Acta Cryst. (1998). D**54**, 1168–1177

Protein Three-Dimensional Structural Databases: Domains, Structurally Aligned Homologues and Superfamilies

R. Sowdhamini,[a]† David F. Burke,[a] Charlotte Deane,[a] Jing-fei Huang,[a,b] Kenji Mizuguchi,[a] Hampapathulu A. Nagarajaram,[a] John P. Overington,[c] N. Srinivasan,[a]‡ Robert E. Steward[a] and Tom L. Blundell[a]*

[a]*Department of Biochemistry, University of Cambridge, 80 Tennis Court Road, Cambridge CB2 1QW, England,* [b]*Kunming Institute of Zoology, The Chinese Academy of Sciences, Eastern Jiaochang Road, Kunming, Yunnan 650223, Peoples Republic of China, and* [c]*Pfizer Central Research, Sandwich, Kent CT13 9NJ, England.*
E-mail: tom@cryst.bioc.cam.ac.uk

(*Received 27 March 1998; accepted 18 May 1998*)

Abstract

This paper reports the availability of a database of protein structural domains (DDBASE), an alignment database of homologous proteins (HOMSTRAD) and a database of structurally aligned superfamilies (CAMPASS) on the World Wide Web (WWW). DDBASE contains information on the organization of structural domains and their boundaries; it includes only one representative domain from each of the homologous families. This database has been derived by identifying the presence of structural domains in proteins on the basis of inter-secondary structural distances using the program *DIAL* [Sowdhamini & Blundell (1995), *Protein Sci.* **4**, 506–520]. The alignment of proteins in superfamilies has been performed on the basis of the structural features and relationships of individual residues using the program *COMPARER* [Sali & Blundell (1990), *J. Mol. Biol.* **212**, 403–428]. The alignment databases contain information on the conserved structural features in homologous proteins and those belonging to superfamilies. Available data include the sequence alignments in structure-annotated formats and the provision for viewing superposed structures of proteins using a graphical interface. Such information, which is freely accessible on the WWW, should be of value to crystallographers in the comparison of newly determined protein structures with previously identified protein domains or existing families.

1. Introduction

The Brookhaven Protein Data Bank (PDB) (Bernstein *et al.*, 1977) currently contains over 7000 entries; after removing the repeated entries of identical proteins (such as the same protein in different complexes or at different resolutions), there remain 1729 proteins (Brenner *et al.*, 1997), including many homologues (see Fig. 1). If only representative structures from the homologous protein 'family' are retained such that no two proteins have more than 25% sequence identity (Hobohm *et al.*, 1992; May 1997 release), the resultant data set still includes 687 proteins. This corresponds to 463 superfamilies of protein domains with 96 superfamilies arising from more than one family (Brenner *et al.*, 1997).

Proteins that have diverged but retain high sequence identity fold into similar three-dimensional structures and usually perform similar functions – these clearly belong to a homologous family (Richardson, 1981; Rossmann & Argos, 1977; Chothia, 1984; Overington *et al.*, 1990, 1993). Proteins or domains of proteins that adopt the same three-dimensional fold despite poor sequence identity and perform remotely similar functions (Blundell & Humbel, 1980; Murzin & Chothia, 1992; Murzin *et al.*, 1995; Murzin, 1996) are termed superfamilies. The identification of new members belonging to pre-existing families and superfamilies is straightforward only when contiguous residues forming a functional motif are conserved, where *PROSITE* searches may be appropriate (Bairoch, 1991). Furthermore these should be distinguished from proteins with no sequence identity and no similarity of functions that nevertheless have the same fold or superfolds (Orengo *et al.*, 1994).

An analysis of protein sequence and structure entries indicates that about 50% of the 'new' sequences could be attributed a previously known function and roughly 20% of the sequences have homologues of known structure (Bork *et al.*, 1992, 1994; Koonin *et al.*, 1994). When the crystal structure of a 'new' protein is determined, it is important to compare its structure with the previously determined structures. This is facilitated by the existence of databases of aligned protein structures and sequences (Overington *et al.*, 1990, 1993; Johnson *et al.*, 1993).

† Address from June 1998: National Centre for Biological Sciences, TIFR Centre, PO Box 1234, Indian Institute of Science Campus, Bangalore 560012, India.
‡ Address from June 1998: Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India.

ISSN 0907-4449

Often homology or structural similarity exists between parts of two different proteins; one or two domains only may be conserved (Wetlaufer, 1973; Richardson, 1981; Wodak & Janin, 1981; Go, 1981). Although algorithms to identify such compact sub-structures have been developed (Schulz, 1977; Crippen, 1978; Rose, 1979; Zehfus & Rose, 1986), it is convenient to use automatic methods so that the information of domain organization can be compiled for the large number of protein structures now available (Islam *et al.*, 1995; Siddiqui & Barton, 1995; Swindells, 1995; Nichols *et al.*, 1995). We have constructed a database of protein structural domains (DDBASE) (Sowdhamini *et al.*, 1996) using the procedure *DIAL* (Sowdhamini & Blundell, 1995).

Structure-based alignment of sequences of related protein domains provides a basis for understanding evolutionary relationships as well as diversity in function and specificity. Such alignments can be used to derive information on amino-acid replacements which are of value also in comparative modelling and fold recognition (Overington *et al.*, 1990). Databases of structural alignments of homologous proteins (HOMSTRAD: HOMologous STRucture Alignment Database) (Overington *et al.*, 1990, 1993; Mizuguchi *et al.*, 1998) and protein superfamilies (CAMPASS: CAMbridge database of Protein Alignments organized as Structural Superfamilies) (*RS*, Sowdhamini *et al.*, 1998) will be described in this paper. Because of the low percentage of sequence identities amongst distantly related proteins, it is difficult, on the basis of sequence alone, to obtain reliable alignments where secondary structures and functionally important residues are aligned correctly. Alignment of proteins in superfamilies, therefore, is based on the conservation of structural features and relationships using the program *COMPARER* (Sali & Blundell, 1990; Zhu *et al.*, 1992). The three databases, described here, are available on the WWW (http://www-cryst.bioc.cam.ac.uk/~ddbase for DDBASE, http://www-cryst.bioc.cam.ac.uk/~homstrad for HOMSTRAD and http://www-cryst.bioc.cam.ac.uk/~campass for CAMPASS).

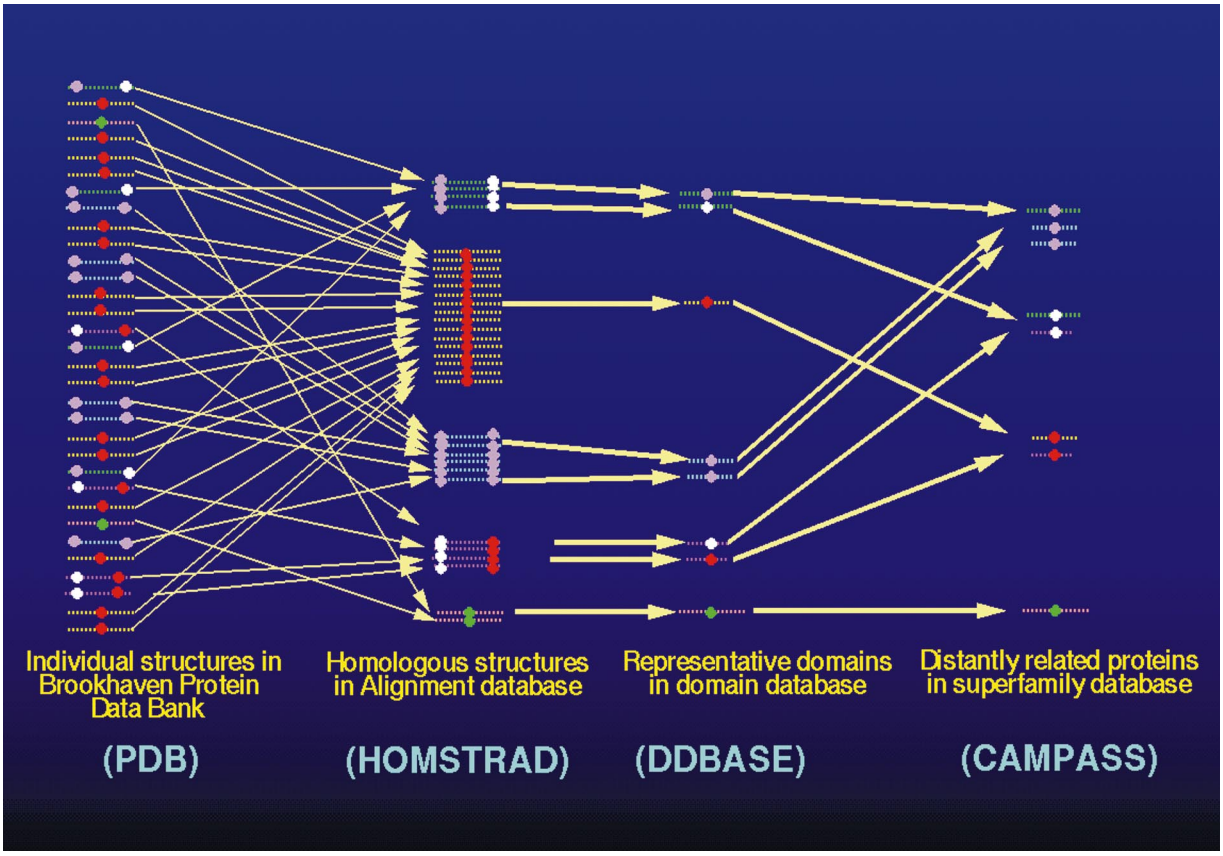

Fig. 1. A cartoon representation of the classification and alignment of proteins at various structural hierarchies. HOMSTRAD database contains alignments of homologous sequences. Some of them exist as multi-domain proteins (denoted by different coloured spheres). DDBASE is a compilation of structural domains found in representatives of homologous proteins. CAMPASS is a database of aligned protein domains belonging to superfamilies.

2. DDBASE

2.1. *Description and availability*

DDBASE is a compilation of the information on structural domains that are present in a representative set of 436 protein chains (Sowdhamini *et al.*, 1996). The identification of structural domains in a protein chain was performed using the program *DIAL* (Sowdhamini & Blundell, 1995), where elements of secondary structure are clustered on the basis of the proximity to each other. This gave rise to 695 structural domains, of which 206 are α-rich, 191 are β-rich and 294 fall under the α-and-β class. 63% of the domains are from multi-domain proteins and 73% of the identified domains have less than 150 residues.

The organization of structural domains in individual protein chains is described on the WWW page assigned to that protein chain; an example is shown in Fig. 2. Secondary-structural dendrograms are provided that correspond to the clustering based on distances between all possible pairs of secondary structures. All possible combinations of nodes in the secondary-structural dendrogram are automatically examined for compactness of putative domains corresponding to clusters and listed with their disjoint-factor values (see Sowdhamini & Blundell, 1995, for details). It is possible for the user to extract the domain boundary corresponding to any situation by clicking on that entry. However, the 'best' domain boundaries, defined by the program, have been identified and the domain organization may be viewed

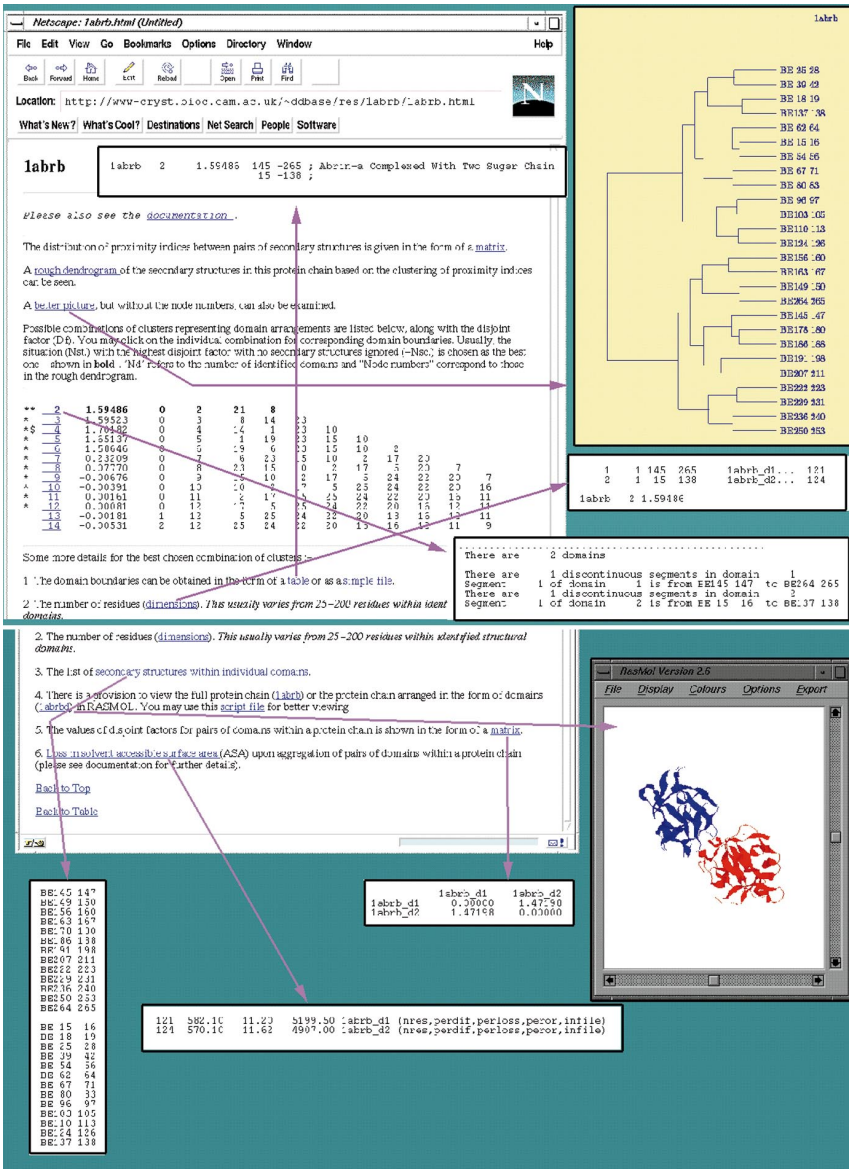

Fig. 2. Domain database (DDBASE) WWW page for the B chain of abrin (PDB code, 1abr) as an example. Domains have been identified using the program *DIAL* (Sowdhamini & Blundell, 1995). The organization of structural domains can be viewed as secondary structural dendrograms where helices and extended strands have been clustered on the basis of intersecondary structural inter-Cα distances. Various combinations of nodes, corresponding to secondary-structural clusters, have been examined for structural compactness and listed along with their disjoint factor (see Sowdhamini & Blundell, 1995, for details). Domain boundaries for all these possibilities can be accessed by clicking on that entry. Further, detailed outputs can be accessed for the 'best' combination. The 'best' combination is usually the one with the highest disjoint factor (Df) without any secondary structures being ignored (-Nst. column shows the number of secondary structures that are ignored while examining various nodes in the dendrogram). The protein chain can be viewed using *RASMOL* (Sayle & Milner-White, 1995) where domains are coloured differently in the case of multi-domain proteins.

Table 1. *Proteins in superfamily and homologous databases*

N_{mem} is the number of members in the superfamily. The first four characters of the member codes correspond to the PDB code, the fifth to the chain identifier and the last character to the domain number. Superfamily name is as defined in SCOP (Murzin *et al.*, 1995). In a few cases where there is considerable functional similarity, we have considered a broader class of proteins under one superfamily (marked as fold). In a few other cases, we have restricted our choice of superfamily members to a group of proteins, defined as a family in SCOP (marked as family), to permit reliable structural superposition and structure-based sequence alignment. N_{hom} is the number of homologous proteins in this family. Many of them are single member families.

Superfamily code (N_{mem})	Member codes	Superfamily name	Homologous family name	N_{hom}
4helud (3)	256ba0	Cytochromes	Cytochrome *b*562	1
	11bbha0, 2ccya0		Cytochrome c'	2
FAD-binding-like (13)	1gal-1, 1pbe-2†, 3cox-1	FAD/NAD(P)-binding domain	Cholesterol oxidase (full protein)	3
	1gnd-2		Guanine nucleotide dissociation inhibitor	1
	1npx-2, 1fcda2†, 1fcda1†		Disulfide oxidoreductase	10
	1trb-1†, 1trb-2†, 3grs-1		As above	
	3grs-2, 3lada2		As above	
	2tmda2		Trimethylamine dehydrogenase	1
FMN_typeI (2)	2tmda1†, 1oyb-0†	FMN-linked oxidoreductases	Flavin-binding beta-barrel	2
PH (3)	btn-0, 1dyna0, 1mai-0	PH domain-like	Pleckstrin-homology domain	7‡
SH3(2)	1lck-2, 1pht-0	SH3 domain	SH3 domain	7
ab5_toxins (5)	1bova0, 1chbd0, 1ptob2, 1ptod0, 1ptof0	Bacterial enterotoxins	Bacterial AB5 toxins	8‡
ab_hydrolases (8)	1broa0	Alpha/beta-hydrolases	Bromoperoxidase A2	1
	2had-0		Haloalkane dehalogenase	1
	1thta0		Thioesterases	1
	1gpl-0		Lipase	2
	1tca-0, 2ace-0		alpha beta-hydrolase	3
	1din-0		Dienelactone hydrolase	1
	1whta0		Serine carboxypeptidase	3
actinIA (3)	1atna3, 3hsc-2	Actin-like ATPase domain	Actin	2
	1glcg1		Glycerate kinase	1
actinIIA (3)	1atna1, 3hsc-3	Actin-like ATPase domain	See actinIA	
	1glcg2		See actinIA	
actin_binding (2)	1vil-0, 1svq-0	Actin depolymerizing proteins	Gelsolin-like	3‡
adk (2)	2ak3a1, 1gky-1	Nucleotide and nucleoside kinases	Nucleotide kinase	5
adp (4)	1ddt-3, 1dmaa0, 1ltaa0	ADP-ribosylation	ADP-ribosylating toxins	6‡
	1ptoa0		As above	
animal_viral (5)	1bbt30, 2rhn3m, 1cov1m	Animal virus proteins (family)	Picornavirus coat proteins	6
	61bbt10, 1bbt2m		As above	
anticodon_binding (2)	1asya2, 1lyla2	An anticodon-binding domain (family)	An anticodon-binding domain	
asp_hiv (3)	1hiva0	Acid proteases	Retroviral proteinase	4‡
	45pep-2, 5pep-1		Aspartic proteinase	11
bacteriophage (2)	1gpc-0, 2gva-0	Bacteriophage ssDNA-binding proteins	Bacteriophage ssDNA-binding proteins	3‡
beta-gamma-crystallin_like (3)	4gcr-1, 1prs-1	Crystallins/protein S/killer toxin	Crystallin	5
	1wkt-0		Yeast killer toxin	1
bgt-gpb (2)	2bgu-0	Beta-glucosyltransferase & glycosyltransferase	Beta-glucosyltransferase	1
	1gpb-0		Oligosaccharide phosphorylase	3a
cbp (7)	3cln-2, 2scpa2, 2scpa1	EF-hand	Calcium binding protein – calmodulin-like	6
	2sas-1, 2sas-2, 1rec-1†		As above	5
	1rro-m†		Parvalbumin	5
ccperoxy (3)	1lgaa0, 1scha0†, 2cyp-0	Heme-dependent peroxidases	Peroxidase	4
creatinase (2)	1chma2, 1mat-0	Creatinase/methionine aminopeptidase	Creatinase/methionine aminopeptidase	3‡
ctt (2)	1ctt-1, 1ctt-2	Cytidine deaminase	Cytidine deaminase	1
cys (2)	2act-0, 1gcb-1†	Papain-like	Cysteine proteinase	5
cystineknot (6)	1bet-0	Cystine-knot cytokines	Neurotrophin	3‡
	a1aoca2		Coagulogen	1
	1pdga0		Platelet-derived growth factor	1
	1hcna0, 1hcnb0		Gonadotropin	1

Table 1 (*cont.*)

Superfamily code (N_{mem})	Member codes	Superfamily name	Homologous family name	N_{hom}
	2tgi-0		Transforming growth factor β	4‡
cytc (3)	351c-0,1cyi-0†	Monodomain cytochrome c (family)	Cytochrome-c5	5
	1ycc-0		Cytochrome-c	9
cytokine (2)	1i1b-0, 4fgf-0 (2fgf)	Cytokine	Interleukin 1-β-like growth factor	5
exopeptidase (3)	1amp-0	Zn-dependent exopeptidases	Bacterial aminopeptidases	2‡
	1lcpa1		Leucine aminopeptidase, C-domain	1
	2ctb-0		Pancreatic carboxypeptidases	3‡
ferredoxin_reductases (3)	2pia-3	Ferredoxin reductase-like C-terminal domain	Phthalate dioxygenase reductase	1
	1ndh-2, 1fnc-2		Reductases	5‡
flav (7)	1bmta1	Flavodoxin-like(fold)	Methionine synthase C-	1
	1orda4		Ornithine decarboxylase N-domain	1
	1cus-m		Cutinase	1
	3chy-0		CHEY-like	5‡
	1scua2		Succinyl-CoA synthetase-α -chain C-domain	1
	4fxn-0		Flavodoxin	6
	1qora1		Alcohol/glucose dehydrogenase, C-domain	2
globins (7)	1flp-0, 1ithb0, 3sdha0, 2gdm-0, 1mbc-0, 2hbg-0, 1ash-0	Globin-like	Globin	23
glucoamylase_like (3)	1gai-0	Glycosyltransferases of the superhelical fold	Glucoamylase	1
	1clc-1, 1cem-0		Cellulase catalytic domain	3‡
glucosyltransferases (18)	1bgl-2, 1ecea0, 1edg-0	Glycosyltransferases	beta-glycanases	11‡
	1ghsa0, 1xyza0, 1cec-0		As above	
	1byb-0		beta-amylase	1
	1cbg-0		Family 1 of glycosyl hydrolase	4‡
	1cgt-1, 1bpla1†, 1ppi-1†		Amylase (full protein)	6
	2amg-1†		As above	
	1ctn-1, 2ebn-0, 2hvm-0		Type II chitinase	6‡
	1nar-0		As above	
	1qba-1		Bacterial chitobiase ca. domain	1
	4xiaa1		Xylose isomerase	5
gshase_2 (4)	1gsh-3, 2dln-2	Glutathione synthetase ATP-binding-like	Peptide synthetases C-domain	2‡
	1scub3		Succinyl-CoA synthetase beta- N-	1
	1dik-2		Pyruvate phosphate dikinase N-	1
gshase_3 (5)	1gsh-2	Glutathione synthetase ATP-binding like	See gshase_2	
	2dln-1		See gshase_2	
	1scub2		See gshase_2	
	1bnca3		Biotin carboxylase	1
	1dik-3		See gshase_2	
ig (12)	1cid-2, 1vcaa2, 3 cd4-1	Immunoglobulin	Immunoglobulin domain – C2 set	2
	1hsaa2, 1vabb0		Histocompatibility antigen-binding domain	5
	1nct-0, 1tit-0, 1tlk-0		I set domains	7‡
	1vcaa1, 1wit-0		As above	
	2fbjl2, 3hflh1		Immunoglobulin domain C1 set – constant immunoglobulin	17
il8_like (2)	1huma, 1ikl- (1il8)	Interleukin 8-like chemokines	Interleukin 8-like protein	5
kinases (3)	1atpe0, 1csn-0, 1irk-0	Protein kinases (PK) ca. core	kinase(1apm)	7
lectins (6)	1saca0	ConA-like lectins/glucanases	Pentraxin	2‡

Table 1 (*cont.*)

Superfamily code (N_{mem})	Member codes	Superfamily name	Homologous family name	N_{hom}
	1ayh-m		Bacillus 1-3,1-4-β-glucanase (2ayh)	3‡
	2ltn-m		Plant lectin	7
	1slt-0		S-lectin	2
	1kit-2, 1kit-3		Vibrio cholerae sialidase, N-	1
lipocalin (5)	1icm-0 (1ifb), 1mup-0	Lipocalins	Lipocalin	12
	1epba0 (1bbp), 1bbpa0		As above	
	1fel-0 (1rbp)		As above	
methyltransferases (5)	1vpt-1	S-adenosyl-L-methionine -dependent methyltransferases	Polymerase regulatory subunit VP39	1
	2adma2, 1hmy-1		DNA methylases	3‡
	1vid-1		Catechol O-methyltransferase COMT	1
	1xvaa1		Glycine N-methyltransferase	1
muconate_lactonizing (3)	1muca1, 2mnr-1	Enolase & muconate-lactonizing C-domain	Muconate lactonizing enzyme-like	3‡
	4enl-1		Enolase	2‡
nip (3)	1dts-0, 1adea1, 1nipb0	P-loop containing nucleotide triphosphate hydrolases	Nitrogenase iron protein-like	3‡
p450 (4)	2cpp-0, 2hpda0, 1cpt-0	Cytochrome P450	Cytochrome p450	3
	1oxa-0		As above	
pbgd1 (4)	1pda-1, 1sbp-2, 1omp-1	Periplasmic binding II	Phosphate binding protein-like	12‡
	1lfg-1		Transferrin	5‡
pbgd2 (4)	1pda-2, 1omp-2, 1sbp-1	Periplasmic binding II	See pbgd1	
	1lfg-3		See pbgd1	
phospholipase (2)	1bp2-0	Phospholipase A2	Phospholipase A2	7
	1poc-m		Insect phospholipase A2	1
plant_viral (5)	1bmv21, 1cwpam, 1bmv10	Plant virus proteins (family)	Plant virus coat protein (4sbv)	2
	1bmv22, 2stv-m		As above	
plp1 (4)	1ars-2	PLP-dependent transferases	Aspartate aminotransferase (3aat)	2
	1dge-2		omega-Amino acid_pyruvate aminotransferase-like	2‡
	1orda2		Ornithine decarboxylase major domain	1
	1tpla1		Tyrosine phenol-lyase	1
plq (2)	1plq-1, 1plq-2	DNA-clamp	DNA polymerase processivity factor	1
porins (3)	2omf-0, 2por-0†	Porins	Porin	3
	1mal-0		Maltoporin	2‡
ppase1 (3)	1spia2	Sugar phosphatases	Fructose-1,6-bisphosphatase	3‡
	2hhma1		Inositol monophosphatase	1
	1inp-1		Inositol polyphosphate 1-phosphatase	1
ppase2 (3)	1spia1	Sugar phosphatases	See ppase1	
	2hhma2		See ppase1	
	1inp-2		See ppase1	
ras (4)	5p21-0, 1eft-1 (1etu)	G proteins(family)	GTP-binding protein	4
	1tada1†, 1hura0†		As above	
repressor_like (4)	1copd0, 1r69-0, 1neq-0†	Lambda repressor-like DNA-binding domains	DNA-binding repressor (2cro)	5
	1octc0		Oct-1 POU-specific domain	1
ribonucleaseh_like (5)	1bco-1	Ribonuclease H-like	Mu transposase core domain	1
	1kfd-1		Exonuclease domain of DNA polymerase KF	2‡
	1hjra0		RuvC resolvase	1
	2rn2-0		Ribonuclease H (1rnh)	3
	1itg-0		Retroviral integrase	2‡
rubredoxins (3)	8rxna0	Rubredoxin-like(fold)	Rubredoxin (7rxn)	5
	4at1b2		Aspartate carbamoyl transferase_RC	1
	1tfi-0		A transcriptional factor domain	2‡

Table 1 (*cont.*)

Superfamily code (N_{mem})	Member codes	Superfamily name	Homologous family name	N_{hom}
serineproteases1 (5)	1sgt-1	Trypsin-like serine proteases	Serine proteinase, mammalian	16
	1hava1		picornain	2‡
	2alp-2, 1arb-1†		Serine proteinase, bacterial	4
	1svpa1		Viral proteases	2‡
serineproteases2 (4)	2alp-1, 1arb-2	Trypsin-like serine proteases	See serineproteases1	
	1hava		See serineproteases1	
	1svpa2		See serineproteases1	
sial_neur (3)	1eus-0 (1nsb), 1dim-0	Sialidases (neuraminidases)	Neuraminidase	4
	1nsca0	As above		
sslipid (2)	1hyp-0	Bifunctional inhibitor/lipid-transfer Seed storage 2S albumin	Plant lipid-transfer and hydrophobic proteins	4‡
	1bip-0		Bifunctional proteinase	1
strep (2)	1sria0	Avidin/streptavidin	Avidin (1pts)	2
	1smpi0		Metalloprotease inhibitor	1
superantigen_toxins (2)	1tssa1, 1se2-1	Superantigen toxins N-domain (family)	Superantigen toxins N-domain	4‡
thiamin_binding (6)	1pyda1, 1pyda2, 1powa1	Thiamin-binding	Pyruvate oxidase and decarboxylase	3‡
	1powa2		As above	
	1trka1, 1trka2		Transketolase	1
thioredoxin (6)	1erv-0, 1thx-0, 1aba-0	Thioredoxin-like	Thioredoxin (3trx)	4
	1dsba1		Disulfide-bond formation facilitator	2
	2gsta1		Glutathione S-transferase (5gst)	7
	1gp1a0		Glutathione peroxidase	1
trp-biosynthesis (3)	1igs-0, 1pii-2, 1wsya0	Tryptophan biosynthesis enzymes	Tryptophan biosynthesis enzyme	2
tyrosine_phosphatases (3)	2hnq-0, 1ypta0	Phosphotyrosine protein phosphatases I	Higher molecular-weight phosphotyrosine	3‡
	1vhra0		Dual-specificity phosphatase	1
viral_coat (3)	2bbva0	Viral coat and capsid proteins	Insect virus proteins	1
	2tbva2		Plant virus coat protein	2
	2cas1m†		Picornavirus coat proteins	7

† This entry is yet to be added in one of the existing families in the homologous alignment database. ‡ This family is yet to be added in the homologous alignment database.

on graphics using *RasMol* (Sayle & Milner-White, 1995). Each domain can be identified by its unique six-character code (the first four characters correspond to the PDB code of the protein, the fifth to the chain identifier and the sixth, as a subscript, corresponds to the domain numbering as in the individual domain pages).

2.2. *Application*

DDBASE can be used to trace similarities where particular domains are shared between proteins. It is especially useful where there are discontinuous domains. 400 large (with seven or more secondary structures) domains can be grouped into 30 classes on the basis of the structural similarity estimated from structural environments of individual secondary structures (Rufino & Blundell, 1994; Sowdhamini *et al.*, 1996). The clustering of individual protein domains into structurally similar classes can also be examined on the DDBASE WWW page.

3. HOMSTRAD and CAMPASS

3.1. *Description and availability*

HOMSTRAD and CAMPASS are databases of structure-based alignments of protein sequences, grouped into homologous families and superfamilies, respectively. Aligned sequences of families of homologous protein structures are available in HOMSTRAD (Overington *et al.*, 1990, 1993) and categorized according to the secondary-structural classes. There are 130 homologous protein families with at least two members in the March 1998 version. The sequences of homologous proteins within a family are initially aligned using the rigid-body superposition program *MNYFIT* (Sutcliffe *et al.*, 1987) or *COMPARER* (Sali & Blundell, 1990; Zhu *et al.*, 1992) and later subjected to a careful manual examination. Similar types of information are available for CAMPASS, the database of protein (domain)s belonging to superfamilies (*RS*, Sowdhamini *et al.*,

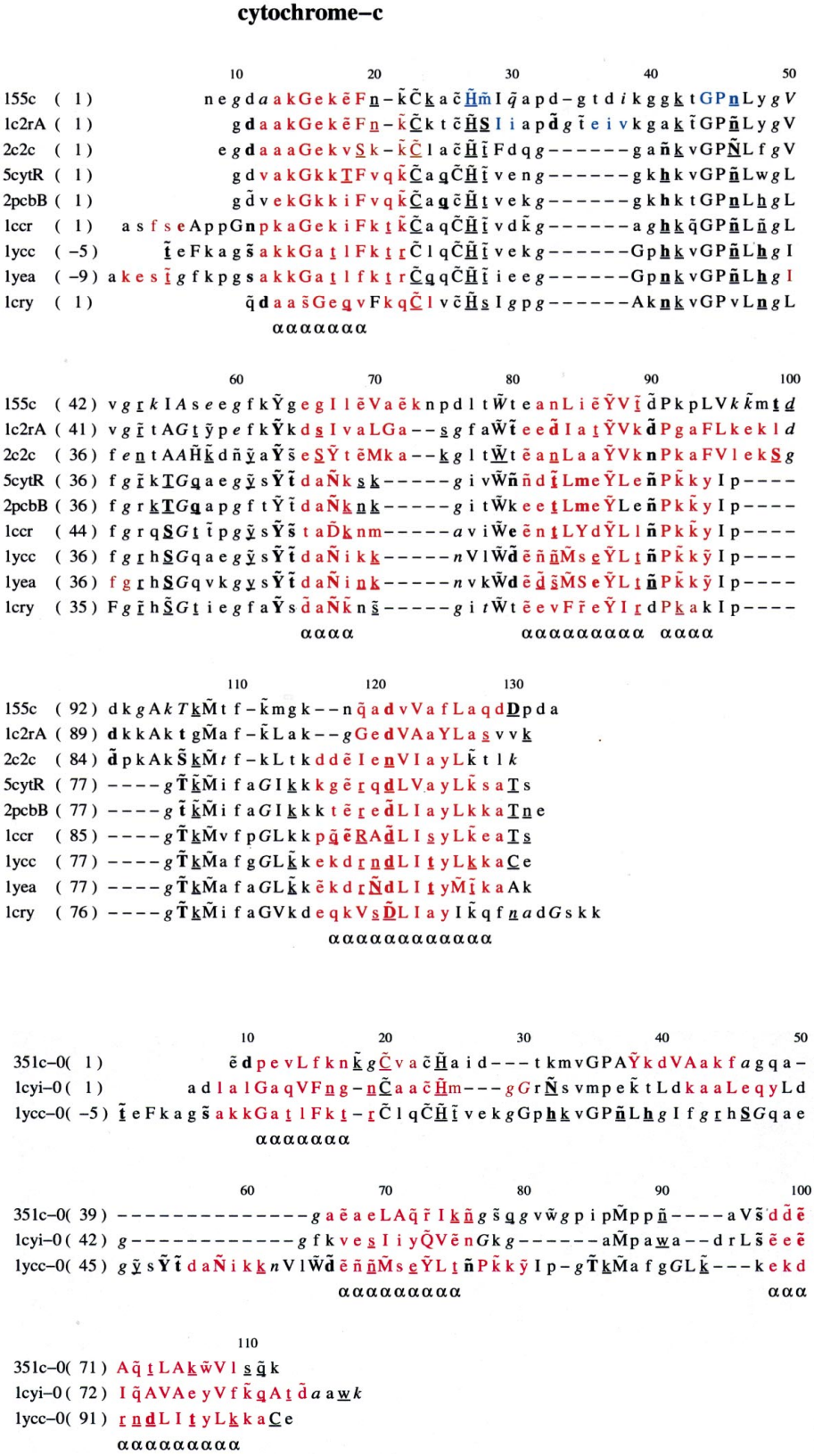

Fig. 3. HOMSTRAD database. Structure-based alignment of proteins in the family of cytochrome *c*. The first four characters of the code of the protein corresponds to the PDB code. Numbers in brackets correspond to residue numbers and residues are shown in single letter code. The alignment has been formatted using *JOY* (Overington *et al.*, 1990). The conserved helices are important to the structural integrity of the proteins; functionally important residues (for example C*XX*CH, residue number 13 of 1ycc) are conserved. Residues are classified into two categories: those which are in the interior and those which are solvent-exposed (with solvent accessibility (ASA) values more than 7% (Hubbard & Blundell, 1987). In the sequence alignment, the solvent-exposed and solvent-buried residues are shown in lower case and upper case, respectively. Residues which have a positive φ value and a *cis*-peptide bond in their backbone conformation are shown in italics and with a breve accent on top, respectively. Disulfide-bonded cystine residues are shown by a cedilla symbol. Hydrogen bonding to other side chains, main-chain amides and main-chain carbonyl groups are shown by a tilde (indicated in non-HTML files), in bold and underlined, respectively. Residues in β-strands, α-helices and 3(10)-helices are shown in blue, red and maroon, respectively.

Fig. 4. CAMPASS database. Structure-based alignment of the cytochrome superfamily including distantly related proteins such as *c*550. Helix 2 of 1ycc, conserved within the homologues (see Fig. 3), occurs as an insertion in this alignment. Despite poor sequence identity, the functionally important residues (C*XX*CH) are conserved amongst the members in this superfamily.

1998). Superfamilies of structural domains were selected initially on the basis of structural environment at secondary structural units (Rufino & Blundell, 1994; Sowdhamini *et al.*, 1996). The selection of superfamilies has been extended by referring to SCOP (Murzin *et al.*, 1995) and by including smaller domains like the cystine-knots, not considered earlier in the clustering analysis since they were not easy to compare using automatic structure-based procedures. 367 of 451 superfamilies annotated in SCOP have single families (Brenner *et al.*, 1997; the more recent February 1998 release of SCOP has 419 of the 571 superfamilies with single families). Superfamily members were chosen such that no two domains within a superfamily share more than 25% sequence identity (alignments of closely related proteins are available in HOMSTRAD). This cut-off is consistent with the DDBASE definition in choosing representative protein chains. A rigorous sequence-alignment program, *COMPARER* (Sali & Blundell, 1990; Zhu *et al.*, 1992), was used to align the members of a superfamily on the basis of structural features and relationships, which are equivalenced using simulated annealing. Table 1 lists protein superfamilies, with at least two members within the above-defined cut-off of sequence identity, whose alignments have been compiled in the March 1998 version. This includes 67 multi-member superfamilies which involves 293 domains representing 464 homologous proteins. There are a further 357 superfamilies, annotated in SCOP, which have single members (Murzin *et al.*, 1995; Brenner *et al.*, 1997). A few other multi-member superfamilies included in SCOP, such as the DNA-binding HMG box, pheromones, annexins and insulin-superfamily, were excluded from CAMPASS as members exhibited more than 25% sequence identity.

3.2. *Availability*

The WWW site of HOMSTRAD (Mizuguchi *et al.*, 1998) provides a page for each of the families. The name of the protein, source, resolution and *R* factor are given for each family member corresponding to a PDB entry. The alignment of sequences is formatted in *JOY* (Overington *et al.*, 1990) which highlights the conservation of local-residue structural features such as secondary structure, solvent accessibility and hydrogen bonding. Fig. 3 shows the alignment of cytochrome *c* from different sources and its homologues (cytochrome *c*2 and cytochrome *c*550), as an example.

CAMPASS, on the WWW, provides information on the superfamilies: for each superfamily member, the name, source, resolution and domain boundaries are given. The beginning and end residue numbers for each segment of discontinuous domains are recorded. The pairwise percentage identity matrix of the members is provided. The structure-based alignment in the *JOY*-annotated form (Overington *et al.*, 1990), similar to that described in HOMSTRAD, is shown and also available for extraction in the form of PostScript files, or as LATEX or HTML files or as a plain text file. Fig. 4 shows the alignment of the cytochrome superfamily as an example. A single representative (1ycc) of the nine cytochrome homologues (see above and Fig. 3) has been aligned with rather distantly related cytochromes such as cytochrome *c*6 and *c*551. The structures of the proteins within a family/superfamily have been superposed using *MNYFIT* (Sutcliffe *et al.*, 1987), where the equivalent residues correspond to the final alignment. These superposed structures can be viewed on the WWW using the *RASMOL* graphics interface (Sayle & Milner-White, 1995).

Fig. 5 shows the distribution of pairwise percentage identities in the two alignment databases. Protein pairs in HOMSTRAD have a broad range of pairwise sequence identities with a slightly bimodal distribution (237 pairs have sequence identities between 25 and 30% and 121 pairs have sequence identities between 60 and 65% out of a total of 1962 pairs). However, the majority of homologous proteins in the database have sequence identities between 15 and 65%. The distribution of pairwise sequence identity of members within superfamilies (CAMPASS) is restricted to a maximum of 25%. A vast majority of protein pairs (449 out of 665) have pairwise percentage identities between 5 and 15%.

4. Conclusions

HOMSTRAD and CAMPASS are distinct from but complementary to other databases. SCOP (Murzin *et al.*, 1995) has classified the entire Protein Data Bank at different levels of structural hierarchy and structural domains are defined. There is emphasis on functionality in the clustering of folds. SCOP does not attempt to perform or present sequence or structural alignments. CATH (Orengo *et al.*, 1993, 1994) was originally designed and developed for whole proteins where the

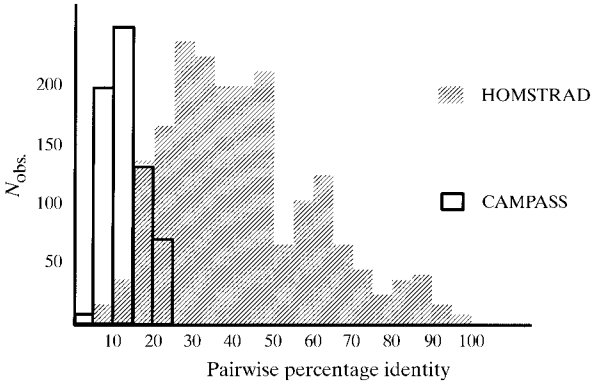

Fig. 5. Distribution of pairwise percentage sequence identities amongst members in the homologue alignment database (HOMSTRAD) and superfamily alignment database (CAMPASS).

authors had taken particular caution to exclude multi-domain proteins. Subsequently, the structures have been systematically classified at the level of domains (Orengo *et al.*, 1997). CATH does not include structure-based alignments of sequences. *FSSP* (Holm & Sander, 1994) is most similar to HOMSTRAD and CAMPASS due to the fact that *FSSP* also provides structure-based sequence alignments, even incorporating remote homologues. However, the alignments do not distinguish homologues and superfamilies from those which only share a similar fold. The databases described in this paper contain structure-based alignments that have been specially annotated to describe the structural environment at residue positions. This should provide extra information useful in the comparison of protein structures.

References

Bairoch, A. (1991). *Nucleic Acids Res.* **19**, 2013–2018.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L. & Humbel, R. E. (1980). *Nature (London)*, **287**, 781–787.

Bork, P., Ouzounis, C. & Sander, C. (1994). *Curr. Opin. Struct. Biol.* **4**, 393–403.

Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). *Nature (London)* **358**, 287–287.

Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997). *Curr. Opin. Struct. Biol.* **7**, 369–376.

Chothia, C. (1984). *Ann. Rev. Biochem.* **53**, 537–572.

Crippen, G. M. (1978). *J. Mol. Biol.* **126**, 315–332.

Go, M. (1981). *Nature (London)*, **291**, 90–92.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). *Protein Sci.* **1**, 409–417.

Holm, L. & Sander, C. (1994). *Nucleic Acids Res.* **22**, 3600–3609.

Hubbard, T. J. P. & Blundell, T. L. (1987). *Protein Eng.* **1**, 159–171.

Islam, S. A., Luo, J. & Sternberg, J. E. (1995). *Protein Eng.* **8**, 513–525.

Johnson, M. S., Overington, J. P. & Blundell, T. L. (1993). *J. Mol. Biol.* **231**, 735–752.

Koonin, E. V., Bork, P. & Sander, C. (1994). *EMBO J.* **13**, 493–503.

Mizuguchi, K., Deane, C., Overington, J. P. & Blundell, T. L. (1998). *Protein Sci.* In the press.

Murzin, A. G. (1996). *Curr. Opin. Struct. Biol.* **6**, 386–394.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Murzin, A. G. & Chothia, C. (1992). *Curr. Opin. Struct. Biol.* **2**, 895–903.

Nichols, W. L., Rose, G. D., Eyck, L. F. T & Zimm, B. H. (1995). *Proteins*, **23**, 38–48.

Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993). *Protein Eng.* **6**, 485–500.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). *Nature (London)*, **372**, 631–634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Overington, J. P., Johnson, M. S., Sali, A. & Blundell, T. L. (1990). *Proc. R. Soc. London Ser. B*, **241**, 132–145.

Overington, J. P., Zhu, Z.-Y., Sali, A., Johnson, M. S., Sowdhamini, R., Louie, G. V. & Blundell, T. L. (1993). *Biochem. Soc. Trans.* **21**, 597–604.

Richardson, J. S. (1981). *Adv. Protein Chem.* **34**, 167–339.

Rose, G. D. (1979). *J. Mol. Biol.* **134**, 447–470.

Rossmann, M. G. & Argos, P. (1977). *J. Mol. Biol.* **109**, 99–129.

Rufino, S. D. & Blundell, T. L. (1994). *Comput. Aided Mol. Design*, **8**, 5–27.

Sali, A. & Blundell, T. L. (1990). *J. Mol. Biol.* **212**, 403–428.

Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.

Schulz, G. E. (1977). *Angew. Chem. Intl Ed.* **16**, 23–33.

Siddiqui, A. S. & Barton, G. J. (1995). *Protein Sci.* **4**, 872–884.

Sowdhamini, R. & Blundell, T. L. (1995). *Protein Sci.* **4**, 506–520.

Sowdhamini, R., Burke, D. F., Huang, J.-F., Mizuguchi, K., Nagarajaram, H. J., Srinivasan, N., Steward, R. E. & Blundell, T. L. (1998). *Structure*. In the press.

Sowdhamini, R., Rufino, S. D. & Blundell, T. L. (1996). *Folding Design*, **1**, 209–220.

Sutcliffe, M. J., Haneef, I., Carney, D. & Blundell, T. L. (1987). *Protein Eng.* **1**, 377–384.

Swindells, M. B. (1995). *Protein Sci.* **4**, 103–112.

Wetlaufer, D. B. (1973). *Proc. Natl Acad. Sci. USA*, **70**, 697–701.

Wodak, S. J. & Janin, J. (1981). *Biochemistry*, **20**, 6544–6553.

Zehfus, M. H. & Rose, G. D. (1986). *Biochemistry*, **25**, 5759–5765.

Zhu, Z.-Y., Sali, A. & Blundell, T. L. (1992). *Protein Eng.* **5**, 43–51.