Structural similarity to link sequence space: New potential superfamilies and implications for structural genomics

PATRICK ALOY,¹ BALDOMERO OLIVA,² ENRIQUE QUEROL,²
FRANCESC X. AVILES,² AND ROBERT B. RUSSELL^{1,3}

¹EMBL, Meyerhofstrasse, 1. D-69117, Heidelberg, Germany

²Institut de Biotechnologia i Biomedicina, Universitat Autònoma de Barcelona, Bellaterra 08193, Barcelona, Spain

(RECEIVED September 28, 2001; FINAL REVISION January 29, 2002; ACCEPTED January 29, 2002)

Abstract

The current pace of structural biology now means that protein three-dimensional structure can be known before protein function, making methods for assigning homology via structure comparison of growing importance. Previous research has suggested that sequence similarity after structure-based alignment is one of the best discriminators of homology and often functional similarity. Here, we exploit this observation, together with a merger of protein structure and sequence databases, to predict distant homologous relationships. We use the Structural Classification of Proteins (SCOP) database to link sequence alignments from the SMART and Pfam databases. We thus provide new alignments that could not be constructed easily in the absence of known three-dimensional structures. We then extend the method of Murzin (1993b) to assign statistical significance to sequence identities found after structural alignment and thus suggest the best link between diverse sequence families. We find that several distantly related protein sequence families can be linked with confidence, showing the approach to be a means for inferring homologous relationships and thus possible functions when proteins are of known structure but of unknown function. The analysis also finds several new potential superfamilies, where inspection of the associated alignments and superimpositions reveals conservation of unusual structural features or co-location of conserved amino acids and bound substrates. We discuss implications for Structural Genomics initiatives and for improvements to sequence comparison methods.

Keywords: Protein structure; sequence; function; homology; structural genomics

Structural genomics initiatives attempt to infer details of protein function by way of 3D structure determination (e.g., Eisenberg et al. 2000; Shapiro and Harris 2000), and efforts already have produced structures for proteins of unknown function (e.g., Yang et al. 1998; Zarembinski et al. 1998; Boggon et al. 1999; Christendat et al. 2000). Structure can provide insights into function in a number of different ways. For example, aspects of molecular function can be revealed

if the act of solving the structure reveals details of bound non-protein atoms (e.g., Zarembinski et al. 1998). Alternatively, if a new protein structure adopts a previously observed fold, then it sometimes is possible to infer functional details by considering the function of other proteins adopting the same fold (e.g., Murzin et al. 1995; Artymiuk et al. 1997; Holm and Sander 1997). If fold similarities are ambiguous (e.g., the fold performs many functions) or if a protein adopts a new fold, it still is possible to infer function by comparison of key active site residues (e.g., Wallace et al. 1997; Russell 1998; Aloy et al. 2001) or by similarities between protein-binding sites or surfaces (e.g., Russell et al. 1998; Boggon et al. 1999).

For instances where a new protein structure adopts a previously observed fold, a major goal of structural genomics initiatives is to provide a structural link between sequence

³Reprint requests to: Robert B. Russell, EMBL, Biocomputing, Meyerhofstrasse, 1. D-69117, Heidelberg, Germany; e-mail: russell@embl-heidelberg.de; fax: 49-6221-387-517.

Abbreviations: 3D, three dimensional; Ig, immunoglobulin; RMSD, root mean square deviation; PDB, Protein Data Bank; ATP, adenosine triphosphate; SCOP, structural classification of proteins; NCBI, National Center for Biotechnology Information; URL, universal resource locator.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.3950102>.

families that are not detectably similar when only sequences are compared (e.g., Brenner et al. 1998). A key issue is whether protein structure similarity per se can be used to infer that proteins are homologous and/or whether they are likely to show similarities in their molecular function. If proteins sharing similar 3D structures can be said, with confidence, to share a common ancestor, then a similarity in molecular function is more likely (e.g., Hegyi and Gerstein 1999).

For groups of proteins sharing a common fold in the absence of strong sequence similarity, previous work has focussed on the distinction between “remote homologs” and “analogs” (proteins sharing a similar fold in the absence of convincing evidence of a common ancestor). Studies were aided greatly by the creation of structural classification databases, which produced a reliable set of homologous proteins lacking significant sequence identity. Proteins now are generally classified as belonging to “superfamilies” if evidence for homology comes from evidence apart from structure similarity. Such evidence can include the conservation of key active site or structural residues, common functions, or unusual structural characteristics unlikely to have arisen by chance (e.g., left-handed $\beta\alpha\beta$ motifs). Even in the absence of detectable sequence similarity, proteins are considered to share a common ancestor based on such evidence. Probably the best source of these data is the Structural Classification of Proteins (SCOP) database (Murzin et al. 1995). Here, proteins that adopt a similar fold with little or no sequence similarity are placed in the same superfamily if there is evidence that they share a common ancestor. SCOP has been used in numerous studies on the distinction between remote homology and analogy (e.g., Russell and Barton 1994; Russell et al. 1997; Matsuo and Bryant 1999). Although an automated means to predict superfamily relationships remains elusive, limited success in discerning homology has come from analysis of features such as sequence similarity calculated from structure-based alignments (Murzin 1993b; Russell et al. 1997; Ponting and Russell 2000), structurally conserved core residues (Matsuo and Bryant 1999) or combinations of multiple features (e.g., Holm 1998; Dietmann and Holm 2001).

In parallel with the above developments has been the emergence of databases that classify protein sequences. Genome projects have produced vast numbers of protein sequences, which are frequently grouped into aligned domain families. Accurate alignment of protein sequences permits the construction of sensitive profiles, or hidden Markov models (HMMs; e.g., Eddy 1998) that can be used to detect other remote members of the family. In addition, sequence comparisons enable long protein sequences to be divided into discrete functional or structural domains. Simple modular architecture research tool (SMART; Schulz et al. 2000), protein families (Pfam; Bateman et al. 2000), protein fingerprints (PRINTS; Attwood et al. 2000), clusters of or-

thologous groups (COGs; Tatusov et al. 2000) and BLOCKS (Henikoff et al. 2000) are examples of protein domain sequence alignment databases.

We describe here a method to detect superfamily relationships based on the statistical significance of an inferred structure-based sequence alignment. We first construct a database that merges sequence alignments from SMART and Pfam according to folds and superfamilies defined within the SCOP database. After finding overlapping sequences within the structure and sequence databases, we construct structural alignments and use these to merge different SMART/Pfam alignments. We then apply the statistical P-value described by Murzin (1993b) to assess the significance of sequence identities between pairs of sequences aligned by structure comparison. We discuss interesting new potential superfamilies in addition to implications for protein sequence comparison and structural genomics.

Results

Linking SMART domains via structure

A total of 193 out of 419 SMART domains could be matched to one or more domains in SCOP, and a total of 20,447 sequences out of 30,050 in the database can be matched in whole or part to a domain of known three-dimensional structure. There were several partial overlaps between the databases implying differences in the way in which they define domains. This is perhaps not surprising as SCOP frequently divides protein structures into domains that could not readily be detected by sequence comparison, and which are only apparent upon analysis of the 3D structure. SMART also sometimes divides domains to account for domain insertions that can hinder sequence analysis (e.g., Ponting and Russell 1998).

A total of 30 SCOP superfamilies could be matched to more than one SMART domain, producing a total of 183 potential pairings of SMART domains via structurally similar superfamilies. If one considers “fold” level similarities within SCOP (i.e., where proteins are in the same fold, but in different superfamilies; where evidence for an evolutionary relationship has not yet been found), an additional 120 pairings are added from a total of 11 folds. Details of all of the pairings are given in Table 1. A table showing the results for all the potential pairings, at both superfamily and fold level can be found at http://www.embl-heidelberg.de/~aloy/struct_align.

Superfamily links

For the superfamily pairs, 82 out of 183 had $P_{3D} \leq 5 \times 10^{-3}$ (see Materials and Methods), 31 had $P_{3D} \leq 10^{-5}$ and 13 had $P_{3D} \leq 10^{-10}$. These are instances where SMART domains are clearly sequence similar, but have been split to aid analysis of new protein sequences (C. Ponting, pers. comm.) or where similarities that are known are yet to be merged

Table 1. Potential pairings for SMART domains at fold and superfamily level

SCOP fold	N	SMART domains	SCOP fold	N	SMART domains	SCOP fold	N	SMART domains
Knottins	10	e-FOLN, e-EGF_CA, e-EGF_Lam, e-FU, e-IB, e-PTI, e-fCBD, e-ChtBD, e-WAP, e-BowB	DNA/ RNA-binding 3-helical bundle	9	n-HTH_ARSR, n-ETS, n-FH, n-HTH_CRP, n-IRF, n-HSF, n-SANT, n-HOX, n-PAX	Immunoglobulin fold	7	e-IGc1, e-IGc2, e-IGIPT, e-PKD, e-CA, e
OB fold	4	n-CSP, o-S1, o-SNc, e-TIMP,	SAM domain-like	4	n-HhH2, n-HhH1, o-SAM_PNT, s-SAM	β trefoil	4	e-IL1, e-FGF, e-RIC STI
β -grasp	3	s-RBD, s-RA, s-UBQ	Long α -hairpin	2	o-DnaJ, s-HR1	IL8-like	2	e-SCY, n-CHROMO
Four helical bundle	2	s-HPT, s-TarH	Leucine-rich repeat	2	o-LRRcap, e-LRR_RI			
SCOP superfamily	N	SMART domains	SCOP superfamily	N	SMART domains	SCOP superfamily	N	SMART domains
P-loop containing nucleotide hydrolases	11	s-MYSc, o-AAA, s-RAN, s-RAB, s-SAR, s-RAS, s-RHO, s-ARF, s-G-alpha, s-GuKc, s-KISc	Four helical cytokines	7	e-IL4_13, e-IL10, e-IFabD, e-CSF2, e-IL2, e-LIF_OSM, e-IL6	Winged helix DNA-binding domain	6	n-HTH_ARSR, n-ETS, n-HTH_CRP, n-IRF, n-I
Immunoglobulin	5	e-IGc1, e-IGc2, e-IGv, n-IPT, e-PKD	Nucleosome core histones	5	n-H3, n-H2A, n-H4, n-H2B, n-AHL	PH domain-like	5	s-PTB, s-RanBD, s-WHPH, s-BTK
Cystine-knot cytokines	4	e-GHA, e-PDGF, e-TGFB, e-NGF	Homeodomain-like	3	n-SANT, n-HOX, n-PAX	DEATH domain	3	s-DED, s-DEATH, s-CA
Periplasmic binding protein-like II	3	e-PBPb, e-PBPc, e-TR_FER	Protein tyrosine phosphatase I-like	3	s-PTPc, s-DSPc, o-RHOD	EGF/Laminin	3	e-EGF_CA, e-FOLN, e-EGF_Lam
GTPase activation domain (GAP)	2	s-RasGAP, s-RhoGAP	λ repressor-like DNA binding dom.	2	n-POU, n-HTH_LACI	Actin de-polarizing proteins	2	e-GEL, s-ADF
ER Hand-like	2	s-EH, s-Efh	SAM/pointed dom.	2	o-SAM_PNT, s-SAM	ATPase dom. HSP90/ topoisomerase II/his kinase	2	s-HATPase_c, n-TOP2c
5'-3' exonuclease, C-term dom.	2	n-HhH2, n-HhH1	TNF-like	2	e-C1Q, e-TNF	IGF binding dom.	2	e-IB, e-FU
ConA-like lectins	2	e-PTX, e-GLECT, e-LamG	Cold shock protein	2	n-CSP, o-S1	Ser/Thr phosphatase 2C	2	s-PP2Cc, o-PP2C_SIG
C-type lectin	2	e-LINK, e-CLECT	Kringle-like	2	e-KR, e-FN2	Ras-binding dom.	2	s-RBD, s-RA
DNA binding dom	2	n-MBD, n-AP2	Glucocorticoid receptor-like	2	n-Znf_GATA, n-Znf_C4	β trefoil cytokines	2	e-IL1, e-FGF

into a single SMART domain. Considering the P values corrected for multiple observations, 52 still showed $MP_{3D} \leq 5 \times 10^{-3}$. At least one significant link was found for 20 out of the 30 SCOP superfamilies assigned to more than one SMART domain. It is important to emphasize that many of the similarities are not easily found by sequence methods, with evolutionary relationships only inferred once 3D structures are available.

The SCOP P-loop containing nucleotide triphosphate hydrolases (e.g., Gay and Walker 1983; Saraste et al. 1990) are a large superfamily containing 11 SMART domains. Similarities between ATP/GTP-binding and other motifs provided significant links for 38 pairs out of 55. This is despite the fact that proteins within this superfamily frequently show topological differences, such as variations in β -strand order and direction (Murzin et al. 1995).

We also linked several ancient DNA-binding protein families. For example, the nucleosome core histone superfamily in SCOP contains at least five different SMART domains, and six of the 10 possible pairings were found to be significant. Similarly, the winged helix DNA-binding proteins, characterized by a 3-helix DNA-binding bundle and a small β -sheet (wing), contains six SMART domains and 15 potential pairs, five of which could be linked with confidence.

The method also links the tumor necrosis factor (TNF) and collagen-1q (C1Q) domains from SMART. This similarity was first reported after the structure of ACRP30, a homolog of C1Q domains, was determined by Shapiro and Scherer (1998), who inferred an evolutionary relationship based on similarities in key motifs and in the trimeric structures. Here, the sequence of the low P_{3D} value links human CD40 ligand (PDB code 1aly; a distant TNF homolog) and human C1q B (C1QB_HMAN). The alignment of TNFs with C1qs shows that there are numerous conserved hydrophobic positions as well as conservation of key residues around the trimer interface as has been discussed previously (Shapiro and Scherer 1998). In the absence of functional knowledge for one of these families, a functional similarity might be inferred by this P_{3D} calculation.

Fold links

The situation is different for fold-level similarities, where the SCOP database does not consider the pairs necessarily to be homologous. Here, 35 of a possible 120 pairings had $P_{3D} \leq 5 \times 10^{-3}$, five had $P_{3D} \leq 10^{-5}$, and none had $P_{3D} \leq 10^{-10}$. However, six out of the 11 different SCOP folds assigned to more than one SMART domain contain at least one significant pair, but only 10 links remain significant at when the MP_{3D} is calculated. Figure 1 shows the SMART domains linked at fold level.

We found significant links, at either superfamily or fold level, for all the possible pairs within the DNA/RNA-binding 3-helical bundle fold, thus all nine SMART domains adopting this fold could be linked together. The most significant links (Fig. 1) join the Paired Box (PAX), Arsenical Resistance Operon Repressor, helix-turn-helix (HTH_ARSR), SW13 ADA2 N-CoR TFIIB (SANT), homeodomain (HOX), helix-turn-helix and cAMP regulatory protein (HTH_CRP) all with $P_{3D} \leq 10^{-4}$ and $MP_{3D} \leq 5 \times 10^{-3}$. This suggests that they may share a common origin, like other proteins adopting this fold.

The method found several links between proteins adopting an Ig-like fold. Common ancestors have been proposed for many of these, particularly between the Ig and fibronectin domains (e.g., Bazan 1990). The most significant link is that between the Cadherin (CA) and polycystic kidney disease (PKD) domains (P_{3D} -value = 3.9×10^{-5} , MP_{3D} = 1.1×10^{-3}). Both of these domains are extracellular and both are thought to be involved in cell-cell contacts (Hughes

Fig. 1. Significant links ($P_{3D} \leq 5 \times 10^{-3}$) at fold level between SMART domains identified by the method discussed in the text. Thick continuous lines indicate $MP_{3D} \leq 5 \times 10^{-3}$ and dashed lines $MP_{3D} > 5 \times 10^{-3}$.

et al. 1995; Shapiro et al. 1995), thus the P_{3D} -value could be indicating a common ancestor that also is associated with a common function.

Within the OB fold (Murzin 1993c), the method found links between two SMART domains in the superfamily containing ribosomal protein S1-like RNA-binding domain (S1) and cold-shock proteins (CSP) and the superfamily containing staphylococcal nuclease (SNC). MP_{3D} between these families were also low ($\leq 5.7 \times 10^{-3}$). Figure 2 shows a superimposition and structure-based sequence alignment of the known structures and the sequences providing the best link between these two families. For SNC, the location of the nucleotide-binding site is known (the location of a bound nucleotide analog inhibitor thymidine 3',5'-bisphosphate is shown in Fig. 2). Several of the residues within the nucleases that are in contact with the nucleotide also are found in S1 and comprise a D[RK]xxGR motif that shows good (but not perfect) conservation in both families. These residues occur in an unusual loop within the most C-terminal β -hairpin found in the OB fold. In addition, an aspartate from a

Fig. 2. (a) Molscript (Kraulis 1991) figures showing *Staphylococcus aureus* nuclease (staphylococcal nuclease; left; PDB code 1kdc) and *Escherichia coli* S1 RNA-binding protein (right; 1sro) in a similar orientation. Structural equivalent regions (identified by the method of Russell and Barton 1992) are labeled with arrows (β-strands) or ribbons (α-helices) or coils, with nonequivalent regions shown as Cα trace. Residues common to both structures are shown in ball-and-stick format. The coloring scheme moves through the spectrum from blue to red from N- to C-terminus. Linkage details: RMSD = 2.0 Å in 38 Cα atoms; 13 identities in 32 equivalent residues; P_{3D} -value = 4.7×10^{-6} MP_{3D} = 8.2×10^{-5} . (b) Alscript (Barton 1993) figure showing the structural alignment of the two structures in a with secondary structures (below). The best linking sequence is also shown for the nucleases (NUC_SFLX; *S. flexneri* nuclease); the best link was between this sequence and the *E. coli* S1 RNA-binding domain. Positions within the alignment showing conservation of residue character are colored as follows: yellow background, conserved hydrophobic; blue background, conserved small; red text, conserved polar. Identical residues are boxed. Secondary structures are shown as arrows (β-strands) and cylinders (α-helices) below the alignment and colored as for a.

different region in both structures that is in contact with a bound calcium atom within SNe also shows good conservation within each of these diverse families.

Links were found within the β-grasp fold between ubiquitin (UBQ) and two members of the Ras-binding superfamily: the Ras association domain (RA) and the Raf-like Ras-binding domain (RBD). Ubiquitin proteolysis plays a crucial role in protein degradation that controls the timed destruction of cellular regulatory proteins, including cyclins, tumor suppressor p53, or transcription factors. RA and RBD families bind Ras-like proteins and are involved in signaling processes. The related functions (signaling and cell-cycle control) of these superfamilies together with the low P_{3D} -values may indicate that they share a common ancestor.

The procedure also found low P_{3D} -values linking several members of the cysteine-knot fold, including the insulin growth-factor-binding protein (IGF-BP; e.g., Clemmons 1993), chitin-binding domains (ChtBD), and the epidermal growth

factor. Although functional similarities have been reported for the EGF family (Mas et al. 1998; Blanco-Aparicio et al. 1998), there is no obvious similarity in the functions of these domains, and the low P_{3D} -values may be an artefact of being small cysteine-rich domains. The similarities involve only a few structurally equivalent residues (<10) and the identities consist mostly of the cysteines forming the disulfides that define the fold.

The method also found linkages within the β-trefoil fold. Previously, analysis of 3D structures and sequences showed that the interleukin-1 (IL1), fibroblast growth factor (FGF), Ricin-type (RICIN), and the soybean trypsin inhibitor (Kunitz) families share a degree of sequential and functional similarities (Ponting and Russell 2000).

Linking Pfam families via structure

A total of 711 out of 2008 Pfam families could be assigned to one or more SCOP domains (117,017 out of 178,110

sequences). As for SMART, there were some disagreements between domain definitions. A total of 61 Pfam domains could be aligned to multiple SCOP domains in nonoverlapping regions. Table 2 shows the 113 SCOP superfamilies and 26 SCOP folds that could be matched to more than one Pfam family. Associated with this are 954 potential pairings within superfamilies and 1406 potential pairings within folds. A table showing the results for all the significant pairings at superfamily and fold level can be found at http://www.embl-heidelberg.de/~aloy/struct_align.

Superfamily links

We found $P_{3D} \leq 5 \times 10^{-3}$ for 231 out of the 954 potential pairs within superfamilies. Eighty-five of the pairings had $P_{3D} \leq 10^{-5}$ and four had $P_{3D} \leq 10^{-10}$. The last were clear cases of sequence-detectable homologs residing in separate domains within Pfam. In addition, 157 of the significant links showed corrected $MP_{3D} \leq 5 \times 10^{-3}$. Structural alignment was able to identify at least one significant pair for 81 out of the 113 SCOP superfamilies assigned to more than one Pfam family.

The results are fully consistent with those obtained with SMART (above). For example, the P-loop nucleotide triphosphate hydrolases superfamily also contains the most different Pfam families. However, the wider coverage of Pfam means that many more relationships were found, several of which are discussed below.

For example, within the NAD(P)-binding Rossmann-fold domains we found significant links for 24 of the 136 potential pairs and a total of 12 out of the 17 Pfam families belonging to this superfamily. This NAD/NADP-binding domain is present in a large number of families, including dehydrogenases, synthetases, reductases, and methylases, and the difficulty of detecting similarities using only sequence analysis has been discussed recently (Kunin et al. 2001).

We also found several links within the four helical cytokine superfamily. This superfamily contains, at least, one pair with a significant P_{3D} -value for eight out of its 10 Pfam families. The relationship between these proteins is well known, though the similarities frequently are not detectable by sequence comparison alone.

Fold links

The situation is again different for families residing in the same fold but in different superfamilies. With the same cutoff ($P_{3D} \leq 5 \times 10^{-3}$) we found 152 significant pairs out of the potential 1406. Eleven pairs had $P_{3D} \leq 10^{-5}$ and only one had $P_{3D} \leq 10^{-10}$. However, we found at least one significant link for 16 out of the 26 different SCOP folds containing two or more Pfam families. When the family diversity score is considered, 51 out of the 152 significant links still show $MP_{3D} \leq 5 \times 10^{-3}$. Figure 3 shows the Pfam families linked at fold level. As for the superfamily links, all overlapping links

found in SMART domains also were present involving the Pfam equivalents, and again, because of the larger coverage of sequence space by Pfam, additional links were between families not present in SMART, many of which are discussed below.

The $(\beta\alpha)_8$ (TIM) barrel fold is the largest, comprising a total of 38 different Pfam families. Of the 703 potential pairings, 84 showed to be significant. These pairings effectively mean that 33 out of the 38 families assigned to this fold can be linked. Moreover, the five families that were not linked to the others (Glyco_hydro_1, Glyco_hydro_14, PI-PLC-X, ALAD, and AP_endonucleas2) belong to the same SCOP superfamily as other linked families, meaning that there are homologous links within SCOP between these and the others. The results are consistent with the work of Copley and Bork (2000) who found that all but one enzyme family adopting a TIM-barrel fold can be linked to the others using sequence search methods like PSI-blast (Altschul et al. 1990), suggesting a common ancestor for this fold that performs many different biochemical functions. They were unable to link the dihydroorotase family (proposed to adopt a TIM barrel by Holm and Sander 1997). No structures of the dihydroorotase family are known, meaning that our study could not suggest an evolutionary link between this family and any of the others in this fold.

Six out of the eight Pfam families within the flavodoxin-like fold could be linked into a single group, with the most significant link being that between Response regulator receiver domain (response_reg) and flavodoxin families (P_{3D} -value = 3.85×10^{-5} ; $MP_{3D} = 1.1 \times 10^{-5}$). The Response regulator family includes CheY (Volz 1993), the receiver domain among two-component signal transduction systems. In response to a specific stimulus, these proteins are phosphorylated leading to a conformational change that is detected by an effector domain. The flavodoxins (Vervoort et al. 1994) are small proteins that bind FMN and serve as redox centers and electron transport proteins. Two of the other three-linked families (DHquinase_II and Ligase-CoA) are ATP/GTP-binding proteins. Our results indicate that the function of the five families may have diverged from common nucleotide-binding site. Though the CheY family does not bind nucleotides, structural and functional analogies between the CheY-like receiver domains and small GTP-binding proteins have been noted and a common ancestry has been proposed (Artymiuk et al. 1990; Lukat et al. 1991).

We found several links between superfamilies within the SCOP ferredoxin-like fold, some of which appeared to be associated with functional similarities (see below). This fold comprises a repeat of a split $\alpha\beta\alpha$ motif that forms an antiparallel β -sheet flanked on one side by two α -helices. It is one of the most populated in SCOP, with 31 different superfamilies, and also performs a large number of different functions. In total, 12 out of the 17 Pfam families assigned to this fold were linked with significant P_{3D} -values.

Table 2. Potential pairings for Pfam families at fold and superfamily level

SCOP fold	N	Pfam family	SCOP fold	N	Pfam family	SCOP fold	N	Pfam family
TIM beta/ alpha-barrel	35	TGT, MM_CoA_mutase, UPF0001, PI-PLC-Y, DAHP_synth_1, F_bp_aldolase, Glyco_hydro_14, PI-PLC-X, IGPS, oxidored_FMN, QRPTase, enolase, PRAI, Transaldolase, DHPS, chitinase_2, DHDPs,	DNA/ RNA-binding 3-helical bundle	15	HTH_5, Methyltransf_1, HTH_2, Fork_head, myb_DNA-binding, trans_reg_C, HSF_DNA-bind, linker_histone, LexA_DNA_bind, PAX, IRF, recombinase, Ets, homeobox	Ferredoxin-like	15	Ribosomal_S6, HPPK, Thr_dehydrat_C, Transl; FTR, guanylate_cyc, EF E2_C, PyrI, HMA, NDK Ribosomal_L30, rrm, fe
			Immuno- globulin-like beta-sandwich	14	MSP_domain, Transglutamin_C, PKD, arrestin, Filamin, pili_assembly, Neocarzinostat, oxidored_molyb, Tissue_fac, TIG, sodcu, fn3, RHD, ig	Knottins (small inhibitors, toxins, lectins)	12	wap, Colipase, Hirudin-like, Gamma-thionin, CI squash, Bowman-Birk_1, chitin_binding, toxin_3, toxin_2, EGF
			OB-fold	10	Ribosomal_S17, SNase, Pyrophosphatase, Enterotoxin_B, SSB, TIMP, CSD, S1, Stap_toxin, eIF-5a	Common fold of diphtheria toxin/ transcription factors/ cytochrome f	8	STAT, Fimbrial, CBD_2, Cohesin, T-box, Diphtheria_tox, P53, RE
Ribonuclease H-like motif	7	Creatinase_N, actin, HSP70, maseH, hexokinase, DNA_pol_B, integrase	beta-Grasp (ubiquitin-like)	7	IgG_binding_B, IF3, RA, ubiquitin, Stap_Strp_toxin, fer2, gln-synt	Phosphorylase/ hydrolase-like	6	Pept_tRNA_hydro, Peptidase_M17, Mtap_P, Peptidase_M20, Peptidase C15, Zn carb
SH3-like barrel	5	CcdB, Peptidase_S24, SH3, eIF-5a, integrasae	Rubredoxin-like	5	Desulfoferredox, COX5B, TFIIS, rubredoxin, PyrI	Flavodoxin-like	5	DHquinase_II, ligase-Cc, Cutinase, flavodoxin, response reg
Barrel-sandwich hybrid	4	GCV_H, PTS_EIIA_1, biotin_lipoyl, CPSase_L_chain	beta-Trefoil	4	interleukin-1, Ricin_B_lectin, FGF, Kunitz_legume	Double-stranded beta-helix	4	PMI_typeI, Seedstore_7 Fe_Asc_oxidored, cNMP binding
Prealbumin-like	4	VHL, Dioxygenase, Transthyretin, CBD 4	Toxins' membrane translocation domains	4	Bcl-2, endotoxin, Diphtheria tox, Colicin	Long alpha-hairpin	3	DnaJ, GreA_GreB, sodf
alpha-alpha superhelix	3	I4-3-3, PPTA, Sec7	Acyl carrier protein-like	3	Colicin_Pyocin, pp-binding, gag p24	ATP-grasp	3	CPSase_sm_chain, cpn60 TCP1, PEP-utiliz
Four-helical up-and-down bundle	3	Hemerythrin, TMV_coat, Cytochrome C 2						

(continued)

One of the more intriguing similarities within the ferredoxin-like fold is that between the RNA recognition motif (rrm) and heavy-metal-associated (HMA) domain with a

P_{3D} -value = 4.7×10^{-5} MP_{3D} = 9.0×10^{-4} . The RNA recognition motifs are found in a variety of proteins, including proteins implicated in regulation of alternative splicing, and

Table 2. Continued

SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family
P-loop containing nucleotide triphosphate hydrolases	16	PRK, TK_herpes, Guanylate_kin, Sulfotransfer, adenylatekinase, recA, GTP_EFTU, kinesin, G-alpha, arf, ras, fer4_NifH, SRP54, myosin_head, UvrD-helicase, IQ	NAD(P)-binding Rossmann-fold domains	15	THF_DHG_CYH, IlvC, Semialdehyde_dh, GFO_IDH_MocA, adh_zinc, GLFV_dehydrog, DapB, gpdh, Idh, 3HCDH, malic, adh_short, Epimerase, 3Beta_HSD, adh short C2	4-helical cytokines	10	LIF_OSM, IL4, GM-CS, EPO_TPO, IL2, interfer hormone, IFN-gamma, I IL-6
Periplasmic binding protein-like II	8	PstS, lig_chan, SBP_bac_5, SBP_bac_3, Sulphate_bind, Porphobil_deam, transferrin, SBP_bacterial_1	ConA-like lectins/ glucanases	8	Glyco_hydro_12, Glyco_hydro_16, pentaxin, Glyco_hydro_7, Gal-bind_lectin, Glyco_hydro_11, lectin_legB, lectin_legA	Glycosyltrans- ferases	8	Glyco_hydro_14, chitina Glyco_hydro_10, Glyco_hydro_17, Glyco_hydro_1, cellulas alpha-amylase, Glyco_h
PLP-dependent transferases	8	OKR_DC_1, Cys_Met_Meta_PP, aminotran_1, aminotran_2, SHMT, aminotran_3, aminotran 5, Beta elim lyase	Membrane all-alpha	7	COX3, MscL, photoRC, COX2, COX1, cytochrome_b_C, cytochrome_b_N	Winged helix DNA-binding domain	7	HTH_5, Fork_head, trans_reg_C, HSF_DNA linker_histone, LexA_DNA_bind, IRF,
alpha/beta- Hydrolases	7	serine_carbpept, Lipase_3, DLH, Peptidase_S9, lipase, COesterase, abhydrolase	Viral coat and capsid proteins	7	Bromo_CP, Peptidase_A6, Polyoma_coat, Parvo_coat, Viral coat, Tymo_coat, rhv	Immunoglobulin	6	arrestin, Filamin, oxidored_molyb, TIG, R
FAD/NAD(P)- binding domain	6	CHOD, GMC_oxred, pyr_redox, GDI, FAD_binding_3, Monooxygenase	Thioredoxin-like	6	Calsequestrin, GSHPx, DSBA, thioered, glutaredoxin, GST	S-adenosyl-L- methionine- dependent methyltrans- ferases	6	PARP_regulatory, N6_N4_Mtase, CheR, Methyltransf_3, DNA methylase, RrnaA
Aldolase	5	DAHP_synth_1, F_bP_aldolase, Transaldolase, DHDPS, glycolytic enzy	Metalloproteases ("zincins"), catalytic domain	5	Peptidase_M27, Peptidase_M8, Reprolysin, Peptidase_M4, Peptidase_M10	Homeo- domain-like	5	HTH_2, myb_DNA-bin, PAX, recombinase, hom
Membrane all-alpha	5	COX2, Cu-oxidase, copper-bind, COX1, COX3	Nucleotidylyl transferase	5	tRNA-synt_id, tRNA-synt_1b, tRNA-synt_1c, Cytidylyltransf, tRNA-synt 1	Nucleic acid-binding proteins	5	Ribosomal_S17, SSB, CS1, eIF-5a
Cystine knot cytokines	5	hormone6, TGF-beta, NGF, PDGF, Cys knot	Cyclin-like	4	RB_A, RB_B, cyclin, transcript fac2	Lysozyme-like	4	Phage_lysozyme, Glyco hydro 46, SLT, 1
Glutathione synthetase ATP-binding domain-like	4	GARS, Dala_Dala_ligas, CPSase_L_chain, PPK N term	DHS-like NAD/ FAD-binding domain	4	DS, ETF_alpha, TPP_enzymes, ETF_beta	ATPase domain of HSP9	4	signal, HSP90, DNA_to DNA_mis_repair
Galactose- binding domain-like	4	XRCC1_N, EPH_Ibd, Glyco hydro 2, endotoxin	p53-like transcription factors	4	STAT, T-box, P53, RHD	DNA/RNA polymerases	4	RNA_dep_RNA_pol, rv DNA pol B, DNA pol

(continued)

Table 2. Continued

SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family
PH domain-like	4	Ran_BP1, PID, WH1, PH	Actin-like ATPase domain	3	actin, HSP70, hexokinase	(Phosphotyrosine protein) phosphatases II	3	Rhodanese, DSPc, Y phosphatase
Ferritin-like	3	FA_desaturase, ribonuc_red, ferritin	DNA/RNA polymerases	3	maseH, DNA_pol_B, integrase	NAD(P)-binding Rossmann-fold domains	3	THF_DHG_CYH, malic GLFV dehydrog
Zn-dependent exopeptidases	3	Peptidase_M17, Peptidase_M20, Zn_carbOpept	lambda repressor-like DNA-binding domains	3	lacI, HTH_3, pou	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin	3	Seedstore_2S, tryp_alpha_amyl, LTP
DEATH domain	3	DED, CARD, death	P-loop containing nucleotide triphosphate hydrolases	3	G-alpha, arf, ras	Thiolase-like	3	ketoacyl-synt, thiolase, Chal_stil_synt
Six-hairpin glycosyl-transferases	3	Glyco_hydro_8, Glyco_hydro_15, Glyco_hydro_9	N-terminal nucleophile aminohydrolases (Ntn hydrolases)	3	Penicil_amidase, proteasome, GATase_2	FAD/NAD(P)-binding domain	3	CHOD, GDI, GMC_oxr
MHC antigen-recognition domain	3	MHC_I, MHC_II_alpha, MHC_II_beta	Cysteine proteinases	3	Peptidase_C5, UCH, Peptidase_C1	FKBP-like	3	Rotamase, GreA_GreB,
ADP-ribosylation	3	Enterotoxin_A, Diphtheria_tox, PARP	Tetrahydrobiopterin biosynthesis enzymes-like	3	GTP_cyclohydrol, PTPS, Uricase	Glyceraldehyde-3-phosphatase dehydrogenase-like, C-terminal domain	3	G6PD, Semialdehyde_dh
Class II aaRS and biotin synthetases	3	tRNA-synt_2b, tRNA-synt_2d, tRNA-synt 2	beta-Lactamase/D-ala carboxypeptidase	3	Transpeptidase, Peptidase_S11, beta-lactamase	FMN-linked oxidoreductases	3	oxidored_FMN, FMN_c Glu synthase
ETFP adenine nucleotide-binding domain-like	3	Usp, ETF_alpha, ETF_beta	Glutathione synthetase ATP-binding domain-like	3	CPSase_L_chain, Dala_Dala_ligas, GARS	Thiamin diphosphate-binding fold (THDP-binding)	3	PRO_N, TPP_enzymes, transketolase
Scorpion toxin-like	3	Gamma-thionin, toxin_3, toxin_2	Glucocorticoid receptor-like (DNA-binding domain)	3	zf-C4, LIM, GATA	Defensin-like	3	Defensin_beta, defensin toxin_4
Bet v1-like	2	Ring hydroxyl A, Bet v 1	S15/NS	2	Flu NS1, Ribosomal S15	Phosphatase/sulphatase	2	alk phosphatase, Sulfata
Histone-fold	2	CBFD_NFYB_HMF, histone	CoA-dependent acyltransferases	2	2-oxoacid_dh, CAT	Phosphoglycerate mutase-like	2	acid_phosphat, PGAM
Terpenoid synthases	2	polyprenyl_synt, Terpene synth	GTPase activation domain, GAP	2	RhoGAP, RasGAP	Di-copper centre-containing domain	2	hemocyanin, tyrosinase
NAD(P)-binding Rossmann-fold domains	2	livC, 3HCDH	PapD-like	2	MSP_domain, pili_assembly	Ribosomal protein S	2	Ribonuclease_P, Ribosomal S5

(continued)

Table 2. *Continued*

SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family	SCOP superfamily	N	Pfam family
Rubredoxin-like	2	COX5B, rubredoxin	Arginase/deacetylase	2	Hist deacetyl, arginase	Snake toxin-like	2	UPAR LY6, toxin
DNA clamp	2	DNA pol3 beta, PCNA	Fibronectin type III	2	Tissue fac, fn3	HIT-like	2	GalP UDP tranf, HIT
Segmented RNA-genome viruses' proteins	2	Hemagglutinin, Orbi_VP7	Bacterial enterotoxins	2	Enterotoxin_B, Stap_Strp_toxin	Ferredoxin reductase-like, FAD-linked (N-terminal) domain	2	Pyridox_oxidase, Cyt_reductase
C-type lectin-like	2	lectin c, Xlink	SAICAR synthase-like	2	SAICAR synt, PIP5K	GroES-like	2	cpn10, adh zinc
Carbohydrate-binding domain	2	CBD_2, Cohesin	2Fe-2S ferredoxin-related	2	fer2, fer4	DNA breaking-rejoining enzymes	2	Topoisomerase_I, Phage integrase
TNF-like	2	TNF, Clq	Cytokine	2	interleukin-1, FGF	Acid proteases	2	rvp, asp
dsRNA-binding domain-like	2	dsrm, Ribosomaal_S5	Trypsin-like serine proteases	2	Flavi_helicase, trypsin	Enolase C-terminal domain-like	2	enolase, MR_MLE
Pectin lyase-like	2	pec lyase, Glyco hydro 28	Single hybrid motif	2	GCV H, biotin lipoyl	Tautomerase/MIF	2	Tautomerase, MIF
Actin depolymerizing proteins	2	cofilin_ADF, Gelsolin	Ribulose-phosphate binding barrel	2	IGPS, PRAI	Prokaryotic type I DNA topoisomerase	2	Toprim, Topoisom_bac
Phosphatidylinositol-specific phospholipase C (PI-PLC)	2	PI-PLC-Y, PI-PLC-X	Inosine monophosphate dehydrogenase (IMPDH)	3	aldo_ket_red, IMPDH_C, IMPDH_N	Porins	2	Gram-ve_porins, TonB_
ClpP/crotonase	2	CLP protease, ECH	Kringle-like	2	fn2, kringle	Sugar phosphatases	2	inositol P, FBPase
Caspase-like	2	ICE_p20, ICE_p10	Globin-like	2	Phycobilisome, globin			

function to bind single-stranded RNA. The heavy-metal-associated proteins are implicated in the regulation of cytoplasmic metal concentration, and although they are known to be ATP dependent, the location of the ATP-binding site has not yet been determined. Figure 4 shows two representative structures from these superfamilies and the associated alignment of the best link. The conserved residues are found in the nucleotide-binding site of the RNA recognition protein, which suggests that it also may correspond to the ATP-binding site for the HMA proteins. Interestingly, the location of this binding site corresponds with the ferredoxin-like fold “supersite” proposed by Russell et al. (1998). This fold shows a tendency to bind a diversity of different ligands at a common location, which is on the side of the β -sheet not flanked by α -helices. It is worth noting that both rrm and HMA domains occur as tandem repeats, and an oligomeric structure has been proposed for the rrm domains to aid the binding of single-stranded RNA.

We detected links between different superfamilies within the barrel-sandwich hybrid fold, the best being that between

the glycine cleavage H protein (GCV_H) and phosphoenolpyruvate-dependent sugar phosphotransferase system, EIIA 1 (PTS_EIIA_1). The core of this fold comprises seven or eight strands in two β -sheets, and structural representatives from these two families can be superimposed such that seven β strands are equivalent (35 C α atoms superimpose with RMS = 2.0 Å and percent sequence identity of 4%). However, the initial structural alignment identified a shorter region of high structure and local sequence similarity (Fig. 5), which identifies a repeat found at the N-terminus of transit peptide H protein of the Gly cleavage system (20–65 of PDB 1hpca) and toward the C-terminus of glucose permease domain IIA (75–120 of 1gpr). Investigating the similarity further reveals that these families are better superimposed if one considers a circular permutation (45 C α s; RMS = 1.3; %I = 33%) as is shown in Figure 5. This similarity includes two unusual loop structures containing several identities that are absent from the unpermuted alignment. This fold is composed of repeats of three β strands, of which the Gly cleavage system protein contains two and the

Fig. 3. Significant links at fold level between Pfam domains identified by the method discussed in the text. Lines are drawn as for Figure 1.

Fig. 4 (a) Molscript (Kraulis 1991) figures showing the Poly(A)-binding protein (*left*; 1cvj, chain F) and the Copper transporter ATPase (*right*; 1aw0) in a similar orientation. Details for the figures are as for Figure 2. Linkage details: RMSD = 1.8 Å in 39 C α atoms; seven identities in 12 equivalent residues; P_{3D} -value = 4.7×10^{-5} MP $_{3D}$ = 9.1×10^{-4} . (b) Alscript (Barton 1993) figure showing the structural alignment of the two structures in *a* with secondary structures (below). The best linking sequences are also shown (COPA_ENTHR and RO33_NICSY). Conserved positions and secondary structures are shown as described in Figure 2.

permase contains four (bottom of Fig. 5). The permuted structure may have been the result of a partial duplication of a longer protein as has been described previously (Ponting and Russell 1998). The repeat structure within the Gly cleavage system family has been discussed recently (Anantharaman et al. 2001).

Hypothetical protein ybl036c: Output from a Structural Genomics initiative

A result from a Structural Genomics initiative provides an interesting test for the method. The three dimensional structure of a member of the UPF0001 Pfam family (hypothetical protein ybl036c; PDB codes 1ct5, 1b54) was solved recently as part of the BNL Human Proteome Project (<http://proteome.bnl.gov>), and was found to adopt a $(\beta\alpha)_8$ -TIM-barrel structure. The 3D structure also showed the protein to bind pyridoxal-phosphate (PLP).

The structure shows a striking similarity to members of the alanine racemase family (Ala_racemase). This similarity is detectable by sequence comparison alone (e.g., Psi-blast with an E-value of 7×10^{-31}) and also shows a very significant pairwise P_{3D} (1×10^{-44}), meaning that they are placed within the same SCOP superfamily. Ignoring this obvious similarity, the method also finds significant links to five

other families adopting the $(\beta\alpha)_8$ -(TIM)-barrel fold: aldo/keto reductases (aldo_ket_red), indol-3-glycerol phosphate synthase (IGPS), dihydrodipicolinate synthetase (DHDPS), dihydropteroate synthase (DHPS), and fructose-bisphosphate aldolase class II (F_bp_aldolase). As discussed above, this fold performs many different biochemical functions, making attempts to assign function from structure difficult. Moreover, there also is growing evidence that many of the $(\beta\alpha)_8$ -barrels have evolved from a common ancestor (see above; Copley and Bork 2000).

Inspection of the alignments and superimpositions with the Pfam families having the most significant P_{3D} -values suggests that probably the best functional match is that with the second lowest value: the IGPS Pfam family, the second best score (P_{3D} -value = 1.94×10^{-3}), rather than the aldolase/ketolase reductases (aldo_ket_red; P_{3D} -value = 4.84×10^{-4}). Inspection showed the length of the structurally equivalent regions to be longer when comparing UPF0001 and IGPS, in addition to overlap many of the residues involved in function in IGPS and PRAI (Wilmanns et al. 1992), some of which are fully conserved in the merged alignments (K55 and G236). IGPS, together with the PRAI and Trp_synthase families, belong to the Ribulose phosphate-binding barrel superfamily. Most of the enzymes in

Fig. 5. (a) Molscript (Kraulis 1991) figures showing the N-terminus of transit peptide H protein of the Gly cleavage system (*left*; 1hpc, chain a) and the C-terminus of glucose permease domain IIA (*right*; 1gpr) in a similar orientation. Details for the figures are as for Figure 2. Linkage details: RMSD = 1.4 Å in 45 C_α atoms; 15 identities in 45 equivalent residues; P_{3D}-value = 1.3×10⁻⁴ MP_{3D} = 2.5×10⁻³. (b) Alscript (Barton 1993) figure showing the structural alignment of the two structures in *a* with secondary structures (below). The best linking sequences are also shown (O59049 and PTBA_ERWCH). Conserved positions and secondary structures are shown as described in Figure 2. The numbers within the alignment denote the start and end of the aligned segments (note, in particular, that 1hpc is permuted relative to 1gpr). (c) Figure showing the similarity in a topology diagram. β-strands are denoted as triangles; α-helices as circles, colored in an analogous fashion to *a*.

this superfamily are involved in amino-acid synthesis pathways, and some also bind PLP. Moreover, three of the other Pfam families that could be linked to UPF0001 are involved in tryptophan or lysine synthesis. All these results indicate that the 26 members of the UPF0001 Pfam family could play a role in amino-acid synthesis.

Implications for structural genomics

The results above show that the method often is able to identify superfamily relationships, and possible functional similarities between domains that have been linked by a structure similarity. The accuracy of identifying such relationships is important for Structural Genomics projects that target domains of unknown structure from databases such as SMART or Pfam. Using P_{3D} ≤ 5×10⁻³ to predict superfamily relationships gives a sensitivity (the percentage of correct relationships identified) of 45% for SMART domains and 24% for those from Pfam. For the more stringent MP_{3D} ≤ 5×10⁻³ the values are 27% and 16%, respectively. We suspect that the lower sensitivity for the Pfam domains is because the alignments are less diverse than their SMART counterparts, meaning that one is less likely to find a significant link when considering all sequences.

Quoting an associated specificity (the percentage of incorrect relationships identified) is problematic, as there is no

definitive set of “false positives.” Ideally false positives would consist of proteins with different folds, but for these it would not be possible to build meaningful structure-based alignments. An alternative is to use pairs of proteins in the same fold that are definitely “not” in the same superfamily (i.e., fold level links). Though many folds in SCOP contain multiple superfamilies, new evidence often emerges (i.e., like the examples above) that permits them to be merged together. Thus calculating specificity in this way will give an underestimate of the correct value. With this caveat in mind, the specificities for P_{3D} ≤ 5×10⁻³ are 80% for SMART domains and 89% for Pfam, and those for MP_{3D} ≤ 5×10⁻³ are 95% and 96%, respectively.

These values of sensitivity and specificity would be applicable for situations where a new structure has been determined for a domain from SMART or Pfam and adopts a fold known previously. We anticipate that our method will often be able to identify many superfamily relationships and thus place a new structure into the correct evolutionary and often functional context.

Discussion

We have demonstrated how a merger of protein structure and sequence databases can suggest likely evolutionary links between protein domain families. We have also iden-

tified potentially new homologous relationships that may be associated with similarities in molecular function.

This study has gone some way toward quantifying how Structural Genomics projects will gradually link sequence families together, ultimately providing 3D structure and additional functional information for all protein families with at least one member amenable to structure determination by X-ray crystallography or nuclear magnetic resonance. A protocol like that described here will permit evolutionary and functional similarities to be uncovered automatically as the number of known structures and sequences continues to increase.

It is intriguing that current sensitive sequence searching methods apparently fail to detect some similarities that are quite clearly associated with a degree of sequence conservation (e.g., TNF/C1Q). It may be that sequence profiles are too specific to one family to detect more distantly related sequences, even when key sequence motifs are conserved. Another possible explanation comes from inspection of aligned segment lengths. We found that aligned segments (i.e., those not containing a gap in any sequence) are typically shorter within the SCOP-linked alignments than in alignments derived only by a comparison of sequence (results not shown). This may mean that the model for aligned segments and gaps currently used by sequence comparison methods is too strict to permit alignments such as those obtained by structure comparison.

The P-value first described by Murzin (1993b) attempts to assess the likelihood that a pair of proteins, aligned based on their three-dimensional structures, will have a certain degree of sequence similarity. The prior probability of amino acid identity is based on the abundance of the amino acids, and accommodates the assumption that certain features of common protein structures, such as burial in the hydrophobic core or surface exposure, will increase the chances of amino-acid identities. What it does not take into account is the possibility that certain folds will have strict requirements for particular amino acids at certain positions, which could well be the result of convergent evolution to a stable fold. It has been argued that this is the case for the β -trefoils (including the FGFs and IL-1s; Murzin 1993a; Ponting and Russell 2000), and could well be the case for other folds. It is impossible to discern such occurrences at present, thus this possibility should be remembered when considering the links proposed.

Structural Genomics initiatives provide structures for proteins that often are of unknown function. In the absence of further experiments, the ability to place a new structure in the correct evolutionary context is currently the best method for predicting details regarding molecular function. Methods such as that described here and elsewhere (Copley and Bork 2000; Todd et al. 2001; Landgraf et al. 2001; Aloy et al. 2001) will thus be of growing importance in this new age of 3D structural annotation.

Materials and methods

Data

We obtained aligned sequence data from the SMART (version 4.1, <http://smart.embl-heidelberg.de/>) and Pfam (release 5.0, <http://www.sanger.ac.uk/Software/Pfam/>) world wide web (WWW) pages. We extracted SCOP classifications from the WWW page (release 1.50, <http://scop.mrc-lmb.cam.ac.uk/scop/>), and converted them into protein sequence data via the STAMP package (Russell and Barton 1992; <http://barton.ebi.ac.uk/manuals/stamp.html>).

Merging SMART/Pfam alignments with SCOP sequences

SMART and Pfam contain cross-references to appropriate PDB identifiers. However, the PDB identifier alone is not sufficient to identify a domain in SCOP, since such identifiers can contain both multiple polypeptide chains and multiple domains. Accordingly, we constructed a hidden Markov model for each SMART and Pfam alignment using the HMMBUILD program from the HMMer package (S.R. Eddy, unpubl.; <http://hmm.wustl.edu>), and used this to search the SCOP database (HMMSEARCH) to provide links between the two databases. SCOP sequences were considered to reside in the SMART or Pfam domain if they had HMMSEARCH E-values $\leq 10^{-3}$ and if the alignment covered 60% of either the SCOP sequence or the SMART/Pfam domain. Once identified, we aligned the sequences for SCOP entries with those from SMART/Pfam using HMMALIGN. This resulted in a set of alignments containing the original sequences from SMART/Pfam in addition to those from SCOP.

Combining SMART/Pfam alignments from the same SCOP classifications

When we found that different SMART/Pfam entries contained domains from the same SCOP fold or superfamily, we merged the alignments via an alignment of structures. We aligned structures using the STAMP package for protein structure alignment and superimposition. All alignments were checked and, if necessary, edited manually to avoid situations where structural alignment was ambiguous, or lead to erroneous results owing to distortions of the structure as a result of bound substrates or poorly determined/missing residues. In two cases, STAMP found several alignments with good scores. These comparisons were manually edited and the best alignment was selected. The structural alignments then were used to merge all associated sequence data. A summary of these linkages can be found in Table 1. For folds/superfamilies containing more than two SMART/Pfam families, we constructed all "pairwise" alignments. This was done to ensure maximum alignment quality.

It is difficult to do a direct comparison between alignments derived by consideration of protein sequence alone with those derived from three-dimensional structure. The main difficulty is that structural alignment methods either do not necessarily give a meaningful alignment of those regions that are different between protein structures, or they do not attempt to align them at all. Accordingly, we processed the structural alignments prior to merging them with sequence alignments. Sequences outside of the structural conserved regions were shortened to the minimum possible length. This mimics what would likely happen during a sequence alignment of the same proteins, assuming that, in the best

possible sequence alignment for the proteins, only the structurally equivalent regions would be accurately aligned.

All alignments are available via the WWW (http://www.embl-heidelberg.de/~aloy/struct_align).

Statistical significance of SMART/Pfam families linked by SCOP structures

Murzin (1993b) proposed a P-value to suggest the likelihood that a “sequence” identity found after “structure”-based alignment could occur by chance (hereafter called P_{3D}). Given n structurally conserved sites (i.e., C_α positions) between two similar protein 3D structures, he suggested that the probability that m of these sites would contain identical amino acids would be:

$$P_{3D} = P(m) \leq \frac{n!}{(m! - m)!} \bar{p}^m (1 - \bar{p})^{n-m} \quad (\text{provided that } m > m_0 + \sigma)$$

where $\bar{p}$ is the mean probability of finding identical residues and structurally equivalent sites, $m_0 = n\bar{p}$ (where the binomial has its maximum), and $\sigma = \sqrt{n\bar{p}(1 - \bar{p})}$ (the half-width of the approximating distribution). All that is required is to approximate the probability that structurally equivalent sites in two similar protein structures will have identical residues by chance. Murzin suggested that the value would be larger than 1/20, probably about 1/15 but certainly smaller than 1/10. These values attempt to account for the tendency for buried residues to be hydrophobic and exposed residues to be hydrophilic. In other words, the probability would be $>1/20$ as buried or exposed sites are more likely to contain a smaller subset of the 20 amino acids (i.e., hydrophobic and polar residues, respectively). This calculation was originally applied to the cystatin-monellin similarity, where an evolutionary relationship was inferred based on a P_{3D} of $\sim 10^{-3}$. For more details, we refer the reader to Murzin (1993b).

Here, we calculated P_{3D} for all pairs of sequences coming from different SMART/Pfam alignments as aligned according to the structures of one or more members of the alignment. We defined structurally equivalent regions by the method of Russell and Barton (1992), and extrapolated these positions to all sequences in the aligned SMART/Pfam families. This calculation assumes that the alignment of sequences is correct and that the known structures for SMART/Pfam families are good models for the remaining sequences. Owing to the high quality of both the sequence alignments within individual SMART/Pfam domains and the structure-based alignments, we do not suspect that the calculation would differ greatly if all proteins were of known structure.

Here, we used the most stringent value of 1/10 meaning that P_{3D} values are higher than if we had used 1/15. We also considered significant links those with $P_{3D} \leq 5 \times 10^{-3}$. This value is an order of magnitude lower than those calculated, and argued to be biologically significant, with a more lenient P_{3D} calculation (1/15) for β -trefoil proteins (Ponting and Russell 2000). Thus, we are confident that all the linkages reported are biologically relevant when considering single pairs of protein structures. Selecting the more lenient value of 1/15 has the effect of identifying more of both fold and superfamily links, which we suspect is lowering the specificity of the approach, as many more links between folds are found that may not be true superfamily relationships. For more details, see Implications for structural genomics in Results, above.

P_{3D} was originally described for the comparison of a single pair of protein structures. Because here we are looking for the mini-

mum value from a (sometimes large) number of pairs of proteins, there is a statistical tendency that means low values are more likely to arise by chance (i.e., akin to the difference between a P-value and E-value in database searches such as BLAST; Altschul et al. 1990). A drastic over estimation of the correction needed would be to multiply the lowest pairwise P-value by the number of possible pairs. However, this assumes that all observed pairs are independent observations, which is certainly not the case for sequences that are highly similar. We thus sought a quantity that measures the “effective sequence number,” giving more weight to unique sequences (i.e., those without close homologs in the alignments) and less to those with many similar sequences. We used a diversity score for a multiple alignment described by Rychlewski et al. (2000):

$$D = \sum_{i=1}^n \frac{1}{1 + \sum_{j=1, j \neq i}^n S_{ij}^2}$$

where S_{ij} is the sequence identity between sequences i and j , and n is the number of sequences in the alignment.

If all sequences in the alignment are very similar, D tends to one, otherwise it increases as a function of the diversity of the sequences, with the total number of sequences in the alignment the upper limit. We thus define the P_{3D} for a multiple set of sequences between two families as:

$$MP_{3D} = \max(P_{3D}) \times (D_A D_B)$$

where D_A and D_B are the diversity scores for alignments A and B .

Acknowledgments

We thank Chris Ponting (Functional Genetics Unit, Oxford) and Rich Copley (EMBL, Heidelberg) for helpful discussions. This work was supported by grants BIO2000-0647, BIO2001-2064, BIO98-0362, and FEDER-2FD97-0872 from the CICYT, by CERBA and by C4-CESCA (Barcelona, Spain).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Aloy, P., Querol, E., Aviles, F.X., and Sternberg, M.J.E. 2001. Automated structure-based prediction of functional sites in proteins—Application to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **311**: 395–408.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Anantharaman, V., Koonin, E.V., and Aravind, L. 2001. Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J. Mol. Biol.* **307**: 1271–1292.
- Artymiuk, P.J., Rice, D.W., Mitchell, E.M., and Willet, P. 1990. Structural resemblance between the families of bacterial signal-transduction proteins and G proteins revealed by graph theoretical techniques. *Protein Eng.* **4**: 39–43.
- Artymiuk, P.J., Poirrette, A.R., Rice, D.W., and Willet, P. 1997. A polymerase I palm in adenyllyl cyclase? *Nature* **388**: 33–34.
- Attwood, T.K., Croning, M.D., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., and Wright, W. 2000. PRINTS-S: The database formerly known as PRINTS. *Nucleic Acids Res.* **28**: 225–227.

- Bairoch, A. and Apweiler, R. 1999. The SWISSPROT protein sequence data bank and its new supplement TrEMBL in 1999. *Nucleic Acids Res.* **27**: 49–54.
- Barton, G.J. 1993. ALSCRIPT: A tool to format multiple sequence alignments. *Protein Eng.* **6**: 37–40.
- Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. 2000. The Pfam protein families database. *Nucleic Acids Res.* **28**: 263–266.
- Bazan, J.F. 1990. Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci.* **87**: 6934–6938.
- 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.
- Blanco-Aparicio, C., Molina, M.A., Fernandez-Salas, E., Frazier, M.L., Mas, J.M., Querol, E., Aviles, F.X., and de Llorens, R. 1998. Potato carboxypeptidase inhibitor, a T-knot protein, is an epidermal growth factor antagonist that inhibits tumor cell growth. *J. Biol. Chem.* **273**: 12370–12377.
- Boggon, T.J., Shan, W.S., Santagata, S., Myers, S.C., and Shapiro, L. 1999. Implication of tubby proteins as transcription factors by structure-based functional analysis. *Science* **286**: 2119–2125.
- Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95**: 6073–6078.
- Christendat, D., Yee, A., Dharamsi, A., Kluger, Y., Savchenko, A., Cort, J.R., Booth, V., Mackereth, C.D., Saridakis, V., Ekiel, I., Kozlov, G., Maxwell, K.L., Wu, N., McIntosh, L.P., Gehring, K., Kennedy, M.A., Davidson, A.R., Pai, E.F., Gerstein, M., Edwards, A.M., and Arrowsmith, C.H. 2000. Structural proteomics of an archaeon. *Nat. Struct. Biol.* **7**: 903–909.
- Clemmons, D.R. 1993. IGF binding proteins and their functions. *Mol. Reprod. Dev.* **35**: 368–374.
- Copley, R.R. and Bork, P. 2000. Homology among (β/α) (8) barrels: Implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**: 627–641.
- Dietmann, S. and Holm, L. 2001. Identification of homology in protein structure classification. *Nat. Struct. Biol.* **8**: 953–957.
- Eddy, S.R. 1998. Profile hidden Markov models. *Bioinformatics* **14**: 755–763.
- Eisenberg, D., Marcotte, E.M., Xenarios, I., and Yeates, T.O. 2000. Protein function in the post-genomic era. *Nature* **405**: 823–826.
- Flores, T.P., Orengo, C.A., Moss, D.S., and Thornton, J.M. 1993. Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**: 1811–1826.
- Gay, N.J. and Walker, J.E. 1983. Homology between human bladder carcinoma oncogene product and mitochondrial ATP-synthase. *Nature* **301**: 262–264.
- Hegyi, H. and Gerstein, M. 1999. The relationship between protein structure and function: A comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**: 147–164.
- Henikoff, J.G., Greene, E.A., Pietrokovski, S., and Henikoff, S. 2000. Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.* **28**: 228–230.
- Holm, L. 1998. Unification of protein families. *Curr. Opin. Struct. Biol.* **8**: 372–379.
- Holm, L. and Sander, C. 1997. Decision support system for the evolutionary classification of protein structures. *Ismb* **5**: 140–246.
- Hughes, J., Ward, C.J., Peral, B., Aspinwall, R., Clark, K., San Millan, J.L., Gamble, V., and Harris, P.C. 1995. The polycystic kidney disease 1 (PKD1) gene encodes a novel protein with multiple cell recognition domains. *Nature Genet.* **10**: 151–160.
- Kraulis, P.J. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**: 946–950.
- Kunin, V., Chan, B., Sitbon, E., Lithwick, G., and Pietrokovski, S. 2001. Consistency analysis of similarity between multiple alignments: Prediction of protein function and fold structure from analysis of local sequence motifs. *J. Mol. Biol.* **307**: 939–949.
- Landgraf, R., Xenarios, I., and Eisenberg, D. 2001. Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. *J. Mol. Biol.* **307**: 1487–1502.
- Lukat, G.S., Lee, B.H., Mottonen, J.M., Stock, A.M., and Stock, J.B. 1991. Roles of the highly conserved aspartate and lysine residues in the response regulator of bacterial chemotaxis. *J. Biol. Chem.* **266**: 8348–8354.
- Mas, J.M., Aloy, P., Marti-Renom, M.A., Oliva, B., Blanco-Aparicio, C., Molina, M.A., de Llorens, R., Querol, E., and Aviles, F.X. 1998. Protein similarities beyond disulphide bridge topology. *J. Mol. Biol.* **284**: 541–548.
- Matsuo, Y. and Bryant, S.H. 1999. Identification of homologous core structures. *Proteins* **35**: 70–79.
- Murzin, A.G. 1993a. Can homologous proteins evolve different enzymatic activities? *Trends Biochem. Sci.* **18**: 403–405.
- Murzin, A.G. 1993b. Sweet-tasting protein monellin is related to the cystatin family of thiol proteinase inhibitors. *J. Mol. Biol.* **230**: 689–694.
- Murzin, A.G. 1993c. OB (oligonucleotide/oligosaccharide binding)-fold: Common structural and functional solution for non-homologous sequences. *EMBO J.* **12**: 861–867.
- 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.
- Ponting, C.P. and Russell, R.B. 1998. Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**: 364–371.
- Ponting, C.P. and Russell, R.B. 2000. Identification of distant homologues of FGFs suggests a common ancestor for all β -trefoil proteins. *J. Mol. Biol.* **302**: 1041–1047.
- Russell, R.B. 1998. Detection of protein three-dimensional side-chain patterns: New examples of convergent evolution. *J. Mol. Biol.* **279**: 1211–1227.
- Russell, R.B. and Barton, G.J. 1992. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins* **14**: 309–323.
- Russell, R.B. and Barton, G.J. 1994. Structural features can be unconserved in proteins with similar folds. An analysis of side-chain to side-chain contacts secondary structure and accessibility. *J. Mol. Biol.* **244**: 332–350.
- Russell, R.B., Sasienski, P.D., and Sternberg, M.J.E. 1998. Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**: 903–918.
- Russell, R.B., Saqi, M.A., Sayle, R.A., Bates, P.A., and Sternberg, M.J.E. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269**: 423–439.
- Rychlewski, L., Jaroszewski, L., Weizhong, L.I., and Godzik, A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**: 232–241.
- Saraste, M., Sibbald, P.R., and Wittinghofer, A. 1990. The P-loop—a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.* **15**: 430–434.
- Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28**: 231–234.
- Shapiro, L. and Harris, T. 2000. Finding function through Structural Genomics. *Curr. Opin. Biotechnol.* **11**: 31–35.
- Shapiro, L. and Scherer, P.E. 1998. The crystal structure of a complement-1q family protein suggests an evolutionary link to tumor necrosis factor. *Curr. Biol.* **8**: 335–338.
- Shapiro, L., Kwong, P.D., Fannon, A.M., Colman, D.R., and Hendrickson, W.A. 1995. Considerations on the folding topology and evolutionary origin of cadherin domains. *Proc. Natl. Acad. Sci.* **92**: 6793–6797.
- Tatusov, R.L., Galperin, M.Y., Natale, D.A., and Koonin, E.V. 2000. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **28**: 33–36.
- Todd, A.C., Orengo, C.A., and Thornton, J.M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**: 1113–1143.
- Vervoort, J., Heering, D., Peelen, S., and van Berkel, W. 1994. Flavodoxins. *Methods Enzymol.* **243**: 188–203.
- Volz, K. 1993. Structural conservation in the CheY superfamily. *Biochemistry* **32**: 11741–11753.
- Wallace, A.C., Borkakoti, N., and Thornton, J.M. 1997. TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Sci.* **6**: 2308–2323.
- Wilmanns, M., Priestle, J.P., Niermann, T., and Jansonius, J.N. 1992. Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: Indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 Å resolution. *J. Mol. Biol.* **223**: 477–507.
- Yang, F., Gustafson, K.R., Boyd, M.R., and Wlodawer, A. 1998. Crystal structure of *Escherichia coli* HdeA. *Nat. Struct. Biol.* **5**: 763–764.
- Zarembinski, T.I., Hung, L.W., Mueller-Dieckmann, H.J., Kim, K.K., Yokota, H., Kim, R., and Kim, S.H. 1998. Structure-based assignment of the biochemical function of a hypothetical protein: A test case of Structural Genomics. *Proc. Natl. Acad. Sci.* **95**: 15189–15193.