# Chapter 8

# Evolution of Protein Domain Architectures

Kristoffer Forslund and Erik L.L. Sonnhammer

## Abstract

This chapter reviews the current research on how protein domain architectures evolve. We begin by summarizing work on the phylogenetic distribution of proteins, as this directly impacts which domain architectures can be formed in different species. Studies relating domain family size to occurrence have shown that they generally follow power law distributions, both within genomes and larger evolutionary groups. These findings were subsequently extended to multidomain architectures. Genome evolution models that have been suggested to explain the shape of these distributions are reviewed, as well as evidence for selective pressure to expand certain domain families more than others. Each domain has an intrinsic combinatorial propensity, and the effects of this have been studied using measures of domain versatility or promiscuity. Next, we study the principles of protein domain architecture evolution and how these have been inferred from distributions of extant domain arrangements. Following this, we review inferences of ancestral domain architecture and the conclusions concerning domain architecture evolution mechanisms that can be drawn from these. Finally, we examine whether all known cases of a given domain architecture can be assumed to have a single common origin (monophyly) or have evolved convergently (polyphyly).

**Key words:** Protein domain, Protein domain architecture, Superfamily, Monophyly, Polyphyly, Convergent evolution, Domain evolution, Kingdoms of life, Domain co-occurrence network, Node degree distribution, Power law, Parsimony

## 1. Introduction

### 1.1. Overview

By studying the domain architectures of proteins, we can understand their evolution as a modular phenomenon, with high-level events enabling significant changes to take place in a time span much shorter than required by point mutations only. This research field has become possible only now in the -omics era of science, as both identifying many domain families in the first place and acquiring enough data to chart their evolutionary distribution require access to many completely sequenced genomes. Likewise, the conclusions drawn generally consider properties averaged for entire species or organism groups or entire classes of proteins, rather than properties of single genes.

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*,
Methods in Molecular Biology, vol. 856, DOI 10.1007/978-1-61779-585-5_8,

We begin by introducing the basic concepts of domains and domain architectures, as well as the biological mechanisms by which these architectures can change. The remainder of the chapter is an attempt at answering, from the recent literature, the question of which forces shape domain architecture evolution and in what direction. The underlying issue concerns whether it is fundamentally a random process or whether it is primarily a consequence of selective constraints.

### 1.2. Protein Domains

Protein domains are high-level parts of proteins that either occur alone or together with partner domains on the same protein chain. Most domains correspond to tertiary structure elements, and are able to fold independently. All domains exhibit evolutionary conservation, and many either perform specific functions or contribute in a specific way to the function of their proteins. The word domain strictly refers to a distinct region of a specific protein, an instance of a domain family. However, domain and domain family are often used interchangeably in the literature.

### 1.3. Domain Databases

By identifying recurring elements in experimentally determined protein 3D structures, the various domain families in structural domain databases, such as SCOP (1) and CATH (2), were gathered. New 3D structures allow assignment to these classes from semiautomated inspection. The SUPERFAMILY (3) database assigns SCOP domains to all protein sequences by matching them to Hidden Markov Models (HMMs) that were derived from SCOP superfamilies, i.e., proteins whose evolutionary relationship is evidenced structurally. The Gene3D (4) database is similarly constructed, but based on domain families from CATH.

This approach resembles the methodology used in pure sequence-based domain databases, such as Pfam (5). In these databases, conserved regions are identified from sequence analysis and background knowledge to make multiple sequence alignments. From these, HMMs are built that are used to search new sequences for the presence of the domain represented by each HMM. All such instances are stored in the database. The HMM framework ensures stability across releases and high quality of alignments and domain family memberships. The stability allows annotation to be stored along with the HMMs and alignments. The INTERPRO database (6) is a metadatabase of domains combining the assignments from several different source databases, including Pfam. The Conserved Domain Database (CDD) is a similar metadatabase that also contains additional domains curated by the NCBI (7). SMART (8) is a manually curated resource focusing primarily on signaling and extracellular domains. ProDom (9) is a comprehensive domain database automatically generated from sequences in UniProt (10). Likewise, ADDA (11) is automatically generated by clustering subsequences of proteins from the major sequence databases. It is currently being used for generating Pfam-B families, low-fidelity sets of putative domains which may provide starting points for new Pfam-A families. Such automatic approaches, however, inevitably produce low-quality domain definitions and alignments, and lack annotation.

Since the domain definitions from different databases only partially overlap, results from analyses often cannot be directly compared. In practice, however, choice of database appears to have little effect on the main trends reported by the studies described here.

### 1.4. Domain Architectures

The term "domain architecture" or "domain arrangement" generally refers to the domains in a protein and their order, reported in N- to C-terminal direction along the amino acid chain. Another recurring term is domain combinations. This refers to pairs of domains co-occurring in proteins, either anywhere in the protein (the "bag-of-domains" model) or specifically pairs of domains being adjacent on an amino acid chain, in a specific N- to C-terminal order (12). The latter concept is expanded to triplets of domains, which are subsequences of three consecutive domains, with the N- and C-termini used as "dummy" domains. A domain X occurring on its own in a protein, thus, produces the triplet N-X-C (13).

### 1.5. Mechanisms for Domain Architecture Change

Most mutations are point mutations: substitutions, insertions, or deletions of single nucleotides. While conceivably enough of these might create a new domain from an old one or noncoding sequence or remove a domain from a protein, in practice we are interested in mechanisms, whereby the domain architecture of a protein changes instantly or nearly so. Figure 1 shows some examples of ways in which domain architectures may mutate. In general, adding or removing domains requires genetic recombination events. These can occur either through errors made by systems for repairing DNA damage, such as homologous (14, 15) or nonhomologous (illegitimate) (16, 17) recombination, or through the action of mobile genetic elements, such as DNA transposons (18) or retrotransposons (19, 20). Recombination can cause loss or duplication of parts of genes, entire genes, or much longer chromosomal regions.

In organisms that have introns, exon shuffling (21, 22) refers to the integration of an exon from one gene into another, for instance through chromosomal crossover, gene conversion, or mobile genetic elements. Exons could also be moved around by being brought along by mobile genetic elements, such as retrotransposons (22, 23).

Two adjacent genes can be fused into one if the first one loses its transcription stop signals. Point mutations can cause a gene to lose a terminal domain by introducing a new stop codon, after which the "lost" domain slowly degrades through point mutations as it is no longer under selective pressure (24). Alternatively, a multidomain gene might be split into two genes if both a start

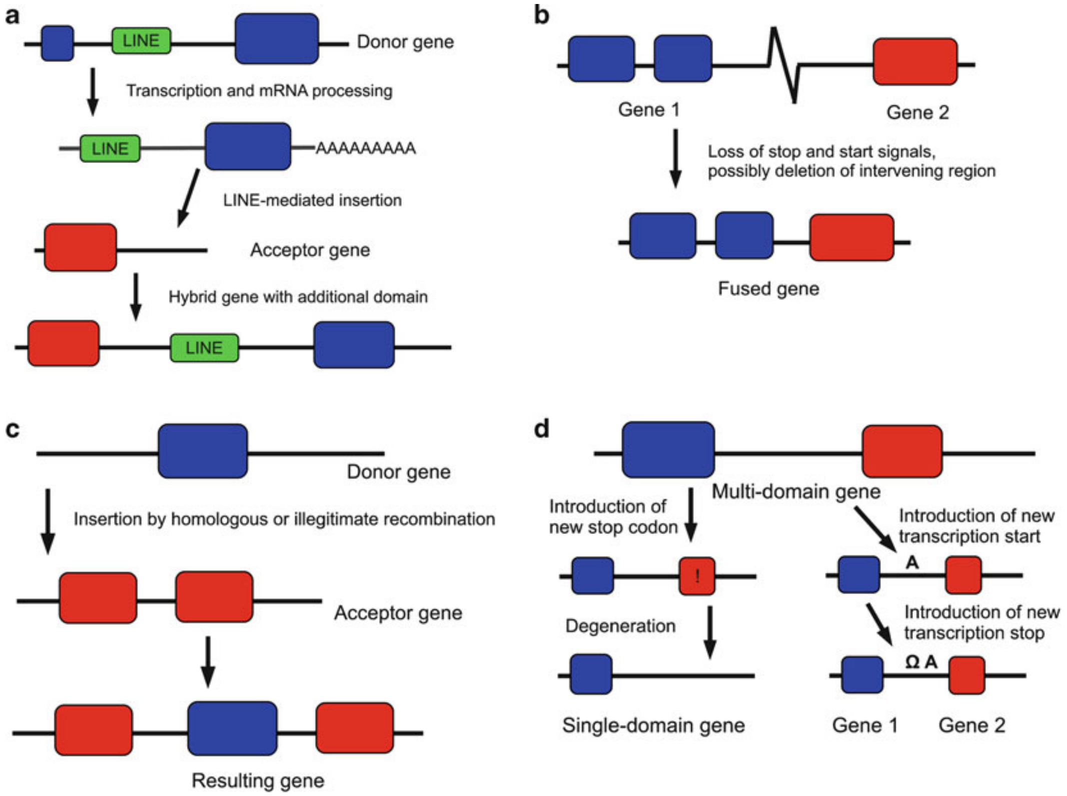
a
LINE
Donor gene
Transcription and mRNA processing
LINE
AAAAAAAAA
LINE-mediated insertion
Acceptor gene
Hybrid gene with additional domain
LINE
b
Gene 1
Gene 2
Loss of stop and start signals, possibly deletion of intervening region
Fused gene
c
Donor gene
Insertion by homologous or illegitimate recombination
Acceptor gene
Resulting gene
d
Multi-domain gene
Introduction of new stop codon
Introduction of new transcription start
A
!
Degeneration
Introduction of new transcription stop
Ω A
Single-domain gene
Gene 1
Gene 2

Fig. 1. Examples of mutations that can change domain architectures. Adapted from Buljan and Bateman (BioMed Central, 2010). (**a**) Gene fusion by a mobile element. LINE refers to a Long Interspersed Nuclear repeat Element, a retrotransposon. The reverse transcriptase encoded within the LINE causes its mRNA to be reverse transcribed into DNA and integrated into the genome, making the domain-encoding blue exon from the donor gene integrate along with it in the acceptor gene. (**b**) Gene fusion by loss of a stop signal or deletion of much of the intergenic region. Genes 1 and 2 are joined together into a single, longer gene. (**c**) Domain insertion through recombination. The blue domain from the donor gene is inserted within the acceptor gene by either homologous or illegitimate recombination. (**d**) *Right*: Gene fission by introduction of transcription stop (the letter Ω) and start (the letter A). *Left*: Domain loss by introduction of a stop codon (exclamation mark) with subsequent degeneration of the now untranslated domain.

and a stop signal are introduced between the domains. Novel domains could arise, for instance, through exonization, whereby an intronic or intergenic region becomes an exon, after which subsequent mutations would fine tune its folding and functional properties (23, 25).

## 2. Distribution of the Sizes of Domain Families

Domain architectures are fundamentally the realizations of how domains combine to form multidomain proteins with complex functions. Understanding how these combinations come to be requires first that we understand how common the constituent domains of those architectures are, and whether there are selective pressures determining their abundances. Because of this, the body of work concerning the sizes and species distributions of domain families becomes important to us.

Comprehensive studies of the distributions and evolution of protein domains and domain architectures are possible as genome sequencing technologies have made many entire proteomes available for bioinformatic analysis. Initial work (26–28) focused on the number of copies that a protein family, either single domain or multidomain, has in a species. Most conclusions from these early studies appear to hold true for domains, supradomains (see below), and domain architectures (29–31). In particular, these all exhibit a "*dominance of the population by a selected few*" (28), i.e., a small number of domain families are present in a majority of the proteins in a genome, whereas most domain families are found only in a small number of proteins.

Looking at the frequency $N$ of families of size $X$ (defined as the number of members in the genome), in the earliest studies, this frequency was modeled as the power law

$$N = cX^{-a},$$

where $a$ is a slope parameter. The power law is a special case of the generalized Pareto distribution (GPD) (32):

$$N = c(i + X)^{-a}.$$

Power law distributions arise in a vast variety of contexts: from human income distributions, connectivity of Internet routers, word usage in languages, and many other situations ((27, 28, 34, 35), see also ref. 36 for a conflicting view). Luscombe et al. (28) described a number of other genomic properties that also follow power law distributions, such as the occurrence of DNA "words," pseudogenes, and levels of gene expression. These distributions fit much better than the alternative they usually are contrasted against, an exponential decay distribution. The most important difference between exponential and power law distributions in this context concerns the fact that the latter has a "fat tail," that is, while most domain families occur only a few times in each proteome, most domains in the proteome still belong to one of a small number of families.

Later work ((32, 37), see also ref. 38) demonstrated that proteome-wide domain occurrence data fit the general GPD better than the power law, but that it also asymptotically fits a power law as $X \gg i$. The deviation from strict power law behavior depends on proteome size in a kingdom-dependent manner (37). Regardless, it is mostly appropriate to treat the domain family size distribution as approximately (and asymptotically) power law like, and later studies typically assume this.

The power law, but not the GPD, is scale free in the sense of fulfilling the condition

$$f(ax) = g(a)f(x),$$

where $f(x)$ and $g(x)$ are some functions of a variable $x$, and $a$ is a scaling parameter, that is, studying the data at a different scale does not change the shape of function. This property has been extensively studied in the literature and is connected to other attributes, notably when it occurs in network degree distributions (i.e., frequency distributions of edges per node). Here, it has been associated with properties, such as the presence of a few central and critical hubs (nodes with many edges to other nodes), the similarity between parts and the whole (as in a fractal), and the growth process called preferential attachment, under which nodes are more likely to gain new links the more links they already have. However, the same power law distribution may be generated from many different network topologies with different patterns of connectivity. In particular, they may differ in the extent that hubs are connected to each other (36). It is possible to extend the analysis by taking into account the distribution of degree pairs along network edges, but this is normally not done.

What kind of evolutionary mechanisms give rise to this kind of distribution of gene or domain family sizes within genomes? In one model by Huynen and van Nimwegen (26), every gene within a gene family is more or less likely to duplicate, depending on the utility of the function of that gene family within the particular lineage of organisms studied, and they showed that such a model matches the observed power laws. While they claimed that any model that explains the data must take into account family-specific probabilities of duplication fixation, Yanai and coworkers (39) proposed a simpler model using uniform duplication probability for all genes in the genome, and also reported a good fit with the data.

Later, more complex birth–death (37) and birth–death-and-innovation models (BDIM) (27, 32) were introduced to explain the observed distributions, and from investigating which model parameter ranges allow this fit the authors were able to draw several far-ranging conclusions. First, the asymptotic power law behavior requires that the rates of domain gain and loss are asymptotically equal. Karev et al. (32) interpreted this as support for a punctuated equilibrium-type model of genome evolution, where domain family size distributions remain relatively stable for long periods of time but may go through stages of rapid evolution, representing a shift between different BDIM evolutionary models and significant changes in genome complexity. Like Huynen and van Nimwegen (26), they concluded that the likelihood of fixated domain duplications or losses in a genome directly depends on family size. The family, however, only grows as long as new copies can find new functional niches and contribute to a net benefit for survival, i.e., as long as selection favors it.

Aside from Huynen and van Nimwegen's, none of the models discussed depend very strongly on family-specific selection to explain the abundances of individual gene families, nor do they exclude such selection. Some domains may be highly useful to their host organism's lifestyle, such as cell–cell connectivity domains to an organism beginning to develop multicellularity. Expansion of these domain families might, therefore, become more likely in some lineages than in others. To what extent these factors actually affect the size of domain families remains to be fully explored. Karev et al. (32) suggested that the rates of domain-level change events themselves—domain duplication and loss rates, as well as the rate of influx of novel domains from other species or de novo creation—must be evolutionarily adapted, as only some such parameters allow the observed distributions to be stable. van Nimwegen (40) investigated how the number of genes increases in specific functional categories as total genome size increases. He found that the relationship matches a power law, with different coefficients for each functional class remaining valid over many bacterial lineages. Ranea et al. (41) found similar results. Also, Ranea et al. (42) showed that, for domain superfamilies inferred to be present in the last universal common ancestor (LUCA), domains associated with metabolism have significantly higher abundance than those associated with translation, further supporting a connection between the function of a domain family and how likely it is to expand.

Extending the analysis to multidomain architectures, Apic et al. (30) showed that the frequency distribution of multidomain family sizes follows a power law curve similar to that reported for individual domain families. It, therefore, seems likely that the basic underlying mechanisms should be similar in both cases, i.e., duplication of genes, and thus their domain architectures, is the most important type of event affecting the evolution of domain architectures.

Have the trends described above stood the test of time as more genomes have been sequenced and more domain families have been identified? We considered the 1,503 complete proteomes in version 24.0 of Pfam, and plotted the frequency $Y$ of domain families that have precisely $X$ members as a function of $X$, and fit a power law curve to this. Figure 2a shows the resulting plots for three representative species, one complex eukaryote (*Homo sapiens*), one simple eukaryote (*Saccharomyces cerevisiae*), and one prokaryote (*Escherichia coli*). Figure 2b shows the corresponding plots for all domains in all complete eukaryotic, bacterial, and archaeal proteomes. The power law curve fits decently well, with slopes becoming less steep for the more complex organisms, whose distributions have relatively more large families. The power law-like behavior suggests that complex organisms with large proteomes were formed by heavily duplicating domains from relatively few families. Figure 3a and b show equivalent plots, not for single

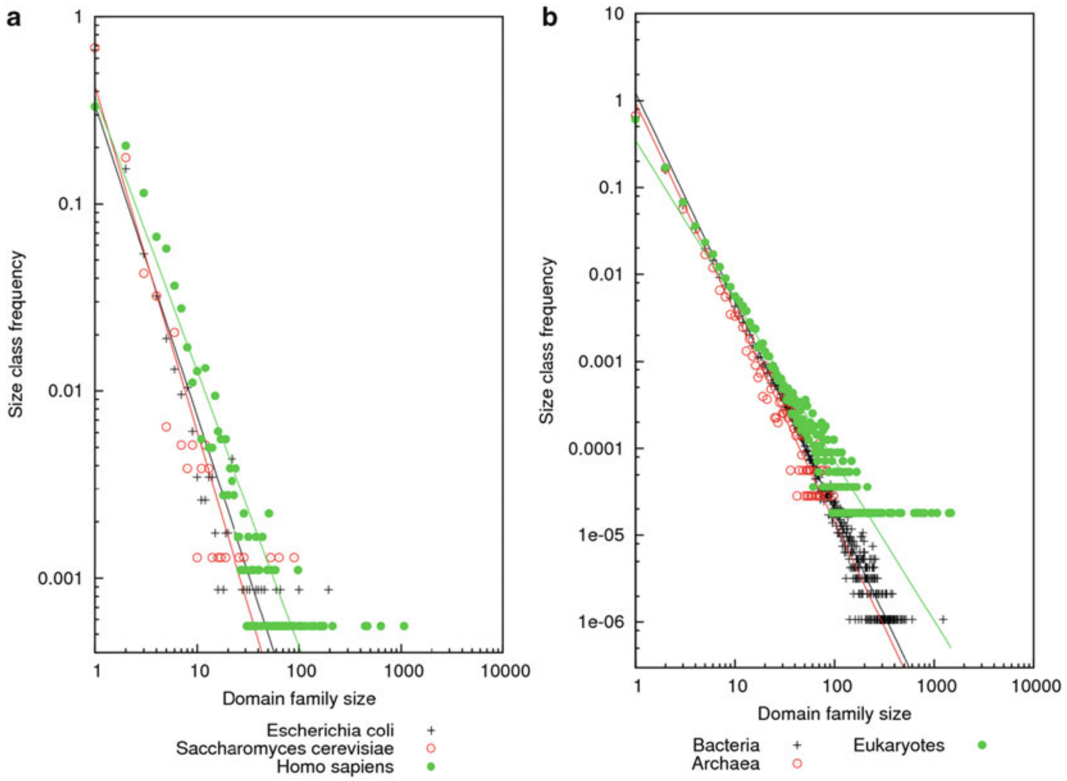

Fig. 2. (**a**) Distribution of domain family sizes in three selected species. Power law distributions were fitted to these curves such that, for frequency *f* of families of size *X*, $f = cX^a$. For *Saccharomyces cerevisiae*, $a = -1.8$, for *Escherichia coli*, $a = -1.7$, and for *Homo sapiens*, $a = -1.5$. (**b**) Distribution of domain family sizes across the three kingdoms. Power law distributions were fitted to these curves such that, for frequency *f* of families of size *X*, $f = cX^a$. For bacteria, $a = -2.4$, for archaea, $a = -2.4$, and for eukaryotes, $a = -1.8$.

domains but for entire multidomain architectures. The curve shapes as well as the relationship between both species and organism groups are similar, indicating that the evolution of these distributions have been similar.

## 3. Kingdom and Age Distribution of Domain Families and Architectures

How old are specific domain families or domain architectures? With knowledge of which organism groups they are found in, it is possible to draw conclusions about their age, and whether lineage-specific selective pressures have determined their kingdom-specific abundances. Domain families as well as their combinations have arisen throughout evolutionary history, presumably by new combinations of preexisting elements that may have diverged beyond recognition or by processes, such as exonization. We can estimate the age of a domain family by finding the largest clade of organisms within which it is found, excluding organisms with only xenologs,

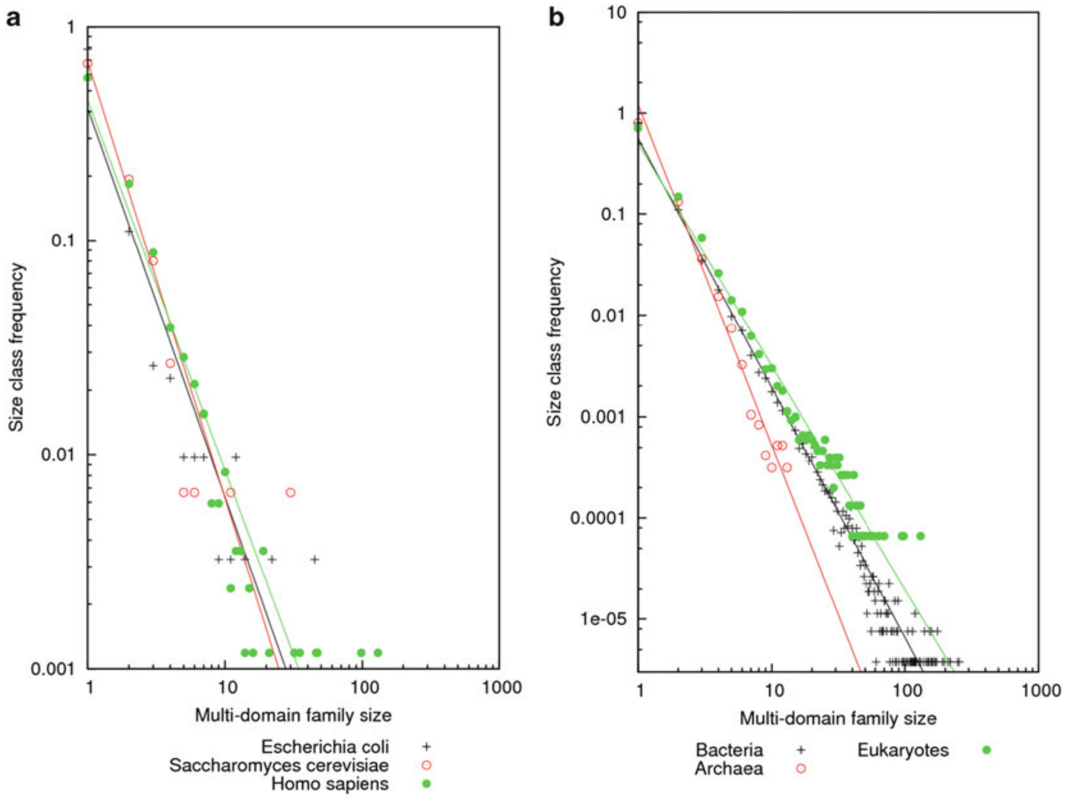

Fig. 3. (**a**) Distribution of multidomain (architecture) family sizes in three selected species. Power law distributions were fitted to these curves such that, for frequency *f* of families of size *X*, $f = cX^a$. For *Saccharomyces cerevisiae*, $a = -2.0$, for *Escherichia coli*, $a = -1.8$, and for *Homo sapiens*, $a = -1.7$. (**b**) Distribution of multidomain (architecture) family sizes across the three kingdoms. Power law distributions were fitted to these curves such that, for frequency *f* of families of size *X*, $f = cX^a$. For bacteria, $a = -2.5$, for archaea, $a = -3.4$, and for eukaryotes, $a = -2.2$.

i.e., horizontally transferred genes (13). The age of this lineage's root is the likely age of the family. The same holds true for domain combinations and entire domain architectures. This methodology allows us to determine how changing conditions at different points in evolutionary history, or in different lineages, have affected the evolution of domain architectures.

Apic et al. (29) analyzed the distribution of SCOP domains across 40 genomes from archaea, bacteria, and eukaryotes. They found that a majority of domain families are common to all three kingdoms of life, and thus likely to be ancient. Kuznetsov et al. (37) performed a similar analysis using INTERPRO domains, and found that only about one-fourth of all such domains were present in all three kingdoms, but a majority was present in more than one of them. Lateral gene transfer or annotation errors can cause a domain family to be found in one or a few species in a kingdom without actually belonging to that kingdom. To counteract this, one can require that a family must be present in at least a reasonable fraction of the species within a kingdom for it to be considered anciently present there. For instance, using Gene3D assignments of CATH domains to 114 complete genomes, mainly bacterial, Ranea et al. (42) isolated protein superfamily domains that were present in at least 90% of all the genomes and also at least 70% of the archaeal and eukaryotic genomes. Under these stringent cutoffs for considering a domain to be present in a kingdom, 140 domains, 15% of the CATH families found in at least 1 prokaryote genome, were inferred to be ancient. Chothia and Gough (43) performed a similar study on 663 SCOP superfamily domains evaluated at many different thresholds, and found that while 516 (78%) superfamilies were common to all three kingdoms at a threshold of 10% of species in each kingdom only 156 (24%) superfamilies were common to all three kingdoms at a threshold of 90%. They also showed that for prokaryotes a majority of domain instances (i.e., not domain families but actual domain copies) belong to common superfamilies at all thresholds below 90%.

Extending to domain combinations, Apic et al. (29) reported that a majority of SCOP domain pairs are unique to each kingdom, but also that more kingdom-specific domain combinations than expected were composed only of domain families shared between all three kingdoms. This would imply a scenario, where the independent evolution of the three kingdoms mainly involved creating novel combinations of domains that existed already in their common ancestor.

Several studies have reported interesting findings on domain architecture evolution in lineages closer to ourselves: in metazoa and vertebrates. Ekman et al. (44) claimed that new metazoa-specific domains and multidomain architectures have arisen roughly once every 0.1–1 million years in this lineage. According to their results, most metazoa-specific multidomain architectures are a combination of ancient and metazoa-specific domains. The latter category are, however, mostly found as novel single-domain proteins. Much of the novel metazoan multidomain architectures involve domains that are versatile (see below) and exon bordering (allowing for their insertion through exon shuffling). The novel domain combinations in metazoa are enriched for proteins associated with functions required for multicellularity—regulation, signaling, and functions involved in newer biological systems, such as immune response or development of the nervous system, as previously noted by Patthy (21). They also showed support for exon shuffling as an important mechanism in the evolution of metazoan domain architectures. Itoh et al. (45) added that animal evolution differs significantly from other eukaryotic groups in that lineage-specific domains played a greater part in creating new domain combinations.

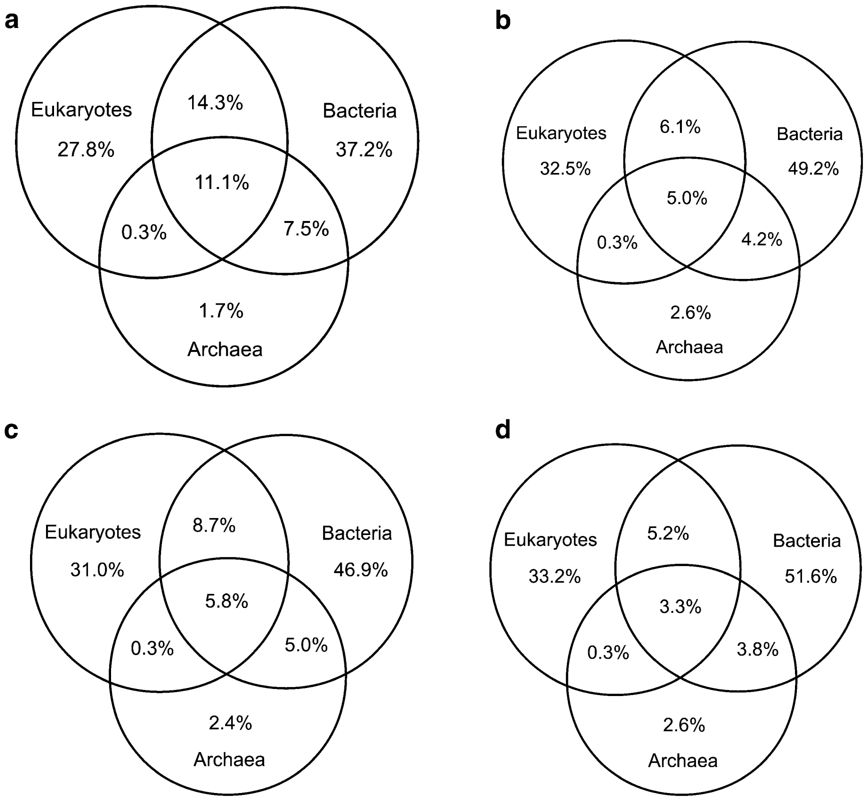

Fig. 4. (**a**) Kingdom distribution of unique domains. Values are given as percentages of the total 7,270 domains. (**b**) Kingdom distribution of unique domain pairs. Values are given as percentages of the total 6,270 domain pairs. (**c**) Kingdom distribution of unique domain triplets. Values are given as percentages of the total 20,396 domain triplets. (**d**) Kingdom distribution of unique multidomain architectures. Values are given as percentages of the total 7,862 multidomain architectures.

In the most recent datasets, what is the distribution of domains and domain combinations across the three kingdoms of life? Looking at the set of complete proteomes in version 24.0 of Pfam, the distribution of domains across the three kingdoms is as displayed in the Venn diagram of Fig. 4a. Figure 4b and c shows the equivalent distributions of immediate neighbors and triplets of domains, respectively, and Fig. 4d shows the distribution of multidomain architectures across kingdoms. The numbers are somewhat biased toward bacteria as 90% of the complete proteomes are from this kingdom. However, with this high coverage of all kingdoms (76 eukaryotic, 68 archaeal, and 1,359 bacterial proteomes), the results should be robust in this respect. Compared to most previous reports, we see a striking difference in that a much smaller portion of domains are shared between all kingdoms. There are some potential artifacts which could affect this analysis. If lateral gene transfer is very widespread, we may overestimate the number of families present in all three kingdoms. Moreover, there are cases, where separate Pfam families are actually distant homologs of each other, which could lead to underestimation of the number of ancient families. To counteract this, we make use of Pfam clans, considering domains in the same clan to be equivalent. While not all distant homologies have yet been registered in the clan system, performing the analysis on the clan level reduces the risk of such underestimation.

Our finding that 11% of all Pfam-A domains are present in all kingdoms is strikingly lower than in the earlier works, and is even lower than reported by Ranea et al. (42), who used very stringent cutoffs. However, a direct comparison of statistics for Pfam domains/clans and CATH superfamilies is difficult. The decrease in ancient families that we observe may be a consequence of the massive increase in sequenced genomes and/or that the recent growth of Pfam has added relatively more kingdom-specific domains. We further found that only 2–3% of all domains or domain combinations are unique to archaea, suggesting that known representatives of this lineage have undergone very little independent evolution and/or that most archaeal gene families have been horizontally transferred to other kingdoms. The trend when going from domain via domain combinations to whole architectures is clear—the more complex patterns are less shared between the kingdoms. In other words, each kingdom has used a common core of domains to construct its own unique combinations of multidomain architectures.

## 4. Domain Co-occurrence Networks

A multidomain architecture connects individual domains with each other. There are several ways to derive these connections and quantify the level of co-occurrence. The simplest method is to consider all domains on the same amino acid chain to be connected, but we can also limit the set of co-occurrences we consider to, e.g., immediate neighbor pairs or triplets. Regardless of which method is used, the result is a domain co-occurrence network, where nodes represent domains and where edges represent the existence of proteins in which members of these families co-occur. Figure 5 shows an example of such a network and the set of domain architectures which defines it. This type of explicit network representation is explored in several studies, notably by Itoh et al. (45), Przytycka et al. (46), and Kummerfeld and Teichmann (12). It is advantageous as it allows the introduction of powerful analysis tools developed within the engineering sciences for use with artificial network structures, such as the World Wide Web. The patterns of co-occurrences that we observe should be a direct consequence of the constraints and conditions under which

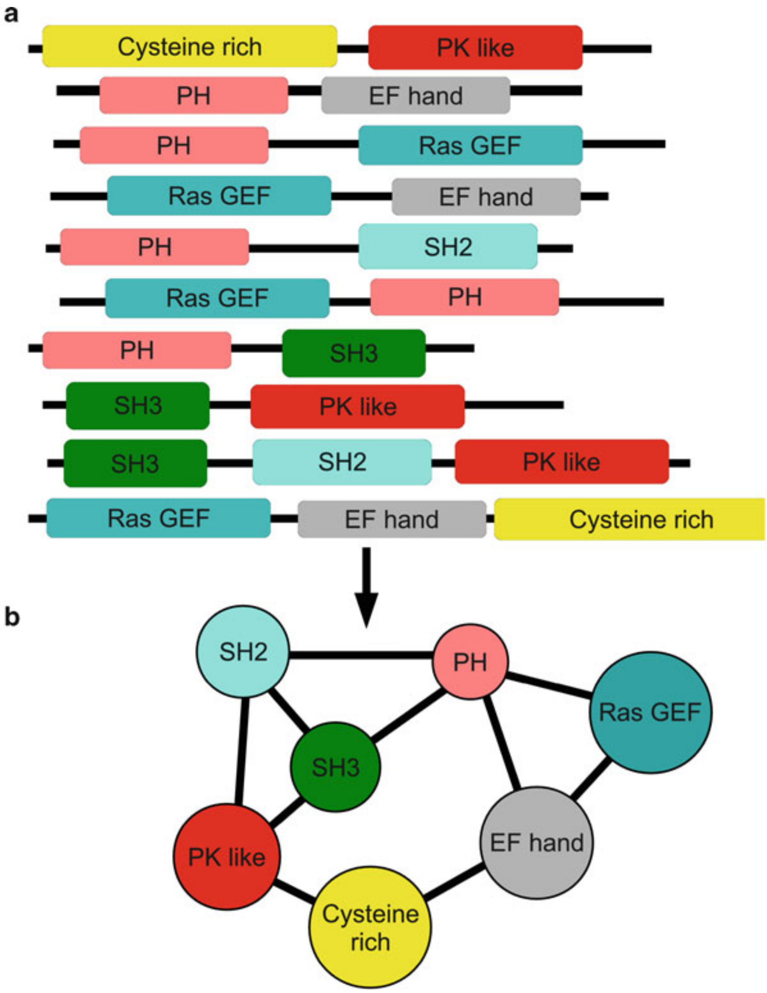

Fig. 5. Example of protein domain co-occurrence network, adapted from Kummerfeld and Teichmann (BioMed Central, 2009). (**a**) Sample set of domain architectures. *The lines* represent proteins, and *the boxes* their domains in N- to C-terminal order. (**b**) Resulting domain co-occurrence (neighbor) network. Nodes correspond to domains, and are linked by an edge if at least one domain exists, where the two domains are found adjacent to each other along the amino acid chain.

domain architectures evolve, and because of this the study of these patterns becomes relevant for understanding such factors.

The frequency distribution of node degrees in the domain co-occurrence network has been fitted to a power law (29) and a more general GPD as well (34). The closer this approximation holds, the more the network will have the scale-free property. This property can be thought of as a hierarchy in the network, where the more centrally connected nodes link to more peripheral nodes with the same relative frequency at each level. In the context of domains, this means that a small number of domains co-occur with a high number of other domains, whereas most domains only have a few neighbors—usually, some of the highly connected hubs. The most highly connected domains are referred to as promiscuous (47), mobile, or versatile (13, 48, 49). Many such hub domains are involved in intracellular or extracellular signaling, protein–protein interactions and catalysis, and transcription regulation. In general, these are domains that encode a generic function, e.g., phosphorylation, that is reused in many contexts by additional domains that confer substrate specificity or localization. Table 1 shows the domains (or clans) with the highest numbers of immediate neighbors in Pfam 24.0.

One way of evolving a domain co-occurrence network that follows a power law is by "preferential attachment" (33, 46). This means that new edges (corresponding to proteins, where two domains co-occur) are added with a probability that is higher the more edges these nodes (domains) already have, resulting in a power law distribution.

Apic et al. (30) considered a null model for random domain combination, in which a proteome contains domain combinations with a probability based on the relative abundances of the domains only. They showed that this model does not hold, and that far fewer domain combinations than expected under it are actually seen. If most domain duplication events are gene duplication events that do not change domain architecture—or at the very least, do not disrupt domain pairs—then this finding is not unexpected, nor does it require or exclude any particular selective pressure to keep these domains together in proteins. There is growing support for the idea that separate instances of a given domain architecture in general descend from a single ancestor with that architecture (50), with polyphyletic evolution of domain architectures occurring only in a small fraction of cases (46, 51, 52).

Itoh et al. (45) performed reconstruction of ancestral domain architectures using maximum parsimony, as described in the next section. This allowed them to study the properties of the ancestral domain co-occurrence network, and thus explore how network connectivity has altered over evolutionary time. Among other things, they found increased connectivity in animals, particularly of animal-specific domains, and suggest that this phenomenon explains the high connectivity for eukaryotes reported by Wuchty (34). For nonanimal eukaryotes, they reported a correlation between connectivity and age such that older domains had relatively higher connectivity, with domains preceding the divergence of eukaryotes and prokaryotes being the most highly connected, followed by early eukaryotic domains. In other words, early eukaryotic evolution saw the emergence of some key hub proteins while the most prominent eukaryotic hubs emerged in the animal lineage.

**Table 1**
**The 20 most densely connected hubs with regards to immediate domain neighbors, according to Pfam 24.0**

| Identifier | Name | Number of different immediate neighbors |
|---|---|---|
| CL0123 | Helix-turn-helix clan | 202 |
| CL0023 | P-loop containing nucleoside triphosphate hydrolase superfamily | 166 |
| CL0063 | FAD/NAD(P)-binding Rossmann fold Superfamily | 155 |
| CL0159 | Ig-like fold superfamily (E-set) | 71 |
| CL0036 | Common phosphate-binding site TIM barrel superfamily | 71 |
| CL0016 | Protein kinase superfamily | 62 |
| CL0172 | Thioredoxin like | 52 |
| CL0202 | Galactose-binding domain-like superfamily | 50 |
| CL0058 | Tim barrel glycosyl hydrolase superfamily | 50 |
| CL0125 | Peptidase clan CA | 46 |
| CL0028 | Alpha/beta hydrolase fold | 45 |
| CL0304 | CheY-like superfamily | 44 |
| CL0137 | HAD superfamily | 42 |
| PF00571 | CBS domain | 41 |
| CL0219 | Ribonuclease H-like superfamily | 41 |
| CL0010 | Src homology-3 domain | 41 |
| CL0300 | Twin-arginine translocation motif | 40 |
| CL0261 | NUDIX superfamily | 40 |
| CL0025 | His Kinase A (phospho-acceptor) domain | 39 |
| CL0183 | PAS domain clan | 38 |

What is the degree distribution of current domain co-occurrence networks? We again used the domain architectures from all complete proteomes in version 24.0 of Pfam, and considered the network of immediate neighbor relationships, i.e., nodes (domains) have an

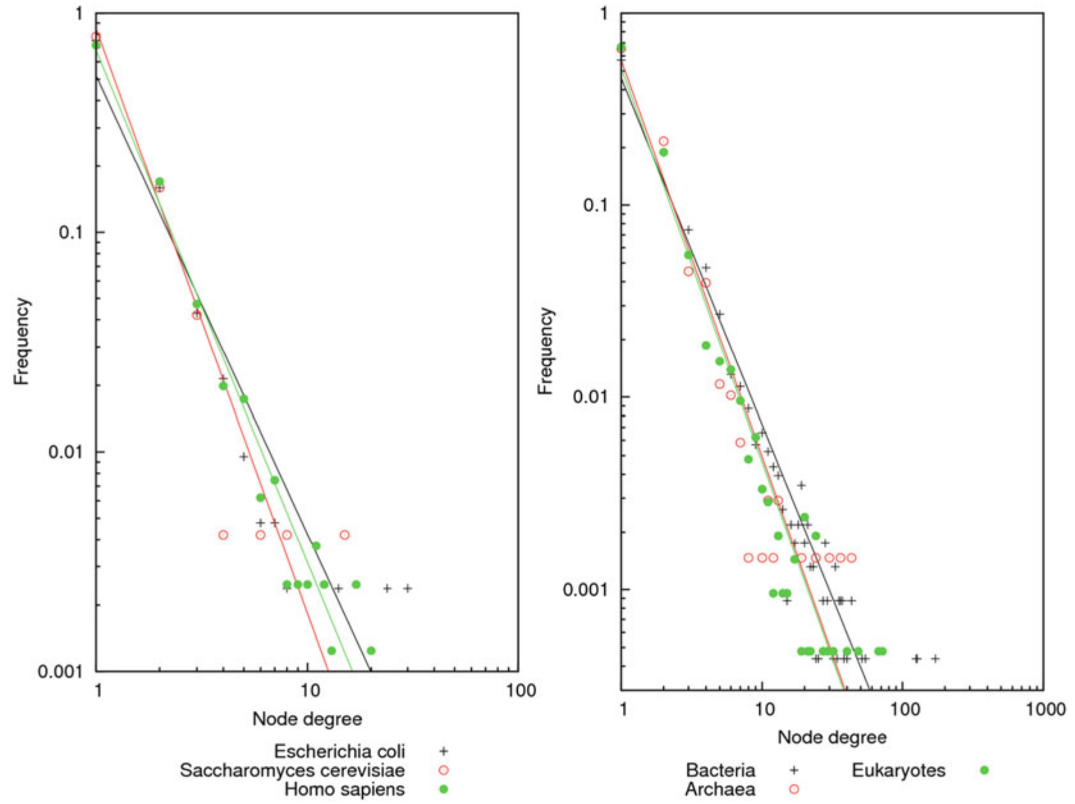

Fig. 6. (**a**) Distribution of domain co-occurrence network node degrees in three selected species. Power law distributions were fitted to these curves such that, for frequency $f$ of nodes of degree $X$, $f = cX^a$. For *Saccharomyces cerevisiae*, $a = -2.7$, for *Escherichia coli*, $a = -2.1$, and for *Homo sapiens*, $a = -2.3$. (**b**) Distribution of domain co-occurrence network node degrees across the three kingdoms. This corresponds to a network, where two domains are connected if any species within the kingdom has a protein, where these domains are immediately adjacent. Power law distributions were fitted to these curves such that, for frequency $f$ of nodes of degree $X$, $f = cX^a$. For bacteria, $a = -1.8$, for archaea, $a = -2.1$, and for eukaryotes, $a = -2.1$.

edge between them if there is a protein, where they are adjacent. Each domain was assigned a degree as its number of links to other domains. We then counted the frequency with which each degree occurs in the co-occurrence network. Figure 6a shows this relationship for the set of domain architectures found in the same species as for Fig. 2a, and Fig. 6b shows the equivalent plots for the three kingdoms as found among the complete proteomes in Pfam. Regressions to a power law have been added to the plots. The presence of a power law-like behavior of this type implies that few domains have very many immediate neighbors while most domains have few immediate neighbors. Note that the observed degrees in our dataset were strongly reduced by removing all sequences with a stretch longer than 50 amino acids lacking domain annotation.

## 5. Supradomains and Conserved Domain Order

As we have seen, whole multidomain architectures or shorter stretches of adjacent domains are often repeated in many proteins. These only cover a small fraction of all possible domain combinations. Are the observed combinations somehow special? We would expect selective pressure to retain some domain combinations but not others, since only some domains have functions that would synergize together in one protein. Often, co-occurring domains require each other structurally or functionally, for instance in transcription factors, where the DNA-binding domain provides substrate specificity, whereas the trans-activating domain recruits other components of the transcriptional machinery (53). Vogel et al. (31) identified series of domains co-occurring as a fixed unit with conserved N- to C-terminal order but flanked by different domain architectures, and termed them supradomains. By investigating their statistical overrepresentation relative to the frequency of the individual domains in the set of nonredundant domain architectures (where "nonredundant" is crucial, as otherwise, e.g., whole-gene duplication would bias the results), they identified a number of such supradomains. Many ancient domain combinations (shared by all three kingdoms) appear to be such selectively preserved supradomains.

How conserved is the order of domains in multidomain architectures? In a recent study, Kummerfeld and Teichmann (12) built a domain co-occurrence network with directed edges, allowing it to represent the order in which two domains are found in proteins. As in other studies, the distribution of node degrees fits a power law well. Most domain pairs were only found in one orientation. This does not seem required for functional reasons, as flexible linker regions should allow the necessary interface to form also in the reversed case (50), but may rather be an indication that most domain combinations are monophyletic. Weiner and Bornberg-Bauer (54) analyzed the evolutionary mechanisms underlying a number of reversed domain order cases and concluded that independent fusion/fission is the most frequent scenario. Although domain reversals occur in only a few proteins, it actually happens more often than was expected from randomizing a co-occurrence network (12). That study also observed that the domain co-occurrence network is more clustered than expected by a random model, and that these clusters are also functionally more coherent than would be expected by chance.

## 6. Domain Mobility, Promiscuity, or Versatility

While some protein domains co-occur with a variety of other domains, some are always seen alone or in a single architecture in all proteomes where they are found. A natural explanation is that some domains are more likely to end up in a variety of architectural contexts than others due to some intrinsic property they possess. Is such domain versatility or promiscuity a persistent feature of a given domain, and does it correlate with certain functional or biological properties of the domain?

Several ways of measuring domain versatility have been suggested. One measure, NCO (34), counts the number of other domains found in any architectures, where the domain of interest is found. Another measure, NN (30), instead counts the number of distinct other domains that a domain is found adjacent to. Yet another measure, NTRP (55), counts the number of distinct triplets of consecutive domains, where the domain of interest is found in the middle. All of these measures can be expected to be higher for common domains than for rare domains, i.e., variations in domain abundance (the number of proteins a domain is found in) can hide the intrinsic versatility of domains. Therefore, three different studies (13, 48, 56) formulated *relative domain versatility* indices that aim to measure versatility independently of abundance. It is worth noting that most studies have considered only immediately adjacent domain neighbors in these analyses, a restriction based on the assumption that those are more likely to interact functionally than domains far apart on a common amino acid chain.

The first relative versatility study was presented by Vogel et al. (56), who used as their domain dataset the SUPERFAMILY database applied to 14 eukaryotic, 14 bacterial, and 14 archaeal proteomes. They modeled the number of unique immediate neighbor domains as a power law function of domain abundance, performed a regression on this data, and used the resulting power law exponent as a relative versatility measure. Basu et al. (48) used Pfam and SMART (8) domains and measured relative domain versatility for 28 eukaryotes as the immediate neighbor pair frequency normalized by domain frequency. They then defined promiscuous domains as a class according to a bimodality in the distribution of the raw numbers of unique domain immediate-neighbor pairs. Weiner et al. (13) used Pfam domains for 10,746 species in all kingdoms, and took as their relative versatility measure the logarithmic regression coefficient for each domain family across genomes, meaning that it is not defined within single proteomes.

To what extent is high versatility an intrinsic property of a certain domain? Vogel et al. (56) only examined large groups of domains together and therefore did not address this question for single domains. Basu et al. (48) and Weiner et al. (13) instead analyzed each domain separately and concluded that there are strong variations in relative versatility at this level. Their results are very different in detail, however, reflected by the fact that only one domain family (PF00004, AAA ATPase family) is shared between the ten most versatile domains reported in the two studies. As they used fairly similar domain datasets, it would appear that the results strongly depend on the definition of relative versatility.

Another potential reason for the different results is that Basu's list was based on eukaryotes only while Weiner's analysis was heavily biased toward prokaryotes. Furthermore, the top ten lists in Basu et al. (48) and their follow-up paper (49) only overlap by four domains; yet the main difference is that in the latter study all 28 eukaryotes were considered while the former study was limited to the subset of 20 animal, plant, and fungal species. The choice of species, thus, seems pivotal for the results when using this method. They also used different methods for calculating the average value of relative versatility across many species, which may influence the results.

Does domain versatility vary between different functional classes of domains? Vogel et al. (56) found no difference in relative versatility between broad functional or process categories or between SCOP structural classes. In contrast to this, Basu et al. (48) reported that high versatility was associated with certain functional categories in eukaryotes. However, no test for the statistical significance of these results was performed. Weiner et al. (13) also noted some general trends, but found no significant enrichment of Gene Ontology terms in versatile domains. This does not necessarily mean that no such correlation exists, but more research is required to convincingly demonstrate its strength and its nature.

Another important question is to what extent domain versatility varies across evolutionary lineages. Vogel et al. (56) reported no large differences in average versatility for domains in different kingdoms. The versatility measure of Basu et al. (48) can be applied within individual genomes, which means that according to this measure domains may be versatile in one organism group but not in another, as well as gain or lose versatility across evolutionary time. They found that more domains were highly versatile in animals than in other eukaryotes. Modeling versatility as a binary property defined for domains in extant species, they further used a maximum parsimony approach to study the persistence of versatility for each domain across evolutionary time, and concluded that both gain and loss of versatility are common during evolution. Weiner at al. (13) divided domains into age categories based on distribution across the tree of life, and reported that the versatility index is not dependent on age, i.e., domains have equal chances of becoming versatile at different times in evolution. This is consistent with the observation by Basu et al. (48) that versatility is a fast-evolving and varying property. When measuring versatility as a regression within different organism groups, Weiner et al. (13) found slightly lower versatility in eukaryotes, which is in conflict with the findings of Basu et al. (48). Again, this underscores the strong dependence of the method and dataset on the results.

Further properties reported to correlate with domain versatility include sequence length, where Weiner et al. (13) found that longer domains are significantly more versatile within the framework of their study while at the same time shorter domains are more abundant, and hence may have more domain neighbors in absolute numbers. Basu et al. (48) further reported that more versatile domains have more structural interactions than other domains. To determine which of these reported correlations genuinely reflect universal biological trends, further comprehensive studies are needed using more data and uniform procedures. This would hopefully allow the results from the studies described here to be validated, and any conflicts between them to be resolved.

Basu et al. (48) further analyzed the phylogenetic spread of all immediate domain neighbor pairs ("bigrams") containing domains classified as promiscuous. The main observation this yielded was that although most such combinations occurred in only a few species most promiscuous domains are part of at least one combination that is found in a majority of species. They interpreted this as implying the existence of a reservoir of evolutionarily stable domain combinations from which lineage-specific recombination may draw promiscuous domains to form unique architectures.

## 7. Principles of Domain Architecture Evolution

What mutation events can generate new domain architectures, and what is their relative predominance? The question can be approached by comparing protein domain architectures of extant proteins. This is based on the likely realistic assumption that most current domain architectures evolved from ancestral domain architectures that can still be found unchanged in other proteins. Because of this, in pairs of most similar extant domain architectures, one can assume that one of them is ancestral. This agrees well with results indicating that most groups of proteins with identical domain architectures are monophyletic. By comparing the most similar proteins, several studies have attempted to chart the relative frequencies of different architecture-changing mutations.

Björklund et al. (57) used this particular approach and came to several conclusions. First, changes to domain architecture are much more common by the N- and C-termini than internally in the architecture. This is consistent with several mechanism for architecture changes, such as introduction of new start or stop codons or mergers with adjacent genes, and similar results have been found in several other studies (23, 24, 58). Furthermore, insertions or deletions of domains ("indels") are more common than substitutions of domains, and the events in question mostly concern just single domains, except in cases with repeats expanding with many domains in a row (59). In a later study, the same group made use of phylogenetic information as well, allowing them to infer directionality of domain indels (44). They then found that domain insertions are significantly more common than domain deletions.

Weiner et al. (24) performed a similar analysis on domain loss and found compatible results—most changes occur at the termini. Moreover, they demonstrated that terminal domain loss seldom involves losing only part of a domain or rather that such partial losses quickly progress into loss of the entire domain.

There is some support (21, 60, 61) for exon shuffling to have played an important part in domain evolution, and there are a number of domains that match intron borders well, for example structural domains in extracellular matrix proteins. While it may not be a universal mechanism, exon shuffling is suggested to have been particularly important for vertebrate evolution (21).

## 8. Inferring Ancestral Domain Architectures

The above analyses, based on pairwise comparison of extant protein domain architectures, cannot tally ancestral evolutionarily events nearer the root of the tree of life. With ancestral architectures, one can directly determine which domain architecture changes have taken place during evolution and precisely chart how mechanisms of domain architecture evolution operate, as well as gauge their relative frequency. A drawback is that since we can only infer ancestral domain architectures from extant proteins, the result depends somewhat on our assumptions about evolutionary mechanisms. On the upside, it should be possible to test how well different assumptions fit the observed modern-day protein domain architecture patterns.

Attempts at such reconstructions have been made using parsimony. Given a gene tree and the domain architectures at the leaves, dynamic programming can be used in order to find the assignment of architectures to internal nodes that requires the smallest number of domain-level mutation events. This simple model can be elaborated by weighting loss and gain differently or requiring that a domain or an architecture can only be gained at most once in a tree (Dollo parsimony) (62).

An early study of Snel et al. (63) considered 252 gene trees across 17 fully sequenced species and used parsimony to minimize the number of gene fission and fusion events occurring along the species tree. Their main conclusion, that gene fusions are more common than gene fissions, was subsequently supported by a larger study by Kummerfeld and Teichmann (64), where fusions were found to be about four times as common as fissions in a most parsimonious reconstruction. Fong et al. (65) followed a similar procedure on yet more data and concluded that fusion was 5.6 times as likely as fission.

Buljan and Bateman (58) performed a similar maximum parsimony reconstruction of ancestral domain architectures. They too observed that domain architecture changes primarily take place at the protein termini, and the authors suggested that this might largely occur because terminal changes to the architecture are less likely to disturb the overall protein structure. Moreover, they concluded from reconciliation of gene and species trees that domain architecture changes were more common following gene duplications than following speciation, but that these cases did not differ with respect to the relative likelihood of domain losses or gains.

Recently, Buljan et al. (23) presented a new ancestral domain architecture reconstruction study which assumed that gain of a domain should take place only once in each gene tree, i.e., Dollo parsimony (62). Their results also support gene fusion as a major mechanism for domain architecture change. The fusion is generally preceded by a duplication of either of the fused genes. Intronic recombination and insertion of exons are observed, but relatively rarely. They also found support for de novo creation of disordered segments by exonization of previously noncoding regions.

## 9. Polyphyletic Domain Architecture Evolution

There appears to be a "grammar" for how protein domains are allowed to be combined. If nature continuously explores all possible domain combinations, one would expect that the allowed combinations would be created multiple times throughout evolution. Such independent creation of the same domain architecture can be called convergent or polyphyletic evolution, whereas a single original creation event for all extant examples on an architecture would be called divergent or monophyletic evolution. This is relevant for several reasons, not least because it determines whether or not we can expect two proteins with identical domain architectures to have the same history along their entire length.

A graph theoretical approach to answer this question was taken by Przytycka et al. (46), who analyzed the set of all proteins containing a given superfamily domain. The domain architectures of these proteins define a domain co-occurrence network, where edges connect two domains both found in a protein, regardless of sequential arrangement. The proteins of such a set can also be placed in an evolutionary tree, and the evolution of all multidomain architectures containing the reference domain can be expressed in terms of insertions and deletions of other domains along this tree to form the extant domain architectures. The question, then, is whether or not all leaf nodes sharing some domain arrangement (up to and including an entire architecture) stem from a single ancestral node possessing this combination of domains. For monophyly to be true for all architectures containing the reference domain, the same companion domain cannot have been inserted in more than one place along the tree describing the evolution of the reference domain. By application of graph theory and Dollo parsimony (62), they showed that monophyly is only possible if the domain co-occurrence network defined by all proteins containing the reference domain is chordal, i.e., it contains no cycles longer than three edges.

Przytycka et al. (46) then evaluated this criterion for all superfamily domains in a large-scale dataset. For all domains where the co-occurrence network contained fewer than 20 nodes (domains), the chordal property held, and hence any domain combinations or domain architectures containing these domains could potentially be monophyletic. By comparing actual domain co-occurrence networks with a preferential attachment null model, they showed that far more architectures are potentially monophyletic than would be expected under a pure preferential attachment process. This finding is analogous to the observation by Apic et al. (30) that most domain combinations are duplicated more frequently (or reshuffled less) than expected by chance. In other words, gene duplication is much more frequent than domain recombination (56). However, for many domains that co-occurred with more than 20 other different domains, particularly for domains previously reported as promiscuous, the chordal property was violated, meaning that multiple independent insertions of the same domain, relative to the reference domain phylogeny, must be assumed.

A more direct approach is to do complete ancestral domain architecture reconstruction of protein lineages and to search for concrete cases that agree with polyphyletic architecture evolution. There are two conceptually different methodologies for this type of analysis. Either one only considers architecture changes between nodes of a species tree or one considers any node in a reconstructed gene tree. The advantage of using a species tree is that one avoids the inherent uncertainty of gene trees, but on the other hand only events that take place between examined species can be observed.

Gough (51) applied the former species tree-based methodology to SUPERFAMILY domain architectures, and concluded that polyphyletic evolution is rare, occurring in 0.4–4% of architectures. The value depends on methodological details, with the lower bound considered more reliable.

The latter gene tree-based methodology was applied by Forslund et al. (52) to the Pfam database. Ancestral domain architectures were reconstructed through maximum parsimony of single-domain phylogenies which were overlaid for multidomain proteins. This strategy yielded a higher figure, ranging between 6 and 12% of architectures depending on dataset and whether or not incompletely annotated proteins were removed. The two different approaches, thus, give very different results. The detection of polyphyletic evolution is in both frameworks dependent on the data that is used—its quality, coverage, filtering procedures, etc. The studies used different datasets which makes it hard to compare. However, given that their domain annotations are more or less comparable, the major difference ought to be the ability of the gene-tree method to detect polyphyly at any point during evolution, even within a single species. It should be noted that domain annotation is by no means complete—only a little less than half of all residues are assigned to a domain (5)—and this is clearly a limiting factor for detecting architecture polyphyly. The numbers may, thus, be adjusted considerably upward when domain annotation reaches higher coverage.

Future work will be required to provide more reliable estimates of how common polyphyletic evolution of domain architectures is. Any estimate will depend on the studied protein lineage, versatility of the domains, and methodological factors. A comprehensive and systematic study using more complex phylogenetic methods than the fairly ad hoc parsimony approach, as well as effective ways to avoid overestimating the frequency of polyphyletic evolution due to incorrect domain assignments or hidden homology between different domain families, may be the way to go. At this point, all that can be said is that polyphyletic evolution of domain architectures definitely does happen, but relatively rarely, and that it is more frequent for complex architectures and versatile domains.

## 10. Conclusions

As access to genomic data and increasing amounts of compute power has grown during the last decade, so has our knowledge of the overall patterns of domain architecture evolution. Still, no study is better than its underlying assumptions, and differences in the representation of data and hypotheses means that results often cannot be directly compared. Overall, however, the current state of the field appears to support some broad conclusions.

Domain and multidomain family sizes, as well as numbers of co-occurring domains, all approximately follow power laws, which implies a scale-free hierarchy. This property is associated with many biological systems in a variety of ways. In this context, it appears to reflect how a relatively small number of highly versatile components have been reused again and again in novel combinations to create a large part of the domain and domain architecture repertoire of organisms. Gene duplication is the most important factor to generate multidomain architectures, and as it outweighs domain recombination only a small fraction of all possible domain combinations is actually observed. This is probably further modulated by family-specific selective pressure, though more work is required to demonstrate to what extent. Most of the time, all proteins with the same architecture or domain combination stem from a single ancestor, where it first arose, but there remains a fraction of cases, particularly with domains that have very many combination partners, where this does not hold.

Most changes to domain architectures occur following a gene duplication, and involves the addition of a single domain to either protein terminus. The main exceptions to this occur in repeat regions. Exon shuffling played an important part in animals by introducing a great variety of novel multidomain architectures, reusing ancient domains as well as domains introduced in the animal lineage.

In this chapter, we have reexamined with the most up-to-date datasets many of the analyses done previously on less data, and found that the earlier conclusions still hold true. Even though we are at the brink of amassing enormously much more genome and proteome data thanks to the new generation of sequencing technology, there is no reason to believe that this will alter the fundamental observations we can make today on domain architecture evolution. However, it will permit a more fine-grained analysis, and also there will be a greater chance to find rare events, such as independent creation of domain architectures. Furthermore, careful application of more complex models of evolution with and without selection pressure may allow us to determine more closely to what extent the process of domain architecture evolution was shaped by selective constraints.

## 11. Materials and Methods

Updated statistics were generated from the data in Pfam 24.0. All Uniprot proteins belonging to any of the full proteomes covered in Pfam 24.0 were included. These include 1,359 bacteria, 76 eukaryotes, and 68 archaea. All Pfam-A domains regardless of type were included. However, as stretches of repeat domains are highly variable, consecutive subsequences of the same domain were collapsed into a single pseudo-domain, if it was classified as type Motif or Repeat, as in several previous works (44, 52, 56, 65).

Domains were ordered within each protein based on their sequence start position. In the few cases of domains being inserted within other domains, this was represented as the outer domain followed by the nested domain, resulting in a linear sequence of domain identifiers. As long regions without domain assignments are likely to represent the presence of as-yet uncharacterized domains, we excluded any protein with unassigned regions longer than 50 amino acids (more than 95% of Pfam-A domains are longer than this). This approach is similar to that taken in previous works (51, 52, 57).

Other studies (44, 59) have instead performed additional more sensitive domain assignment steps, such as clustering the unassigned regions to identify unknown domains within them.

Pfam domains are sometimes organized in clans, where clanmates are considered homologous. A transition from a domain to another of the same clan is, thus, less likely to be a result of domain swapping of any kind, and more likely to be a result of sequence divergence from the same ancestor. Because of this, we replaced all Pfam domains that are clan members with the corresponding clan.

The statistics and plots were generated using a set of Perl, R, and GnuPlot scripts, which are available upon request. Power law regressions were done using the Marquardt–Levenberg nonlinear least squares algorithm as implemented in GnuPlot and allowed to continue until the convergence criterium (for least squares sum $X_i$ following the $i$th iteration, $(X_i - X_{i+1})/X_i$ should not exceed $10^{-5}$) was met. For reasons of scale, the regression for a power law relation, such as

$$N = cX^{-a},$$

was performed on the equivalent relationship

$$\log(X) = (1/a)(\log(c) - \log(N)),$$

for the parameters $a$ and $c$, with the exception of the data for Fig. 6, where instead the relationship,

$$\log(N) = \log(c) - a\log(X),$$

was used. Moreover, because species or organism group datasets were of very different size, raw counts of domains were converted to frequencies before the regression was performed.

## 12. Online Domain Database Resources

For further studies or research into this field, the first and most important stop will be the domain databases. Table 2 presents a selection of domain databases in current use.

## 13. Exercises/ Questions

- Which aspects of domain architecture evolution follow from properties of nature's repertoire of mutational mechanisms, and which follow from selective constraints?
- What trends have characterized the evolution of domain architectures in animals?

### Table 2
### A selection of protein domain databases

| Database | URL | Notes |
|---|---|---|
| ADDA | http://ekhidna.biocenter.helsinki.fi/sqgraph/pairsdb | Automatic clustering of protein domain sequences |
| CATH | http://www.cathdb.info | Based solely on experimentally determined 3D structures |
| CDD | http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml | Metadatabase joining together domain assignments from many different sources, as well as some unique domains |
| Gene3D | http://gene3d.biochem.ucl.ac.uk | Bioinformatical assignment of sequences to CATH domains using hidden Markov models |
| INTERPRO | http://www.ebi.ac.uk/interpro | Metadatabase joining together domain assignments from many different sources |
| Pfam | http://pfam.sanger.ac.uk | Domain families are defined from manually curated multiple alignments, and represented using Hidden Markov Models |
| PRODOM | http://prodom.prabi.fr | Automatically derived domain families from proteins in UniProt |
| SCOP | http://scop.mrc-lmb.cam.ac.uk | Based solely on experimentally determined 3D structures |
| SMART | http://smart.embl-heidelberg.de | Domain families are defined from manually curated multiple alignments, and represented using Hidden Markov Models |
| SUPERFAMILY | http://supfam.cs.bris.ac.uk | Bioinformatical assignment of sequences to SCOP domains using Hidden Markov Models trained on the sequences of domains in SCOP |

- Discuss approaches to handle limited sampling of species with completely sequenced genomes. How can one draw general conclusions or test the robustness of the results? Apply, e.g., to the observed frequency of domain architectures that have emerged multiple times independently in a given dataset.
- Describe the principle of “preferential attachment” for evolving networks. In what protein domain-related contexts does this seem to model the evolutionary process, and what distribution of node degrees does it produce?
- What protein properties correlate with domain versatility? Can the versatility of a domain be different in different species (groups) and change over evolutionary time?
- What protein domain-related properties differ between prokaryotes and eukaryotes?

## References

1. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C and Murzin AG. (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res. 36(Database issue):D419–425.
2. Cuff AL, Sillitoe I, Lewis T, Redfern OC, Garratt R, Thornton J and Orengo CA. (2009) The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res. 37(Database issue):D310-314.
3. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C and Gough J. (2009) SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res. 37(Database issue): D380-386.
4. Lees J, Yeats C, Redfern O, Clegg A and Orengo C. (2010) Gene3D: merging structure and function for a Thousand genomes. Nucleic Acids Res. 38(1):D296-D300.
5. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunesekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR and Bateman A. (2010) The Pfam protein families database. Nucleic Acids Research, Database Issue 38:D211–222.
6. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, Finn RD, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Laugraud A, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, Mistry J, Mitchell A, Mulder N, Natale D, Orengo C, Quinn AF, Selengut JD, Sigrist CJ, Thimma M, Thomas PD, Valentin F, Wilson D, Wu CH and Yeats C. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res. 37(Database issue):D211-5
7. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N and Bryant SH. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37(Database issue):D205-210.
8. Letunic I, Doerks T and Bork P. (2009) SMART 6: recent updates and new developments. Nucleic Acids Res. 37(Database issue): D229–232.
9. Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S and Kahn D. (2005) The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. 33(Database issue):D212–215.
10. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. 38(Database issue):D142–148.
11. Heger A, Wilton CA, Sivakumar A and Holm L. (2005) ADDA: a domain database with global coverage of the protein universe. Nucleic Acids Res. 33(Database issue): D188–191.
12. Kummerfeld SK and Teichmann SA. (2009) Protein domain organisation:adding order. BMC Bioinformatics 10 (39). BioMed Central 2010.
13. Weiner J 3$^{rd}$, Moore AD and Bornberg-Bauer E. (2008) Just how versatile are domains? BMC Evolutionary Biology 8(285).
14. del Carmen Orozco-Mosqueda M, Altamirano-Hernandez J, Farias-Rodriguez R, Valencia-Cantero E and Santoyo G. (2009) Homologous recombination and dynamics of rhizobial genomes. Research in Microbiology 160(10):733–741.
15. Heyer WD, Ehmsen KT, and Liu J. (2010) Regulation of Homologous Recombination in Eukaryotes. Annu. Rev. Genet. 44:113–139.
16. Brissett NC and Doherty AJ. (2009) Repairing DNA double-strand breaks by the prokaryotic non-homologous end-joining pathway. Biochemical Society Transactions 37:539–545.
17. van Rijk A and Bloemendal H. (2003) Molecular mechanisms of exon shuffling: illegitimate recombination. Genetica 118:245-249.
18. Feschotte C and Pritham EJ. (2007) DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 41:331-368.
19. Cordaux R and Batzer MA. (2009) The impact of retrotransposons on human genome evolution. Nature Reviews Genetics 10:691–703.
20. Gogvadze E and Buzdin A. (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci. 66 (23):3727–3742.
21. Patthy L. (2003) Modular assembly of genes and the evolution of new functions. Genetica. 2003 Jul;118(2–3):217–31.
22. Liu M and Grigoriev A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of exon shuffling? Trends Genet. 20(9):399–403.

23. Buljan M, Frankish A and Bateman A. (2010) Quantifying themechanisms of domain gain in animal proteins. Genome Biol. 11(7):R74. BioMed Central 2010.
24. Weiner J 3rd, Beaussart F and Bornberg-Bauer E. (2006) Domain deletions and substitutions in the modular protein evolution. FEBS Journal 273: 2037–2047.
25. Schmidt EE and Davies CJ. (2007) The origins of polypeptide domains. Bioessays. 29(3): 262–270.
26. Huynen MA and van Nimwegen E. (1998) The Frequency Distribution of Gene Family Sizes in Complete Genomes. Mol. Biol. Evol. 15(5):583–589.
27. Qian J, Luscombe NM and Gerstein M (2001) Protein Family and Fold Occurrence in Genomes: Power-law Behaviour and Evolutionary Model. J. Mol. Biol. 313:673–681.
28. Luscombe NM, Qian J, Zhang Z, Johnson T and Gerstein M. (2002) The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties. Genome Biol 3: RESEARCH0040.
29. Apic G, Gough J and Teichmann SA. (2001) Domain Combinations in Archaeal, Eubacterial and Eukaryotic Proteomes. J. Mol. Biol. 310:311–325.
30. Apic G, Huber W and Teichmann SA. (2003) Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. Journal of Structural and Functional Genomics 4:67–78.
31. Vogel C, Berzuini C, Bashton M, Gough J and Teichmann SA. (2004) Supra-domains: Evolutionary Units Larger than Single Protein Domains. J. Mol. Biol. 336:809–823.
32. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS and Koonin EV. (2002) Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evol Biol. 2 (1):18.
33. Barabási AL and Albert R. (1999) Emergence of scaling in random networks. Science. 286 (5439):509–512.
34. Wuchty S. (2001) Scale-free Behavior in Protein Domain Networks. Mol. Biol. Evol. 18(9):1694–1702.
35. Rzhetsky A and Gomez SM. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome. Bioinformatics. 17(10):988–996.
36. Li L, Alderson D, Tanaka R, Doyle JC and Willinger W. (2005) Towards a Theory of Scale-Free Graphs: Definition, Properties, and Implications. Internet Mathematics 2 (4): 431–523.
37. Kuznetsov V, Pickalov V, Senko O and Knott G. (2002) Analysis of the evolving proteomes: Predictions of the number of protein domains in nature and the number of genes in eukaryotic organisms. J. Biol. Syst. 10(4):381–407.
38. Koonin EV, Wolf YI and Karev GP. (2002) The structure of the protein universe and genome evolution. Nature 420:218-223.
39. Yanai I, Camacho CJ and DeLisi C. (2000) Predictions of Gene Family Distributions in Microbial Genomes: Evolution by Gene Duplication and Modification. Phys. Rev. Let. 85 (12):2641–2644.
40. van Nimwegen E. (2005) Scaling laws in the functional content of genomes. Annu. Rev. Biochem. 74:867–900.
41. Ranea JAG, Buchan DWA, Thornton JM and Orengo CA (2004) Evolution of Protein Superfamilies and Bacterial Genome Size. J. Mol. Biol. 336:871–887.
42. Ranea JAG, Sillero A, Thornton JM, and Orengo CA. (2006) Protein superfamily evolution and the last universal common ancestor (LUCA). Journal of Molecular Evolution 63(4):513-525.
43. Chothia C and Gough J. (2009) Genomic and structural aspects of protein evolution. Biochem. J. 419:15–28.
44. Ekman D, Björklund ÅK and Elofsson A. (2007) Quantification of the Elevated Rate of Domain Rearrangements in Metazoa. J. Mol. Biol. 372:1337–1348.
45. Itoh M, Nacher JC, Kuma K, Goto S and Kanehisa M. (2007) Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. Genome Biol. 8(6):R121.
46. Przytycka T, Davis G, Song N and Durand D. (2006) Graph theoretical insights into evolution of multidomain proteins. J Comput Biol. 13(2):351–363.
47. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D. (1999). Detecting protein function and protein-protein interactions from genome sequences. Science. 285 (5428):751–753.
48. Basu MK, Carmel L, Rogozin IB, and Koonin EV. (2008) Evolution of protein domain promiscuity in eukaryotes. Genome Res. 18:449–461.
49. Basu MK, Poliakov E and Rogozin IB. (2009) Domain mobility in proteins: functional and evolutionary implications. Briefings in Bioinformatics 10(3):205–216.

50. Bashton M and Chothia C. (2002) The Geometry of Domain Combination in Proteins. J. Mol. Biol. 315:927–939.

51. Gough J. (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21(8):1464–1471.

52. Forslund K, Hollich V, Henricson A, and Sonnhammer ELL. (2008) Domain Tree Based Analysis of Protein Architecture Evolution MBE 25:254–264.

53. Brivanlou AH and Darnell JE. (2002) Signal Transduction and the Control of Gene Expression. Science 295(5556):813 – 818.

54. Weiner J 3rd and Bornberg-Bauer E. (2006) Evolution of Circular Permutations in Multidomain Proteins. Mol. Biol. Evol. 23(4):734–743.

55. Tordai H, Nagy A, Farkas K, Bányai L, Patthy L. (2005) Modules, multidomain proteins and organismic complexity. FEBS J 272 (19):5064–5078.

56. Vogel C, Teichmann SA and Pereira-Leal J. (2005) The Relationship Between Domain Duplication and Recombination. J. Mol. Biol. 346:355–365.

57. Björklund ÅK, Ekman D, Light S, Frey-Skött J and Elofsson A. (2005) Domain Rearrangements in Protein Evolution. J. Mol. Biol. 353:911–923.

58. Buljan M and Bateman A. (2009) The evolution of protein domain families. Biochem. Soc. Trans. 37:751–755.

59. Björklund ÅK, Ekman D and Elofsson A. (2006) Expansion of Protein Domain Repeats. PLoS Comput Biol 2(8):114.

60. Doolittle RD and Bork P (1993) Evolutionary mobile modules in proteins. Scient Am Oct:34–40.

61. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E and Elofsson A. (2008) Arrangements in the modular evolution of proteins. Trends Biochem Sci. 33 (9):444–151.

62. Farris JS. (1977). Phylogenetic analysis under Dollo s Law. Systematic Zoology 26: 77–88.

63. Snel B, Bork P and Huynen M. (2000) Genome evolution. Gene fusion versus gene fission. Trends Genet. 16(1):9–11.

64. Kummerfeld SK and Teichmann SA. (2005) Relative rates of gene fusion and fission in multi-domain proteins. Trends in Genetics 21 (1):25–30.

65. Fong JH, Geer LY, Panchenko AR and Bryant SH. (2007) Modeling the Evolution of Protein Domain Architectures Using Maximum Parsimony. J Mol Biol. 366(1):307–315.