doi:10.1016/j.jmb.2003.12.026 J. Mol. Biol. (2004) 336, 809–823

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Supra-domains: Evolutionary Units Larger than Single Protein Domains

Christine Vogel[1]*, Carlo Berzuini[2,3], Matthew Bashton[1], Julian Gough[4,5] and Sarah A. Teichmann[1]*

[1]*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK*

[2]*MRC Biostatistics Unit Institute of Public Health Cambridge CB2 2SR, UK*

[3]*Dipartimento di Informatica e Sistemistica, University of Pavia, 27100 Pavia, Italy*

[4]*Genome Exploration Research Group, RIKEN Genomic Sciences Centre, W121 1-7-22 Suehiro-cho, Tsurumi-ku Yokohama 230-0045, Japan*

[5]*Department of Structural Biology, Fairchild bldg, D109 Stanford, CA 94305-5126 USA*

**Corresponding authors*

Domains are the evolutionary units that comprise proteins, and most proteins are built from more than one domain. Domains can be shuffled by recombination to create proteins with new arrangements of domains. Using structural domain assignments, we examined the combinations of domains in the proteins of 131 completely sequenced organisms. We found two-domain and three-domain combinations that recur in different protein contexts with different partner domains. The domains within these combinations have a particular functional and spatial relationship. These units are larger than individual domains and we term them "supra-domains". Amongst the supra-domains, we identified some 1400 (1203 two-domain and 166 three-domain) combinations that are statistically significantly over-represented relative to the occurrence and versatility of the individual component domains. Over one-third of all structurally assigned multi-domain proteins contain these over-represented supra-domains. This means that investigation of the structural and functional relationships of the domains forming these popular combinations would be particularly useful for an understanding of multi-domain protein function and evolution as well as for genome annotation. These and other supra-domains were analysed for their versatility, duplication, their distribution across the three kingdoms of life and their functional classes. By examining the three-dimensional structures of several examples of supra-domains in different biological processes, we identify two basic types of spatial relationships between the component domains: the combined function of the two domains is such that either the geometry of the two domains is crucial and there is a tight constraint on the interface, or the precise orientation of the domains is less important and they are spatially separate. Frequently, the role of the supra-domain becomes clear only once the three-dimensional structure is known. Since this is the case for only a quarter of the supra-domains, we provide a list of the most important unknown supra-domains as potential targets for structural genomics projects.

Keywords: domain combination; protein family; domain architecture; multi-domain protein; functional annotation

Introduction

Domains are the evolutionary and structural units that form proteins: they can occur on their own in single-domain proteins, or in combination with different partner domains making multi-domain proteins. The domains within a protein are often also structurally and functionally independent. Domains that are related to each other by descent from a common ancestor, are members of the same superfamily as described in the structural classification of proteins database, SCOP.[1] The SCOP definition of a domain is evolutionary: SCOP domains exist either on their own and/or in combination with other domains.

Abbreviations used: HMM, hidden Markov model; SCOP, structural classification of proteins.

E-mail addresses of the corresponding authors: cvogel@mrc-lmb.cam.ac.uk; sat@mrc-lmb.cam.ac.uk

Though many small proteins consist of a single domain, such proteins represent only roughly one-third of the proteins in a prokaryote proteome,[2] and even less in a eukaryote proteome.[3] Given that the vast majority of proteins have two or more domains, and that domains adjacent on a protein chain can interact[4] and determine protein function, understanding the way domains combine in proteins is central to our knowledge of proteomes.

Investigation of pairwise domain combinations in multi-domain proteins found that a few superfamilies are highly versatile and have many different partner domains, while most domain superfamilies are observed only with one or two other different partner superfamilies.[5,6] While most domain superfamilies occur in all three kingdoms of life, domain combinations are more kingdom-specific.[5] Furthermore, most domain superfamilies form far fewer domain combinations than statistically expected from their abundance.[7]

Here, we investigate those two-domain and three-domain combinations that are re-used in different protein contexts with different partner domains. One such example is the P-loop containing nucleotide triphosphate hydrolase domain and the translation protein domain that occur as one combination in several different translation factors, shown in Figure 1. Another example was reviewed recently: the combination of an SH3, SH2 and protein kinase domain is found in many different receptors involved in signal transduction.[8] We call these domain combinations supra-domains to illustrate the greater degree of conservation and higher-order nature of these domain combinations. A supra-domain is defined as a domain combination in a particular N-to-C-terminal orientation that occurs in at least two different domain architectures in different proteins with: (i) different types of domains at the N and C-terminal end of the combination; or (ii) different types of domains at one end and no domain at the other.

Note that this definition is stricter in terms of the degree of recombination required compared to the SCOP definition of a domain, because a SCOP domain need occur only on its own, or with one other domain, provided that the other domain occurs with a different partner domain.

Given the definition of a supra-domain above, the two or three domains in a supra-domain could have recombined as a unit to form new domain architectures, or the individual domains could have assembled by some other route to end up adjacent to each other in different domain architectures. Either way, the combination of domains is selected to occur in different proteins due to the functional advantages of having that particular combination. There are several pieces of evidence that support the former scenario of recombination as one evolutionary unit, which we list in order of decreasing strength.

First, three-dimensional structural analyses of individual protein families such as the Rossmann-domain superfamily[9] have shown that proteins with the same domain architecture are related by descent, in other words they have evolved from one common ancestor. N. Kerrison, C. Chothia & S.A.T. (unpublished results) have shown that this is true for over one-half of all two-domain protein families of known structure in the current

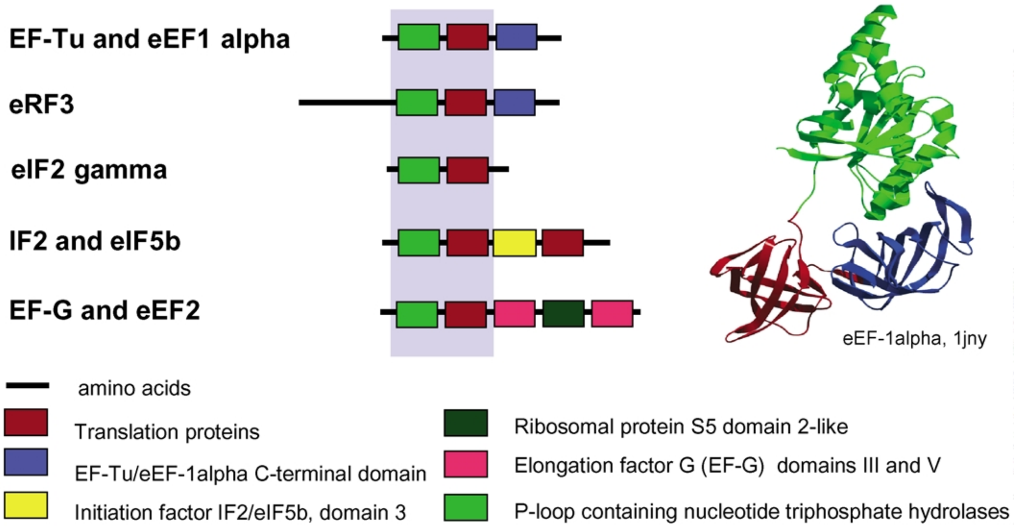

EF-Tu and eEF1 alpha
eRF3
eIF2 gamma
IF2 and eIF5b
EF-G and eEF2
eEF-1alpha, 1jny
amino acids
Translation proteins
EF-Tu/eEF-1alpha C-terminal domain
Initiation factor IF2/eIF5b, domain 3
Ribosomal protein S5 domain 2-like
Elongation factor G (EF-G) domains III and V
P-loop containing nucleotide triphosphate hydrolases

Figure 1. Supra-domain in translation factors: the P-loop containing nucleotide triphosphate hydrolases domain and the translation proteins domain. This supra-domain occurs in prokaryotic and eukaryotic translation factors that hydrolyse GTP. GTP hydrolysis in the P-loop domain drives the conformational change in the translation proteins domain, which is then transmitted onto the ribosome. The supra-domain occurs in 35 different domain architectures, and five of these are given here. The inset shows a protein of known structure, which contains the supra-domain (PDB: 1jny).[27]

databases. Second, domain pairs occur in only one N-to-C-terminal order in structural assignments to genome sequences, with only a small fraction of exceptions.[5] This conservation of domain order is likely to be evolutionary instead of purely functional, as the same interface and functional sites could be formed by two domains irrespective of their order, given a long linker between domains allowing for the same spatial relationship to be achieved. Third, proteins sharing series of domains tend to have the same or similar functions,[10] which is not the case if domain order is switched.[9]

The conserved functional relationship of the domains within supra-domains means that characterising these domain combinations and their functions can be a useful resource for annotation of unknown genome sequences. As mentioned above, Gerstein and co-workers showed that annotation transfer is more reliable for domain combinations in multi-domain proteins[10] than for individual domains.[11] Domain combinations have been shown to be useful for annotation of subcellular localizations.[12] Therefore, a comprehensive survey of supra-domains across proteomes contributes to prediction of protein function as well as understanding protein evolution.

In our study, we focus on the two-domain (duplet) and three-domain (triplet) supra-domains in 131 genomes, and investigate their characteristics. First, we describe the repertoires of duplet and triplet supra-domains. Next, we concentrate on the supra-domains that are over-represented with respect to the abundance of their individual component domains, as examples of supra-domains that have tightly coupled component domains. Having considered the component domains within each supra-domain, we then look at the supra-domain as a whole. We investigate the versatility, that is the number of different N and C-terminal partner domains of a supra-domain as a result of recombination, as well as the duplication of supra-domains. We describe the distribution of supra-domains across the three kingdoms of life. Finally, we examine the functions and structures of a subset of the supra-domains in order to better understand why these domain combinations play a special role within multi-domain proteins. Combining the results of our analysis, we can suggest supra-domains whose structure and exact function is still unknown and which represent interesting targets for experimental analysis because of their importance in multi-domain proteins.

Identifying Supra-domains in Multi-domain Proteins

Domain architectures of proteins in genomes

For the analysis of conserved domain combinations, or supra-domains, across the three kingdoms of life, we used the sets of predicted proteins from 131 completely sequenced genomes: 16 eukaryotes, 17 archaea and 98 bacteria. (The genomes are listed in the Background information section of the Supplementary website†.) Domain assignments to these proteins were taken from the SUPERFAMILY database[13] version 1.63. The SUPERFAMILY database contains a library of hidden Markov models (HMMs) based on the domains of known three-dimensional structure, and the assignments made by these HMMs to the predicted proteins of all completely sequenced genomes. The SUPERFAMILY database takes the definition of domains from the structural classification of proteins (SCOP) database,[1] using the superfamily level of classification, which groups together domains sharing a common evolutionary ancestor.

The individual assignments of SCOP domains to genome sequences were converted into a linear string of superfamily domains for each protein; the ordering is defined by the sequence of the centre-points of each domain. The assignments of known structural domains by SUPERFAMILY do not always cover the entire sequence. Thus, as well as the ordering of the assigned domains, it was necessary to determine the presence of unknown domains (unassigned regions, probably without a homologue of known structure) which can be N-terminal or C-terminal to, or between, assigned domains. It is not possible to determine the number of unknown domains in any given unassigned region, so each region was simply labelled as a gap containing one or more unknown domains. These strings define what we term the domain architecture, and the composition and interaction of the domains of the architecture determine the function of the protein. Both the presence of gaps and the process of forcing domains into a simple linear string mean that the string will not always be a complete description of the architecture. The procedure cannot account for the few cases where the domains are inserted into each other (roughly 9% in the RCSB Protein Data Bank (PDB),[14] R. Aroul Selvam, T. Hubbard & R. Sasidharan, unpublished results). However, more importantly for this work, each architecture will almost always be assigned a single unique string. The method for generating domain architectures from structural assignments is described in more detail elsewhere‡.

In the process of identifying supra-domains, we do not take into account the number of proteins in each kingdom of life that have a given unique domain architecture. Our assumption is that proteins with the same architecture have evolved from a common ancestor *via* gene duplication, so we effectively group all proteins with the same domain architecture into a single family described

† http://www.mrc-lmb.cam.ac.uk/genomes/cvogel/SupraDomains/

‡ http://supfam.org/SUPERFAMILY/combs.html

by that domain architecture. There is evidence from sequence, structural and functional analyses to suggest that this assumption is largely true, as described in the Introduction with reference to the conservation of domains within a supra-domain. Conservation of complete domain architecture in multi-domain proteins is implied in the recent search tools based on domain architectures. The conserved domain architecture retrieval tool (CDART)[15] connected with the conserved domain database (CDD),[16] the sequence analysis database SMART,[17] as well as search tools of the Pfam database,[18] allow retrieval of proteins with particular domain architectures.

Grouping of the 261,344 multi-domain sequences from the genomes yields 28,387 unique domain architectures. Please refer to Figure 2 for an overview of the grouping of architectures.

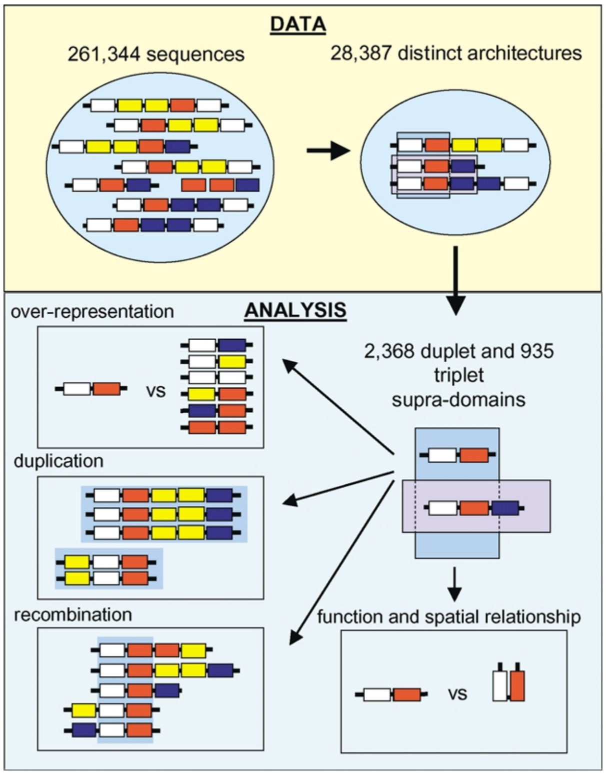

DATA
261,344 sequences
28,387 distinct architectures
over-representation
ANALYSIS
vs
2,368 duplet and 935 triplet supra-domains
duplication
recombination
function and spatial relationship
vs

Figure 2. Overview of the extraction and analysis of supra-domains. The coloured boxes represent domains of different superfamilies and the black lines represent protein sequences. The representation is a schematic of the procedures used and the different characteristics of supra-domains analysed. The data set comprises 261,344 multi-domain sequences, symbolised in the first diagram. Of these sequences, 140,425 are composed of two domains, 73,224 are composed of three domains and 47,695 are composed of four and more domains. This grouping of proteins with the same architecture results in 2804, 10,035 and 19,991 distinct domain architectures with at least two domains in Archaea, Bacteria and Eukaryotes, respectively, that is a total of 28,387 distinct domain architectures as shown in the second part of Figure 2. These distinct domain architectures contained 9398 two-domain and 4323 three-domain combinations out of which 2368 and 935, respectively, qualify as supra-domains. We analysed these supra-domains with respect to their duplication and recombination in different proteins, the over-representation of the supra-domains relative to their component domains and the supra-domain functions.

The repertoire of supra-domains

In the set of 131 genomes studied here, the proteins have domains from 1216 different superfamilies. Approximating the number of superfamilies to 1000, 10^{3n} n-domain combinations can be formed from these different superfamilies, where n is the number of domains in a combination: that is, as a rough estimate, 10^6 and 10^9 for two-domain and three-domain combinations, respectively. Given the set of about 261,344 multi-domain sequences, many of which contain more than two domains, a considerable fraction of the possible two-domain and three-domain combinations could, in theory, be observed in these sequences. However, only a tiny fraction is actually seen, presumably due to selection for their function: 9398 different combinations with two domains and 4323 different combinations with three domains. Table 1 gives an overview of our observations across all three kingdoms of life.

About a quarter of the duplet and triplet combinations qualify as supra-domains according to our definition of a domain pair or triplet in one particular N-to-C-terminal orientation that acts as a recombinatorial unit: 2368 of the 9398 two-domain combinations (25%) and 935 of the 4323 three-domain combinations (22%) (Table 1). N-to-C-terminal order must be conserved, so that there is not recombination within the group of domains of the supra-domain, though the supra-domain itself may duplicate and recombine with other domains. This recombinatorial unit of two or three domains, that is the supra-domain, occurs either with at least two different partner domains in different proteins, or with different partner domains at one end and no partner domain at the other. We require evidence of at least one recombination event in which the supra-domain is conserved. If there is a gap in a domain architecture, this effectively splits the sequence, and no supra-domain can have partly unassigned regions.

Both the duplet and triplet supra-domains can be divided into two types of supra-domains: complex, if adjacent domains belong to different superfamilies, and repetitive, if two or more adjacent domains belong to the same superfamily. The latter case represents potential intra-sequence domain duplication. There are more complex supra-domains than repetitive ones: The majority of duplet supra-domains is complex (2064 *versus* 304), while the distribution is more even for triplet supra-domains (424 *versus* 511) (Table 1).

Table 1. Sequences, domain architectures and domains per kingdom of life

	Archaea	Bacteria	Eukarya	Total (union across archaea, bacteria and eukarya)
Different multi-domain sequences	12,444	103,307	145,593	261,344
Different domain architectures	2804	10,035	19,991	28,387
Different domain superfamilies	812	1099	1171	1216
Duplet supra-domains				
Different duplet combinations	1292	4327	5407	9398
Duplet supra-domains	593	1301	1874	2368 2064/304
Over-represented supra-domains of which complex/repeat	307[a] (405)	621[a] (697)	1017[a] (1,080)	1203 951/252
Triplet supra-domains				
Different triplet combinations	436	1649	2678	4323
Triplet supra-domains	167	388	727	935 424/511
Over-represented supra-domains of which complex/repeat	0[a] (18)	45[a] (53)	137[a] (149)	166 87/89

Overview of the proteins, domain and domain combinations in our analysis. The total number of two-domain (duplet) and three-domain (triplet) combinations, and the subset of these that qualify as supra-domains, and as statistically significant over-represented supra-domains, are provided.

[a] Supra-domains that are significant in this kingdom; the number of supra-domains that are significant in any of the three kingdoms and present in the particular kingdom is denoted in parentheses.

Characterising the Repertoire of Supra-domains in Terms of Domain Patterns in Multi-domain Proteins

Finding over-represented supra-domains by statistical analysis

Practically all the supra-domains, like most other domain combinations in multi-domain proteins, have undergone millions of years of evolution and selection, which means that each of the domains in a supra-domain has a defined role in the proteins that contain the supra-domain. Given that a supra-domain represents a self-contained unit that occurs in different domain architectures, we expect that the two or three component domains within a supra-domain have a functional relationship such that this domain combination is of wide use in different contexts. However, it is possible that some domain combinations have ended up adjacent to each other in multiple different domain architectures without necessarily having a specific, conserved relationship. We explore this idea by extracting those supra-domains that are over-represented relative to the abundance of the component domains. The domains that make these over-represented supra-domains are more likely to have a close, conserved functional or structural relationship. In the following, we first define a measure of association between the constituent domains relative to the abundance of the individual component domains, called r_{ij}^2, and then extract significantly over-represented supra-domains by p-value calculations.

We start by considering a generic duplet supra-domain (i,j), where superfamily i is followed by superfamily j in N-to-C-terminal order. These two superfamilies have frequencies p_i and p_j in the set of non-redundant domain architectures in the particular kingdom. Similarly, p_{ij} denotes the frequency of the specific domain pair (i,j). Then the squared correlation coefficient:

$$r_{ij}^2 = \frac{(p_{ij} - p_i p_j)^2}{p_i p_j (1 - p_i)(1 - p_j)}$$

provides a measure of over-representation for the duplet (i,j) relative to the abundance of individual component domains i and j, as the numerator compares the observed frequency of the duplet with the corresponding value expected under an assumption that the components are independent. This measure can be interpreted in terms of the information provided by the superfamily of one of the two domains about the superfamily of the other. Large values of r_{ij}^2 indicate that domains of type i strongly "prefer" domains of type j at their C terminus (and/or *vice versa*). This idea was recently applied to domain prediction.[19] The r_{ij}^2 is in the range (0, 1) as long as p_i and p_j differ from 0 and 1. These properties make r_{ij}^2 a convenient way of classifying and ranking duplet supra-domains in terms of the internal association of the domain pair.

Besides assigning an r_{ij}^2 value to each duplet supra-domain (i,j), we compare the observed duplet frequency p_{ij} with its corresponding distribution under the null hypothesis, H_0, that the component domains are independent (method described in the legend to Table 2), resulting in a (one-sided) p-value for the test of the null hypothesis. If this p-value falls below a certain significance threshold, indicating strong evidence against H_0, we call the (i,j) supra-domain over-represented. The method requires that we associate the duplet (i,j) with a table of counts described in Table 2, called the duplet association table $t^{[ij]}$. In order to illustrate the biological meaning of these tables and their implications for the significance of over-representation of supra-domains, we now consider three examples of this type of table for eukaryotic duplets:

(a) $t_{11}^{[ij]} = 23$, $t_{12}^{[ij]} = 9$, $t_{21}^{[ij]} = 208$, $t_{22}^{[ij]} = 31,665$ (*i*, DBL homology; *j*, PH-domain-like superfamilies)

(b) $t_{11}^{[ij]} = 238$, $t_{12}^{[ij]} = 1743$, $t_{21}^{[ij]} = 1281$, $t_{22}^{[ij]} = 28,713$ (*i*, Immunoglobulin domains; *j*, Fibronectin type III)

(c) $t_{11}^{[ij]} = 49$, $t_{12}^{[ij]} = 309$, $t_{21}^{[ij]} = 3111$, $t_{22}^{[ij]} = 28,506$ (*i*, spermadhesin, CUB-domain; *j*, EGF/laminin)

The supra-domain in example (a) occurs at modest frequency, even though one of the individual component supra-domains is abundant on their own. Nevertheless, because the *p*-value for this duplet is almost zero, this supra-domain is highly significant. The squared correlation r_{ij}^2 is 0.28, which reflects a moderate degree of internal association of the domains within the supra-domain. In example (b) both component domains occur in many other combinations, so the r_{ij}^2 is relatively low (0.0077). The supra-domain combination is still frequent enough to have a significant *p*-value. The supra-domain in example (c) is less common as compared to alternative combinations of the individual component domains, especially of the C-terminal domain ($t_{12}^{[ij]}$), so it is not likely an over-represented supra-domain. In fact, the *p*-value for this supra-domain 0.019, turns out to be above the significance threshold, and the squared correlation r_{ij}^2 is almost zero.

The legend for Table 2 describes our procedure for selecting a significant subset of duplet supra-domains in each kingdom, on the basis of

Table 2. Duplet and triplet association tables

	C-terminal superfamily	
N-terminal superfamily	*j*	Non-*j*
A. *Duplet association*		
i	$t_{11}^{[ij]}$	$t_{12}^{[ij]}$
Non-*i*	$t_{21}^{[ij]}$	$t_{22}^{[ij]}$
B. *Triplet association*		
i	$t_{11}^{[ikj]}$	$t_{12}^{[ikj]}$
Non-*i*	$t_{21}^{[ikj]}$	$t_{22}^{[ikj]}$

The occurrence of each supra-domain, and its individual constituent domains, within the data set of different domain architectures in each kingdom can be summarized as a duplet or triplet association table. The structure of these tables, and how they were used to calculate *p*-values, are discussed below. A, *Duplet association*; $t_{11}^{[ij]}$ denotes the number of times family *j* is found C-terminal to family *i* in the data set of different domain architectures in the kingdom under consideration. The symbol $t_{12}^{[ij]}$ represents the number of times superfamily *i* is N-terminal to a superfamily different from *j*, whereas $t_{21}^{[ij]}$ denotes the number of times superfamily *j* is C-terminal to a superfamily different from *i*. Finally, $t_{22}^{[ij]}$ is the number of duplets in the data set of domain architectures that contain neither *i* as the N-terminal domain nor *j* as the C-terminal domain and this value is roughly constant for all duplet supra-domains. Because of large $t_{22}^{[ij]}$ values in conjunction with very small $t_{21}^{[ij]}$ and $t_{12}^{[ij]}$ values, large-scale approximations implicit in standard 2×2 table association tests could not be applied in most of our duplet association tables. To compute the *p*-value for a given duplet association table, we generated a large number (100,000) of replicate versions of the table with equal probability under H_0, preserving the row and column margins, which is equivalent to sampling replicate values of $t_{11}^{[ij]}$ from a suitable hypergeometric distribution. The *p*-value was then estimated as the proportion of replicate tables with the upper left element equal or greater than the observed $t_{11}^{[ij]}$. In some tables, explicit calculation of the *p*-value was avoided by exploiting the following two inequalities. First, given two tables with identical entries except for $t_{11}^{[ij]}$, the one with the higher value for $t_{11}^{[ij]}$ will have lower *p*-value. Second, given tables with identical entries except for one of the off-diagonal elements ($t_{12}^{[ij]}$ or $t_{21}^{[ij]}$), the one with the smaller off-diagonal element will have lower *p*-value. We computed a separate *p*-value for each duplet supra-domain in each kingdom. This gave each duplet as many *p*-values as the number of kingdoms where that duplet is a supra-domain. The next step was to select, in each kingdom, those duplet supra-domains which are over-represented in that kingdom. This we did by applying the idea of False Discovery Rate (FDR), introduced by Benjamini & Hochberg.[20] The procedure proposed by these authors, in our context, can be summarized as follows: (1) order the *p*-values in the selected kingdom and let them be denoted as $q_1 \leq q_2 \leq \ldots \leq q_m$; (2) select each duplet *i* where $i \leq \max\{j\colon q_j \leq j\alpha/m\}$ as an over-represented duplet in that kingdom, where we specify α as the tolerated FDR, that is, the tolerated proportion of duplets classified as over-represented, which do not actually depart from H_0. We set $\alpha = 0.01$ in Eukarya and $\alpha = 0.05$ in the remaining two kingdoms. This turned out, in practice, to be equivalent to establishing the following three kingdom-specific thresholds on *p*-value: 0.008 (Eukarya and Bacteria) and 0.006 (Archaea), which was considerably less conservative than application of the usual Bonferroni rule. Our analysis showed that the squared correlation coefficient r_{ij}^2 had negligible values in all duplets that are not over-represented. This suggests that a classification of duplets in terms of the single measure r_{ij}^2 is useful. B, *Triplet association*: The methodology applied to duplets can be extended in a straightforward way to deal with triplets. Let (i, k, j) denote a generic triplet involving superfamilies *i*, *k* and *j* in consecutive positions in N-to-C-terminal order. We associate each triplet with a triplet association table, as shown here, which effectively conditions on the middle domain belonging to superfamily *k*. In other words, we are focusing on the association of the N-terminal domain *i* and the C-terminal domain *j* given that the middle domain belongs to superfamily *k*. Thus, the triplet association table is concerned with only two superfamilies, *i* and *j*, and is defined in a way very similar to that for the duplet association table. The four cells in this table have exactly the same meaning as in the duplet association table described above, except that the association is between two domains N and C-terminal to a domain from superfamily *k*. So, $t_{11}^{[ikj]}$ denotes the frequency of the triplet in the set of domain architectures under consideration, and $t_{12}^{[ikj]}$ is the number of triplets where a domain from superfamily *k* is C-terminal to a domain from superfamily *i* and N-terminal to a domain from a superfamily other than *j*, and *vice versa* for $t_{21}^{[ikj]}$. It should be noted that in contrast to the duplet association table, where $t_{22}^{[ij]}$ is approximately constant, the value of $t_{22}^{[ikj]}$ can vary. We apply the same procedures for calculation of *p*-values and selection of thresholds on the triplet association tables as for the duplet association tables. An over-represented triplet is a triplet where the N-terminal superfamily label *i* is significantly associated with the C-terminal label *j*, conditionally on being separated by a middle domain from superfamily *k*. This can be interpreted in terms of a deviation of the (i, k, j) triplet from a first-order Markov model of dependence. Finally, the previously squared correlation coefficient r_{ij}^2 is defined on a triplet association table in the same way as in the duplet association table, as previously discussed. The threshold *p*-values for triplets in eukaryotes and bacteria is 0.02. The *p*-values and other information on the duplet and triplet supra-domains are on the website: http://www.mrc-lmb.cam.ac.uk/genomes/cvogel/SupraDomains/

the computed p-values, by applying ideas of false discovery rates.[20] This yielded the set of over-represented supra-domains for each kingdom, with 307, 621 and 1017 over-represented duplet supra-domains in archaea, bacteria and eukaryotes, respectively (see Table 1 for an overview). These represent 1203 over-represented duplet supra-domains in total out of the 2368 supra-domains. Some of these over-represented supra-domains are present in more than one or two kingdoms but not always significant in all kingdoms (see Table 1 for details). In only one over-represented domain combination, the ERF1 domain (N-terminal domain of eukaryotic peptide chain release factor subunit 1) in combination with the Translation Machinery Components domain, the N-terminal domain does not occur in any combination other than with that C-terminal domain. The combination is frequent enough to be significant.

The above methodology can be extended in a straightforward way to deal with triplets (*ikj*), as described in the legend to Table 2. We apply the same procedures for calculation of p-values and selection of thresholds on the triplet association tables as for the duplet association tables. Of the total 935 triplet supra-domains, 166 triplet combinations have p-values lower than the significance threshold. For the triplet supra-domains, the phylogenetic distribution is similar to duplets, but even more biased. There are no over-represented triplet supra-domains in archaea, which again might be due to the small number of completely sequenced archaea, and there are 45 and 137 over-represented triplet supra-domains in bacteria and eukaryotes, respectively. Only some of these eukaryote over-represented supra-domains have a representative of known structure, and we inspect their functions at a later stage.

Duplication and recombination of supra-domains

Above, we considered over-represented supra-domains in terms of the association of the component domains with respect to their individual abundance. However, supra-domains are defined with respect to their partner domains in different domain architectures. Therefore, another important feature is the versatility of the whole supra-domain with respect to N-terminal and C-terminal partner domains in different domain architectures (Figure 2).

It is known that the number of different partner domains for a single domain or for a domain combination follows a power law distribution: many domains or domain combinations have only very few different N-terminal or C-terminal partner domains. Few domains or domain combinations occur with many different domain superfamilies.[5] The recombinatorial properties of supra-domains are similar. Many supra-domains have two or three different N-terminal and C-terminal domain partners, but few supra-domains are highly versatile with up to 158 different partners, for example in the combination of two P-loop hydrolase domains. The ten most versatile supra-domains that occur in all three kingdoms of life are described in Table 3.

By definition, supra-domains are generally somewhat more versatile than all domain combinations: all two-domain combinations have, on average, 3.0 different partner domains, while supra-domains and over-represented supra-domains are, on average, more versatile, with 5.5 and 7.2 different partner domains, respectively. Table S4 on the Supplementary website† compares the average numbers of different N and C-terminal partner domains for single domains and domain combinations.

In terms of N and C-terminal versatility, supra-domains again behave in a manner similar to that of single domains. Most single domains have roughly equal numbers of different N-terminal as compared to C-terminal partner domains, with a Pearson correlation coefficient of 0.989 (see the Supplementary website†). We observe the same for domain combinations and supra-domains, suggesting that the N-to-C-terminal order of domains in a protein is irrelevant for function (as long as the spatial relationship is preserved[9]) so a domain pair could in theory occur in either orientation to carry out its function. However, in evolution the order for each particular domain pair is fixed initially, and is conserved from then onwards, with a small number of exceptions where domains occur in both orientations relative to each other. Interestingly, even though component domains of supra-domains tend to be more versatile than other domains, there is no correlation between the versatility of component domains and the versatility of the domain combination, as shown in the Supplementary Table S4.

Nature can re-use domain combinations by duplicating the sequences, not just by combining the particular domain combination with many other different domains. Like single domains[6] and domain combinations,[5] supra-domains follow a power law distribution with respect to duplication. Many supra-domains occur in ten or fewer sequences, but a few supra-domains are highly duplicated and occur in over 200 sequences. One example is the combination of the NAD(P)-binding Rossmann domain with the glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain that occurs in just under 1400 sequences across the 131 genomes of our data set. Though this supra-domain is presented as a linear combination of two domains according to SUPERFAMILY, it is an example of one domain inserted into a discontinuous domain. This suggests that some domain pairs that are supra-domains might be tightly linked due to insertions, as this impairs the recombination of the component domains.

The duplication, or abundance, and recombination clearly are independent characteristics of

Table 3. Top ten most versatile complex supra-domains common to archaea, bacteria and eukaryotes

No. sequences in the three kingdoms	No. different partner domains	*Domain combination*	
Spatial relationship	Function	Description	Representative in PDB
A_44 B_1699 E_52	38	**Homodimeric domain of signal transducing histidine kinase; ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase**	1b3qA
Separate	Signal transduction	This domain combination consists of a dimerisation domain and a histidine kinase domain of the type that occurs in two-component signal transduction pathways.[28] Most of these proteins have additional N-terminal domains that are frequently involved in small-molecule binding. Some of the proteins have additional C-terminal domains that are equivalent to components downstream in the signal transduction pathway, such as the CheY receiver domain or the histidine phosphotransfer (HPT) domain. All the proteins with this supra-domain are likely to be involved in signal transduction	
A_10 B_95 E_36	30	**P-loop containing nucleotide triphosphate hydrolases; PEP carboxykinase-like**	–
Unknown	Enzyme	There is no known structure for the combination of these two domains, which are both nucleotide triphosphate hydrolase domains	
A_58 B_556 E_216	25	**Acyl-CoA dehydrogenase NM domain-like; acyl-CoA dehydrogenase C-terminal domain-like**	1bucA
Interface	Enzyme	A binding pocket for FAD and acyl-CoA is formed at the interface of the two domains.[29] Many of the enzymes containing this supra-domain are involved in oxidation of fatty acids and amino acid catabolism	
A_60 B_725 E_320	21	**GroES-like; NAD(P)-binding Rossmann-fold domains**	1a71A
Interface	Enzyme	The N-terminal domain is the catalytic, alcohol-binding domain, with two bound zinc ions, one of which is catalytic.[30] The C-terminal domain binds the NAD cofactor, and is involved in dimerisation in some examples. Though the two domains bind distinct molecules, the two molecules have to be positioned carefully for the reaction to be catalysed, so the functions are effectively joint	
A_66 B_626 E_327	20	**P-loop containing nucleotide triphosphate hydrolases; Translation proteins**	1aipA
Separate	Enzyme	This supra-domain is in all prokaryotic and eukaryotic translation factors that hydrolyze GTP. All proteins with this supra-domain for which information is available[31,32] contact the base of the L7/L12 stalk through the N-terminal domain, and the shoulder of the small subunit of the ribosome *via* the second (translation protein) domain. GTP hydrolysis drives conformational change in second domain, which is transmitted to ribosomal proteins. Proteins with this domain are prokaryotic initiation factor 2 and the eukaryotic homologue eIF5B, elongation factors Tu and G, equivalent to eukaryotic EF1A and 2, and prokaryotic release factor 3 and selenocysteinyl-tRNA specific translation factor	
A_22 B_104 E_281	19	**Ribonuclease H-like; DNA/RNA polymerases**	1d5aA
Separate	Enzyme	The N-terminal domain is the exonuclease proof-reading domain, while the C-terminal domain polymerises DNA, as discussed by Doublie *et al*.[33]	
A_28 B_342 E_175	19	**Riboflavin synthase domain-like; ferredoxin reductase-like, C-terminal NADP-linked domain**	1a8p
Interface	Enzyme	The N-terminal domain binds FAD and the C-terminal domain binds NADPH.[23] The FAD acts as an intermediate in electron transfer between NADPH and substrate, and this domain combination is used by many different enzymes	
A_65 B_639 E_175	18	**NAD(P)-binding Rossmann-fold domains; 6-phosphogluconate dehydrogenase C-terminal domain-like**	1bg6
Interface	Enzyme	The Rossmann domain binds NAD, while the C-terminal domain is the catalytic domain that binds the main substrate in the dehydrogenase reaction, and is known to be involved in dimerisation in some instances.[34] The cofactor and main substrate have to be positioned precisely relative to each other for the reaction to be carried out, so the activity of the two domains is effectively joint	

(*continued*)

Table 3 Continued

No. sequences in the three kingdoms Spatial relationship	No. different partner domains Function	*Domain combination* Description	Representative in PDB
A_63 B_630 E_149 Interface	17 Enzyme	**PreATP-grasp domain; glutathione synthetase ATP-binding domain-like** Lots of different enzymes forming carbon–nitrogen bonds have this combination of domains. Both domains contribute to substrate binding and the active site, and the C-terminal domain binds ATP as well as the other substrate; e.g. see Thoden *et al.*[35]	1gsh
A_10 B_492 E_17 Interface	15 Enzyme	**MurD-like peptide ligases, catalytic domain; MurD-like peptide ligases, peptide-binding domain** The catalytic domain binds ATP, and the peptide-binding domain binds glutamate in the enzymes of this ADP-forming amide bond ligase family.[36] A third variable substrate is bound at the interface of the two domains	1fgs

The number of sequences per kingdom, the spatial relationship of the domains, the total number of N-terminal and C-terminal partner domains, an overall functional classification as well as a detailed description of the domain functions are provided for each of the ten most versatile, universal two-domain combinations. We omitted three combinations with very low counts in one kingdom.

domains or domain combinations. Though highly duplicated supra-domains generally also have several different N and C-terminal partner domains, there is no obvious correlation between the extent of duplication and recombination of supra-domains (Figure 3). Similarly, the over-representation of a supra-domain is independent of its number of duplicates or its versatility. Thus all three characteristics, the duplication, versatility and over-representation of a supra-domain, are independent of each other. However, we do observe a larger number of over-represented supra-domains amongst the more versatile supra-domains, as shown in Figure 3.

Distribution of supra-domains across the three kingdoms of life

Over half of the individual domain superfamilies are common to all three kingdoms of life (64%, Figure 4(a)). In contrast to individual domains, pairwise domain combinations are much more specific in their phylogenetic distribution,[5] with only 4% common to archaea, bacteria and eukaryotes (Figure 4(a)). The supra-domains characterised in our analysis represent a subset of the pairwise domain combinations, and are intermediate between individual domains and domain pairs in terms of the fraction common to all three kingdoms of life (15%).

The fraction of common combinations is even higher for over-represented supra-domains (27%). In fact, almost all two-domain combinations shared by the three kingdoms of life qualify as over-represented supra-domains, while this is not the case for kingdom-specific combinations. Since there is a larger fraction of shared supra-domains, the fraction of kingdom-specific supra-domains or over-represented supra-domains is lower than for all two-domain combinations. Many of these kingdom-specific supra-domains consist of component domains that are shared across all kingdoms (Figure 4(a); and see the Supplementary website†).

Similar to the situation for single domains and for all pairwise domain combinations, eukaryotes

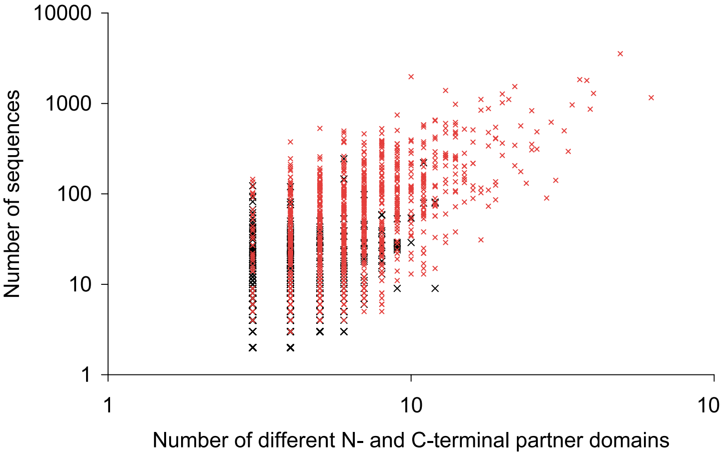


Figure 3. Relationship between versatility, abundance and over-representation of supra-domains. For all two-domain supra-domains the number of different N-terminal and C-terminal partner domains (versatility) is plotted against the total number of sequences in which the combination occurs (abundance). Thus there can be up to four N and C-terminal partners in two sequences. Over-represented supra-domains are plotted in red, non-over-represented supra-domains in black crosses.

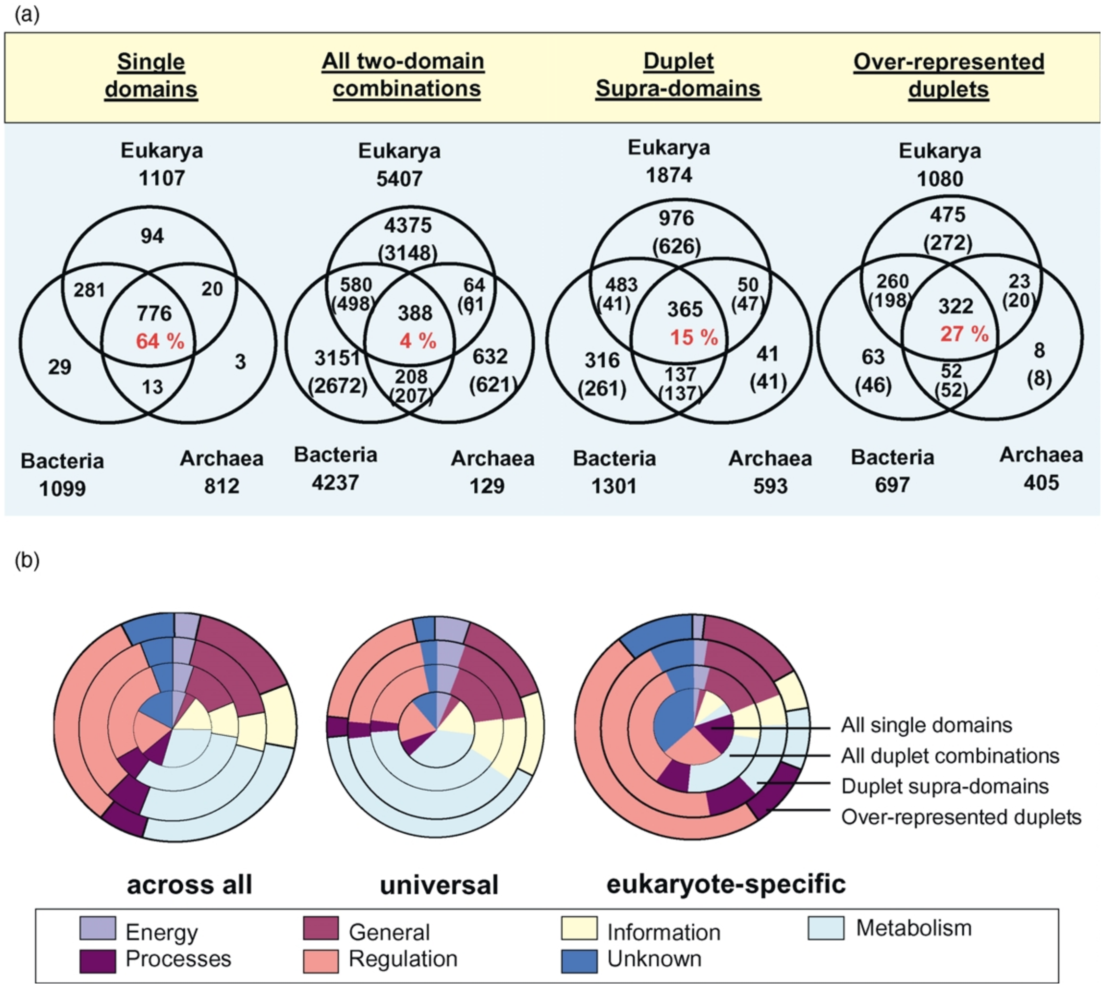


Figure 4. The distribution of domains and domain combinations across the three kingdoms of life and across broad functional classes. (a) The distribution of single-domains, all two-domain, duplet supra-domain and over-represented duplet supra-domain combinations across the three kingdoms of life. The numbers in parentheses denote the subset of the domain combinations that consist of component domains that occur in all three kingdoms of life. The red numbers denote the percentage of the total. (b) The distribution of individual domains (innermost circle), all two-domain combinations (second circle from centre), two-domain supra-domains (third circle) and over-represented two-domain supra-domains (outermost circle) across broad functional classes. From left to right, the distributions represent all entities (the union across the three kingdoms of life), the entities shared by the three kingdoms and the eukaryote-specific entities. The functional classes are: (i) information (storage, replication, transcription/translation); (ii) regulation; (iii) metabolism; (iv) energy (part of metabolism); (v) processes (cell motility, transport and so forth); (vi) general (enzymatic reactions, protein interaction etc); and (vii) unknown or unclassifiable. Across SCOP and across all three kingdoms of life, the largest fraction, one-third, are domains involved in metabolism. For example, many proteins are oxidases and reductases (left-hand panel). These are followed by domains in the general category, like for example P-loop NTP hydrolases, Rossmann domains or SAM methyltransferases, and domains in the regulatory category.

have the largest fraction of supra-domains, and of kingdom-specific supra-domains. This is due partly to the bias in the dataset towards eukaryotic and bacterial domains of known structure, and the lack of archaeal protein structures.

Characterising the Repertoire of Supra-domains in Terms of Function

Though supra-domains are defined by their characteristics of recombination, they have been selected to occur in different domain architectures with different partner domains due to their functional features. Thus, the domains in a supra-domain have a combined function that is useful in different contexts. We first discuss the distribution of single domains and domain combinations across broad and specific functional categories, and then consider the detailed functions and spatial relationships of the domains within supra-domains.

Functions of supra-domains

We assigned 1108 of the SCOP domains that occur amongst the supra-domains to one of 41 functional categories (see the Supplementary website†). The functional categories are based on the scheme used in the COGs[21] database and extended as needed. The functional categories were then grouped into seven broad functional classes, which are: (i) information (storage, replication, transcription/translation); (ii) regulation; (iii) metabolism; (iv) energy (part of metabolism); (v) processes (cell motility, transport etc); (vi) general (enzymatic reactions, protein interaction, etc.); and (vi) unknown or unclassifiable. For details, please refer to the Supplementary website†. Since functional classifications always have pitfalls and limitations, our discussion of the functional categories of supra-domains will be qualitative rather than quantitative.

Within the 1203 over-represented supra-domains, the majority of domains used belong to the following categories: signal transduction (208); DNA-binding/transcription factors (204); small-molecule binding (156); and unknown functions (144). Other highly populated functional categories comprise domains of: general function (125); cell adhesion (117); proteases, peptidases and their inhibitors (115); and other enzymes (115).

We compared this distribution of supra-domain functions to that of the individual domains in SCOP as well as of all two-domain combinations. The functions of the individual domains in SCOP are obviously biased by the selection of proteins for structure determination, and this distribution is shown in Figure 4(b), left-hand panel. The distribution of functional classes of pairwise domain combinations and supra-domains is very similar, only with fewer domains of unknown function. Therefore, supra-domains occur across all functional classes in an unbiased manner.

We also examined the eukaryote-specific distribution of functions (Figure 4(b), right-hand panel). Eukaryotes employ more regulatory domains than bacteria or archaea. Examples of these eukaryote-specific domains are the growth-factor receptor domain, SH2 or SH3 domains, or zinc fingers.

Two types of spatial relationships of supra-domains

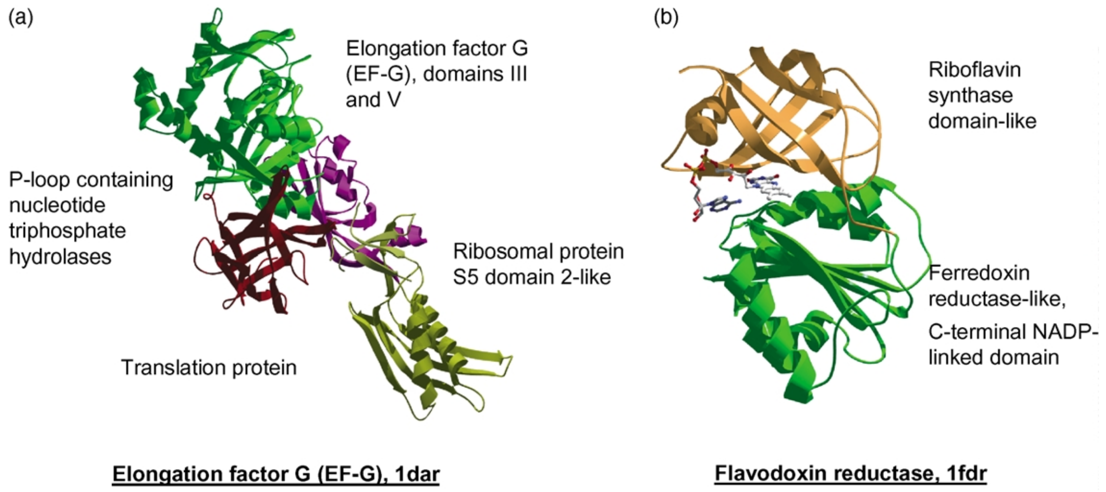
The way in which the individual domains within all these supra-domains interact to form an evolutionary unit that is useful in different domain architectures becomes apparent if a three-dimensional structure of the supra-domain is available. There are two different basic ways in which a supra-domain can function within a protein. Case A, "separate": the domains have separate activities, such that the linkage between the two domains can be flexible within the proteins and/or in the evolution of the supra-domain. Case B, "interface": the interface between the domains is essential to the activity of the protein, and the interface is likely to be conserved across different proteins with this supra-domain.

Both ways provide clear reasons for the conservation and re-use of these domain combinations in different contexts. Once the structural and functional mode of interaction of the component domains of a supra-domain are known, the information can be used to provide a general functional annotation of genome sequences of unknown function that contain such supra-domains.

An example in which the two domains have separate, distinct activities that are linked within one protein is the supra-domain with the P-loop nucleotide triphosphate hydrolase domain and the translation proteins domain (Table 3, Figures 1 and 5(a)). This supra-domain is ubiquitous across all 131 genomes in our data set. The supra-domain is found in several translation factors, with different domain partners at the N and C terminus, as shown in Figure 1. The structure has been solved for a number of proteins with this supra-domain, and an example is shown in Figure 1, for the archaeal elongation factor eEF-1alpha.[22] In this protein, the P-loop domain binds and hydrolyses GTP, which drives a conformational change that is transmitted to the other domain. The translation proteins domain interacts with the ribosome.

The riboflavin synthase domain-like domain followed by the ferredoxin reductase-like, C-terminal NADP-linked domain is an example of the closer type of functional relationship described in case B above, in which an activity is created at the interface between the two domains so that they physically interact with each other (Table 3 and Figure 5(b)). This domain combination is found in several enzymes, e.g. oxidoreductases, which transfer electrons from NADPH onto flavodoxin or ferredoxin using FAD as an intermediate. The structure has been solved for several of these enzymes, one of which is shown here.[23] The two cofactors FAD and NADPH are held in a fixed orientation relative to each other by the N and C-terminal domains, respectively. The orientation of the domains is the same in other examples of solved structures of this supra-domain (data not shown). It is this particular orientation that allows electron transfer to occur between the two cofactors, and so the function of the supra-domain relies on close interaction at the interface of the two domains.

In the manner described for the two examples given above, we analysed the spatial relationships of the domains in the over-represented duplet supra-domains shared by Archaea, Bacteria and Eukaryotes. Of the 230 over-represented duplet supra-domains shared by the three kingdoms, there are 179 with a known three-dimensional structure. Amongst the known supra-domains, 139 have complex domain compositions, while the remaining 41 have repetitive domain compositions. Detailed examination of 116 supra-domains of known structure, summarized in Supplementary Table S2, showed that there are about equal numbers of supra-domains for the two types of relationships. Amongst the 75 complex


(a)
Elongation factor G (EF-G), domains III and V
P-loop containing nucleotide triphosphate hydrolases
Ribosomal protein S5 domain 2-like
Translation protein
(b)
Riboflavin synthase domain-like
Ferredoxin reductase-like, C-terminal NADP-linked domain

Elongation factor G (EF-G), 1dar

Flavodoxin reductase, 1fdr

Figure 5. Examples of supra-domains. The two structures are examples of the two different ways of interaction of component domains within a supra-domain: in (a) the three domains have largely independent functions and are spatially separate, while in (b) the two domains interact closely and have a tight interface. (a) Elongation factor G (EF-G), 1dar.[22] The supra-domain consists of a P-loop nucleotide triphosphate hydrolases domain (light green) in combination with a translation proteins domain (red). The supra-domain is combined with an elongation factor G (EF-G), domains III and V domain and a ribosomal protein S5 domain 2-like domain shown in purple and grey, respectively. (b) Ferredoxin reductase, 1fdr.[23] This supra-domain is a combination of a riboflavin synthase-like domain (brown) and a ferredoxin reductase domain (green). The cofactor bound is FAD.

supra-domains, 28 supra-domains have functionally independent, spatially separate component domains (case A, above), while 36 have component domains where the interface is important for the function (case B, above), 11 are unclassifiable. This suggests there is no preference for supra-domains with one type of spatial relationship or another in nature. The r_{ij}^2 statistic described above fails to provide a measure for the spatial relationship: The two groups (supra-domains of the separate type (case A) and of the interface type (case B) do not segregate clearly according to their recombinatorial versatility or the r_{ij}^2 statistic of internal association (data not shown). Both types of spatial relationship occur in very versatile supra-domains and supra-domains with tight association of the constituent domains.

One of the duplet supra-domains studied here illustrates that there can be exceptions to the classification into separate and interface-type spatial relationships. This is the N-terminal nucleophile aminohydrolases followed by a SIS domain (PDB 1jxa[24]). In this supra-domain, the two domains are linked by a channel through which ammonia is passed from the N-terminal domain to the C-terminal domain, which imposes a constraint (case B, above) on the interface. However, the two domains have independent functions: the N-terminal domain catalyses hydrolysis of glutamine to glutamate, and the C-terminal domain is a sugar isomerase. The second reaction catalysed by the C-terminal domain relies on the ammonia produced by glutamine hydrolysis as a nitrogen donor for the sugar isomerisation.

We describe the spatial relationship for the domains in the ten complex supra-domains that are the most versatile in terms of their N and C-terminal partner domains (Table 3). These supra-domains occur in all three kingdoms of life. Nine of the supra-domains are of known structure with four supra-domains that have spatially separate domains (case A, above), and four in which the interface between the component domains is important (case B, above), and one that is unclear. Thus, there is again an even distribution of the two spatial relationships. Eight of the supra-domains are enzymatic. The other two supra-domains all have functions in signal transduction, and it is likely that they both have distinct but linked functions between the domains, though one is of unknown structure.

The eukaryotic triplet supra-domains

Besides the duplet supra-domains common to all three kingdoms, the triplet supra-domains specific to eukaryotes are of particular interest. This is because many classes of proteins specific to eukaryotes, including those associated with multicellularity, are long proteins with three or more domains. These multi-domain proteins function in signal transduction, cell adhesion, development or the nervous system. Therefore, one would also expect triplet supra-domains specific to eukaryotes to be involved in these important functional categories.

There are 113 triplet over-represented supra-domains specific to eukaryotes, and only 19 of these have homologues of known structure. Ten out of 19 known supra-domains are repeats of extra-cellular domains involved in cell adhesion and signalling: LDL receptor-like modules, glucocorticoid receptor-like domains, EGF/laminin domains, complement control module/SCR domain, integrin, immunoglobulin or fibronectin type III domains. Another supra-domain consists of extracellular domains: two growth factor receptor domains with a leucine-rich L domain in the middle. This domain architecture is observed, for example, in the extracellular portion of the EGF receptor structure, which has two growth factor receptor-L-domain repeats.[25] Another supra-domain is present in a receptor: the transferrin receptor, apical domain followed by a Zn-dependent exopeptidase domain and a transferrin receptor ectodomain, C-terminal domain. This is involved in iron uptake in the protein of known three-dimensional structure, the human transferrin receptor.[26]

One of the best known eukaryote-specific triplet supra-domain of known structure is associated with signal transduction proteins in the cytoplasm: the SH3-SH2-protein kinase supra-domain that recurs in a variety of domain architectures, as discussed by Harrison.[8] The SH3-SH2-kinase supra-domain represents an ensemble of domains that create coordinated regulatory properties of the kinase activity, depending on the binding of other proteins to the SH3 and SH2 domains, and the phosphorylation state of the supra-domain, as described in detail by Harrison.[8] This triplet supra-domain is a good example of domains acting independently, as in case A described above.

As more structures of multi-domain proteins are available, an analysis beyond these few examples will be feasible.

Targets for Structure Determination

From the above analysis, it should be clear that knowledge of the structure of a supra-domain representative provides important insights into the functions of these domain combinations. While all 2368 duplet supra-domains occur in 40% of all multi-domain sequences with domain assignments, the 200 most duplicated duplet supra-domains occur in 28%, or more than 75,000 sequences. (For more details, please refer to Supplementary Table S1.) Knowledge of these 200 supra-domains provides information on about one-fifth of all structurally assigned multi-domain sequences that occur in Archaea, Bacteria and Eukaryotes (see the Supplementary website). Of these 200 most duplicated duplet supra-domains, 161 have a known structure.

An example of such a highly duplicated supra-domain of known structure is the homodimeric domain of signal transducing histidine kinase in combination with the ATPase domain of HSP90, described in Table 3. This supra-domain occurs in about 1700 sequences, most of which are uncharacterised: three out of four sequences in *Helicobacter pylorii*, or 45 out of 56 sequences in *Pseudomonas aeruginosa* are annotated as hypothetical proteins. The fact that these proteins contain this supra-domain tells us that the proteins are involved in signal transduction, and that the domains in the supra-domain are likely to have the same structure and function as the homologous supra-domain of known structure.

Similarly, the top 200 most duplicated triplet supra-domains occur in more than a tenth (11% or almost 30,000) of all sequences with at least three domains, as described in Supplementary Table S1. Of these 200 triplet supra-domains, 107 have homologues of known structure. The most duplicated supra-domains of unknown structure are of particular interest as targets for structure elucidation, and are listed in the Supplements.

Functional annotation is more reliable for proteins that consist of several domains,[10] because the protein functions tend to be more conserved if they have more domains in common. This should be especially true of supra-domains, as they represent evolutionary units conserved throughout recombination events. Thus, knowledge of the structure and therefore exact function of the most prevalent supra-domains should be immensely useful for genome annotation.

Our analysis of different characteristics of domain combinations enables us to rank these according to their importance. The ordered lists could be useful in target structure selection: amongst the top 200 of the duplet supra-domains, 39 of the most duplicated supra-domains do not have a known structure (PDB,[14] February 2003). These 39 unknown domain combinations alone occur in more than 11,300 sequences (4%) throughout the three kingdoms of life. Two interesting examples of duplet supra-domains of unknown structure are the homeodomain-like domain in combination with the ribonuclease H-like domain, which is nucleic acid binding, and the winged helix DNA-binding domain with the periplasmic binding protein II domain. The inverted form of the latter supra-domain is present in PDB, but not the duplet in this N-to-C-terminal order. This supra-domain alone occurs in almost 2000 sequences, which are probably mostly transcription factors.

A list of the top 200 supra-domains and information on whether supra-domains have homologues of known structure can be obtained from the Supplementary website†.

Discussion and Conclusion

We have introduced the concept of supra-domains, which are evolutionary units in the same sense as individual domains: they can exist on their own in a protein, or in combination with several different domains at their N and C termini. Supra-domains consist of domains that interact in a manner that is useful in the different contexts, and have therefore been selected in evolution to be an essential part of many different proteins.

They occur across all functional categories without a bias in the distribution of functions of individual domains as compared to all domain combinations. We have shown that there are two different types of spatial relationships between the domains of a supra-domain: either the two domains have separate activities such that their interface is not crucial, or the activities of the two domains are intimately linked such that the interface is important. Judging from examination of a subset of supra-domains of known structure, both types occur roughly equally.

The importance of supra-domains becomes clear upon examination of their characteristics within the repertoire of multi-domain proteins across genomes. Some supra-domains are highly duplicated and occur in many sequences, while others are versatile and occur with many different domain partners. We extracted the 1203 duplet and 166 triplet supra-domain that are significantly over-represented, i.e. where the association of the component domains is tighter than in other supra-domains in a statistical sense.

There is no simple relationship between the duplication, versatility and over-representation of domain combinations; these characteristics seem to be independent of each other for many supra-domains. More versatile supra-domains, however, tend to be over-represented and highly abundant. Our analysis of these characteristics has shed light on the relationships of domains that form the repertoire of multi-domain proteins.

We ranked the supra-domains with respect to these different criteria and list those few hundred domain combinations that are components of one-third of the whole multi-domain protein repertoire with structural assignments. These ordered lists are useful for target selection in structural genomics projects. An understanding of this very small number of the most popular supra-domains in terms of their function and three-dimensional structure will provide information about a large number of all multi-domain proteins. This, in turn, will be a useful tool in the annotation of the vast numbers of sequences generated by the genome projects.

Acknowledgements

We thank Cyrus Chothia for helpful discussions, and Martin Madera for development and maintenance of the SUPERFAMILY database. We are grateful to Mary Pacold for her involvement in SUPERFAMILY post-processing. C.V. has a predoctoral fellowship from the Boehringer Ingelheim Fonds.

References

1. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536–540.
2. Teichmann, S. A., Park, J. & Chothia, C. (1998). Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements. *Proc. Natl Acad. Sci. USA*, **95**, 14658–14663.
3. Gerstein, M. (1998). How representative are the known structures of the proteins in a complete genome? A comprehensive structural census. *Fold. Des.* **3**, 497–512.
4. Park, J., Lappe, M. & Teichmann, S. A. (2001). Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* **307**, 929–938.
5. Apic, G., Gough, J. & Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* **310**, 311–325.
6. Wuchty, S. (2001). Scale-free behavior in protein domain networks. *Mol. Biol. Evol.* **18**, 1694–1702.
7. Apic, G., Huber, W. & Teichmann, S. A. (2003). Multidomain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics*, **4**, 67–78.
8. Harrison, S. C. (2003). Variation on an Src-like theme. *Cell*, **112**, 737–740.
9. Bashton, M. & Chothia, C. (2002). The geometry of domain combination in proteins. *J. Mol. Biol.* **315**, 927–939.
10. Hegyi, H. & Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res.* **11**, 1632–1640.
11. Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233–249.
12. Mott, R., Schultz, J., Bork, P. & Ponting, C. P. (2002). Predicting protein cellular localization using a domain projection method. *Genome Res.* **12**, 1168–1174.
13. Gough, J. (2002). The SUPERFAMILY database in structural genomics. *Acta Crystallog. sect. D*, **58**, 1897–1900.
14. Westbrook, J., Feng, Z., Chen, L., Yang, H. & Berman, H. M. (2003). The Protein Data Bank and structural genomics. *Nucl. Acids Res.* **31**, 489–491.
15. Geer, L. Y., Domrachev, M., Lipman, D. J. & Bryant, S. H. (2003). CDART: protein homology by domain architecture. Conserved domain architecture retrieval tool. *Genome Res.* **12**, 1619–1623.
16. Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y. & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucl. Acids Res.* **30**, 281–283.
17. Letunic, I., Goodstadt, L., Dickens, N. J., Doerks, T., Schultz, J., Mott, R. *et al.* (2002). Recent improvements to the SMART domain-based sequence annotation resource. *Nucl. Acids Res.* **30**, 242–244.
18. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R. *et al.* (2002). The Pfam protein families database. *Nucl. Acids Res.* **30**, 276–280.
19. Coin, L., Bateman, A. & Durbin, R. (2003). Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl Acad. Sci. USA*, **100**, 4516–4520.
20. Benjamini, Y. & Hochberg, Y. (1995). Controlling the

false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. ser. B*, **57**, 289–300.

21. Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V. *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
22. al-Karadaghi, S., Aevarsson, A., Garber, M., Zheltonosova, J. & Liljas, A. (1996). The structure of elongation factor G in complex with GDP: conformational flexibility and nucleotide exchange. *Structure*, **4**, 455.
23. Ingelman, M., Bianchi, V. & Eklund, H. (1997). The three-dimensional structure of flavodoxin reductase from *Escherichia coli* at 1.7 Å resolution. *J. Mol. Biol.* **268**, 147–157.
24. Teplyakov, A., Obmolova, G., Badet, B. & Badet-Denisot, M. (2001). Channeling of ammonia in glucosamine-6-phosphate synthase. *J. Mol. Biol.* **313**, 1093–1102.
25. Ogiso, H., Ishitani, R., Nureki, O., Fukai, S., Yamanaka, M., Kim, J. H. *et al.* (2002). Crystal structure of the complex of human epidermal growth factor and receptor extracellular domains. *Cell*, **110**, 775–787.
26. Lawrence, C. M., Ray, S., Babyonyshev, M., Galluser, R., Borhani, D. W. & Harrison, S. C. (1999). Crystal structure of the ectodomain of human transferrin receptor. *Science*, **286**, 779–782.
27. Vitagliano, L., Masullo, M., Sica, F., Zagari, A. & Bocchini, V. (2001). Crystal structure of *Sulfolobus solfataricus* elongation factor 1 alpha in omplex with GDP. *EMBO J.* **20**, 5305.
28. Bilwes, A. M., Alex, L. A., Crane, B. R. & Simon, M. I. (1999). Structure of CheA, a signal-transducing histidine kinase. *Cell*, **96**, 131–141.
29. Battaile, K. P., Molin-Case, J., Paschke, R., Want, M., Bennett, D., Vockley, J. & Lim, J.-J. P. (2002). Crystal structure of rat short chain acyl-coA dehydrogenase complexed with acetoacetyl-coA. *J. Mol. Chem.* **277**, 12200–12207.
30. Davis, G. J., Bosron, W. F., Stone, C. L., Owusu-Dekyi, K. & Hurley, T. D. (1996). X-ray structure of human beta3beta3 alcohol dehydrogenase. The contribution of ionic interactions to coenzyme binding. *J. Biol. Chem.* **271**, 17057–17061.
31. Song, H., Parsons, M. R., Rowsell, S., Leonard, G. & Phillips, S. E. (1999). Crystal structure of intact elongation factor EF-Tu from *Escherichia coli* in GDP conformation at 2.05 Å resolution. *J. Mol. Biol.* **285**, 1245–1256.
32. Roll-Mecak, A., Shin, B. S., Dever, T. E. & Burley, S. K. (2001). Engaging the ribosome: universal IFs of translation. *Trends Biochem. Sci.* **26**, 705–709.
33. Doublie, S., Sawaya, M. R. & Ellenberger, T. (1999). An open and closed case for all polymerases. *Structure*, **7**, R31–R35.
34. Thomazeau, K., Dumas, R., Halgand, F., Forest, E., Douce, R. & Biou, V. (2000). Structure of spinach acetohydroxyacid isomeroreductase complexed with its reaction product dihydroxymethylvalerate, manganese and (phospho)-ADP-ribose. *Acta Crystallog. sect. D*, **56**, 389–397.
35. Thoden, J. B., Kappock, J., Stubbe, H. & Holden, H. M. (1999). Three-dimensional structure of N5-carboxyaminoimidazole ribonucleotide synthetase: a member of the ATP grasp protein superfamily. *Biochemistry*, **38**, 15480–15492.
36. Sun, X., Cross, J. A., Bognar, A. L., Baker, E. N. & Smith, C. A. (2001). Folate-binding triggers the activation of folylpolyglutamate synthase. *J. Mol. Biol.* **310**, 1067–1078.

Edited by G. von Heijine

(Received 1 September 2003; received in revised form 4 December 2003; accepted 5 December 2003)