Full text provided by www.sciencedirect.com

# Structure, function and evolution of multidomain proteins

Christine Vogel*, Matthew Bashton, Nicola D Kerrison,
Cyrus Chothia and Sarah A Teichmann

Proteins are composed of evolutionary units called domains; the majority of proteins consist of at least two domains. These domains and nature of their interactions determine the function of the protein. The roles that combinations of domains play in the formation of the protein repertoire have been found by analysis of domain assignments to genome sequences. Additional findings on the geometry of domains have been gained from examination of three-dimensional protein structures. Future work will require a domain-centric functional classification scheme and efforts to determine structures of domain combinations.

**Addresses**
MRC Laboratory of Molecular Biology, Hills Road,
Cambridge CB2 2QH, UK
*e-mail: cvogel@mrc-lmb.cam.ac.uk

**Current Opinion in Structural Biology** 2004, **14**:208–216

This review comes from a themed issue on
Theory and simulation
Edited by Joel Janin and Thomas Simonson

0959-440X/$ – see front matter

**DOI 10.1016/j.sbi.2004.03.011**

**Abbreviations**

| | |
|---|---|
| **PDB** | Protein Data Bank |
| **RMSD** | root mean square deviation |
| **SCOP** | Structural Classification of Proteins |
| **SH** | Src homology |
| **WHD** | Winged helix domain |

## Introduction

There are various uses of the word domain with respect to proteins. Here, we define a protein domain as an independent, evolutionary unit that can form a single-domain protein or be part of one or more different multidomain proteins. The domain can either have an independent function or contribute to the function of a multidomain protein in cooperation with other domains. The definition of a domain as an evolutionary unit is used in the Structural Classification of Proteins (SCOP) database [1].

In SCOP, domains that have a common ancestor based on sequence, structural and functional evidence are grouped into superfamilies. There are more than 1200 domain superfamilies in the current version of the database [2], though estimates of the total number of superfamilies vary from a few to several thousand [3–5]. Domains from the superfamilies in SCOP can be assigned to 40–60% of the residues in the proteins of completely sequenced genomes using homology-based methods. These include the profile hidden Markov models in the SUPERFAMILY database [6,7•], the structural profiles of the PSSM server [8], the PSI-BLAST profiles in the Gene3D database [9] or combined approaches [10•]. From the assignment of structural domains to genome sequences, it is clear that some two-thirds of proteins consist of two or more domains in prokaryotes [11] and an even larger fraction in eukaryotes [12].

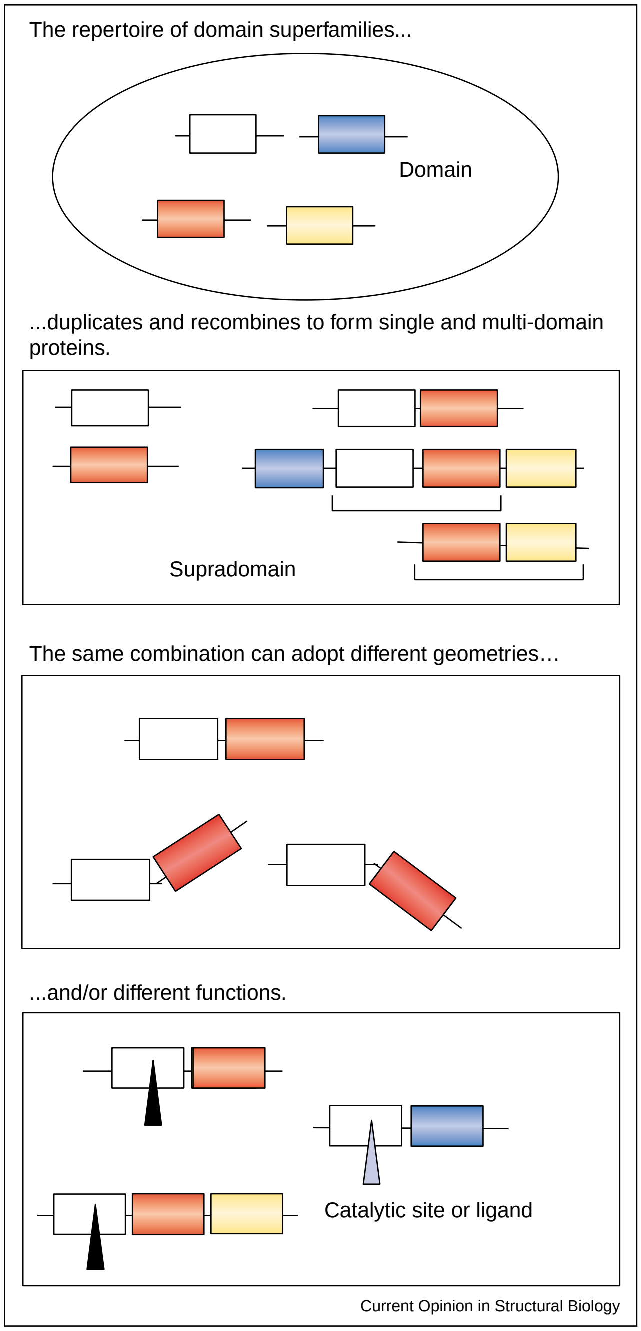
As most proteins consist of multiple domains, and domains determine the function and evolutionary relationships of proteins, it is important to understand the principles of domain combinations and interactions. In this review, we discuss how domain superfamilies form the repertoire of multidomain proteins via duplication and recombination (Figure 1). We then describe the principles and extent of conservation of the N- to C-terminal order of domains, their three-dimensional geometry and their functional relationships. This will illustrate the importance of domain combinations to an understanding of protein evolution, structure and function, and to target selection in structural genomics.

## Proteins are formed by duplication, divergence and recombination of domains

In order to understand how multidomain proteins function, it is useful to know how they are created in evolution and how they are related to each other. Duplication is one of the main sources for creation of new whole genes [13] and this is also true at the level of domains: at least 58% of the domains in *Mycoplasma* [11] and 98% of the domains in humans [6,7•,14•] are duplicates. The domains of different superfamilies are duplicated to different extents and this results in a distribution of superfamily sizes in genomes that follows a power law [15–17]. This means that there are a few highly abundant superfamilies, for example, the P-loop NTP hydrolases, NAD(P)-binding Rossmann domains and certain kinase families. The expansion of superfamilies in a particular phylogenetic group can deliver one explanation for the characteristics of the organisms in that group (e.g. the immunoglobulin proteins in metazoa) [18–22]. Once a domain or protein has duplicated, it can evolve a new or modified function either by sequence divergence or by combining with other domains to form a multidomain protein with a new series of domains. The N- to C-terminal series of domains in a protein is its ‘domain architecture’.

We will consider the recombination of domains in order to form different domain architectures in more detail.

**Figure 1**


The repertoire of domain superfamilies...
Domain
...duplicates and recombines to form single and multi-domain proteins.
Supradomain
The same combination can adopt different geometries…
...and/or different functions.
Catalytic site or ligand
Current Opinion in Structural Biology

The role of domains in protein evolution. Overview of different aspects of multidomain proteins: the repertoire of domain superfamilies and their role in the formation of multidomain proteins by duplication and recombination, and the geometry and functional relationships of domains within these combinations. Domains belonging to the same superfamily are represented as rectangles of the same colour. Supradomains are two- or three-domain combinations that occur in different domain architectures with different N- and C-terminal neighbours, as shown in the second panel. These short series of domains form functional units that are reused in different protein contexts.

The major molecular mechanism that leads to multidomain proteins and novel combinations is non-homologous recombination, sometimes referred to as ‘domain shuffling’. In eukaryotes, there is evidence that there is a tendency for exon boundaries to coincide with domain boundaries, which suggests that proteins may be formed by intronic recombination (e.g. [23,24]). Another important recombinatorial mechanism is the fusion of genes [25,26], which is more common than the splitting or fission of a gene [27,28].

## The properties of the repertoire of domain combinations

Our knowledge of the domain architectures of proteins stems from the assignment of structural domains to the whole or part of 40–60% of the predicted proteins from completely sequenced genomes [7•]. From examination of these proteins, it has become clear that the formation of new domain combinations is an important mechanism in protein evolution. The proteins from more than 100 different organisms contain several thousand different combinations of two superfamilies [29,30], but this is far fewer (less than 0.5%) than would be possible given the total number of superfamilies [31] or the number of multidomain proteins per proteome [29]. This number is likely to decrease even more if membrane proteins are included [32]. The limited repertoire of domain combinations that are observed in proteins indicates that all combinations have been under strong selection.

A few domain superfamilies are highly versatile and have neighbouring domains from many superfamilies, whereas most superfamilies are little versatile [14•,17,31,33]. The distribution of the number of partner superfamilies per superfamily follows a power law, like the distribution of superfamily sizes mentioned above. Despite these general principles, each domain superfamily has its own story. Some superfamilies are highly versatile, some are highly abundant and some superfamilies are both [29,30]. It is the structure and function of the domains and domain combinations that determine why they have been selected.

The properties observed for single domains are similar to those for combinations of two or more domains [29–31]. A few two-domain combinations, for example, are highly versatile and occur with many different additional domains, but most two-domain combinations occur in only one or two different protein contexts. Important examples of the reuse of particular domains and domain combinations come from signal transduction. For instance, the combination of SH3 and SH2 domains recurs in several different signal transduction proteins [34••,35•], and this versatility of recombination qualifies the SH3–SH2 domain pair as what we have called a ‘supradomain’ [30,36]. Supradomains are two- or three-domain combinations that occur in different domain architectures with different N- and C-terminal neighbours, a concept illustrated in Figure 1.

## The sequential order of domains is conserved

If the same domain combination is observed in two different proteins, one possibility is that they have evolved by duplication rather than assembled independently by different recombinatorial routes. Most instances of the same two-domain combination or domain architecture have evolved from the same ancestor; there are several lines of evidence that support this. First, three-dimensional structural analyses of individual protein families, such as the Rossmann domains [37••], have shown that proteins with the same domain architecture are related by descent (i.e. evolved from one common ancestor). Unpublished data by Kerrison, Chothia and Teichmann has shown that this is true for most two-domain protein families of known structure in the current databases [2]. Second, with only a small fraction of exceptions (less than 10%), two domains occur in only one N- to C-terminal order in structural assignments to genome sequences [31]. This conservation of domain order is likely to be historical instead of functional, as a very similar interface and functional sites could be formed by two domains in either order, for instance, given a long linker [37••]. The conserved order of domains can thus be exploited to improve domain assignments to protein sequences [38•]. Last, proteins sharing the same series of domains tend to have the same function [39], which is rarely the case if domain order is switched ([37••]; J Gough, personal communication).

## The geometry of domain combinations

Above, we discussed how the sequential order of domains within multidomain proteins of the same domain architecture is largely conserved and suggests homology. We will now describe the combinations and interactions of domains on a different level, that is, with respect to their three-dimensional arrangement or geometry. The geometry of Rossmann domains and their partner domains on one protein chain is conserved whenever the partner domains are from the same superfamily [37••]. Studies of small numbers of families of protein complexes, for which the interdomain geometry occurs across different polypeptide chains, have also revealed extensive conservation of geometry [40•,41]. These results suggested that, for proteins of unknown structure, their quaternary structure and complex geometry can usually be modelled based on homologous polypeptide(s) of known structure. Impressive examples of such structural models of complexes include the yeast ribosome [42,43] and exosome [44].

However, in order to thoroughly understand the evolution of interdomain interfaces and assess the true extent of conservation, a survey of all proteins of known structure is necessary. Aloy, Russell and co-workers [45•] found that there is a tendency for the geometry of interaction of protein domains to be more conserved the more similar the domain sequences are. They assessed similarity of domain geometry by comparing the average shift of a group of points in each of two domains. In order to understand the changes in geometry of domain combinations in more detail, Kerrison *et al.* (see also [14•]) studied the rotation, shift, interface size and residue contacts of related two-domain combinations (N Kerrison *et al.*, unpublished). They analysed 143 pairs of homologous proteins with two domains, extracted from SCOP version 1.63 [2]. They found that, when one pair of homologous domains are superposed, the positions of the two second domains mostly differ by shifts and rotations of less than 5 Å and 20°.

Although automatic approaches are useful to gain first insights into general relationships, manual inspection and alignment of the residues at the interface of domains can reveal the precise nature of the changes. This is illustrated by the examples in Figure 2. As shown in Figure 2a, homodimeric transcriptional repressors of the Iron-dependent repressor protein superfamily consist of a Winged helix domain (WHD) and a dimerisation domain. The two proteins shown have conserved interdomain geometry. In contrast, Figure 2b,c shows two different proteins for which the domain geometry has changed. Both proteins consist of a Homeodomain-like DNA-binding domain and a Tetracyclin repressor-like C-terminal domain. The rotation and shift of the latter domain become clear when the interface residues of the N-terminal domains are aligned (Figure 2c).

## Functional relationships of domains in multidomain proteins

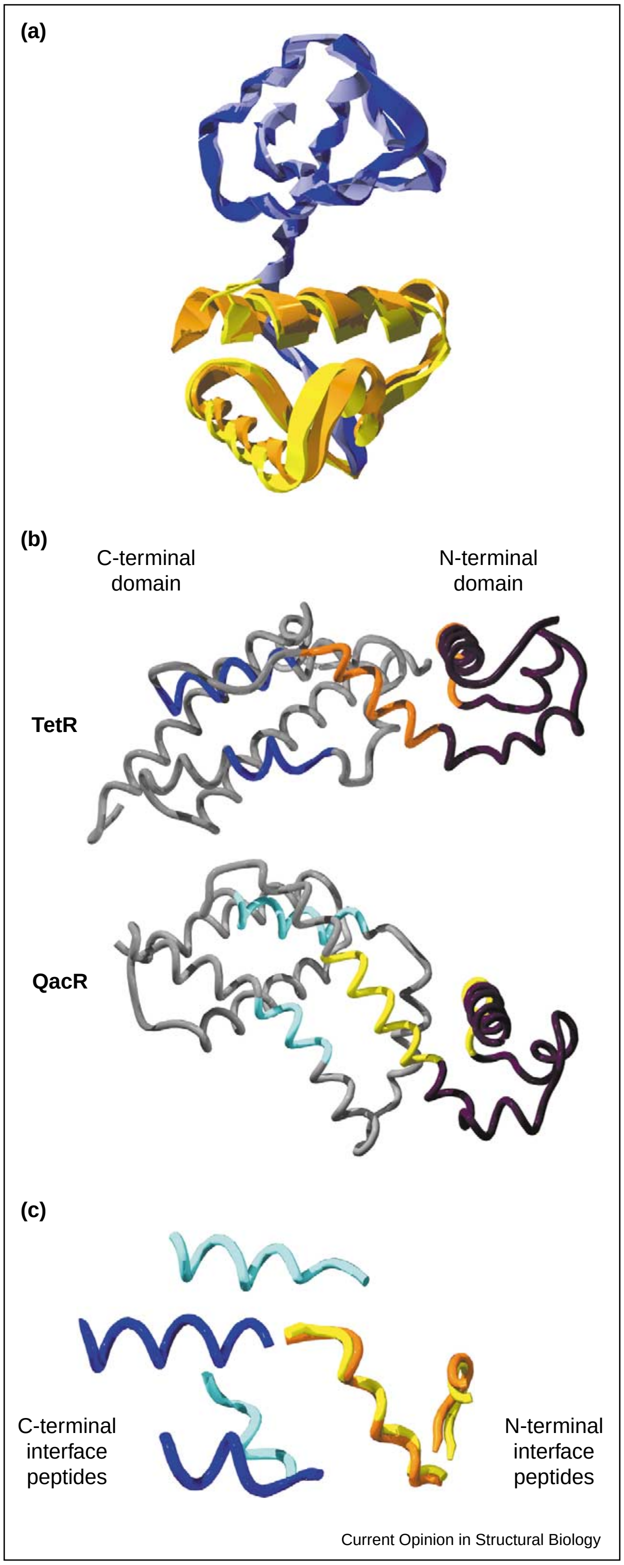
In order to delineate the functional relationships of domains in single- and multi-domain proteins, one needs a systematic understanding of the domain functions in different contexts, that is to say, the range of functions of a particular domain depending on its different partner domains. Existing functional classification schemes, such as GenProtEC [46], GO [47], MIPS [48], that used in COGs [49] and the EC (Enzyme Commission) classification [50], operate at the level of the whole protein and are thus inadequate to describe the contribution of the individual domains to protein function.

In order to understand the molecular roles of individual domains, it is vital to know their three-dimensional structure. Todd *et al.* [51,52] and Bartlett *et al.* [53••] have studied domain function and evolution at a detailed structural level, but were primarily concerned with individual superfamilies as opposed to domains in the context of their combinations. The role of 'ancillary' or 'accessory' domains is alluded to briefly in their work.

Bashton and Chothia (M Bashton, C Chothia, unpublished) have developed a domain-centric scheme that emphasises domain function in the context of domain neighbours in multidomain proteins, providing functional annotations for a subset of SCOP domains. The annotation is based on detailed examination of the protein structures, which is essential for understanding the precise molecular function of the domain and its contribution to the function of the whole protein. In this domain-centric functional classification scheme, domains are classified into seven categories that encompass catalytic activity, cofactor binding, responsibility for subcellular localisation, protein–protein interaction and so forth.

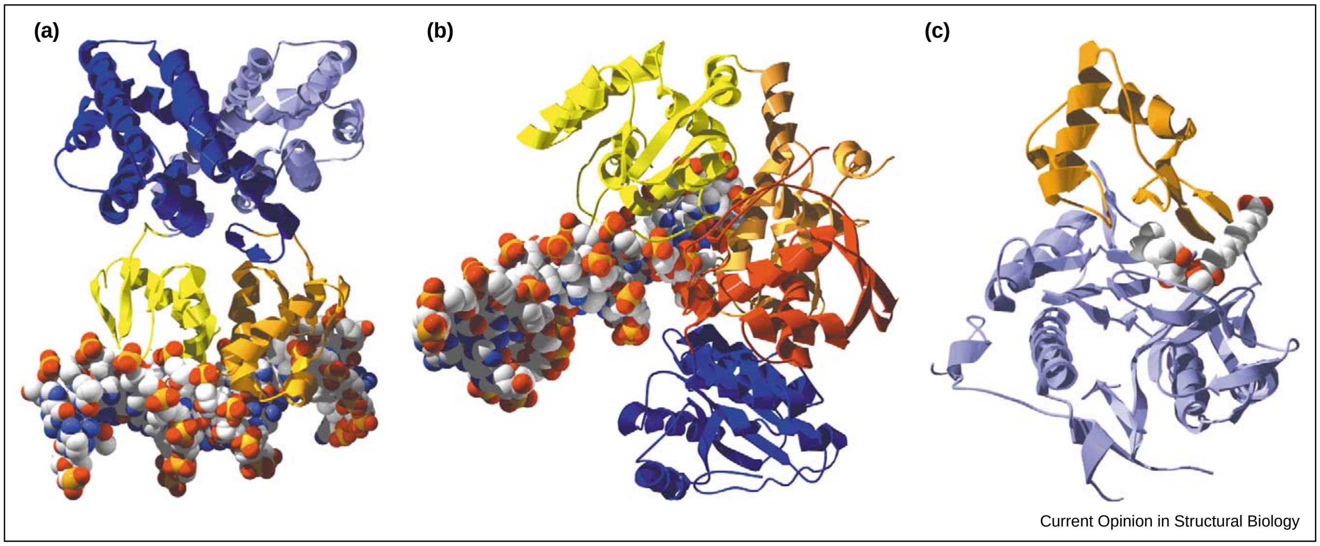
Two generic principles emerge. First, a domain can perform the same function, but in different protein contexts (i.e. with different partner domains). This is illustrated by Figures 2a and 3a,b. The WHDs in these examples combine with different sensory, regulatory and enzymatic domains, but maintain their function in that they target the protein to a specific sequence. In contrast, the WHD in Figure 3c acts as a substrate specificity pocket and has no DNA-binding activity at all. This domain has diverged and acquired a novel or modified function. In an analogy to linguistics, one can describe the two different fates of a

**Figure 2**


(a)
(b)
C-terminal domain
N-terminal domain
TetR
QacR
(c)
C-terminal interface peptides
N-terminal interface peptides
Current Opinion in Structural Biology

Geometry of domains in different transcriptional regulators. **(a)** Superposition of two chains shows that the geometry of the domain pair is conserved in the two proteins. The two structures are homodimeric transcriptional repressors consisting of a WHD and a dimerisation domain of the Iron-dependent repressor protein superfamily. **(b)** Two transcriptional repressors composed of a Homeodomain-like domain and a Tetracyclin repressor-like C-terminal domain are shown in such a way that the N-terminal domains (in black with orange and yellow interface peptides) are in the same orientation. The C-terminal domains are clearly rotated relative to each other. In this representation, the difference in geometry is apparent as a downward tilt of the C-terminal domain of TetR (dark blue helices) relative to the QacR C-terminal domain (light blue helices). **(c)** Residues forming contacts at the interdomain interfaces of the structures in (b). The orange and yellow interface peptides of the N-terminal domains are superposed, and the difference in the positions of the blue C-terminal interface peptides is clear. Again it appears as a downward tilt of the dark blue TetR helices compared with the light blue QacR interface peptides. Structural information. (a) Iron-dependent regulatory protein IdeR from *Mycobacterium tuberculosis* (PDB code 1b1b, chain A [61]) structurally aligned with diphtheria toxin repressor DtxR from *Corynebacterium diphtheriae* (PDB code 1g3t, chain B [62]). The N-terminal WHDs are shown in orange and yellow, and the C-terminal dimerisation domains are in different shades of blue. They have about 80% sequence identity to each other and the interface between the two domains is about 1400 Å$^2$ in both structures. (b) Tet repressor D from *Escherichia coli* (PDB code 2tct [63]) and the *Staphylococcus aureus* multidrug-binding protein QacR (PDB code 1jt6, chain A [64]). The N-terminal domains are DNA-binding Homeodomain-like domains and are shown in black, with the peptides forming the interdomain interface in orange and yellow. The C-terminal domains are Tetracyclin repressor-like and are shown in grey, with interface peptides in different shades of blue. The chains shown here are both in the ligand-bound state. The two proteins have about 10% sequence identity to each other both across the entire sequence and in the residues that form contacts at the interdomain interfaces. The interface between the domains is around 2100 Å$^2$ in both structures. (c) The residues from the N-terminal domains of the structures in (b) that form interface contacts, shown in orange and yellow, were superimposed and are shown in the same orientation. The difference in position of the C-terminal interface residues, shown in shades of blue, is apparent and corresponds to a shift of 10 Å and a rotation of 40°.

Figure 3



Variation in function of the Winged helix domain in different proteins. These three proteins each contain WHDs and illustrate how a superfamily can undergo syntactical and semantic shifts in protein function in different domain contexts. Many transcription factors are made by combining a WHD with a sensory or regulatory domain, as in FadR [65], shown in **(a)**, and in the proteins shown in Figure 2a. The WHD can also be found in enzymatic proteins, such as restriction endonucleases, where it combines with a catalytic domain that nicks DNA, as in the FokI protein [66], shown in **(b)**. In (a,b), the WHD performs the same role (i.e. it targets the protein to a specific sequence), but the range of functions is achieved by combining the WHD with different partner domains, so it is exhibiting a syntactical shift. **(c)** A semantic shift is found in human methionine aminopeptidase 2 [67], in which the WHD acts as a substrate specificity pocket with no DNA-binding activity at all. Structural information. (a) FadR (PDB code 1hw2 [65]): for the $\alpha$ chain, the WHD is orange and the oligomerisation/CoA-binding domain (fatty acid responsive transcription factor FadR, C-terminal domain) is dark blue; for the $\beta$ chain, the colours are yellow and light blue, respectively. (b) Restriction endonuclease FokI (PDB code 1fok [66]): the three WHDs are shown in yellow, orange and red (N- to C-terminal order), and the catalytic domain (Restriction endonuclease-like) is in blue. (c) Human methionine aminopeptidase 2 (PDB code 1boa [67]): the Creatinase/aminopeptidase domain is shown in light blue and the WHD is in orange. The WHDs of FadR and human methionine aminopeptidase 2 superpose with RMSD = 0.995 Å (not shown).

particular domain superfamily as syntactical and semantic shift, respectively.

## Domain combinations of known and unknown structure

The discussion of domain function above illustrates how knowledge of three-dimensional structures of proteins is key to a detailed understanding of how they work. For this reason, enormous efforts are currently underway in structural genomics projects to obtain complete structural coverage of all domain superfamilies and folds as far as possible [54]. However, beyond the structure determination of single domains, the structure determination of a wide range of domain combinations is crucial for a deeper understanding of protein relationships, functions and interactions, for the reasons we have discussed in the previous sections. Furthermore, for many domain architectures, multidomain proteins of known structure can be used confidently as a template for the domain geometry of homologous proteins of unknown structure ([44,45•]; N Kerrison *et al.*, unpublished).

Domain combinations can be prioritised for target selection according to different criteria: their abundance, their distribution across the three kingdoms of life or their versatility with respect to other combination partners [29]. Our concept of versatile domain combinations, or supradomains (Figure 1), introduces another useful filter [30]. As mentioned above, supradomains are two- or three-domain combinations that occur in different domain architectures with different N- and C-terminal neighbours. The 200 most duplicated two-domain supradomains, for example, occur in more than 75 000 sequences (28% of the sequences with domain assignments) from 113 archaeal, bacterial and eukaryotic completely sequenced genomes. Knowledge of their structure can hence provide insights into the function of almost one-third of the sequences in this data set. Table 1 lists the ten most abundant combinations of these 200 supradomains that do not have homologues of known structure. Almost all of these combinations occur in biochemically characterised proteins. However, they also occur in many other uncharacterised proteins. One example is the above-mentioned winged helix DNA-binding domain in combination with the periplasmic binding protein II domain (Table 1). This domain combination alone occurs in almost 2000 sequences of unknown function, many of which could be regulators like the two examples in the table. Exact knowledge of

**Table 1**

**The most duplicated two-domain combinations – targets for structure determination.**

| Domain combination | In how many sequences | Functional category (possible) | Known proteins with the domain combinations (examples) |
|---|---|---|---|
| Winged helix DNA-binding domain and periplasmic binding protein-like II | 1972[a] | DNA binding | OXYR_ECOLI: OxyR is a positive regulator of hydrogen-peroxide-inducible genes in *E. coli* and other bacteria, and is homologous to other regulatory proteins<br>NODD_RHILE: NodD is responsible for activating transcription of the Nod genes in the bacterium in response to plant inducers |
| Homeodomain-like and ribonuclease-H-like | 745[a] | DNA binding | TC3A_CAEEL: Transposase in *Caenorhabditis elegans* |
| Glucocorticoid-receptor-like (DNA-binding domain) and nuclear receptor ligand-binding domain | 576 | Signal transduction | PRGR_HUMAN: The human progesterone receptor is involved in the regulation of gene expression, and affects cellular proliferation and differentiation in target tissues |
| PYP-like sensor domain (PAS domain) and homodimeric domain of signal-transducing histidine kinase | 483[a] | Signal transduction | NTRB_ECOLI: NTRB acts as a signal transducer involved in nitrogen regulation in *E. coli*<br>TORS_ECOLI: The TorS sensor protein in *E. coli* is part of a two-component regulatory system |
| ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase and CheY-like | 414[a] | Signal transduction | ARCB_ECOLI: ArcB is a member of the two-component regulatory system arcB/arcA. Sensor-regulator protein for anaerobic repression of the arc modulon |
| Actin-like ATPase domain and heat shock protein 70 kDa (HSP70), C-terminal substrate-binding fragment | 344[a] | Chaperone | HSCC_ECOLI: Hsc62 is a DnaK homologue of *E. coli*<br>HS7F_CAEEL: The protein is a member of the Hsp70 multi-gene family of mitochondrial chaperones in *C. elegans* |
| Calcium ATPase, transmembrane domain M and HAD-like | 321[a] | Transport | PMA1_CANAL: Plasma membrane H-ATPase from *Candida albicans*. The H-pump produces a proton gradient that is used for active nutrient transport |
| Growth factor receptor domain and EGF/laminin | 282 | Signal transduction | MTN3_HUMAN: Matrilin-3 is an extracellular matrix protein and a major component of cartilage<br>FBL2_HUMAN: Fibulin-2 binds, depending on calcium, to fibronectin and other ligands<br>EGF_HUMAN: Human epidermal growth factor receptor |
| Winged helix DNA-binding domain and phosphosugar isomerase | 252 | DNA binding | GATR_ECOLI: A repressor of the GAT operon for galacticol transport and metabolism in *E. coli* |
| TRAF domain-like and POZ domain | 248 | Signal transduction | SPOP_HUMAN: Speckle-type POZ protein is an antigen recognised by serum from a scleroderma patient<br>The domain combination occurs in no other protein in SwissProt |

This table lists those two-domain combinations without homologues of known structure in decreasing order of occurrence in proteins of completely sequenced genomes. Only combinations of two different domains are shown; repetitions of the same domain are not listed. The last column provides one or more examples of biochemically characterised proteins from SwissProt (version 41.20 [68]) that contain the particular domain combination. It should be noted that some of these domain pairs that are common in genome sequences could be flexible instead of having a rigid interdomain geometry. If this is the case, structure determination is more difficult and less meaningful in terms of the domain geometry, though the domains are still functionally linked of course. [a] The domain combination occurs in all three kingdoms of life.

the properties of these domain combinations will contribute enormously to annotation in terms of protein function [39] and structure.

## Conclusions

We have provided an overview of the role of domain combinations in the formation of the protein repertoire. From the fairly comprehensive domain assignments that are available for completely sequenced genomes, it has become clear that the majority of proteins, even in simple genomes, are multidomain. Though the domain combinations observed are only a small fraction of all possible combinations of the repertoire of protein families, the emergence of new combinations is linked to speciation and specific phylogenetic groups: whereas more than half of all domain superfamilies are common to archaea, bacteria and eukaryotes, this is the case for only about 5% of two-domain combinations [7•,30,31]. Domain combinations and expansions of domain superfamilies, as well as other processes such as alternative splicing [55], play important roles in the emergence of more complex organisms [14•,20,22,56].

Proteins that contain the same domain combination or have the same domain architecture tend to have a common ancestor and common functional features. For instance, supradomains are combinations of domains that adopt a function that is useful within a variety of different domain architectures. Several domain assignment servers now offer tools for searches for particular domain combinations and architectures (e.g. SUPERFAMILY [7•], SMART [57], Pfam [58] and the Conserved Domain Architecture Retrieval Tool [CDART] [59] connected with the Conserved Domain Database [CDD] [60]).

In order to understand the molecular details of the functions of domain combinations, the three-dimensional structure of the domain architecture is needed. Structural genomics projects could therefore have novel combinations of domains, in addition to the structures of individual domains, as their aim. Target selection of domain combinations could be based on the number of proteins containing the domain combination, as well as the versatility of the domain combination in terms of occurrence in different domain architectures.

Knowledge of this kind could be integrated with genome-scale data of different types and contribute towards a more comprehensive understanding of the evolution of the structure and function of the protein repertoire.

## Acknowledgements

We are grateful to Jung-Hoon Han for help with Figure 2. CV has a pre-doctoral fellowship from the Boehringer Ingelheim Fonds.

## References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP - a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

2. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data**. *Nucleic Acids Res* 2004, **32**:D226-D229.

3. Chothia C: **Proteins - 1000 families for the molecular biologist**. *Nature* 1992, **357**:543-544.

4. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds**. *Nature* 1994, **372**:631-634.

5. Coulson AF, Moult J: **A unifold, mesofold, and superfold model of protein fold use**. *Proteins* 2002, **46**:61-71.

6. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919.

7. • Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements**. *Nucleic Acids Res* 2004, **32**:D235-D239.

The SUPERFAMILY database provides assignments of the over 1200 domain superfamilies, as defined in the SCOP database, to proteins using highly sensitive hidden Markov models. Close to 60% of all proteins have at least one match and one half of all residues are covered by assignments. The database is located at http://www.supfam.org and updated twice a year. SUPERFAMILY is now part of InterPro.

8. Kelley LA, MacCallum RM, Sternberg MJ: **Enhanced genome annotation using structural profiles in the program 3D-PSSM**. *J Mol Biol* 2000, **299**:499-520.

9. Buchan DW, Rison SC, Bray JE, Lee D, Pearl F, Thornton JM, Orengo CA: **Gene3D: structural assignments for the biologist and bioinformaticist alike**. *Nucleic Acids Res* 2003, **31**:469-473.

10. • Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P *et al.*: **The InterPro Database, 2003 brings increased coverage and new features**. *Nucleic Acids Res* 2003, **31**:315-318.

InterPro is a metaserver that combines several domain assignment servers, including PRINTS, PROSITE, Pfam, ProDom, SMART, TIGRFAMs and also SUPERFAMILY, and integrates information on protein families, domains and functional sites.

11. Teichmann SA, Park J, Chothia C: **Structural assignments to the *Mycoplasma genitalium* proteins show extensive gene duplications and domain rearrangements**. *Proc Natl Acad Sci USA* 1998, **95**:14658-14663.

12. Gerstein M: **How representative are the known structures of the proteins in a complete genome? A comprehensive structural census**. *Fold Des* 1998, **3**:497-512.

13. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**:1151-1155.

14. • Muller A, MacCallum RM, Sternberg MJ: **Structural characterization of the human proteome**. *Genome Res* 2002, **12**:1625-1641.

This large-scale study of the human and three other eukaryote proteomes, as well as several bacterial and archaeal proteomes, focuses on domain superfamilies rather than whole proteins. The authors describe the duplication and expansion of specific domain superfamilies in the human genome compared with other organisms. They also discuss transmembrane and disease-related proteins, and domain superfamilies in the human genome.

15. Wolf YI, Karev G, Koonin EV: **Scale-free networks in biology: new insights into the fundamentals of evolution?** *Bioessays* 2002, **24**:105-109.

16. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model**. *J Mol Biol* 2001, **313**:673-681.

17. Wuchty S: **Scale-free behavior in protein domain networks**. *Mol Biol Evol* 2001, **18**:1694-1702.

18. Hill E, Broadbent ID, Chothia C, Pettitt J: **Cadherin superfamily proteins in *Caenorhabditis elegans* and *Drosophila melanogaster***. *J Mol Biol* 2001, **305**:1011-1024.

19. Vogel C, Teichmann SA, Chothia C: **The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity**. *Development* 2003, **130**:6317-6328.

20. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T *et al.*: **Comparison of the complete protein sets of worm and yeast: orthology and divergence**. *Science* 1998, **282**:2022-2028.

21. Aravind L, Subramanian G: **Origin of multicellular eukaryotes - insights from proteome comparisons**. *Curr Opin Genet Dev* 1999, **9**:688-694.

22. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

23. Kaessmann H, Zollner S, Nekrutenko A, Li WH: **Signatures of domain shuffling in the human genome**. *Genome Res* 2002, **12**:1642-1650.

24. Patthy L: **Genome evolution and the evolution of exon-shuffling–a review**. *Gene* 1999, **238**:103-114.

25. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events**. *Nature* 1999, **402**:86-90.

26. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function**. *Nature* 1999, **402**:83-86.

27. Snel B, Bork P, Huynen M: **Genome evolution. Gene fusion versus gene fission**. *Trends Genet* 2000, **16**:9-11.

28. Yanai I, Wolf YI, Koonin EV: **Evolution of gene fusions: horizontal transfer versus independent events**. *Genome Biol* 2002, **3**:research0024.

29. Apic G, Huber W, Teichmann SA: **Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination**. *J Struct Funct Genomics* 2003, **4**:67-78.

30. Vogel C, Berzuini C, Bashton M, Gough J, Teichmann SA: **Supra-domains - evolutionary units larger than single protein domains**. *J Mol Biol* 2004, in press.

31. Apic G, Gough J, Teichmann SA: **Domain combinations in archaeal, eubacterial and eukaryotic proteomes**. *J Mol Biol* 2001, **310**:311-325.

32. Liu Y, Gerstein M, Engelman DM: **Evolutionary use of domain recombination: a distinction between membrane and soluble proteins**. *Proc Natl Acad Sci USA* 2004, in press.

33. Park J, Lappe M, Teichmann SA: **Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast**. *J Mol Biol* 2001, **307**:929-938.

34. Harrison SC: **Variation on an Src-like theme**. *Cell* 2003,
•• **112**:737-740.

This review presents a good example of domains that reappear in different domain contexts: the SH3, SH2 and kinase domains recombine with various other domains in signal transduction multidomain proteins.

35. Pawson T, Nash P: **Assembly of cell regulatory systems**
• **through protein interaction domains**. *Science* 2003, **300**:445-452.

This review illustrates the modularity of proteins in a colourful manner. It describes the reuse of protein interaction domains, such as SH2 and SH3 domains and others, in the regulation of different cellular processes. The authors focus on the properties of single domains, but also point out the increase in dimensionality of functions and interactions when domains are combined to form multidomain proteins.

36. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire**. *Science* 2003, **300**:1701-1703.

37. Bashton M, Chothia C: **The geometry of domain combination in**
•• **proteins**. *J Mol Biol* 2002, **315**:927-939.

This study presents a detailed analysis of the structures of proteins containing Rossmann fold domains in combination with other domain superfamilies. It demonstrates that, in all the cases studied, the N- to C-terminal order of the domains is conserved because the proteins have descended from a common ancestor. For pairs of proteins in the PDB in which the order is reversed, the interface and functional relationships of the domains are altered.

38. Coin L, Bateman A, Durbin R: **Enhanced protein domain**
• **discovery by using language modeling techniques from speech recognition**. *Proc Natl Acad Sci USA* 2003, **100**:4516-4520.

The technique presented in this paper formalises the use of domain associations for the improvement of domain assignment to protein sequences. In analogy to speech recognition methods that use context information to improve recognition of words, assignment of domains is improved using information on their domain combination context in multidomain proteins.

39. Hegyi H, Gerstein M: **Annotation transfer for genomics: measuring functional divergence in multi-domain proteins**. *Genome Res* 2001, **11**:1632-1640.

40. Aloy P, Russell RB: **Interrogating protein interaction networks**
• **through structural biology**. *Proc Natl Acad Sci USA* 2002, **99**:5896-5901.

The authors describe a method to assess the likelihood of an interface forming between two proteins when the components are modelled on complexes of known structure.

41. Prabu MM, Suguna K, Vijayan M: **Variability in quaternary association of proteins with the same tertiary fold: a case study and rationalization involving legume lectins**. *Proteins* 1999, **35**:58-69.

42. Spahn CM, Beckmann R, Eswar N, Penczek PA, Sali A, Blobel G, Frank J: **Structure of the 80S ribosome from saccharomyces cerevisiae–tRNA-ribosome and subunit-subunit interactions**. *Cell* 2001, **107**:373-386.

43. Beckmann R, Spahn CM, Eswar N, Helmers J, Penczek PA, Sali A, Frank J, Blobel G: **Architecture of the protein-conducting channel associated with the translating 80S ribosome**. *Cell* 2001, **107**:361-372.

44. Aloy P, Ciccarelli FD, Leutwein C, Gavin AC, Superti-Furga G, Bork P, Bottcher B, Russell RB: **A complex prediction: three-dimensional model of the yeast exosome**. *EMBO Rep* 2002, **3**:628-635.

45. Aloy P, Ceulemans H, Stark A, Russell RB: **The relationship**
• **between sequence and interaction divergence in proteins**. *J Mol Biol* 2003, **332**:989-998.

Using RMSD as a simple measure to compare interactions between domains, the authors found that homologues with more than 30% sequence identity usually conserve the geometry of their interaction.

46. Serres MH, Goswami S, Riley M: **GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins**. *Nucleic Acids Res* 2004, **32**:D300-D302.

47. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C *et al.*: **The Gene Ontology (GO) database and informatics resource**. *Nucleic Acids Res* 2004, **32**:D258-D261.

48. Mewes HW, Amid C, Arnold R, Frishman D, Guldener U, Mannhaupt G, Munsterkotter M, Pagel P, Strack N, Stumpflen V *et al.*: **MIPS: analysis and annotation of proteins from whole genomes**. *Nucleic Acids Res* 2004, **32**:D41-D44.

49. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN *et al.*: **The COG database: an updated version includes eukaryotes**. *BMC Bioinformatics* 2003, **4**:41.

50. Bairoch A: **The ENZYME database in 2000**. *Nucleic Acids Res* 2000, **28**:304-305.

51. Todd AE, Orengo CA, Thornton JM: **Evolution of protein function, from a structural perspective**. *Curr Opin Chem Biol* 1999, **3**:548-556.

52. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**:1113-1143.

53. Bartlett GJ, Borkakoti N, Thornton JM: **Catalysing new reactions**
•• **during evolution: economy of residues and mechanism**. *J Mol Biol* 2003, **331**:829-860.

A detailed analysis of catalytic sites and residues in homologous enzymes of different function reveals the economy of evolution: the residue types, functions and mechanistic steps are frequently conserved.

54. Brenner SE: **Target selection for structural genomics**. *Nat Struct Biol* 2000, **7(suppl)**:967-969.

55. Kriventseva EV, Koch I, Apweiler R, Vingron M, Bork P, Gelfand MS, Sunyaev S: **Increase of functional diversity by alternative splicing**. *Trends Genet* 2003, **19**:124-128.

56. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes**. *Genome Res* 2002, **12**:1048-1059.

57. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource**. *Nucleic Acids Res* 2002, **30**:242-244.

58. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database**. *Nucleic Acids Res* 2002, **30**:276-280.

59. Geer LY, Domrachev M, Lipman DJ, Bryant SH: **CDART: protein homology by domain architecture. Conserved Domain Architecture Retrieval Tool**. *Genome Res* 2003, **12**:1619-1623.

60. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure**. *Nucleic Acids Res* 2002, **30**:281-283.

61. Pohl E, Holmes RK, Hol WG: **Crystal structure of the iron-dependent regulator (IdeR) from *Mycobacterium tuberculosis***

**shows both metal binding sites fully occupied**. *J Mol Biol* 1999, **285**:1145-1156.

62. Pohl E, Goranson-Siekierke J, Choi MK, Roosild T, Holmes RK, Hol WG: **Structures of three diphtheria toxin repressor (DtxR) variants with decreased repressor activity**. *Acta Crystallogr* 2001, **57**:619-627.

63. Kisker C, Hinrichs W, Tovar K, Hillen W, Saenger W: **The complex formed between Tet repressor and tetracycline-Mg$^{2+}$ reveals mechanism of antibiotic resistance**. *J Mol Biol* 1995, **247**:260-280.

64. Schumacher MA, Miller MC, Grkovic S, Brown MH, Skurray RA, Brennan RG: **Structural mechanisms of QacR induction and multidrug recognition**. *Science* 2001, **294**:2158-2163.

65. Xu Y, Heath RJ, Li Z, Rock CO, White SW: **The FadR.DNA complex. Transcriptional control of fatty acid metabolism in *Escherichia coli***. *J Biol Chem* 2001, **276**:17373-17379.

66. Wah DA, Hirsch JA, Dorner LF, Schildkraut I, Aggarwal AK: **Structure of the multimodular endonuclease FokI bound to DNA**. *Nature* 1997, **388**:97-100.

67. Liu S, Widom J, Kemp CW, Crews CM, Clardy J: **Structure of human methionine aminopeptidase-2 complexed with fumagillin**. *Science* 1998, **282**:1324-1327.

68. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al.*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**:365-370.