The Generation of New Protein Functions by the Combination of Domains

Matthew Bashton[1,2,*] and Cyrus Chothia[1]
[1] MRC Laboratory of Molecular Biology, Hills Road, Cambridge, CB2 2QH, United Kingdom
[2] EMBL—European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, United Kingdom
*Correspondence: bashton@ebi.ac.uk

DOI 10.1016/j.str.2006.11.009

SUMMARY

During evolution, many new proteins have been formed by the process of gene duplication and combination. The genes involved in this process usually code for whole domains. Small proteins contain one domain; medium and large proteins contain two or more domains. We have compared homologous domains that occur in both one-domain proteins and multidomain proteins. We have determined (1) how the functions of the individual domains in the multidomain proteins combine to produce their overall functions and (2) the extent to which these functions are similar to those in the one-domain homologs. We describe how domain combinations increase the specificity of enzymes; act as links between domains that have functional roles; regulate activity; combine within one chain functions that can act either independently, in concert or in new contexts; and provide the structural framework for the evolution of entirely new functions.

INTRODUCTION

During the course of evolution, the process of gene duplication, sequence divergence, and gene combination has produced many proteins that have new or modified functions. A number of previous studies have described in detail how gene duplication and sequence divergence produce proteins with new properties. Gerlt and Babbitt (2001) have reviewed their work on enzymes that belong to mechanistically diverse superfamilies (those that conserve some part of their catalytic mechanism) and to functionally distinct superfamilies (those that do not conserve a common mechanism). Nahum and Riley (2001) described how sequence divergence in families of *E. coli* proteins has produced proteins with different functions. Todd et al. (2001) in a detailed study of enzymes in 31 superfamilies described point mutations and local structural changes that occur within homologous domains to modify their functions. They also studied 22 pairs of homologous proteins in which one protein is an enzyme and one is not (Todd et al., 2002). Bartlett et al. (2003) made a detailed analysis of catalytic residue conservation and variation in 27 pairs of enzymes that are homologous but have very different functions.

Although there have been detailed investigations of the roles that duplication and sequence divergence play in producing proteins with new functions, far less attention has been paid to the role of gene combination. Hegyi and Gerstein (2001) compared the enzyme commission (EC) numbers and SWISS-PROT keywords of one-domain proteins with those of multidomain proteins that included homologs of the domain in the one-domain protein. They did not however discuss how separate functions of domains combine to create the functions of a whole protein.

The units that are usually involved in combination are the genes, or gene segments, that code for protein domains. Small proteins are formed by a single domain. Most proteins are formed by two or more domains.

Here, we consider 45 sets of proteins. Each set contains one or more one-domain protein and a protein with two or more domains. The one-domain proteins are homologous to one or more of the domains in the multidomain member of the set. For the proteins in these sets, the functions and structures of both the one-domain and multidomain proteins have been well characterized. In all cases, the multidomain protein has a function that is more specific or more complex than that of the one-domain protein. By an examination of the functions and structures of each pair, we determine, at least in outline, the main contribution that gene combination makes to the more specific or more complex functions of the multidomain proteins.

The descriptions given here are not, of course, complete descriptions of the contributions that each domain makes to the functions of the multidomain protein. Even if the proteins were characterized well enough for us to do this, a complete description would require a paper on each set of structures. Here, we determine the major contributions and roles made by each domain to the function of the multidomain protein, and how these are the same as, or different from, the functions found in the one-domain homologs.

A LIBRARY OF PROTEINS WITH HOMOLOGOUS DOMAINS IN ONE-DOMAIN AND TWO- OR MORE DOMAIN PROTEINS

The domain definitions used in this study are taken from the Structural Classification of Proteins (SCOP) database

Table 1. The 45 Sets of One-Domain and Multidomain Proteins that Contain Homologous Domains

Number	One-Domain Protein(s)	Multidomain Proteins	Functional Modification/ Conservation
1	Glucoamylas **1ayx** *3.2.1.3*	Endo/exocellulase:cellobiose E-4 **1js4** *3.2.1.4*	1a C
_	b.2.2		_
2	β-amylase **1bfn** *3.2.1.2*	β-amylase **1b90** *3.2.1.2*	1a C
_	b.3.1		_
3	Bacterial phospholipase C **1ah7** *3.1.4.3*	α-toxin **1ca1** *3.1.4.3*	1a C
_	b.12.1		_
4	Guanylate kinase **1ex7** *2.7.4.8*	Adenylate kinase **1ak2** *2.7.4.3*	1a C
_	g.41.2		_
5a	Glycinamide ribonucleotide transformylase (GART) **1jkx** *2.1.2.2*	Methionyl-tRNA^{fmet} formyltransferase **2fmt** *2.1.2.9*	1a C
5b	3-methyladenine DNA glycosylase **1ewn** *3.2.2.21*	Methionyl-tRNA^{fmet} formyltransferase **2fmt** *2.1.2.9*	5 P
6a	Asparagine synthetase **12as** *6.3.1.1*	Aspartyl-tRNA synthetase (AspRS) **1asy** *6.1.1.12*	1a C
6b	ssDNA-binding protein **1eyg**	Aspartyl-tRNA synthetase (AspRS) **1asy** *6.1.1.12*	1b C
7	RNA methyltransferase FtsJ **1ej0** *2.1.1.-*	Chemotaxis receptor methyltransferase **1af7** *2.1.1.79*	1a C
_	a.58.1		_
8	Lysozyme **1lys** *3.2.1.17*	Lytic transglycosylase Slt70 **1qsa** *3.2.1.-*	1a C
_	a.118.5		_
9	Kanamycin nucleotidyltransferase (KNTase) **1kny** *2.7.7.-*	DNA polymerase β **1bpd** *2.7.7.7*	1a C
_	a.60.6		_
10	γ-glutamyl hydrolase **1l9x** *3.4.19.9*	Carbamoyl phosphate synthetase small subunit **1a9x** *6.3.5.5 (Whole complex)*	1a C
_	c.8.3		_
11	Proline iminopeptidase **1azw** *3.4.11.5*	Prolyl oligopeptidase **1e5t** *3.4.21.26*	1a C
_	b.69.7		_
12	Tryptophanyl-tRNA synthetase **1m83** *6.1.1.2*	Methionyl-tRNA synthetase **1a8h** *6.1.1.10*	1a C
_	a.27.1		_
13a	Tryptophanyl-tRNA synthetase **1m83** *6.1.1.2*	Glutaminyl-tRNA synthetase **1euq** *6.1.1.18*	1a C
13b	Ribosomal protein L25 **1dfu**	Glutaminyl-tRNA synthetase **1euq** *6.1.1.18*	1b C
14	Mitochondrial cytochrome c_6 **1c75**	Cytochrome cd_1 **1aof** *1.9.3.2*	1b C
_	b.70.2		_
15a	Soluble, respiratory-type Rieske protein **1nyk**	Naphthalene 1,2-dioxygenase α subunit **1eg9** *1.14.12.12*	1b C
15b	Phoshatidylinositol transfer protein (PITP) **1fvz**	Naphthalene 1,2-dioxygenase α subunit **1eg9** *1.14.12.12*	6 N
16	Plastocyanin **1pcs**	Nitrous oxide reductase **1qni** *1.7.99.6*	1b C
_	b.69.3		_

Table 1. *Continued*

Number	One-Domain Protein(s)	Multidomain Proteins	Functional Modification/ Conservation
17a	Flavodoxin **1ag9**	Ruberdoxin: oxygen oxidoreductase **1e5d**	1b C
17b	Zn metallo-β-lactamase **1dxk** *3.5.2.6*	Ruberdoxin: oxygen oxidoreductase **1e5d**	7 N
18	Ferredoxin II **1fxd**	Iron hydrogenase large (catalytic) subunit **1hfe** *1.18.99.1*	1b C
_	c.96.1		_
19a	Ruberdoxin **1rb9**	Rubrerythrin **1b71**	1b C
19b	Bacterioferritin (cytochrome b_1) **1bcf**	Rubrerythrin **1b71**	6 N
20	Carboxypeptidase A **1f57** *3.4.17.1*	Peptidase T **1fno** *3.4.11.4*	1c C
_	d.58.19		_
21a	β-glucanase **1ghs** *3.2.1.39*	β-glucuronidase **1bhg** *3.2.1.31*	1d C
21b	Fucose-binding lectin **1k12**	β-glucuronidase **1bhg** *3.2.1.31*	4 N
21c	Glactose mutarotase **1nsx** *5.1.3.3*	β-galactosidase **1jz7** *3.2.1.23*	4 N
_	b.1.4		_
22	β-lactamase **1erm** *3.5.2.6*	Penicillin-recognizing enzyme **1ei5** *3.4.11.19*	1d C
_	b.61.3		_
23	Dual-specificity protein phosphatase VHR (vhr) *3.1.3.48*	Tyrosine phosphatase **2shp** *3.1.3.48*	1e C
_	d.93.1		_
24	B12-dependent (class II) ribonucleotide reductase **1l1l** *1.17.4.2*	R1 subunit of ribonucleotide reductase. Catalyzes the synthesis of deoxyribonucleotides **1r1r** *1.17.4.1*	1e C
_	a.98.1		_
25a	MARR antibiotic-resistance repressor **1jgs**	Molybdate-dependent transcriptional regulator ModE 1b9m	1f C
25b	Molybdate/tungstate-binding protein II **1gug**	Molybdate-dependent transcriptional regulator ModE **1b9m**	4 C
26a	Bacteriophage lambda repressor **1mb**	Purine repressor (PurR) **1bdh**	1f C
26b	Ribose-binding protein **2dri**	Purine repressor (PurR) **1bdh**	4 N
27a	Penicillin V acylase **2pva** *3.5.1.11*	Asparagine synthetase B **1ct9** *6.3.5.4*	2 C
27b	NH3-dependent NAD+-synthetase **1kqp** *6.3.5.1*	Asparagine synthetase B **1ct9** *6.3.5.4*	2 C
28a	Xanthine-guanine PRTase **1nul** *2.4.2.22*	Glutamine 5-phospho-ribosyl-1-pyrophosphate (PRPP) amidotransferase **1ecc** *2.4.2.14*	2 C
28b	Putative glutamine amidotransferase **1te5**	Glutamine 5-phospho-ribosyl-1-pyrophosphate (PRPP) amidotransferase **1ecc** *2.4.2.14*	2 C
29a	Trypsin 1 **1trn** *3.4.21.4*	NS3 protease **1cu1** *3.4.21.-*	2 C
29b	Guanylate kinase **1ex7** *2.7.4.8*	NS3 protease **1cu1** *3.4.21.-*	2 N
30a	UDP-N-acetylglucosamine acyltransferase **1lxa** *2.3.1.129*	*N*-Acetylglucosamine-1-PO_4 uridyltransferase (GlmU) **1hv9** *2.3.1.157 and 2.7.7.23*	2 C
30b	Uridyl transferase **1jvd** *2.7.7.23*	*N*-Acetylglucosamine-1-PO_4 uridyltransferase (GlmU) **1hv9** *2.3.1.157 and 2.7.7.23*	2 C

(*Continued on next page*)

Table 1. *Continued*

Number	One-Domain Protein(s)	Multidomain Proteins	Functional Modification/ Conservation
31a	RNase H **1ril** *3.1.26.4*	DNA polymerase II **1d5a** *2.7.7.7*	2 C
31b	Bacteriophage T7 RNA Polymerase **1msw** *2.7.7.6*	DNA polymerase II **1d5a** *2.7.7.7*	2 C
32a	Phosphoribulokinase **1a7j** *2.7.1.19*	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase **1bif** *2.7.1.105 and 3.1.3.46*	2 C
32b	Prostatic acid phosphatase **1nd6** *3.1.3.2*	6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase **1bif** *2.7.1.105 and 3.1.3.46*	2 C
33a	FHIT (fragile histidine triad protein) **1fit** *3.6.1.29*	NitFhit fusion protein **1ems** *3.6.1.29*	2 C
33b	N-carbamoyl-D-amino acid amidohydrolase **1uf5** *3.5.1.77*	NitFhit fusion protein **1ems** *3.6.1.29*	2 N
34a	I-*Cre*I **1bp7**	PI-SceI **1dfa**	2 C
34b	GyrA intein **am2**	PI-SceI **1dfa**	2 C
35a	Cro lambda repressor **5cro**	SinR repressor **1b0n**	3 P
35b	SinI antirepressor **1b0n**	SinR repressor **1b0n**	8 C
36a	*Drosophila* (short chain) alcohol dehydrogenase **1b16** *1.1.1.1*	Alcohol dehydrogenase **1a71** *1.1.1.1*	3 P
36b	Chaperonin-10 (GroES) **1aon**	Alcohol dehydrogenase **1a71** *1.1.1.1*	6 N
37a	EcoRV **1rva** *3.1.21.4*	Restriction endonuclease (FokI) **1fok** *3.1.21.4*	3 C
37b	MARR antibiotic-resistance repressor **1jgs**	Restriction endonuclease (FokI) **1fok** *3.1.21.4*	1b C
38a	Hah1 Metallochaperone **1fe0**	Copper chaperone of superoxide dismutase (CCS) **1qup**	3 P
38b	Cu, Zn superoxide dismutase (SOD) **1cbj**	Copper chaperone of superoxide dismutase (CCS) **1qup**	5 P
39	[2Fe-2S] ferredoxin **1czp**	Iron protein from quinol-fumarate reductase. (Binds iron clusters in the fumarate reductase complex) **1fum** *1.3.99.1*	4 C
_	a.1.2		_
40a	Glycolate oxidase **1al8** *1.1.3.15*	Flavocytochrome b_2 **1fcb** *1.1.2.3*	4 C
40b	Cytochrome b_5 **1cyo**	Flavocytochrome b_2 **1fcb** *1.1.2.3*	4 C
41a	Ribosomal protein L24 **1jj2**	Ribosomal protein L2 **1ffk**	4 C
41b	Single-stranded DNA-binding protein **1eyg**	Ribosomal protein L2 **1ffk**	4 C
42	Ribosomal protein S9 **1j5e**	Ribosomal protein S5 **1j5e**	4 C
_	d.50.1		_
43	FK-506-binding protein (FKBP12) **1fkh** *5.2.1.8*	GreA transcript cleavage factor **1grj**	5 N
_	a.2.1		_
44	Pyruvoyl-dependent aspartate decarboxylase (ADC) **1aw8** *4.1.1.11*	Formate dehydrogenase H **1aa6** *1.2.1.2*	5 N
_	c.81.1		_

Table 1. *Continued*

Number	One-Domain Protein(s)	Multidomain Proteins	Functional Modification/ Conservation
45a	Threonine synthase **1e5x** *4.2.3.1*	Allosteric threonine deaminase **1tdj** *4.2.1.16*	7 P
45b	Putative glycine cleavage system repressor **1u8s**	Allosteric threonine deaminase **1tdj** *4.2.1.16*	4 P

In cases in which there is more than one known single-domain homolog for a domain combination, the different pairs are separated out with the letters a, b, and c. Where no homologous single-domain protein is available, a "_" is noted and the SCOP superfamily is given. The names and PDB codes as well as enzyme commission (EC) numbers (in italics) are given where available. The type of functional modification is given for each pair in the final column and corresponds to those listed in Table 2. Also, the state of functional conservation from the single-domain environment to the multidomain protein for a particular homologous domain is given as "C" for conserved; "N" for nonconserved; and "P" for partial conservation.

(Murzin et al., 1995). In the SCOP classification scheme, a domain is an evolutionary unit, rather than a structural one. This means that for a protein to be split into domains these regions must be seen elsewhere in a different structural context: they must be found in combination with different domains and/or in isolation.

SCOP domains are classified into families (evolutionary relationship shown by residue identity) and superfamilies (evolutionary relationship shown by features of their structure, function, and sequence). Here, we use domains classified at the superfamily level. In SCOP, the superfamilies of domains are identified by labels that take the form of c.2.1: the first character identifies the class of the protein, the second character identifies the fold, and the third character indicates the particular superfamily. Our work is based on the 1.65 version of the SCOP database.

To create our data set, we first collected a list of all proteins in SCOP that contained at least two domains that come from different superfamilies. Proteins have the same domain architecture when they have domains from the same superfamilies in the same sequential order. From this list, one representative structure was chosen to represent each domain architecture: usually the one whose function is the best characterized. This resulted in a list of 172 multidomain proteins. The domains in these proteins belong to 1 of 255 superfamilies. For each of these superfamilies, SCOP was then searched for the presence of a functional one-domain homolog that is not a fragment of a larger protein. Where there were several one-domain homologs, we chose the one whose function is closest to that in the multidomain protein. Of the 255 superfamilies present in the two-domain proteins, 83 had a corresponding one-domain homolog.

We then initially searched for descriptions for the functions of the one-domain proteins and their homologs in the multidomain protein by reading the primary literature listed in the protein data bank PDB entry (Berman et al., 2000) for that structure. This was followed by a literature search to retrieve functional information produced since structural determination. In some cases, details of the function of a domain in the multidomain protein and/or the one-domain protein(s) are not known. However, there were 45 sets of proteins for which we could attribute functions to each domain in the multidomain protein and to the one-domain protein(s) that is homologous to one or more of the domains in multidomain protein.

PRESENTATION OF THE RESULTS

In Table 1, we list the one-domain and multidomain proteins in the 45 sets. All but one of the sets has one multidomain protein that is formed by domains from two superfamilies. Of these, 23 have one-domain proteins that belong to both of the superfamilies that form the multidomain protein. Another 21 sets have a single one-domain protein that belongs to one of the superfamilies in the multidomains proteins: members of the other superfamilies are, at present, only observed in domain combinations. There is one set with two related multidomain proteins whose domains come from four superfamilies (entry 21 in Table 1): one-domain homologs are known for three of these superfamilies.

For one-domain proteins and multidomain proteins of known structure, we give the full name of the protein and the PDB code of a representative structure. For the 21 domains in the multidomain proteins for which no one-domain homolog is currently known, we give the SCOP identifier of its superfamily.

When homologs of one-domain proteins combine with other domains to form multidomain proteins, their functions are modified or changed. The comparison of the functions of the one-domain protein(s) and its homolog in multidomain proteins in the 45 sets allows us to assign their functional modification to 1 of 7 types: see Table 2. Below, we discuss and illustrate examples of the 7 types, as seen in 11 of the 45 sets. In Table 1, we indicate the nature of the changes that occur in all sets; in the Supplemental Data available with this article online, an expanded version of Table 1 (Table S1) gives further details of these changes.

In this paper, we always refer to the multidomain protein as having a function that is changed, modified, or copied from that found in the one-domain protein. We believe that going from the simple to the complex is the most plausible scenario. We do not, of course, believe that the opposite process never occurs. The central interest of this

Table 2. The Seven Types of Functional Modifications Produced by Domain Combinations

Functional Modification	Quantity
1. Proteins whose functions are modified by additional domains:	
a. domains that modify substrate binding	13
b. domains that are catalytic	9
c. domains that modify substrate binding through the formation of oligomers	1
d. domains that link domains that have a direct role in function	2
e. domains that regulate enzyme function	2
f. domains that regulate DNA binding	2
2. Gene fusions that form bifunctional enzymes	16
3. Transfer of a function to an additional domain	4
4. Domains whose combination allows them to function in new contexts	11
5. Loss of catalytic function by a homolog in a domain combination	4
6. Gain of a catalytic activity in a domain combination	3
7. Change in catalyzed reaction in domain combination	2
Domain that conserves its dimerization property (not a modification)	1
Total	70

paper is how combinations of domains have functions that are different or similar to those in one-domain proteins, and the direction in which their evolution has taken place does not qualify these relationships.

ENZYMATIC DOMAINS THAT LARGELY CONSERVE THEIR CATALYTIC PROPERTIES IN DOMAIN COMBINATION BUT HAVE THEIR FUNCTIONS MODIFIED BY OTHER DOMAINS

In this section, we discuss cases in which a one-domain enzyme and a multidomain enzyme both have homologous catalytic domains that largely conserve their function. The additional domains in the multidomain enzyme assist with substrate binding or modulate the function of the catalytic domain in some other way. We consider enzymatic functions to be largely conserved when they meet one of two criteria:

1. the EC numbers are available for both proteins, and at least the first three EC numbers are the same;
2. they carry out the same reaction, e.g., transfer of a NH_2 group, on a related but different substrate(s).

We have chosen to introduce the second criterion, as it is known that proteins with dissimilar EC numbers can have similar reactions (Gerlt and Babbitt, 2001; Nahum and Riley, 2001).

In most of the enzymes discussed in this section, the one-domain protein and its homologous domain in combination with other domains have related substrates, related reaction chemistry, and the same first three EC numbers. We breakdown the cases in this category into separate subcategories depending on the way in which the additional domains modulate the function of the conserved catalytic domain.

Domains that Modify Substrate Binding: Figure 1 and Entry 6

In this large subcategory, there are 13 cases of homologs of one-domain enzymes that catalyze the same or a similar reaction but have their substrate specificity modified or extended by an additional domain: see Table 1, entries 1–13.

Asparagine Synthetase and Aspartyl-tRNA Synthetase

Example of an additional domain extending the substrate-binding properties of a catalytic domain are found in the one-domain enzyme asparagine synthetase (Nakatsu et al., 1998) and the two-domain enzyme aspartyl-tRNA synthetase (Ruff et al., 1991). Both enzymes have homologous catalytic domains that are members of the “Class II aaRS and biotin synthetases superfamily” (d.104.1).

In addition, the synthetase has an N-terminal domain that is a member of the nucleic acid-binding proteins superfamily (b.40.4). This domain is mainly responsible for the specificity of the synthetase for $tRNA^{Asp}$ (Figure 1). Other members of the superfamily bind single-standard DNA and play important roles in DNA replication and recombination.

The one-domain protein asparagine synthetase catalyzes the synthesis of asparagine from aspartic acid via an aminoacyl-adenylate intermediate produced by ATP and ammonia (Nakatsu et al., 1998). The catalytic C-terminal domain of aspartyl-tRNA synthetase also uses an aminoacyl-adenylate intermediate to bind aspartic acid to an NH_2 group of the terminal nucleotide of $tRNA^{Asp}$. Though the reaction of both catalytic domains is very similar, they differ in that aspartyl-tRNA synthetase amidates the α-carboxylate group, while asparagine synthetase amidates the β-carboxylate group. Inspection of the active sites in the two enzymes shows that they have very similar

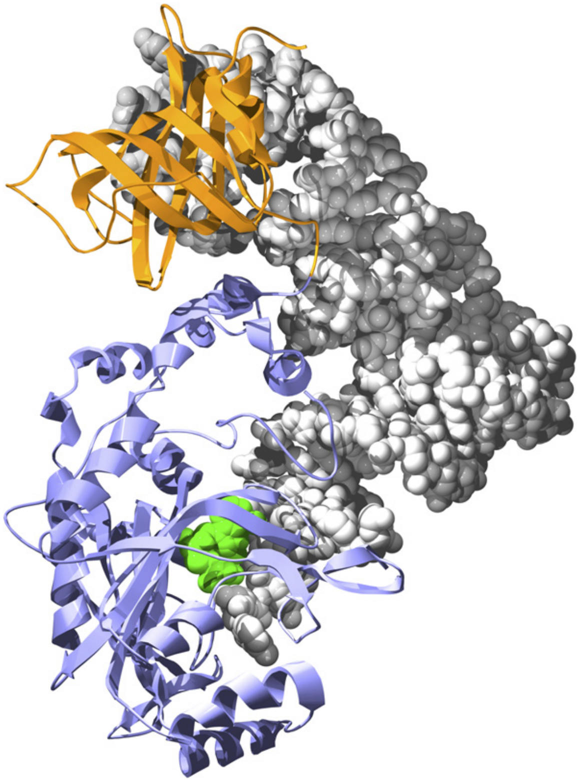

Figure 1. Yeast Aspartyl-tRNA Synthetase 1azs
The N-terminal domain of the "Nucleic acid-binding proteins" superfamily (b.40.4) is shown in orange, and the catalytic C-terminal domain of the "Class II aaRS and biotin synthetases" superfamily (d.104.1) is shown in blue. The $tRNA^{Asp}$ molecule is shown in space-filling representation in gray. ATP bound in the enzyme active site is shown in green. This figure was drawn from 1azs (Ruff et al., 1991).

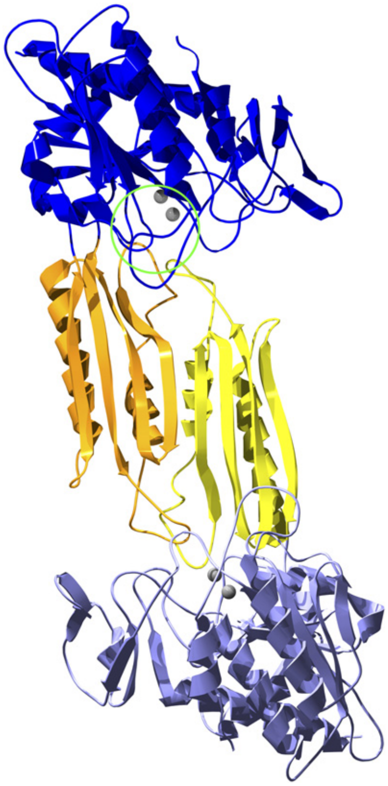

Figure 2. The Peptidase T Dimer from *S. typhimurium* 1fno
The catalytic domains of the "Zn-dependent exopeptidases" superfamily (c.56.5) are shown in light and dark blue, and the dimerization domains of the "Bacterial exopeptidase dimerization domain" superfamily (d.58.19) are shown in orange and yellow. The active site region in one chain containing two zinc ions is circled in green. Substrate binding is contributed to by the catalytic domain and both dimerization domains. This figure was drawn from 1fno (Rees et al., 1983).

structures and carry out their reactions in the same manner, except for small differences in active site residues that direct the amino group to different carboxylates (Nakatsu et al., 1998).

Thus, here the single-stranded nucleic acid recognition function of the N-terminal domain is conserved and is used to give a new specificity to a catalytic domain whose mechanism has been only slightly modified.

Domain that Modulates Function through the Formation of Oligomers and Intersubunit Contacts: Figure 2 and Entry 20

There is one case in our data set (entry 20) where a one-domain enzyme has a homologous domain in a two-domain protein in which the additional domain leads to the formation of a dimer and hence, through interactions between the subunits, a modification of the specificity.

Carboxypeptidase A and Peptidase T

Carboxypeptidase A is a one-domain protein (Rees et al., 1983), and peptidase T has two domains (Hakansson and Miller, 2002). Both enzymes have catalytic domains that are members of the "Zn-dependent exopeptidases" superfamily (c.56.5). The second domain in peptidase T is a member of the "bacterial exopeptidase dimerization domain" superfamily (d.58.19). Pairs of the additional domains bind together to make the protein a dimer (see Figure 2).

Carboxypeptidase A catalyzes the hydrolysis of C-terminal amino acids from polypeptide substrates. The activity of peptidase T is more specific: it will only hydrolyze tripeptides at an unblocked N terminus, and it will not hydrolyze longer polypeptides, unlike carboxypeptidase A. Inspection of the structure of peptidase T shows that the active sites in the two catalytic domains in the dimer are restricted by residues from their own additional domain and by residues from the symmetry-related additional domain. These residues come from the "bottom" of the additional domain in their own monomer and from the "top" of the domain in the symmetry-related monomer (see Figure 2). The presence of these additional residues in the active site of the enzyme reduces the space around the substrate-binding site (Hakansson and Miller, 2002).

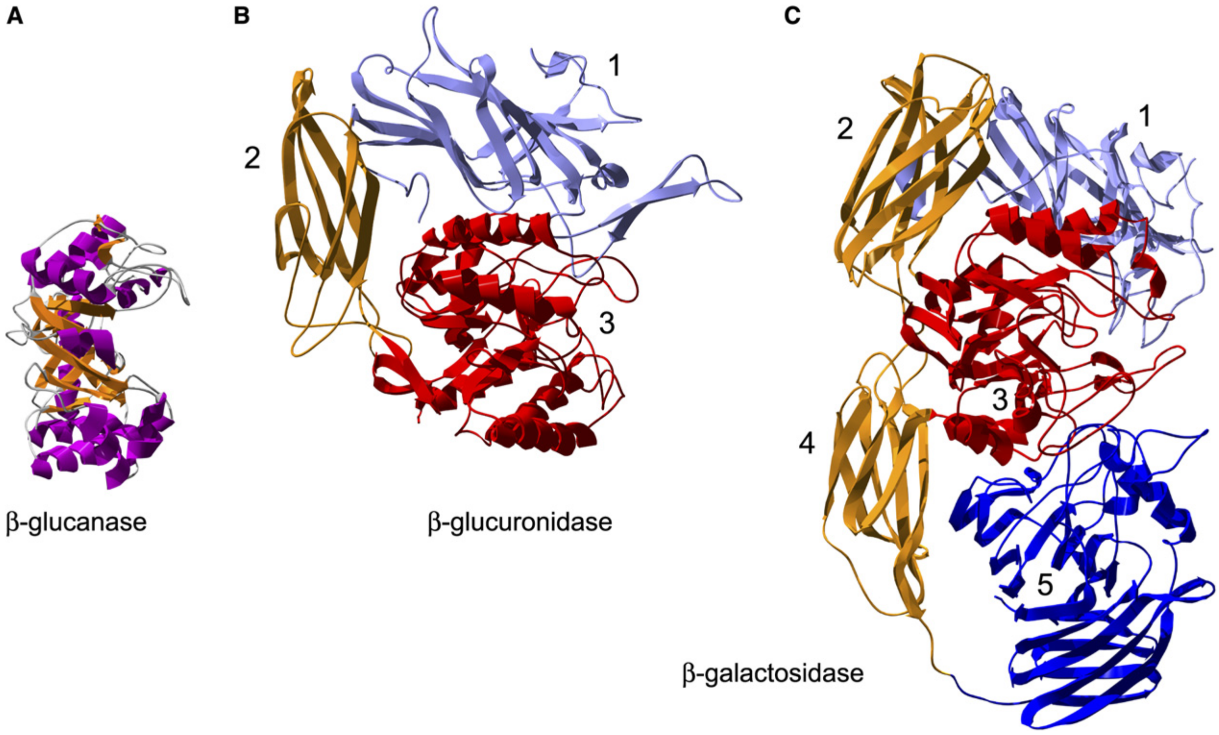

Figure 3. Glycosyl Hydrolases

(A–C) (A) (1→3)-β-glucanase (Varghese et al., 1994) represents the basic ''(Trans) glycosidases'' superfamily (c.1.8). Homologous catalytic domains are found in (B) β-glucuronidase and (C) β-galactosidase. (B) In β-glucuronidase (Jain et al., 1996), the catalytic domain is 3 (in red) and is joined by two other domains: 1 restricts the binding site, and 2 links 1 to 3. (C) β-galactosidase. The first three domains have the same structure as β-glucuronidase (Jacobson et al., 1994). Domain 4 links domain 3 to 5, which contributes to the active site.

Domains that Link Other Domains that Have a Direct Role in Function: Figures 3A–3C and Entry 21

In two-domain proteins, both domains usually play a direct role in function. In proteins with more than two domains, we sometimes find domains that are not directly involved in function but link and orientate the domains that do carry out functions. There are two examples in our data set, entries 21 and 22.

Glycosyl Hydrolases

Examples of domains with linker roles are seen in some of the glycosyl hydrolase proteins. Juers et al. (2000) and Todd et al. (2001) have discussed the relationship of the different functions of these proteins to their different domain structures.

Glycosyl hydrolyzes act on both long-chain polysaccharides and disaccharides. The members of the family that act on polysaccharides usually have just one domain (Figure 3A). These domains belong to the ''Trans-glycosidase'' superfamily (c.1.8) and have a $(\beta/\alpha)_8$ barrel structure. The substrate-binding site is formed by a long groove that extends across the C-terminal end of the barrel, as shown in the drawing of (1→3)-β-glucanase (Varghese et al., 1994) in Figure 3A.

Members of the glycosyl hydrolase family that act on disaccharides have three or more domains that together restrict the size of the substrate-binding site so that it only recognizes disaccharides. One of these is β-glucuronidase, which has three domains (Figure 3B). Another is β-galactosidase, which has five domains (Figure 3C).

In β-glucuronidase, domain 3 is a member of the ''Trans-glycosidase'' superfamily. Domain 1 is a member of the ''Galactose binding'' superfamily (b.18.1) and is homologous to a one-domain fucose-binding lectin. It has loop regions that extend into the active site of domain 3 and thus restricts its size (Jain et al., 1996). These loop regions are not involved in the generic galactose/cellulose-binding function the domain exhibits elsewhere (Juers et al., 1999).

Domain 2 links domain 1 to domain 3, and its packing interactions determine the geometries of these two domains (Figure 3B). Domain 2 has an immunoglobulin-like fold (b.1.4) and is not involved in the active site or subunit interactions.

β-galactosidase has evolved to become a more complex protein in terms of both its structure (five domains) and its function (Jacobson et al., 1994; Juers et al., 2000). It can hydrolyze lactose, or it can convert it to allolactose, which is then hydrolyzed. The first function is similar to that of β-glucuronidase. The latter function is novel and is believed to involve domain 5 (Juers et al., 1999). The active site is formed by domains 1, 3, and 5. Domains 1–3 are homologous to, and have the same arrangement as, the three domains that form β-glucuronidase. Domain 5 is a member of the ''Galactose mutarotase-like''

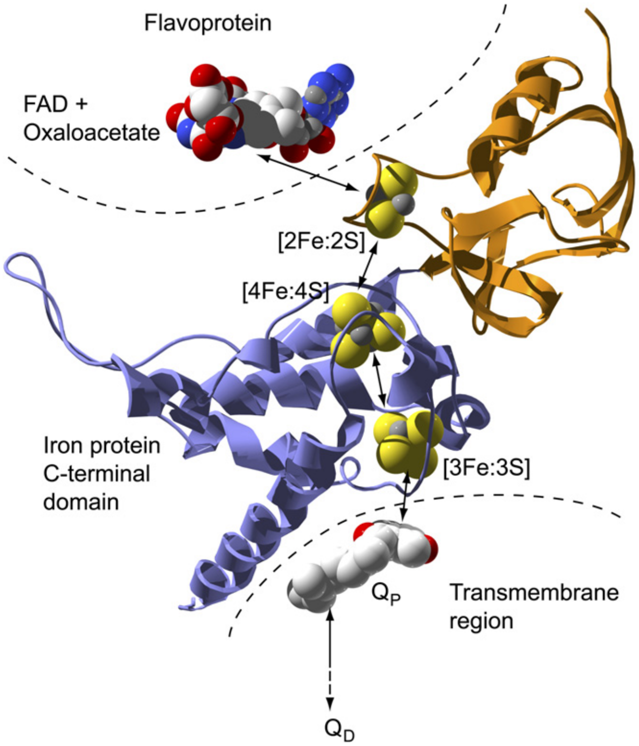

Figure 4. The Iron Subunit of the Quinol-Fumarate Reductase Iron Protein from *E. coli* 1l0v

The diagram shows the electron transport path between FAD and quinol via the three iron-sulfur clusters in the iron subunit. Its N-terminal "2Fe-2S ferredoxin-like" domain is shown in orange, and the C-terminal "α-helical ferredoxin" domain is shown in light blue. For the purpose of clarity, only the FAD of the flavoprotein and proximal quinol (Q_P) transmembrane subunit are shown. This figure was drawn from 1l0v (Iverson et al., 2002).

superfamily (b.30.5) and is found in galactose mutarotase (Thoden et al., 2003) as a single-domain enzyme. Here, it forms part of the active site by packing against the side of domain 3 opposite to that where domain 1 packs (Figure 3C). The residues that domain 5 contributes to the active site are distinct from those in the active site of galactose mutarotase, and they are located in a different part of the fold; thus, these two activities appear to be unrelated.

Domain 4 is a homolog of the immunoglobulin-like domain 2. Its role is to link domains 3 and 5, and it determines the geometry of the two domains: it makes no significant contacts with other subunits and is not part of the active site. Domain 2 in β-galactosidase, however, is not just a linking domain as it is in β-glucuronidase: it contributes to the structure of the active site in an adjacent subunit in the oligomeric complex.

Domains that Regulate Enzyme Function: Entry 23

In our data set, there are two cases (entries 23 and 24) of a one-domain enzyme having a homolog in a domain combination in which additional domains in the combination regulate its function.

VHR and SHP-2 Tyrosine Phosphatases

Example of this subcategory are given by two tyrosine phosphatases—the one-domain dual-specificity tyrosine phosphatase VHR (Yuvaniyama et al., 1996) and the three-domain tyrosine phosphatase SHP-2, which has two SH2 domains (d.93.1) that regulate its activity (Hof et al., 1998). In both enzymes, the homologous catalytic domains (c.45.1) catalyze the removal of a phosphate group from a tyrosine residue.

In SHP-2, the two regulatory SH2 domains are linked to the N terminus of the catalytic domain. In the inactive form of the enzyme, loops from the most N-terminal SH2 domain bind and block the active site of the phosphatase domain. Activation of the enzyme is caused by a phosphopeptide ligand that binds the N-terminal SH2 domain at a site that is quite distinct from that which interacts with the active site of the phosophatase domain. The binding of this peptide produces a conformational change in the structure of the SH2 domain and shifts the relative positions of the loops that bind to the catalytic domain in the inactive form. These shifts give the loops a geometry that does not fit the active site and thus makes it accessible to substrates (Hof et al., 1998).

Domain Combinations that Regulate DNA Binding: Entry 25

There are two cases in our data set (entries 25 and 26) in which the small-molecule recognition function of a single-domain protein is used in a two-domain protein to regulate a transcription factor.

Molybdate/Tungstate-Binding Protein and Molybdate-Dependent Transcription Factor

Example of this are given by two proteins that bind molybdate. The one-domain protein is the molybdate/tungstate-binding protein (MOP, b.40.6), which is probably involved in the storage and homeostasis of these ions (Schuttelkopf et al., 2002). This protein is a hexamer, and the binding site for molybdate is formed at the interface between pairs of subunits.

The multidomain protein ModE is a molybdate-dependent transcription factor that downregulates the molybdate uptake transporter operon modABCD and upregulates production of molybdoenzymes (Schuttelkopf et al., 2003). ModE is inactive in the absence of molybdate and is activated by the binding of molybdate. The protein is a dimer of identical subunits; the first domain of each subunit is a member of the winged helix superfamily (a.4.5), whose members are mostly involved in binding DNA. It is joined by a short linker region to a tandem repeat of MOP domains. Two molybdate-binding sites are formed at the interface between two MOP domains of different subunits.

Upon binding molybdate, the MOP domains undergo conformational changes (Schuttelkopf et al., 2003). These are transmitted to the linker regions, which, in turn, transmit them to the winged helix domains. The effect of these changes is to alter the relative positions of the two DNA recognition helices in the winged helix domain. Relative to the inactive form, the helices are moved 2 Å closer together (Schuttelkopf et al., 2003). This change in their relative orientation allows them to bind to the DNA transcription site.

Gene Fusions that Form Bifunctional Enzymes: Entry 32

Another category in which function is conserved subsequent to domain combination involves pairs of genes that are separate proteins in some organisms fusing to form a bifunctional enzyme in other organisms. The 8 cases (from 16 single-domain proteins) in our data set are found in entries 27–34. Most of these involve combinations of proteins that carry out sequential steps in a pathway. Here, we describe a more complex example.

6-Phosphofructo-2-Kinase/Fructose-2, 6-Bisphosphatase

6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase (Hasemann et al., 1996) has two domains. The reaction catalyzed by the first domain is the reverse of that carried out by the second. The first domain (c.37.1) catalyzes the synthesis of fructose-2,6-bisphosphate, by the addition of phosphate to fructose-6-phosphate, and the second domain (c.60.1) carries out its degradation by the removal of this phosphate. Homologs of both domains are found as one-domain proteins that have similar functions: phosphoribulokinase (Harrison et al., 1998) and prostatic acid phosphatase (Ortlund et al., 2003).

6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase plays a crucial role in glucose homeostasis. In the absence of regulation, one domain in this protein would reverse the action of the other. To prevent this, the actual balance of the two activities is regulated by various metabolites becoming inhibitors at particular concentrations, or by phosphorylation, which affects both kinase and bisphosphatase activities via communication between domains (Pilkis et al., 1995).

Separation in the Multidomain Protein of Functional Roles Found in a One-Domain Protein: Entry 37

There are four cases (entries 35–38) in the data set in which one of the activities found in the one-domain protein is transferred to other domains that combine with the homolog of the one-domain protein.

FokI and EcoRV

Example of this are given by the restriction enzymes EcoRV (Kostrewa and Winkler, 1995) and FokI (Wah et al., 1997). EcoRV is a one-domain protein (c.52.1) from *E. coli* that recognizes the DNA sequence 5′-GATATC-3′ and cleaves it in the center.

FokI (Wah et al., 1997) has three domains that belong to the winged helix superfamily (a.4.5), most of whose members are involved in DNA recognition, and, at the C terminus, a catalytic domain that is homologous to EcoRV. FokI recognizes the DNA sequence 5′-GGATG-3′, but it cleaves DNA nonspecifically at a short distance from that sequence. The structure of a FokI-DNA complex (Wah et al., 1997) shows that the first two of the three winged helix domains are almost entirely responsible for the recognition of the specific DNA sequence. The third winged helix domain acts as a linker to the fourth catalytic domain. In isolation, the FokI catalytic domain only binds DNA weakly and nonspecifically (Li et al., 1992).

Locating DNA-binding specificity and catalysis on different domains allows mutations in the recognition domain to change sequence specificity in a manner that would not affect the existing functionality of the catalytic domain. It also allows recombination with domains that have a different DNA sequence specificity or a different catalytic mechanism. Indeed, Chandrasegaran and coworkers have carried out several experiments using the catalytic domain of FokI in combination with different DNA-binding domains to produce chimeric restriction enzymes (Kim et al., 1996).

Domains Whose Combination Allows Them to Function in New Contexts: Figure 4 and Entry 39

There are 11 cases in our data set (entries 21b, 21c, 25b, 26b, 39–42, and 45b) in which a one-domain protein and a homolog in a multidomain protein conserve their functions but carry them out in different structural contexts.

Ferredoxin and Quinol-Fumarate Reductase

Ferredoxin (d.15.4) (Morales et al., 1999), a soluble one-domain protein, contains a [2Fe-2S] iron cluster that is used to transport electrons between different redox centers. A homolog of ferredoxin, combined with another domain, is found in one of the subunits of the membrane protein quinol-fumarate reductase (Iverson et al., 2002).

Quinol-fumarate reductase (QFR) catalyzes the terminal step of anaerobic respiration. It has four subunits. Two of these reside in the membrane and accept electrons donated by menaquinol. Three iron-sulfur clusters in the third domain then transport these electrons to FAD in the fourth subunit, where they are used to reduce fumarate.

The third subunit has two domains (Figure 4): one domain is a homolog of ferredoxin with one [2Fe-2S] cluster, and the other domain is a member of the α-helical ferredoxin superfamily (a.1.2) and has two [4Fe-4S] clusters. The ferredoxin domain in this subunit does not transport electrons between proteins, as it does in the one-domain homologs, but forms part of a pathway that transports electrons within a protein complex.

Loss of Catalytic Function by a Homolog in a Domain Combination: Figure 5 and Entry 38

There are four cases in the data set in which the catalytic activity present in a one-domain protein is not present in a homolog that is part of a domain combination. In two of these, some other functions are conserved (entries 5b and 38b), and in the other two, there is a complete change in function (entries 43 and 44).

Copper, Zinc Superoxide Dismutase and the Copper Chaperone of Superoxide Dismutase

Examples of this category are given by the one-domain protein superoxide dismutase (SOD) (Tainer et al., 1982) and by the two-domain copper chaperone of SOD (CCS) (Lamb et al., 2001). Both have domains that belong to the "Superoxide dismutase superfamily" (b.1.8). SOD itself is a homodimer that catalyzes the conversion of superoxide radicals to oxygen and hydrogen peroxide. To carry out this function, its active site requires both copper and zinc ions. The copper ions are provided by CCS (Figure 5).

The central region of CCS is formed by a domain that is homologous to SOD, but is not catalytic: it lacks both the metal-binding sites and the catalytic residues. It does, however, retain the ability to form a dimer with a SOD monomer in a manner that is the same as that of a SOD dimmer. At the N terminus of CCS, there is a domain that belongs to the "Heavy metal-associated" superfamily (d.58.17) and is homologous to the one-domain protein Hah1 Metallochaperone that binds and transports copper (Wernimont et al., 2000). At its C terminus, there is a 27 residue extension that is also able to bind copper (Lamb et al., 2001). CCS supplies copper to the SOD enzyme by its central domain first binding to a SOD monomer. The copper ion is then transferred from the N-terminal heavy metal-associated domain to the C-terminal extension, which, in turn, transfers the copper ion to the active site of SOD (Lamb et al., 2001).

Thus, in this example, the domain in CCS that is homologous to the SOD domain does not have the catalytic properties of the latter but has retained its dimerization properties. This property is used to facilitate the process that puts copper into the active site of SOD.

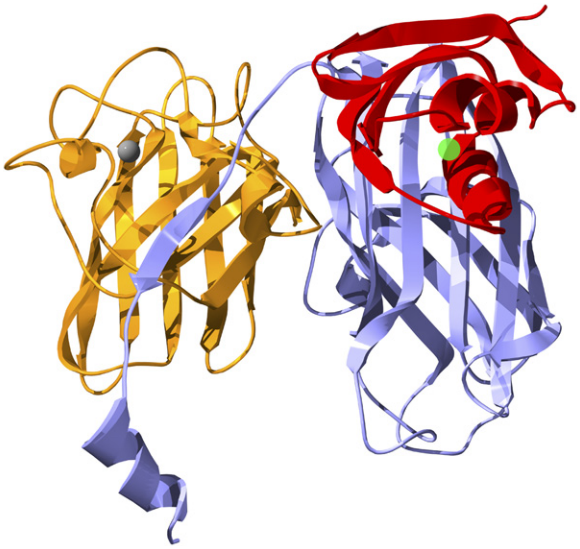

Figure 5. Heterodimeric Structure of Copper, Zinc Superoxide Dismutase in Complex with Its Copper Chaperone 1jk9
The SOD enzyme is shown in orange in complex with its copper chaperone, which is shown in blue and red. The copper chaperone is composed of two domains. The N-terminal domain of the chaperone is shown in red and belongs to the "HMA, heavy metal-associated domain" superfamily. The C-terminal domain of the copper chaperone is homologous to SOD and belongs to the "Cu,Zn superoxide dismutase-like" superfamily. It conserves the ability to form a dimer with SOD. The chaperone also has a C-terminal extension that packs against SOD, which is thought to be involved in the mechanism that loads copper into the SOD active site (Lamb et al., 2001). This figure was drawn from 1jk9 (Lamb et al., 2001).

Gain of Catalytic Activity in a Domain Combination: Entry 15

There are three cases in the data set in which a one-domain protein that is not catalytic has a homolog in a domain combination that is catalytic (entries 15b, 19b, and 36b).

Phosphatidylinsoitol Transfer Protein and the α Subunit of Naphthalene 1,2-Dioxygenase

Examples of this category are provided by the one-domain phoshatidylinsoitol transfer protein (PITP) (Yoder et al., 2001) and the two-domain α subunit of naphthalene 1,2-dioxygenase (NDO) (Kauppi et al., 1998).

PITP is a member of the "Bet v1-like" superfamily (d.129.3) and has no catalytic activity: it mediates the exchange of phoshatidylinositiol and phosphatidylcholine molecules between different membranes. It has an eight-stranded β sheet and three long α helices that pack to form a large, long cavity between the β sheet and the α helices. Phospholipids are accommodated in this cavity, and their phosphate groups make hydrogen bonds to polar residues at the back of the cavity (Yoder et al., 2001).

NDO catalyzes the dihydroxylation of the aromatic compound napthalenene to give *cis*-naphthalene dihydrodiol and is an $\alpha_3\beta_3$ hexamer. There is no evidence for the β subunits being involved in catalysis. The α subunit has two domains. At the N terminus, there is a [2Fe-2S]-binding Rieske domain that belongs to the "Iron-sulphur protein (ISP)" superfamily (b.33.1). It binds a [2Fe-2S] cluster and transports electrons. At the C terminus, there is a catalytic domain that is homologous to PITP. It has a large central cavity that is similar to that in PITP, but it binds a different substrate: napthalenene. At the back of this cavity, there is a mononuclear iron that is linked by hydrogen bonds to the Rieske domain of the adjacent α subunit. The reaction is believed to involve dioxygen binding to the mononuclear iron, followed by its conversion to peroxide by electrons donated from the Rieske domain ion cluster to the active site iron, and the attack of the peroxide on the substrate (Carredano et al., 2000).

The difference in the activity of NDO, compared to PITP, involves the creation of a catalytic active site, domain combination, and the formation of oligomers. The formation of the iron-binding site requires side chains different from those in PITP, and these changes involve residues that are not equivalent to those that bind the phosphate in PITP. Other residue changes alter the shape of the cavity so that naphthalene, rather than phoshatidylinositol, is bound.

Change in the Catalytic Reaction in Domain Combination: Entry 17

In this category, there are two cases in which the reaction catalyzed by the one-domain enzyme is quite different from that of a homologous domain found in the domain combination (entries 17b and 45a).

Zinc-β-Lactamase and Rubredoxin: Oxygen Oxidoreductase

Examples of this category are the one-domain protein Zn-β-lactamase (Carfi et al., 1995) and the two-domain subunit of the dimeric protein Ruberdoxin: oxygen oxidoreductase (ROO) (Frazao et al., 2000). Zn-β-lactamase is a member of the ''Metalo-hydrolase/oxidoreductase'' superfamily (d.157.1). It catalyzes the hydrolysis of the β-lactam ring of the penicillin and cephalosporin antibiotics.

ROO is part of the pathway that protects anaerobic organisms from oxygen by catalyzing its reduction to water. The N-terminal catalytic domain of ROO is homologous to Zn-β-lactamase. In ROO, however, the zinc ion is replaced by di-iron. The second domain of ROO is a member of the same superfamily as the one-domain protein flavodoxin (c.23.5) and similarly binds flavin mononucleotide (FMN). Ruberdoxin supplies electrons to the FMN in the second ROO domain, which then passes them to the di-iron site in the first domain of the adjacent subunit. These electrons, together with H^+ ions, then reduce the oxygen bound by the di-iron to water.

The overall structures of Zn-β-lactamase and the homologous domain in ROO are very similar. The change in their function is mainly produced by structural changes in two regions. The first region is the pocket adjacent to the zinc-binding site that is used by Zn-β-lactamase to bind penicillin and cephalosporin substrates. This pocket is not found in ROO, where the equivalent region is filled by insertions in the ROO structure and by domain-domain contacts (Frazao et al., 2000). The second region concerns the zinc ion bound in the lactamases and the di-iron group bound in ROO. Six residues bind metal ions in both structures, and they come from equivalent sites and have a similar geometry. There are, however, differences in the identity of two side chains: Glu in place of His and Asp in place of Cys:

$$\begin{aligned}\text{ROO: } & \text{H-X-}\mathbf{E}\text{-X-D-X}_{62}\text{-H-X}_{18}\text{-}\mathbf{D}\text{-X}_{60}\text{-H}\\ \text{Zn-}\beta\text{-lactamase: } & \text{H-X-}\mathbf{H}\text{-X-D-X}_{58}\text{-H-X}_{18}\text{-}\mathbf{C}\text{-X}_{41}\text{-H.}\end{aligned}$$

These differences favor the binding of di-iron in ROO and a zinc ion in β-lactamase (Frazao et al., 2000). Note that in ROO, the substitution of zinc for di-iron provides a group that both binds the substrate and carries out catalysis.

CONSERVATION AND CHANGE OF FUNCTION OF HOMOLOGOUS DOMAINS

In the previous sections of the paper, we have described in some detail the extent to which the functions found in one-domain proteins are conserved, modified, or changed in homologous domains found in multidomain proteins. These descriptions cover only one-quarter of the entries in Table 1. Here, we describe the extent of conservation and change in all entries.

Here, we have examined 46 multidomain proteins. Of these, 38 are formed by 2 domains, 6 by 3, 1 by 4, and 1 by 5. These 46 multidomain proteins are composed of 103 domains belonging to 1 of 83 different superfamilies. For 61 of these superfamilies, structures are known for both one-domain and multidomain proteins. At the present time, the domains in the other 22 superfamilies are only known in structures in which they form combinations with domains from other superfamilies. Overall, our data set has a total of 70 unique pairs of one- and multidomain proteins that have a homologous domain in common (see Table 1).

The one-domain proteins that are homologous to those in multidomain proteins have a variety of functions: 44 are enzymes, 8 transport electrons, 6 transport ions or small molecules, 17 bind nucleic acids (including 7 of the enzymes), 1 is a chaperone, and 1 forms dimers. The previous sections of the paper, Table 1, and the Supplemental Data describe the extent of the conservation and/or change in function of the homologous domains found in the one- and multidomain proteins.

Out of a total of 70 unique pairs of single- and multidomain proteins that have a homologous domain in common, the functions found in the one-domain proteins are conserved, entirely or largely, in three-quarters of the domains that form the multidomain proteins; just under one-sixth have functions that are very different from those found in their one-domain homologs, and the remainder have partial conservation of functions (details of the data that provide these proportions are given in the Supplemental Data).

We measured the global sequence identity between all 70 pairs of homologous domains by using the alignment program Needle (Needleman and Wunsch, 1970; Rice et al., 2000) and domain sequences available from the ASTRAL database (Chandonia et al., 2004). We found that the sequence identity of the homologous pairs of domains ranges between 1% and 42%. Most values cluster toward the lower end of this range, and the average value for the 70 pairs is 14.5%. The domains that conserve their functions have an average sequence identity of 15%. If we consider only those pairs in which there has been partial functional conservation, the average sequence identity is 14%. For the cases in which the functions of the homologous domains are not conserved, the average sequence identity is 12.5%. Figure 6 shows the spread of the sequence identities of the pairs whose function is conserved, partially conserved, and not conserved.

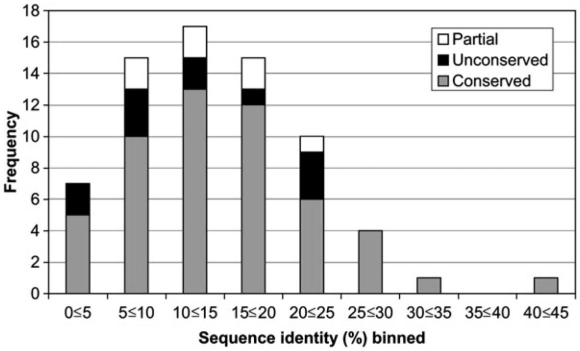

Figure 6. Graph Showing the Extent of the Sequence Identity of Homologous Domains in One- and Multidomain Proteins
Homologous domains falling into the conserved, partial, and nonconserved categories are indicated.

These results suggest that the line between functional conservation and change cannot be attributed on the basis of a simple sequence identity cut-off. Of course, the data set considered here is small, and this conclusion needs to be confirmed by more extensive analysis. However, we should also note that the extent of the functional conservation described here may underestimate its true extent because we have confined our comparisons to proteins for which there are known structures. There may be one-domain proteins whose structures are unknown at present but whose functions are similar to their homologs in the multidomain proteins.

The extent of functional conservation described here is very different from that described by Hegyi and Gerstein (2001). They analyzed 70 superfamilies that occur in both one-domain and multidomain proteins and found that only 14 superfamilies had the "same" function. The reason for the difference is that they used only the conservation of the first three EC numbers and synonymous SWISS-PROT keywords to assign function. In this study, we have determined in molecular detail the function of each domain in the one- and multidomain proteins and were thus able to find similarities that were missed by their procedure.

Conclusion: A Grammar for Domain Combinations

Here, we have discussed how different combinations of domains from different superfamilies produce new functions. In some ways, this process is analogous to how, in language, word combinations function.

Here, we frequently find that the general function of the homolog of the one-domain protein in the multidomain protein has been conserved but has been modified or made more specific. This is achieved by placing the homologous domain into a new domain context or "syntax," in which an additional domain serves to expand, alter, or modulate its functionality. Syntax governs the arrangements in a sentence of words that individually have particular meanings and taken together make "sense"; this can be modified further by the replacement or the addition of other suitable words. The addition of unsuitable words will produce nonsense.

In the small number of cases in which the functions of the domain are totally changed, we find that the common scaffold of the protein domain has been adapted to carry out a different, unrelated reaction or bind a different, unrelated ligand. This is a redesign of the protein's function through progressive mutation of the domain itself to generate quite different functions. It can be thought of as a change in semantics, i.e., that the function or "meaning" of the domain itself has changed. This is found in words that have quite different meanings according to their context: e.g., "she is a red" (i.e., a communist); "the pillar-box is red" (i.e., is painted red).

Another property also observed during this study that is characteristic of the development of natural languages is the generation of discreteness (Senghas et al., 2004). In four cases (entries 35a, 36a, 37a, and 38a), we observed that two distinct functions previously carried out by a one-domain protein have been separated into discrete functions carried out by different domains in the multidomain proteins. This process can be considered advantageous to the domain combination; by the distributing of a specific function to dedicated domains, the protein may achieve greater molecular efficiency in terms of ligand-protein interactions. Moreover, now that previously unified functions are uncoupled in the two-domain state, the possibility of acquiring enhanced, or altered, functionality via separate mutation of domains, without mutual impact on existing functionality, is greatly increased. In addition, the ability to recombine, with synonymous domains carrying out the same function, but being of a different superfamily (as in the case of the DNA-binding domains of FokI), is greatly enhanced by this dichotomy.

These features—change in syntax and semantics and the generation of discreteness—are all properties of a natural grammar of domain combination that determines the assembly of functionally coherent combinations of domains and gives rise to more complex protein functions. This grammar is a consequence of the selection of combinations of domains that make "sense" functionally and deletion of those that are "nonsense."

These considerations are useful in trying to predict the functions of novel domain combinations for which there is no experimental evidence for their function. Thus, a novel combination of two domains that are homologous to a small-molecule-binding protein and to a DNA-binding protein would suggest that in this protein the activity of the DNA-binding domain was being regulated by the other domain binding a small molecule that may or may not be the same as that bound to the one-domain homolog. Conversely, in a novel combination like entry 17 in Table 1, where one domain is homologous to Zn metallo-β-lactamase and the other to flavodoxin, it is difficult to see how, if their functions are conserved, the combination could make coherent use of them. Here, it would be reasonable to assume that at least one domain has changed its function.

To predict the function of novel proteins, by using an approach like that outlined in the previous paragraph, requires a comprehensive description of how the properties of individual domains combine to produce the overall functions of multidomain proteins (Bashton, 2004).

Here, we have analyzed only a very small subset of the different domain combinations found in proteins. New structures will certainly include those that have types of domain combination that go well beyond those listed in Table 2. We would expect, however, that the grammatical trends and functional modifications observed here will hold true for a much larger data set.

Supplemental Data

Supplemental Data include Table S1, detailing the functional conservation and change in all 45 pairs of one-domain and multidomain proteins in full, as well as details of the conservation and change of function of homologous domains as text and are available at http://www.structure.org/cgi/content/full/15/1/85/DC1/.

ACKNOWLEDGMENTS

We are most grateful for the advice and detailed criticisms given to us by Annabel Todd, Siarhei Maslau, and Gemma Holliday.

Received: October 6, 2006
Revised: November 21, 2006
Accepted: November 21, 2006
Published: January 16, 2007

REFERENCES

Bartlett, G.J., Borkakoti, N., and Thornton, J.M. (2003). Catalysing new reactions during evolution: economy of residues and mechanism. J. Mol. Biol. *331*, 829–860.

Bashton, M. (2004). Functional analysis of domain combinations. PhD thesis, University of Cambridge, Cambridge, United Kingdom.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

Carfi, A., Pares, S., Duee, E., Galleni, M., Duez, C., Frere, J.M., and Dideberg, O. (1995). The 3-D structure of a zinc metallo-β-lactamase from *Bacillus cereus* reveals a new type of protein fold. EMBO J. *14*, 4914–4921.

Carredano, E., Karlsson, A., Kauppi, B., Choudhury, D., Parales, R.E., Parales, J.V., Lee, K., Gibson, D.T., Eklund, H., and Ramaswamy, S. (2000). Substrate binding site of naphthalene 1,2-dioxygenase: functional implications of indole binding. J. Mol. Biol. *296*, 701–712.

Chandonia, J.M., Hon, G., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., and Brenner, S.E. (2004). The ASTRAL Compendium in 2004. Nucleic Acids Res. *32*, D189–D192.

Frazao, C., Silva, G., Gomes, C.M., Matias, P., Coelho, R., Sieker, L., Macedo, S., Liu, M.Y., Oliveira, S., Teixeira, M., et al. (2000). Structure of a dioxygen reduction enzyme from *Desulfovibrio gigas*. Nat. Struct. Biol. *7*, 1041–1045.

Gerlt, J.A., and Babbitt, P.C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu. Rev. Biochem. *70*, 209–246.

Hakansson, K., and Miller, C.G. (2002). Structure of peptidase T from *Salmonella typhimurium*. Eur. J. Biochem. *269*, 443–450.

Harrison, D.H., Runquist, J.A., Holub, A., and Miziorko, H.M. (1998). The crystal structure of phosphoribulokinase from *Rhodobacter sphaeroides* reveals a fold similar to that of adenylate kinase. Biochemistry *37*, 5074–5085.

Hasemann, C.A., Istvan, E.S., Uyeda, K., and Deisenhofer, J. (1996). The crystal structure of the bifunctional enzyme 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase reveals distinct domain homologies. Structure *4*, 1017–1029.

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

Hof, P., Pluskey, S., Dhe-Paganon, S., Eck, M.J., and Shoelson, S.E. (1998). Crystal structure of the tyrosine phosphatase SHP-2. Cell *92*, 441–450.

Iverson, T.M., Luna-Chavez, C., Croal, L.R., Cecchini, G., and Rees, D.C. (2002). Crystallographic studies of the *Escherichia coli* quinol-fumarate reductase with inhibitors bound to the quinol-binding site. J. Biol. Chem. *277*, 16124–16130.

Jacobson, R.H., Zhang, X.J., DuBose, R.F., and Matthews, B.W. (1994). Three-dimensional structure of β-galactosidase from *E. coli*. Nature *369*, 761–766.

Jain, S., Drendel, W.B., Chen, Z.W., Mathews, F.S., Sly, W.S., and Grubb, J.H. (1996). Structure of human β-glucuronidase reveals candidate lysosomal targeting and active-site motifs. Nat. Struct. Biol. *3*, 375–381.

Juers, D.H., Huber, R.E., and Matthews, B.W. (1999). Structural comparisons of TIM barrel proteins suggest functional and evolutionary relationships between β-galactosidase and other glycohydrolases. Protein Sci. *8*, 122–136.

Juers, D.H., Jacobson, R.H., Wigley, D., Zhang, X.J., Huber, R.E., Tronrud, D.E., and Matthews, B.W. (2000). High resolution refinement of β-galactosidase in a new crystal form reveals multiple metal-binding sites and provides a structural basis for α-complementation. Protein Sci. *9*, 1685–1699.

Kauppi, B., Lee, K., Carredano, E., Parales, R.E., Gibson, D.T., Eklund, H., and Ramaswamy, S. (1998). Structure of an aromatic-ring-hydroxylating dioxygenase-naphthalene 1,2-dioxygenase. Structure *6*, 571–586.

Kim, Y.G., Cha, J., and Chandrasegaran, S. (1996). Hybrid restriction enzymes: zinc finger fusions to Fok I cleavage domain. Proc. Natl. Acad. Sci. USA *93*, 1156–1160.

Kostrewa, D., and Winkler, F.K. (1995). Mg^{2+} binding to the active site of EcoRV endonuclease: a crystallographic study of complexes with substrate and product DNA at 2 Å resolution. Biochemistry *34*, 683–696.

Lamb, A.L., Torres, A.S., O'Halloran, T.V., and Rosenzweig, A.C. (2001). Heterodimeric structure of superoxide dismutase in complex with its metallochaperone. Nat. Struct. Biol. *8*, 751–755.

Li, L., Wu, L.P., and Chandrasegaran, S. (1992). Functional domains in Fok I restriction endonuclease. Proc. Natl. Acad. Sci. USA *89*, 4275–4279.

Morales, R., Charon, M.H., Hudry-Clergeon, G., Petillot, Y., Norager, S., Medina, M., and Frey, M. (1999). Refined X-ray structures of the oxidized, at 1.3 Å, and reduced, at 1.17 Å, [2Fe-2S] ferredoxin from the cyanobacterium *Anabaena* PCC7119 show redox-linked conformational changes. Biochemistry *38*, 15764–15773.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Nahum, L.A., and Riley, M. (2001). Divergence of function in sequence-related groups of *Escherichia coli* proteins. Genome Res. *11*, 1375–1381.

Nakatsu, T., Kato, H., and Oda, J. (1998). Crystal structure of asparagine synthetase reveals a close evolutionary relationship to class II aminoacyl-tRNA synthetase. Nat. Struct. Biol. *5*, 15–19.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. *48*, 443–453.

Ortlund, E., LaCount, M.W., and Lebioda, L. (2003). Crystal structures of human prostatic acid phosphatase in complex with a phosphate ion and α-benzylaminobenzylphosphonic acid update the mechanistic picture and offer new insights into inhibitor design. Biochemistry *42*, 383–389.

Pilkis, S.J., Claus, T.H., Kurland, I.J., and Lange, A.J. (1995). 6-Phosphofructo-2-kinase/fructose-2,6-bisphosphatase: a metabolic signaling enzyme. Annu. Rev. Biochem. *64*, 799–835.

Rees, D.C., Lewis, M., and Lipscomb, W.N. (1983). Refined crystal structure of carboxypeptidase A at 1.54 Å resolution. J. Mol. Biol. *168*, 367–387.

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. *16*, 276–277.

Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.C., and Moras, D. (1991). Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science *252*, 1682–1689.

Schuttelkopf, A.W., Harrison, J.A., Boxer, D.H., and Hunter, W.N. (2002). Passive acquisition of ligand by the MopII molbindin from *Clostridium pasteurianum*: structures of apo and oxyanion-bound forms. J. Biol. Chem. *277*, 15013–15020.

Schuttelkopf, A.W., Boxer, D.H., and Hunter, W.N. (2003). Crystal structure of activated ModE reveals conformational changes involving both oxyanion and DNA-binding domains. J. Mol. Biol. *326*, 761–767.

Senghas, A., Kita, S., and Ozyurek, A. (2004). Children creating core properties of language: evidence from an emerging sign language in Nicaragua. Science *305*, 1779–1782.

Tainer, J.A., Getzoff, E.D., Beem, K.M., Richardson, J.S., and Richardson, D.C. (1982). Determination and analysis of the 2 Å-structure of copper, zinc superoxide dismutase. J. Mol. Biol. *160*, 181–217.

Thoden, J.B., Kim, J., Raushel, F.M., and Holden, H.M. (2003). The catalytic mechanism of galactose mutarotase. Protein Sci. *12*, 1051–1059.

Todd, A.E., Orengo, C.A., and Thornton, J.M. (2001). Evolution of function in protein superfamilies, from a structural perspective. J. Mol. Biol. *307*, 1113–1143.

Todd, A.E., Orengo, C.A., and Thornton, J.M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. Structure *10*, 1435–1451.

Varghese, J.N., Garrett, T.P., Colman, P.M., Chen, L., Hoj, P.B., and Fincher, G.B. (1994). Three-dimensional structures of two plant β-glucan endohydrolases with distinct substrate specificities. Proc. Natl. Acad. Sci. USA *91*, 2785–2789.

Wah, D.A., Hirsch, J.A., Dorner, L.F., Schildkraut, I., and Aggarwal, A.K. (1997). Structure of the multimodular endonuclease FokI bound to DNA. Nature *388*, 97–100.

Wernimont, A.K., Huffman, D.L., Lamb, A.L., O'Halloran, T.V., and Rosenzweig, A.C. (2000). Structural basis for copper transfer by the metallochaperone for the Menkes/Wilson disease proteins. Nat. Struct. Biol. *7*, 766–771.

Yoder, M.D., Thomas, L.M., Tremblay, J.M., Oliver, R.L., Yarbrough, L.R., and Helmkamp, G.M., Jr. (2001). Structure of a multifunctional protein. Mammalian phosphatidylinositol transfer protein complexed with phosphatidylcholine. J. Biol. Chem. *276*, 9246–9252.

Yuvaniyama, J., Denu, J.M., Dixon, J.E., and Saper, M.A. (1996). Crystal structure of the dual specificity protein phosphatase VHR. Science *272*, 1328–1331.