proteins [2]. Indeed, domain arrangements, even the domain ordering in multidomain proteins, determine their three dimensional arrangements, and therefore, might affect function [25]. In earlier studies, it was shown that most of the domain combinations in multidomain proteins have been formed only once in the evolution, and the domain combinations are inherited rather than formed through convergent evolution [14, 26]. However, in a recent study, Forslund and co-workers claimed that convergent evolution is more prevalent than previously thought [27]. They investigated the prevalence of domain architecture reinvention in 96 genomes with a novel domain tree-based method

that uses maximum parsimony for inferring ancestral protein architectures. They detected multiple origins for 12.4% of the architectures. This result indicates that domain architecture reinvention is a much more common phenomenon than previously thought [27]. Thus, it is possible that the process of convergent domain architecture evolution is driven by functional necessity.

PROMISCUOUS DOMAINS

Domains are present in various combinations in multidomain proteins. While some domains are present in stable configuration, others are present in many different domain milieus. Promiscuous or mobile domains are domains that reside in many different domain combinations [20, 24, 28]. The term promiscuity carries several connotations when applied to a protein domain. In scientific literature, promiscuity can signify domains with higher degree of mobility (as described above), or domains that physically interact with many other domains (protein–protein interactions), or domains that bind different types of molecules. In this article, the term promiscuous domain will be used to mean mobile domains.

Although the reasons why some domains are mobile and others are static are largely unknown, some recent studies indicate the possible properties and thereby hint at the reasons. It has been shown that domains in multidomain proteins are generally smaller in size than those that are present as single domain [20]. This phenomenon is claimed to be due to the fact that domains that are present in different protein environments need to fold independently, and their smaller size facilitates independent folding [20]. It has been shown that the mobility of domains may have a large functional dependence: those required for specific functions tend to get mobile in specific lineages [28].

It has been recently shown that promiscuous domains evolve more slowly compared to non-promiscuous ones [28]. It has also been shown that promiscuous domains identified by their co-occurrence in single polypeptide alone also tend to show a higher number of physical domain–domain interactions [28]. This is true even for promiscuous domains (e.g. SH3 and PDZ, Table 1) that do not bind to other globular domains, but instead to short linear sequence motifs or covalent protein modifications present in the interaction partner. Taking these observations together, it appears that because promiscuous domains need to participate in many different kinds of protein–protein interactions, they tend to evolve slowly than domains that need to participate in specific interactions, where compensatory mutations in the both interaction partners could relax the selection pressure on the sequence.

DOMAIN CO-OCCURRENCE NETWORK

If we plot the frequency distribution of domain in an organism, the plot roughly follows a power-law (Figure 1) [1, 29, 30]. In the power-law, the frequency of an event $f(x)$ is proportional to its rank i with a relation $f(x) \sim i^{-\gamma}$, where γ is a parameter. The power-law has been identified in numerous biological, physical and social contexts,

Table 1: Ten promiscuous domains with the highest average promiscuity in majority of eukaryotes

Domain (ID)	Average promiscuity[a]	Description
PH (smart00233)	680.07	Protein–protein interactions; various signaling processes, in particular, inositol phosphate signaling
AAA+ (smart00382)	637.38	ATPase involved in various functions, including chaperone roles and various forms of signal transduction
SH3 (smart00326)	587.36	Protein–protein interactions; various forms of signaling
CI (smart00109)	442	Small-molecule binding and protein–protein interaction domains present, primarily in protein kinases; various forms of signaling
GATase (pfam00117)	424.69	Glutamine amidotransferase domain found in a variety of metabolic enzymes
PHD (smart00249)	420.38	Protein–protein interactions, primarily in chromatin
PDZ (smart00228)	418.74	Protein–protein interactions; various forms of signaling
BiotinLipoyl (pfam00364)	371.68	Coenzyme-binding domain of various metabolic enzymes
RING (smart00184)	364.35	Ubiquitin signaling: E3 component of ubiquitin ligases
EGF (smart00181)	323.56	Epidermal growth factor domain; various forms of extracellular signaling

[a]Average promiscuity is defined as the mean promiscuity value calculated over 28 eukaryotic species in reference [28].

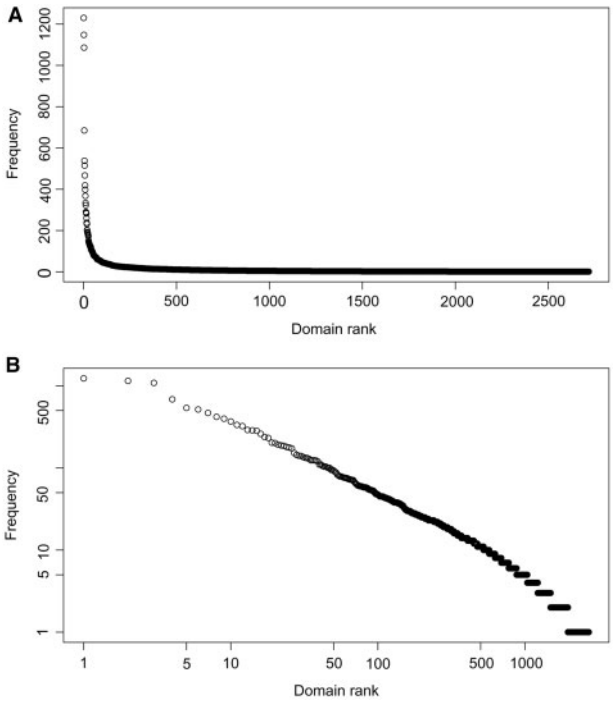

Figure 1: Power-law distribution of domains in human genome. (**A**) Rank of a domain after sorting according the frequency in the genome on *X*-axis is plotted against the frequency on the *Y*-axis. (**B**) Log–log plot of the ranks of domain on *X*-axis is plotted against the frequency on the *Y*-axis.

such as hypertext links in Internet, population distribution is towns, number of reactions in which a particular metabolite is involved, number of pseudogenes in a particular gene family, and many others [31–39]. Two very common versions of the power-law are Zipf's law, which describes the frequency distribution of words in a text [40] and the Pareto distribution, which describes the distribution of people by wealth [41]. Pareto distribution also led to the famous Pareto principle, which says 'few contain many and most contain few' or the so called 80-20 rule. Examples of such rule are 20% of product from a company determines 80% of the return, 20% of the defects caused 80% of the problems, and many others.

The power-law distribution has special mathematical properties related to a type of network called 'scale-free', where the frequency distribution of node degrees (number of nodes to which a given node is connected) follows a power-law [33, 34]. Many biological networks are scale-free in nature, such as metabolic networks, protein–protein interaction networks, and many others [36, 42].

Domain co-occurrence networks also fall under the scale-free category [21]. These networks are graphs in which each node represents a domain, and two nodes are connected by an edge only if they are present is a single protein sequence [20, 21, 43, 44]. In a scale-free network, there are few nodes that are highly connected, but majority of them have low connectivity. Additionally, in a scale-free network, the features of the network and the underlying distribution do not change with the increasing number of nodes. In a protein domain co-occurrence network, promiscuous mobile domains are highly connected nodes or hubs (Figure 2). This type of distribution of connectivity is very different from random network where the connectivity is largely uniform. Moreover, the scale-free nature of such a network is largely assumed to exist due to 'preferential attachment', which dictates that the probability of a node acquiring new connections is proportional to its degree (the number of nodes to which a given node is connected). Thus the implication of such connectivity for a domain co-occurrence network is important in showing that domain combinations in proteins are not random and that promiscuous domains have a tendency to become more promiscuous during evolution.

HOW NEW DOMAIN COMBINATIONS ARE CREATED

To attain promiscuity status a domain needs to create new domain combination, it is, therefore, important to understand how new domain combinations are created in proteins. Although the biological mechanisms that give rise to new domain combinations are largely unknown, several mechanisms have been proposed with anecdotal evidence. Examples of such mechanisms are gene fusion and fission, *de novo* creation of genes from non-coding elements, and recruitment of the mobile genetic elements [45]. Domains are frequently gained by proteins through insertions at the N or C terminus [46, 47]. Repeated domains can also arise through duplication [48]. Novel structure can also arise due to circular permutation of existing domains [49].

It has been shown that the domain boundaries in animal genomes, particularly extracellular portions of animal membrane proteins, coincide with the exons in which the domain resides [50–53]. The idea is that exon-bordering domains may move in

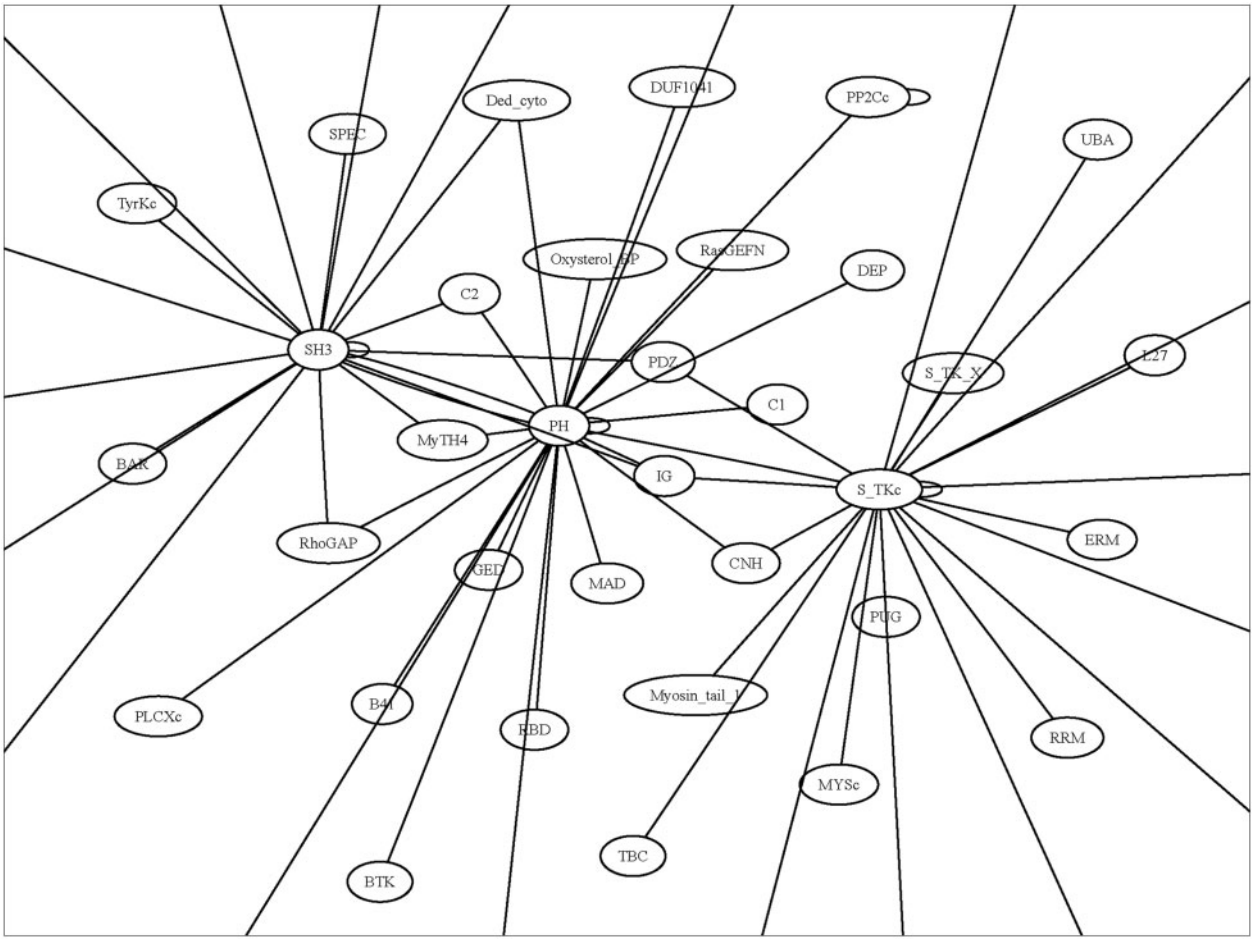

Figure 2: The partial domain co-occurrence graph of promiscuous domains, PH, SH3 and S.TKC in human genome. The nodes represent domains; two nodes are connected by an edge only when the connecting domains are present next to each other on the same protein sequence.

the genome as 'cassette-exon'. The existence of cassette-exons could be explained as a by-product of exon-shuffling, a process where new genes evolve by shuffling of existing exons in a gene. Exon-shuffling has been forwarded as an evidence of the 'intron-early' theory, which proposes that introns were present in the Last Universal Common Ancestor (LUCA) of all extant organisms, and later lost in prokaryotes. In contrast, 'intron-late' proponents believe that they were a late innovation in eukaryotes and prokaryotes never had introns. Evidence of the exon-shuffling has been found in animals [54, 55], whereas in plants and fungi there is no evidence of exon-shuffling [50, 56].

The present diversity of domain combinations in proteins does not differ significantly from stochastic birth, death and innovation models (BDIMs) [1, 30, 39, 57]. These models predict the presence of an equilibrium state of the domain distribution, which is reached exponentially; the death of a domain must be counteracted by 'innovation' or creation of new domains. BDIMs ignore completely the individuality of gene families and the selective forces that make some of them expendable and others indispensable. Despite this obvious over-simplification, BDIMs accurately reproduce the observed family size distributions, suggesting that genome evolution might be largely a stochastic process, which is modulated by natural selection [1, 19].

A QUANTITATIVE MEASUREMENT OF PROMISCUITY

To identify promiscuous domain one needs to consider several parameters. Some of these parameters are as follows: (a) other domains that

co-occur with a particular domain in one protein sequence, (b) number of different multidomain architectures in which a domain participates and (c) the abundance of a domain in the genome. Earlier work relied on the parameter (a) to find promiscuous domains. These works made use of the connectivity parameter of domain co-occurrence network to find out promiscuous domains [21, 44]. Note that by definition promiscuous domains co-occur more with other domains, and therefore, are highly connected nodes or hubs in domain-occurrence network. Works that relied on connectivity parameters simply identified these highly connected nodes. But relying solely the connectivity parameters is largely misleading, because it is known that many domains, though participating in large multidomain architectures, in fact exist in fewer local contexts [20]. It is, therefore, necessary to consider immediate domain neighbors (domains adjacent to a given domain on a polypeptide sequence) to correctly identify promiscuous domains. In a later study, Tordai and co-workers [20] took this fact into account to identify promiscuous domains by considering 'domain triplets', three domains next to each other on a protein sequence. This study identified promiscuous domains as those who participate in many of these triplets. This is akin to using parameter (b). But even this study, which took local environment into account, largely ignored the abundance of domain in the genome [20], a very important criterion to determine domain promiscuity correctly. Promiscuity involves duplication and insertion of a given domain in a new location. Thus it is imperative to differentiate domains that are present with high abundance in the genome and participate in large number of combination as a result of their high abundance, from the true promiscuous domains. This is illustrated in the following example. Consider domain A is present twice in a genome with domains B and C in combinations AB and AC. Now consider another domain P, which is present thrice in the genome, twice as PQ and only once as PR, where Q and R are other two domains. A calculation that ignores the abundance will rank both A and P having same promiscuity. But, in reality, the promiscuity of A should be higher than P because, in spite of having a lower abundance, domain A participates in larger number of combinations.

Recently, we developed a method to objectively measure mobility/promiscuity of a protein domain [28], taking the abundance of a domain into consideration. The method uses techniques from computational linguistics to measure promiscuity from domain co-occurrence. The method, called 'bigram analysis', is generally used to find words with more semantic importance in any language [58]. It has also been employed in finding words that are semantically linked to each other. The idea is to count the number of times a pair of words (bigram) occurs in a text (corpus). If a pair occurs less frequently from the background distribution, it carries more semantic information than the others. Additionally, this analysis also points out the words that, by nature, tend to participate is many bigrams and are, therefore, promiscuous.

We used the whole genome sequence as text (corpus) and each protein as sentence and each domain as word and used the same bigram analysis to statistically identify domains that participate in many bigrams and are therefore promiscuous [28]. This method generates the measured promiscuity value for each domain in the genome. Using this method, we calculated the promiscuity values for each domain in 28 eukaryotic species spanning all the major branches of the eukaryotic tree (see Supplementary Data for details) [28].

It was recently shown that there is a relationship across genomes between the promiscuity of a given domain and its frequency [59]. However, the strength of this relationship differs for different domains. A new index 'domain versatility index' (DVI) was suggested. DVI was defined as the strength of the relationship between the number of occurrences of a domain (N) and the number of bigrams (NN) in which this domain participates. More precisely, the logarithmic regression of NN over N was calculated, and the linear coefficient was taken as DVI. The authors explored links between the versatility of a domain, when unlinked from abundance, and its biological properties. The results suggested that domains occurring as single domain proteins and domains appearing frequently at protein termini have a higher DVI. This is consistent with previous observations that the evolution of domain re-arrangements is primarily driven by fusion of pre-existing arrangements and single domains, as well as by loss of domains at protein termini. Contrary to previous studies, versatility is lower in eukaryotes. It was suggested that a random attachment process is sufficient to explain the observed distribution of domain arrangements [59].

There was also very high correlation (88%) between promiscuity values calculated by DVI and bigram analysis [59].

FUNCTIONAL SIGNIFICANCE OF PROMISCUOUS DOMAINS

The lists of the identified promiscuous domains differ according to the identification methods. However, regardless of identification method it is apparent that the majority of promiscuous domains are involved in signaling [20, 21, 28, 44]. Some domains like PH, SH3, EGF and PDZ are present in the top promiscuous domains in all of these studies. All these domains are involved in cellular signaling one way or another.

According to their functions, promiscuous domains can be classified predominantly into five categories: (a) transcription; (b) signal transduction; (c) extracellular structures/cell–cell signaling; (d) post-translational modification/chaperons/protein turnover and (e) cytoskeleton [28]. Among these categories, signal transduction and extracellular structures/cell–cell signaling are most frequent. If we calculate the number of promiscuous domains in these five categories in all the major branches of eukaryotes (Figure 3), we find that except the category of post-translational modification/chaperones/protein turnover (Figure 3B), other four most frequent categories increase non-linearly with the increase in the number of domains in the genome (Figure 3A and C–E). The linear increase in Figure 3B is largely due to the fact that post-translational modification category includes ubiquitination related domains [28]. It has been recently shown that these domains predominantly are found to be promiscuous in all branches of eukaryotes [28], and therefore, show a uniform increase in promiscuity throughout the eukaryotic kingdom. In other categories (Figure 3A and C–E), there is an initial lag period for low promiscuity, followed by an exponential increase in promiscuity. This entry to the exponential phase with higher promiscuity is due to appearance of specific clades. In the case of extracellular structures/cell–cell signaling (Figure 3D), the entry into exponential promiscuity coincides with the appearance of multicellularity. In other categories (Figure 3A, C and E), the entry into the exponential phase coincides with the appearance of animals. As described in the previous study [28], it is obvious that promiscuity is a feature that has a strong functional component and might be largely dictated by functional requirements of an organism.

ROLE OF PROMISCUOUS DOMAINS IN EVOLUTION

If we observe the distribution of promiscuous domains in three major branches of eukaryotes, animals, plants and fungi, we find that there is a small set of core domains that are promiscuous in all these three branches of life. These core domains are largely involved in biological features that are fundamental to eukaryotic cells, such as chromatin remodeling (PHD, SET, BROMO, CHROMO, BRCT and in part AAA + ATPase) and ubiquitin signaling (RING, UBQ, UCH and UBA) [28]. Moreover, most of these core promiscuous domains are involved in signaling processes in cells (Table 1) [28]. Additionally, there are domains that are promiscuous in specific lineages. Domains that are required for specific biological functions in specific lineage tend to get more promiscuous. Prominent examples are EGF, a domain involved in various forms of extracellular signaling, is promiscuous in animals, and fCBD, a domain involved in cellulose-binding, is promiscuous in fungi [28].

Promiscuity values of the protein domains can be used as an evolutionary character in eukaryotes. Using parsimony, we reconstructed the evolutionary scenario of promiscuity in the major eukaryotic lineage [28]. We found that promiscuity is a volatile character in evolution. Some evolutionary conserved combinations of domains act as a reservoir from which new lineage-specific domain combinations are created [28]. Over all, very few domains have retained their promiscuity status during evolution. Using the unikont-bikont tree topology [60], we found two domains, AAA+ ATPase and BROMO were likely to be promiscuous in the last universal common ancestor of all the analyzed eukaryotic species (LECA; Figure 4). The major gain of promiscuity happened at the base of animals, where 22 domains became promiscuous. In general, there is tendency of increase in promiscuity during eukaryotic evolution [28]. Domain promiscuity can also be used as a genome level feature to reconstruct phylogenetic trees at the genome scale. A phylogenetic tree constructed using promiscuity bears strong resemblance to the existing phylogenetic trees with minor differences [28].

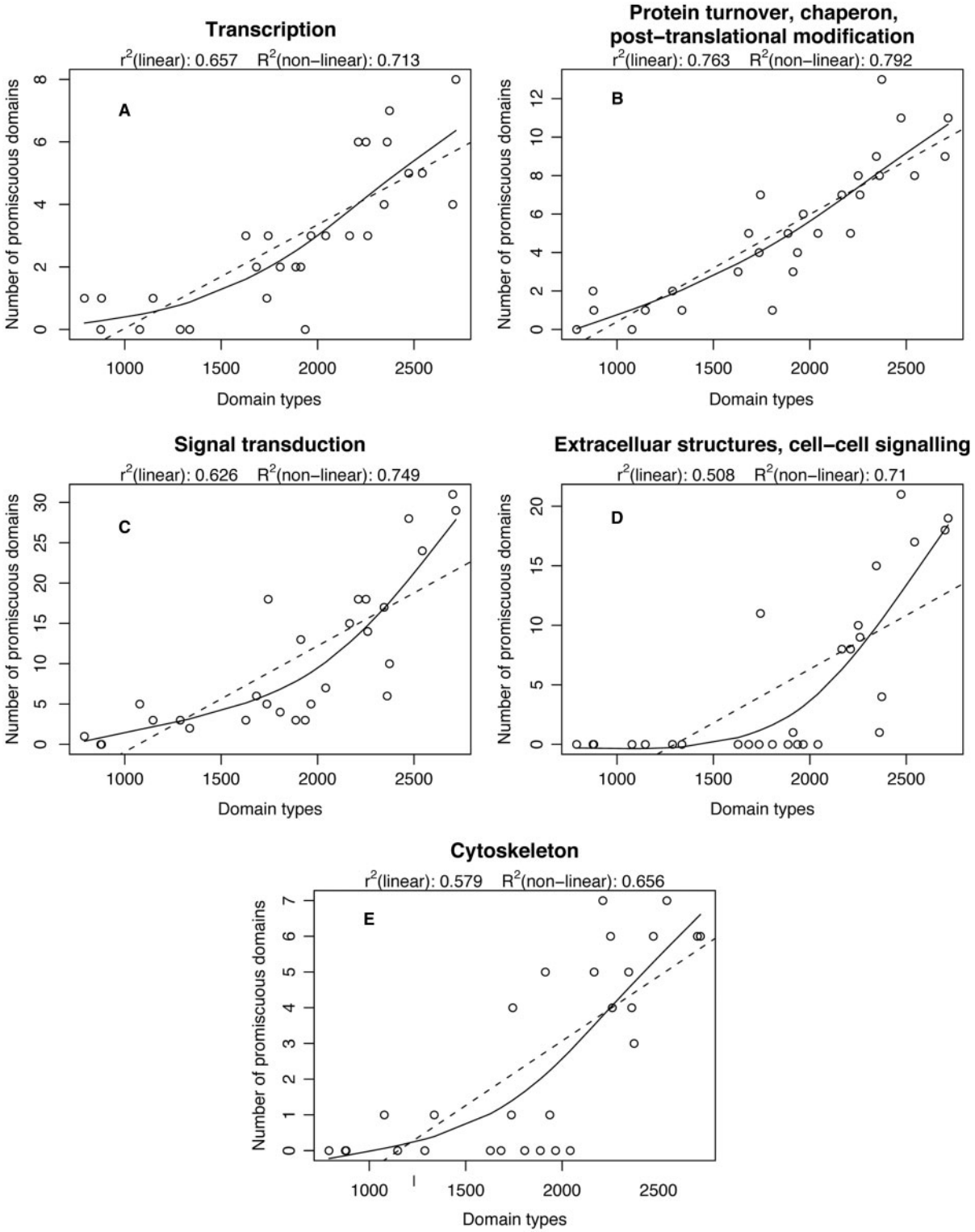

Figure 3: Increase in promiscuous domains in 28 eukaryotic organisms (see [28] for detailed list of the organisms). The organisms are sorted with the increasing number of domain types in the genome and plotted on the *X*-axis. The number of promiscuous domains belonging to the five major categories in each organism is plotted on the *Y*-axis. Each plot represents one category; the category is mentioned on top of each plot. The goodness-of-fit measures for both linear and non-linear fit are also mentioned on top of each plot.

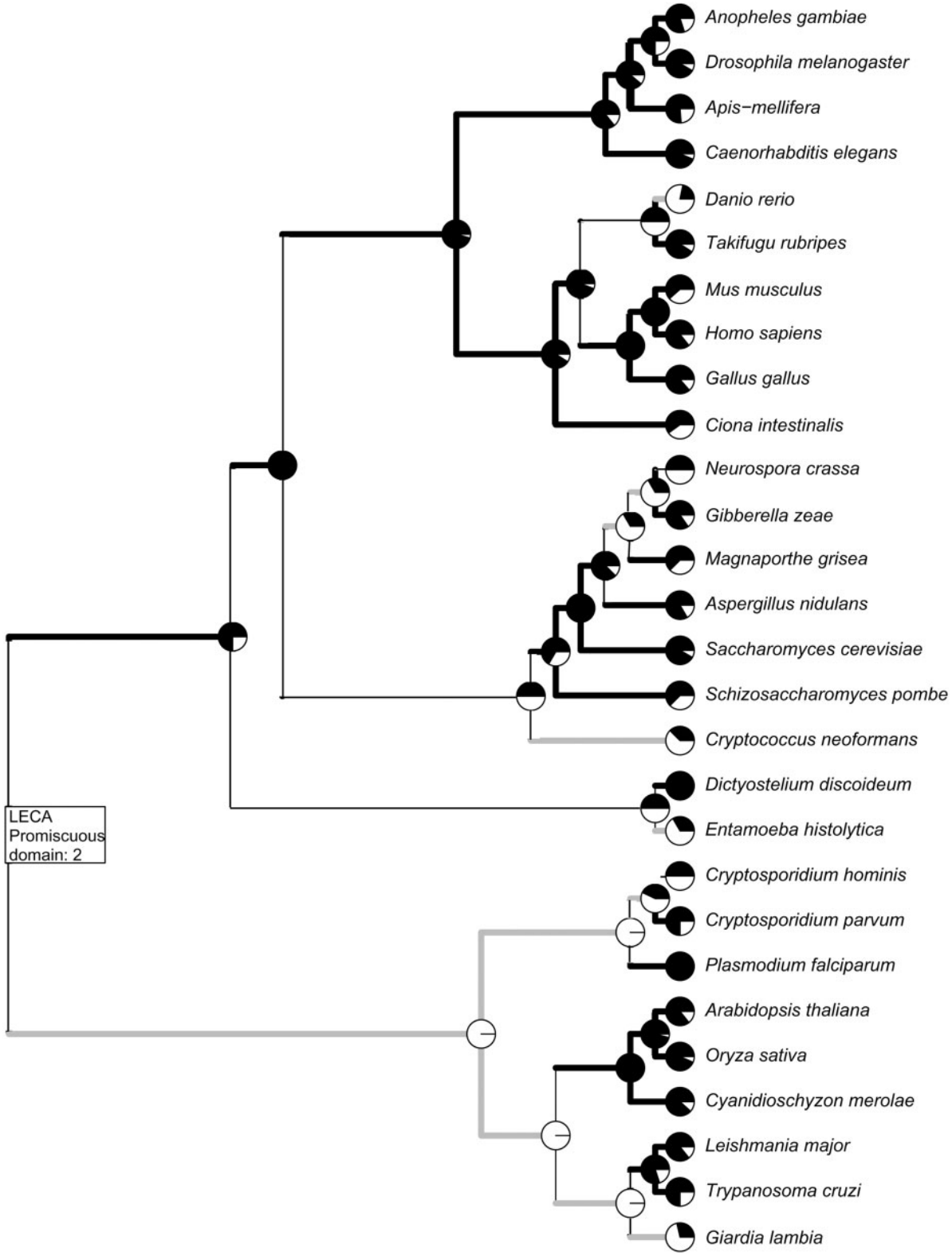

Figure 4: Ancestral reconstruction of domain promiscuity in 28 eukaryotes. The tree topology is from unikont–opisthokont tree [60], and the ancestral reconstruction was created using parsimony with binary character of promiscuity for each domain. Each node is marked with a pie diagram containing gain of promiscuity in black, and loss of promiscuity in white; the gain and loss are relative to the parent node. Each pie diagram shows the fraction of domains that gained or lost promiscuity status. Additionally, each branch is colored according to the overall gain or loss in that branch; thick black lines indicate branches that gained promiscuous domains, and thick grey lines indicate branches that lost promiscuous domains.

CONCLUSIONS

Domain combinations in protein sequences are important biological and evolutionary features. We have only very recently begun to understand the evolution of protein domain architecture. Despite the evidences of domain gain and loss in various organisms, the mechanism through which these dynamics are achieved is largely unknown. Analysis of promiscuous/mobile domains might elucidate the biological mechanisms of how domains are gained in proteins.

There are several genetic mechanisms creating new domain combinations: genetic recombination, exon-shuffling, involvement of transposable elements, etc. We have little evidence of direct involvement of any such mechanism. The contributions of each of these mechanisms are unknown. The probability of joining one given domain type to another largely depends on the probability of genetic change leading to new combinations, and probability of the fixation of the new domain combinations [20]. Minor but a significant portion (up to 12% depending on methods used) [26, 27] of domain combinations in the genome has been shown to be created through convergent evolution, which suggests that selection does play a role in shaping domain combinations. Moreover, we have now moderately good evidence of the functional role of new domain combinations in a lineage specific manner, and therefore, it is not unreasonable to conclude that newly gained domains are fixed though natural selection. More studies are needed before any comprehensive theory of domain combination of protein can be reached. Two independent studies, one from our group [28] and one from Weiner and co-workers [59] taking domain abundance into consideration, came to similar lists of promiscuous domains. This suggests that the identification of promiscuous domains is reliable. However, contradictions remain. For example, Werner and co-workers found that contrary to previously reported findings, the versatility is lower in eukaryotes. The difference is small, but statistically significant [59].

The identification of promiscuous domains has practical applications for comparative and evolutionary genomics. In particular, presence of these domains may be taken into account for sequence comparisons aimed at identification of clusters of orthologous genes, in order to avoid errors in ortholog assignment. For example, the sequences of these domains can be masked. By introducing objective, quantitative measures of domain promiscuity, a rational basis for such a filtering procedure can be designed.

Key Points

- Protein domain promiscuity is a volatile feature in evolution and plays specific functional roles in different phylogenetic lineages.
- Promiscuous domains are, typically, involved in protein–protein interactions and play crucial roles in interaction networks, particularly those that contribute to signal transduction.
- Genetic mechanism(s) shaping domain promiscuity is largely unknown, but we have strong evidence of natural selection shaping promiscuity.

Acknowledgements

We thank Kasturi Mitra, Dr Susan Gentleman and Charlie Shenitz for carefully reading and proof-reading the manuscript. This work was supported in part by the Intramural Research Program of the National Institutes of Health/DHHS.

SUPPLEMENTARY DATA

Lists of promiscuous domains in major eukaryotic lineages and other information are available at http://www.ncbi.nlm.nih.gov/CBBresearch/Koonin/resources/malay/bib2008/.

References

1. Koonin EV, Wolf YI, Karev GP. The structure of the protein universe and genome evolution. *Nature* 2002;**420**:218–23.
2. Doolittle RF. The multiplicity of domains in proteins. *Ann Rev Biochem* 1995;**64**:287–314.
3. Finn RD, Tate J, Mistry J, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2008;**36**:D281–8.
4. Schultz J, Milpetz F, Bork P, *et al*. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci USA* 1998;**95**:5857–64.
5. Marchler-Bauer A, Anderson JB, Derbyshire MK, *et al*. CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* 2007;**35**:D237–40.
6. Mulder NJ, Apweiler R, Attwood TK, *et al*. New developments in the InterPro database. *Nucleic Acids Res* 2007;**35**:D224–8.
7. Murzin AG, Brenner SE, Hubbard T, *et al*. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247**:536–40.
8. Servant F, Bru C, Carrere S, *et al*. ProDom: automated clustering of homologous domains. *Brief Bioinform* 2002;**3**:246–51.

9. Holm L, Sander C. Dictionary of recurrent domains in protein structures. *Proteins* 1998;**33**:88–96.
10. Orengo CA, Michie AD, Jones S, *et al*. CATH–a hierarchic classification of protein domain structures. *Structure* 1997;**5**: 1093–108.
11. Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. *Bioinformatics* 2000;**16**:613–8.
12. Cohen-Gihon I, Lancet D, Yanai I. Modular genes with metazoan-specific domains have increased tissue specificity. *Trends Genet* 2005;**21**:210–13.
13. Vinogradov AE. Compactness of human housekeeping genes: selection for economy or genomic design? *Trends Genet* 2004;**20**:248–53.
14. Apic G, Gough J, Teichmann SA. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 2001;**310**:311–25.
15. Liu J, Rost B. CHOP proteins into structural domain-like fragments. *Proteins* 2004;**55**:678–88.
16. Gerstein M, Levitt M. Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. *Protein Sci* 1998;**7**:445–56.
17. Ekman D, Bjorklund AK, Frey-Skott J, *et al*. Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J Mol Biol* 2005;**348**:231–43.
18. Wolf YI, Brenner SE, Bash PA, *et al*. Distribution of protein folds in the three superkingdoms of life. *Genome Res* 1999;**9**: 17–26.
19. Novozhilov AS, Karev GP, Koonin EV. Biological applications of the theory of birth-and-death processes. *Brief Bioinform* 2006;**7**:70–85.
20. Tordai H, Nagy A, Farkas K, *et al*. Modules, multidomain proteins and organismic complexity. *FEBS J* 2005;**272**: 5064–78.
21. Wuchty S. Scale-free behavior in protein domain networks. *Mol Biol Evol* 2001;**18**:1694–1702.
22. Koonin EV, Aravind L, Kondrashov AS. The impact of comparative genomics on our understanding of evolution. *Cell* 2000;**101**:573–6.
23. Hahn MW, Wray GA. The G-value paradox. *Evol Dev* 2002;**4**:73–5.
24. Marcotte EM, Pellegrini M, Ng HL, *et al*. Detecting protein function and protein–protein interactions from genome sequences. *Science* 1999;**285**:751–3.
25. Bashton M, Chothia C. The geometry of domain combination in proteins. *J Mol Biol* 2002;**315**:927–39.
26. Gough J. Convergent evolution of domain architectures (is rare). *Bioinformatics* 2005;**21**:1464–71.
27. Forslund K, Henricson A, Hollich V, *et al*. Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol* 2008;**25**:254–64.
28. Basu MK, Carmel L, Rogozin IB, *et al*. Evolution of protein domain promiscuity in eukaryotes. *Genome Res* 2008;**18**: 449–61.
29. Kuznetsov VA. *Computational and Statistical Approaches to Genomics*. Kluwer, Boston, 2002.
30. Karev GP, Wolf YI, Rzhetsky AY, *et al*. Birth and death of protein domains: a simple model of evolution explains power law behavior. *BMC Evol Biol* 2002;**2**:18.
31. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science* 1999;**286**:509–512.
32. Bilke S, Peterson C. Topological properties of citation and metabolic networks. *Phys Rev E Stat Nonlin Soft Matter Phys* 2001;**64**:036106.
33. Barabasi A.-L. *Linked: The New Science of Networks*. Perseus Publishing, 2002.
34. Albert R, Barabasi AL. Statistical mechanics of complex networks. *Rev Mod Phys* 2002;**74**:47–97.
35. Gisiger T. Scale invariance in biology: coincidence or footprint of a universal mechanism? *Biol Rev Camb Philos Soc* 2001;**76**:161–209.
36. Jeong H, Tombor B, Albert R, *et al*. The large-scale organization of metabolic networks. *Nature* 2000;**407**:651–4.
37. Huynen MA, van Nimwegen E. The frequency distribution of gene family sizes in complete genomes. *Mol Biol Evol* 1998;**15**:583–9.
38. Luscombe NM, Qian J, Zhang Z, *et al*. The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. *Genome Biol* 2002;**3**:RESEARCH0040.
39. Qian J, Luscombe NM, Gerstein M. Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *J Mol Biol* 2001;**313**:673–681.
40. Zipf GK. Human Behaviour and the Principle of Least Effort. Addison-Wesley, Boston, 1949.
41. Pareto V. Cours d'Economie Politique. Rouge et Cie, Paris, 1897.
42. Jeong H, Mason SP, Barabasi AL, *et al*. Lethality and centrality in protein networks. *Nature* 2001;**411**:41–2.
43. Wuchty S, Almaas E. Evolutionary cores of domain co-occurrence networks. *BMC Evol Biol* 2005;**5**:24.
44. Ye Y, Godzik A. Comparative analysis of protein domain organization. *Genome Res* 2004;**14**:343–53.
45. Long M, Betran E, Thornton K, *et al*. The origin of new genes: glimpses from the young and old. *Nat Rev Genet* 2003;**4**:865–75.
46. Bjorklund AK, Ekman D, Light S, *et al*. Domain rearrangements in protein evolution. *J Mol Biol* 2005;**353**: 911–23.
47. Weiner J, 3rd, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the modular protein evolution. *FEBS J* 2006;**273**:2037–47.
48. Bjorklund AK, Ekman D, Elofsson A. Expansion of protein domain repeats. *PLoS Comput Biol* 2006;**2**:e114.
49. Weiner J, 3rd, Bornberg-Bauer E. Evolution of circular permutations in multidomain proteins. *Mol Biol Evol* 2006; **23**:734–43.
50. Patthy L. Genome evolution and the evolution of exon-shuffling – a review. *Gene* 1999;**238**:103–14.
51. Patthy L. Modular assembly of genes and the evolution of new functions. *Genetica* 2003;**118**:217–31.
52. Liu M, Walch H, Wu S, *et al*. Significant expansion of exon-bordering protein domains during animal proteome evolution. *Nucleic Acids Res* 2005;**33**:95–105.
53. Rogozin IB, Sverdlov AV, Babenko VN, *et al*. Analysis of evolution of exon–intron structure of eukaryotic genes. *Brief Bioinform* 2005;**6**:118–34.
54. Gilbert W. The exon theory of genes. *Cold Spring Harb Symp Quant Biol* 1987;**52**:901–5.

55. Gilbert W, Glynias M. On the ancient nature of introns. *Gene* 1993;**135**:137–44.

56. Patthy L. Exons – original building blocks of proteins? *Bioessays* 1991;**13**:187–92.

57. Rzhetsky A, Gomez SM. Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. *Bioinformatics* 2001;**17**:988–96.

58. Manning CaS, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

59. Weiner 3rd J, Moore AD, Bornberg-Bauer E. Just how versatile are domains? *BMC Evol Biol* 2008;**8**:285.

60. Stechmann A, Cavalier-Smith T. The root of the eukaryote tree pinpointed. *Curr Biol* 2003;**13**:R665–66.