BIOINFORMATICS

Vol. 15 no. 6 1999
Pages 480–500

A comparison of sequence and structure protein domain families as a basis for structural genomics

Arne Elofsson* and Erik L. L. Sonnhammer[1]

Department of Biochemistry, Stockholm University, 106 91 Stockholm, Sweden

Received on November 11, 1998; revised on April 12, 1999; accepted on April 13, 1999

Abstract

Motivation: *Protein families can be defined based on structure or sequence similarity. We wanted to compare two protein family databases, one based on structural and one on sequence similarity, to investigate to what extent they overlap, the similarity in definition of corresponding families, and to create a list of large protein families with unknown structure as a resource for structural genomics. We also wanted to increase the sensitivity of fold assignment by exploiting protein family HMMs.*

Results: *We compared Pfam, a protein family database based on sequence similarity, to Scop, which is based on structural similarity. We found that 70% of the Scop families exist in Pfam while 57% of the Pfam families exist in Scop. Most families that occur in both databases correspond well to each other, but in some cases they are different. Such cases highlight situations in which structure and sequence approaches differ significantly. The comparison enabled us to compile a list of the largest families that do not occur in Scop; these are suitable targets for structure prediction and determination, and may be useful to guide projects in structural genomics. It can be noted that 13 out of the 20 largest protein families without a known structure are likely transmembrane proteins. We also exploited Pfam to increase the sensitivity of detecting homologs of proteins with known structure, by comparing query sequences to Pfam HMMs that correspond to Scop families. For SWISSPROT+TREMBL, this yielded an increase in fold assignment from 31% to 42% compared to using FASTA only. This method assigned a structure to 22% of the proteins in* Saccharomyces cerevisiae, *24% in* Escherichia coli, *and 16% in* Methanococcus jannaschii.

Contact: *arne@biokemi.su.se; Erik.Sonnhammer@cgr.ki.se*
Supplementary information: *http://www.biokemi.su.se/~arne/pfam-scop/*

Introduction

The number of protein sequences in SWISSPROT (Bairoch and Apweiler, 1996) and PIR (George *et al*., 1996) grows at an increasing rate as the genome projects proceed, while at the same time the number of known protein structures in PDB (Abola *et al*., 1987; Bernstein *et al*., 1977) increases at a slower rate (Holm and Sander, 1996). This is widening the gap between known protein sequences and protein structures. However, a large portion of newly determined protein sequences and structures are homologous to previously known proteins, resulting in the accumulation of redundancy in protein databases (Brenner *et al*., 1995; Casari *et al*., 1996; Tatusov *et al*., 1996).

To manage and exploit this redundancy, efforts have been made to classify protein databases into clusters or families of proteins that share certain features, such as sequence similarity, function, structure, or evolutionary origin. Both SWISSPROT and PIR contain family classification, and recently this has also been done systematically based on sequence similarity (Sonnhammer and Kahn, 1994; Wu *et al*., 1996; Linial *et al*., 1997; Sonnhammer *et al*., 1998a). A number of structure-based classification schemes of protein structures in PDB are also available (Murzin *et al*., 1995; Orengo *et al*., 1997; Holm and Sander, 1997). For proteins of known 3D structure, it has proved feasible and advantageous to perform the classification on several hierarchical levels, ranging from nearly identical structures, with high sequence similarity, to ‘common fold’, with virtually no sequence similarity but shared topology of secondary structure elements. In contrast, most sequence-based classification schemes tend to be non-hierarchical, mainly because of the difficulty to define useful levels of similarity for different hierarchical steps, and because of the much larger amount of data to process. Another important issue in all protein classifications is how to define domain boundaries, since each domain in a multi-domain protein may belong to a different family.

*To whom correspondence should be addressed. [1]Present Address: Center for Genomics Research, Karolinska Institutet, S-171 77 Stockholm, Sweden.

Abbreviations: SW+TREMBL, swissprot-35 + trembl 5; Scop, a structural classification of proteins database; HMM, hidden Markov model; Pfam, the Pfam-A 3.3 database; Pfam-3D, the part of Pfam-A 3.3 that is related to a protein of known structure.

Table 1. Correspondence between databases

Query Database	Related to families in	Number of hits	Number of entries in query database	%
Pfam families	Scop	802	1407	57 %
Scop sequences	Pfam	2056	2639	78 %
Scop families	Pfam	642	920	70 %
Scop superfamilies	Pfam	464	641	72 %
Scop folds	Pfam	350	458	76 %
SW+TREMBL sequences	Scop	64665	209668	31 %
SW+TREMBL sequences	Pfam	119964	209668	57 %

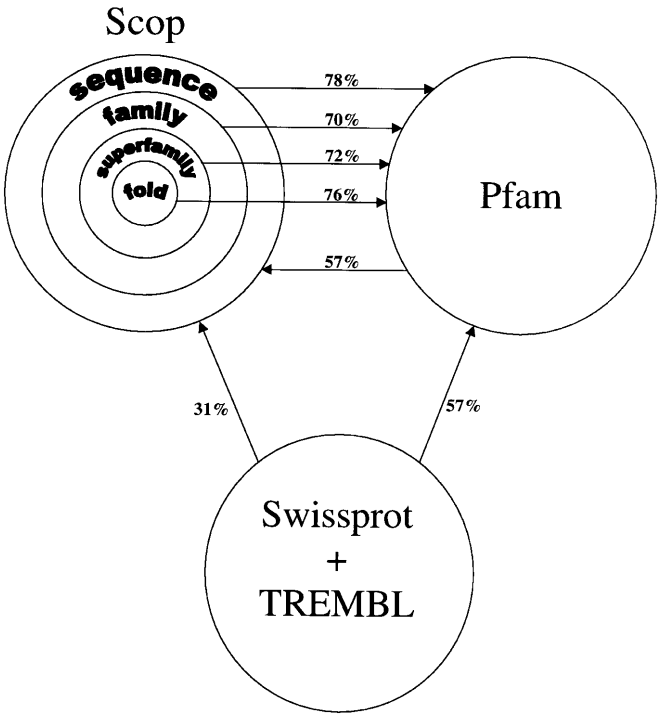

Fig. 1. Overall levels of correspondence between the Scop and Pfam protein domain family databases. The percentages at the arrows indicate the fraction of entries in one database that significantly matches the database the arrow points to.

Most comparisons between classification schemes concentrate on comparing sequence-based approaches to each other or structure-based ones to each other. We wanted instead to compare sequence-based to structure-based classifications, in order to answer a number of questions. First, to what extent do family definitions in sequence- and structure-based classifications overlap? In many cases, the families will overlap perfectly, but frequently they will overlap partly or not at all. Partially overlapping families may reflect differences in philosophy or techniques of the classification schemes, whereas the total absence of a family in one classification may indicate a difference in the underlying source of data. We here exploit the comparison for extracting large protein sequence families that are currently absent from

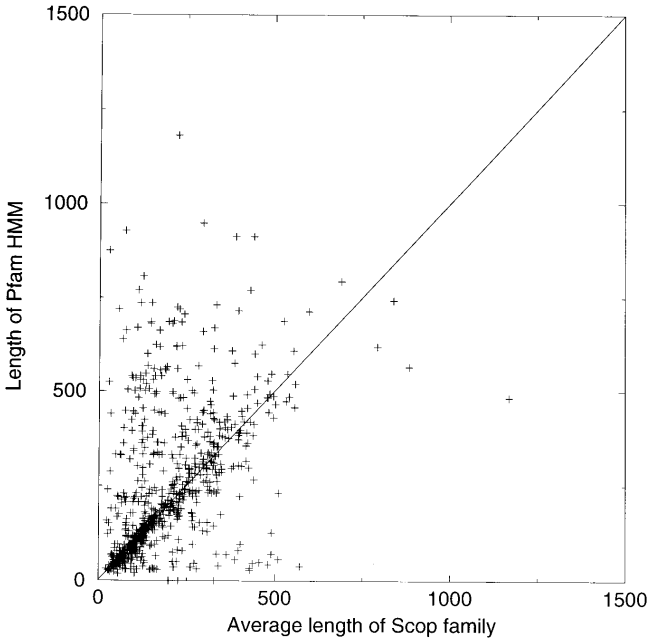

Fig. 2. The lengths of protein domains in Scop (average length of family members) plotted against the lengths of the corresponding Pfam 3.3 families.

PDB, in order to draw attention to them as important targets for structure determination. The emerging field of structural genomics, in which the goal is to determine the structure of all protein domains, would be assisted by such a ranked list.

To this end, we quantified the extent of overlap between a structure-based classification of PDB, Scop (Murzin *et al.*, 1995), and a sequence-based classification of SWISSPROT, Pfam (Sonnhammer *et al.*, 1998a). To be able to compare the clusters on an equal basis, the Scop families of PDB entries were converted to the corresponding clusters of SWISSPROT entries by homology searching. The Scop and Pfam databases were selected because they are both considered to be of high quality; in both databases, domain boundary definitions and family memberships have been verified manually.

Although the basic content of Pfam and Scop are lists of which protein segments belong to which families, the organization of the data is quite different, which makes the comparison somewhat challenging. Scop uses a hierarchical classification scheme at the family, superfamily, fold, and fold class levels, while Pfam has only one. Most Pfam families are clustered at a level corresponding to the Scop family and superfamily levels; in this paper we have focused on Scop on the family level. The higher clustering levels of Scop (fold and fold class) bring proteins together that have so little sequence similarity that they cannot be aligned confidently from the sequence alone. Since Pfam provides a multiple alignment for each family, such high clustering levels would not be feasible in Pfam.

Table 2. List of the largest Pfam families without a match to a Scop family

Pfam-name	Long Pfam-name	Accession Number	Number of members	Comment
cytochrome_b_N	Cytochrome b(N-terminal)/b6/petB	PF00033	2866	TM
Collagen	Collagen triple helix repeat (20 copies)	PF01391	2125	NG 1
7tm_1	7 transmembrane receptor (rhodopsin family)	PF00001	1423	TM
oxidored_q1	NADH-Ubiquinone/plastoquinone (complex I), various chains	PF00361	1315	TM
cytochrome_b_C	Cytochrome b(C-terminal)/b6/petD	PF00032	1288	TM
ABC_tran	ABC transporter	PF00005	1248	
ion_trans	Ion transport protein	PF00520	619	TM
helicase_C	Helicases conserved C-terminal domain	PF00271	603	
Ice_nucleation	Ice nucleation protein repeat	PF00818	506	
E1-E2_ATPase	E1-E2 ATPases	PF00122	502	TM
oxidored_q1_C	NADH-Ubiquinone oxidoreductase (complex I), chain 5 C-terminus	PF01010	450	TM
oxidored_q1_N	NADH-Ubiquinone oxidoreductase (complex I), chain 5 N-terminus	PF00662	428	TM
tubulin	Tubulin	PF00091	405	2
ldl_recept_b	Low-density lipoprotein receptor repeat class B	PF00058	392	TM
sugar_tr	Sugar (and other) transporter	PF00083	363	TM
ABC_membrane	ABC transporter transmembrane region.	PF00664	354	TM
Poty_coat	Potyvirus coat protein	PF00767	350	
signal	Signal carboxyl-terminal domain	PF00512	348	
BPD_transp	Binding-protein-dependent transport systems inner membrane component	PF00528	337	TM
tsp_1	Thrombospondin type 1 domain	PF00090	337	
NADHdh	NADH dehydrogenases	PF00146	331	TM
AAA	ATPases associated with various cellular activities (AAA)	PF00004	331	
late_protein_L1	L1 (late) protein	PF00500	322	
filament	Intermediate filament proteins	PF00038	320	NG
Nebulin_repeat	Nebulin repeat	PF00880	317	
Vif	Retroviral Vif (Viral infectivity) protein	PF00559	298	
phytochrome	Phytochrome, chromophore attachment domain	PF00360	276	
HN	Hemagglutinin-neuraminidase	PF00423	264	TM
myosin_head	Myosin head (motor domain)	PF00063	263	
Alpha_E2_glycop	Alphaviruses E2 glycoprotein	PF00943	261	TM
REV	REV protein (anti-repression transactivator protein)	PF00424	252	
HSP20	Hsp20/alpha crystallin family	PF00011	252	
RNA_dep_RNA_pol	RNA dependant RNA polymerase	PF00680	241	
fusion_gly	Fusion glycoprotein F0.	PF00523	234	TM
Glycos_transf_2	Glycosyl transferases	PF00535	223	
VP7	Glycoprotein VP7	PF00434	223	TM
mito_carr	Mitochondrial carrier proteins	PF00153	214	TM
Flagellin_C	Bacterial flagellin C-terminus	PF00700	211	
Acetyltransf	Acetyltransferase (GNAT) family	PF00583	211	
chloroa_b-bind	Chlorophyll A-B binding proteins	PF00504	211	TM
DEAD	DEAD/DEAH box helicase	PF00270	209	

The Pfam multiple alignments are used to generate hidden Markov model profiles (HMMs), which are used for sensitive detection of family members (Krogh *et al.*, 1994; Eddy, 1997). We show that this can be exploited to find more members of families with a known structure than by pairwise methods such as FASTA. We have further used this method to survey the frac-

Table 2. *Continued*

Pfam-name	Long Pfam-name	Accession Number	Number of members	Comment
Flagellin_N	Bacterial flagellin N-terminus	PF00669	208	
oxidored_q4	NADH-ubiquinone/plastoquinone oxidoreductase, chain 3	PF00507	198	TM
ATP-synt_A	ATP synthase A chain	PF00119	197	TM
UDPGT	UDP-glucoronosyl and UDP-glucosyl transferases	PF00201	196	TM
vMSA	Major surface antigen from hepadnavirus	PF00695	189	TM
flu_virus_nuc	Influenza virus nucleoprotein	PF00506	189	
ketoacyl-synt	Beta-ketoacyl synthase	PF00109	184	
adeno_fiber2	Adenoviral fiber protein (repeat/shaft region).	PF00608	181	
Glycos_transf_1	Glycosyl transferases group 1	PF00534	179	
MIP	Major intrinsic protein	PF00230	177	TM
NTP_transferase	Nucleotidyl transferase	PF00483	175	
laminin_G	Laminin G domain	PF00054	172	
Flavi_M	Flavivirus envelope glycoprotein M	PF01004	169	TM
wnt	wnt family of developmental signaling proteins	PF00110	160	
SRCR	Scavenger receptor cysteine-rich domain.	PF00530	157	
Hepatitis_core	Hepatitis core antigen	PF00906	154	NG
oxidored_q2	NADH-ubiquinone/plastoquinone oxidoreductase chain 4L	PF00420	154	TM
Fimbrial	Fimbrial proteins	PF00419	152	
VP4	Outer Capsid protein VP4 (Hemagglutinin)	PF00426	151	NG
aa_permeases	Amino acid permease	PF00324	150	TM
DUF4	Domain of unknown function	PF00668	149	
PUF	Pumilio-family RNA binding domains (aka PUM-HD, Pumilio homology domain)	PF00806	146	
Chal_stil_synt	Chalcone and stilbene synthases	PF00195	146	
HCV_NS4a	Hepatitis C virus nonstructural protein NS4a	PF01006	144	TM
tubulin-binding	Tau and MAP proteins, tubulin-binding	PF00418	138	
SNF2_N	SNF2 and others N-terminal domain	PF00176	138	
PHD	PHD-finger.	PF00628	137	
DUF5	Domain found in bacterial signal proteins	PF00672	135	TM
oxidored_q3	NADH-ubiquinone/plastoquinone oxidoreductase chain 6	PF00499	135	TM
BTB	BTB/POZ domain	PF00651	128	
HTH_2	Bacterial regulatory helix-turn-helix proteins, araC family	PF00165	128	
RNA_helicase	RNA helicase	PF00910	127	
flg_bb_rod	Flagella basal body rod proteins	PF00460	126	
oxidored_q5_N	NADH-ubiquinone oxidoreductase chain 4, amino terminus	PF01059	125	TM
PKD	PKD domain	PF00801	124	
PAC	Motif C-terminal to PAS motifs	PF00785	124	
Rhabd_glycop	Rhabdovirus spike glycoprotein	PF00974	123	TM
Gram_pos_anchor	Gram positive anchor	PF00746	121	TM
dehydrin	Dehydrins	PF00257	119	NG
lig_chan	Ligand-gated ion channel	PF00060	119	TM
DNA_pol_viral_N	DNA polymerase (viral) N-terminal domain	PF00242	118	
RNA_pol_A	RNA polymerase alpha subunit	PF00623	116	
pyridoxal_deC	Pyridoxal-dependent decarboxylase conserved domain	PF00282	116	

Table 2. *Continued*

Pfam-name	Long Pfam-name	Accession Number	Number of members	Comment
Ribosomal_S4	Ribosomal protein S4	PF00163	114	
HCV_RdRP	Hepatitis C virus RNA dependent RNA polymerase	PF00998	113	
BRCT	BRCA1 C Terminus (BRCT) domain	PF00533	113	
late_protein_L2	Late Protein L2	PF00513	112	NG
RNA_dep_RNApol2	RNA dependant RNA polymerase	PF00978	111	
tetR	Bacterial regulatory proteins, tetR family	PF00440	111	
E6	Early Protein (E6)	PF00518	110	
Flavi_capsid	Flavivirus capsid protein C	PF01003	109	TM
PLDc	Phospholipase D. Active site motif	PF00614	107	TM
RCC1	Regulator of chromosome condensation (RCC1)	PF00415	106	
glycosyl_hydro2	Glycosyl hydrolases family 32	PF00251	105	
S_T_dehydratase	Pyridoxal-phosphate dependant enzymes	PF00291	104	
cys_rich_FGFR	Cysteine rich repeat	PF00839	103	
integrin_A	Integrins alpha chain	PF00357	100	TM

1) collagen - Structure of tubulin was predicted in 1976 (M.H.Miller & Scheraga, 1976)

2) tubulin - Structure of tubulin was solved in 1998 (Downing & Nogales, 1998)

NG) Probably contains non-globularly folding elements and may not be soluble.

TM) Proteins probably contain transmembrane segments.

tion of proteins in three complete genomes, representing Eukarya, Eubacteria and Archaea, that can be assigned a structure by homology, and the fraction that matches a Pfam family.

Materials and methods

Pfam 3.3 (Sonnhammer *et al*., 1997; Bateman *et al*., 1999), containing 1407 families and corresponding HMMs, was used with the HMMER package, version 2.1 (Eddy, 1997). The pdb95d set of Scop version 1.39 was used. In this set only proteins that have less than 95% similarity to any other protein in Scop was included.

It should be mentioned that release 1.39 of Scop (file pdb95d_1.39) was later retracted by the Scop authors due to errors. The current Scop release is thus 1.37, but since many sequences are missing from this release we favored using 1.39 as the existing error did not seem to affect our results.

Each sequence from Scop was matched against the 1407 HMMs of Pfam 3.3, using the family specific GA cutoff defined in Pfam. In some cases, significant similarities between a Scop sequence and a Pfam family was not detected because the Scop entry corresponded to a subdomain while the Pfam family spanned the entire protein. To overcome this problem, we additionally matched all Scop sequences against a set of HMMs that allowed fragmentary matches, using an E-value cutoff of 1.e-5, and used the union of the global and fragment matches for further studies.

To predict if a Pfam family consists of membrane proteins, all sequences in the full Pfam alignment of that family were subjected to two tests: transmembrane helix prediction by TMHMM (Sonnhammer *et al*., 1998b), and scanning for the word 'TRANSMEMBRANE' in the keyword field of the swissprot entry. If more than 25% of the proteins contained at least one predicted transmembrane segment or the 'TRANSMEMBRANE' keyword, the family was annotated as 'probably transmembrane' in Table 2.

The 'non-globular' assignments of families in Table 2 were generated with the program PSEG (Wootton, 1994) as follows. Each sequence segment in the full Pfam alignments were subjected to PSEG complexity analysis in periodicities 1 through 12 with threshold parameters (window = 60; trigger threshold = 3.15; extension threshold = 3.15). These parameters were found optimal in the following test: we required a set of known non-globular domains to be found (myosin, kinesin, tropomyosin, proteoglycan core protein, histone H1, antifreeze protein A, collagen) while detecting as few segments as possible in Scop, assuming PDB to be essentially void of non-globular domains. Pfam families in which more than 5% of the members contained such non-globular segments were annotated as non-globular. In a few cases we noticed that this method assigned transmembrane regions with low compositional complexity as non-globular; these conflicts were resolved manually.

To estimate the 'real' number of members in Scop families we counted the number of significantly (E-value < 1/database size) matching sequences in SWISSPROT 35 + TREMBL 5 (Bairoch and Apweiler, 1996) found by FASTA (Pearson and Lipman, 1988). This is the sequence database used for Pfam 3.3; it was also used for Figure 5 and is referred to in the text as SW+TREMBL.

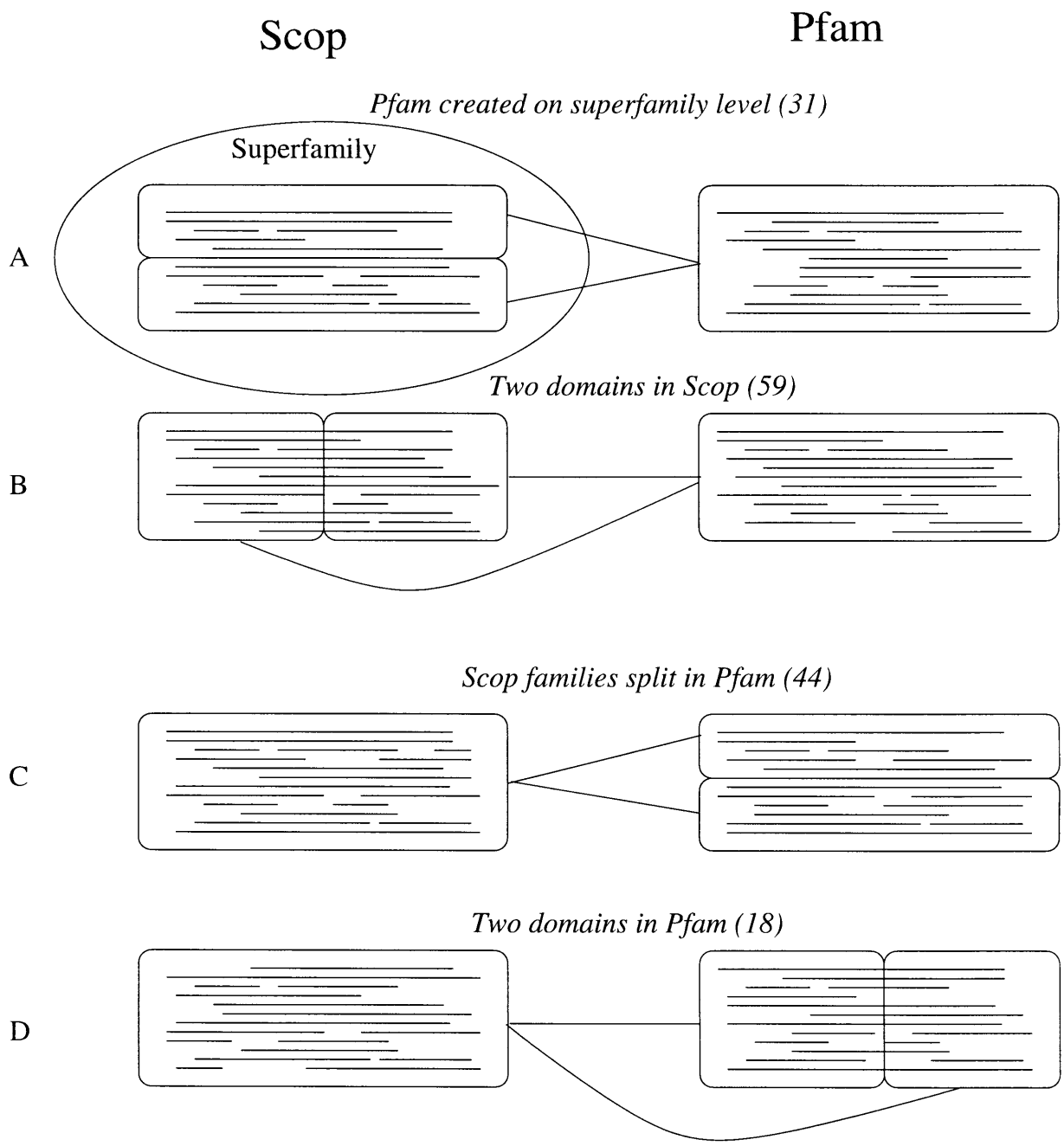

Fig. 3. Illustration of four types of discrepancy between family definitions in Scop and Pfam. The number of occurrences of each type in this study is listed within brackets. (A) One Pfam family corresponds to multiple Scop families because it corresponds to Scop on the superfamily level. (B) One Pfam family corresponds to multiple Scop families because the proteins were split into multiple domains in Scop. (C) One Scop family corresponds to multiple Pfam families because in was split into subfamilies in Pfam. (D) One Scop family corresponds to multiple Pfam families because the proteins were split into multiple domains in Pfam.

Three complete genomes from, *Saccharomyces cerevisiae* (Clayton *et al.*, 1997), *Escherichia coli* (Blattner *et al.*, 1997) and *Methanococcus jannaschii* (Bult *et al.*, 1996) were downloaded from http://www.sanger.ac.uk/Projects/C_elegans/Science98/protein_sets/ (sc and ec), http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/ (mj) and matched to Pfam 3.3 HMMs using the method described above for Scop. We also exploited the possibility to use the subset of Pfam that contains all families with a known structure, Pfam-3D, for fold recognition. A query sequence with a significant match to a Pfam-3D HMM was

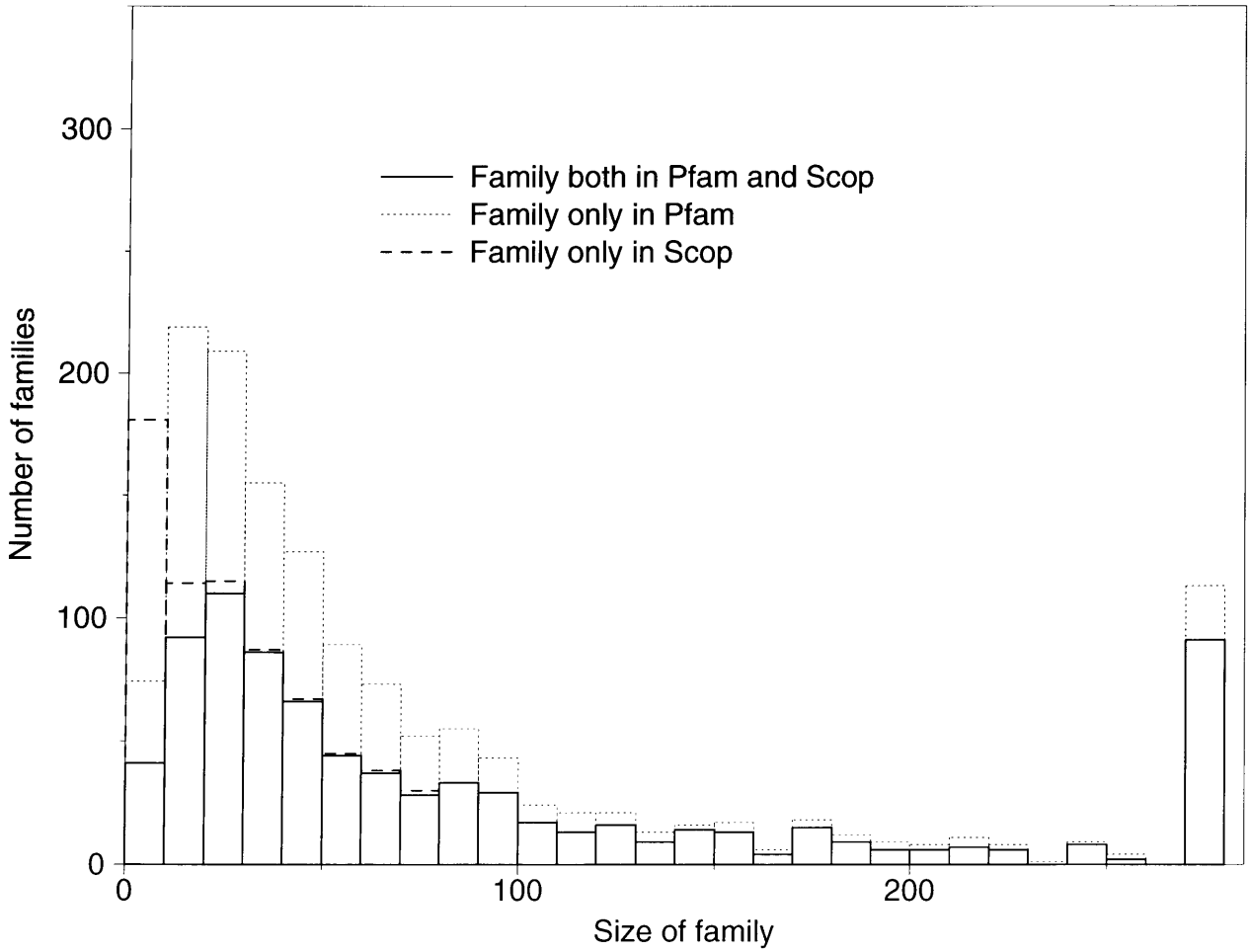

Fig. 4. Size distribution of Scop and Pfam 3.3 families. Families that are present in both Pfam 3.3 and Scop are plotted in thick lines. Stacked on top of these bars are families that are present only in Scop, thin dotted lines, or only in Pfam 3.3, thick dashed lines. The bar furthest to the right represents all families with more than 240 members.

considered to be related to a protein of known structure. The query sequence was also matched directly to the sequence of all proteins of known structure using FASTA and a cutoff that would give one false match for each genome. The union of all query sequences matching either Pfam-3D or a sequence from Scop was considered to be related to a protein of known structure.

Results and discussion

Overall correspondence between Scop and Pfam families

Pfam preferentially contains protein families with many members. A tradeoff for the manually verified quality of Pfam families is that it is not a fully comprehensive collection. The 1407 families in Pfam 3.3 match domains in about half of the proteins in SW+TREMBL. For our analysis, the availability of high-quality HMMs is more important than absolute comprehensiveness. Scop provides a de-facto standard for structural protein domain definitions; unlike other similar databases it is based on manual definitions of domain boundaries and structural and evolutionary relationships.

It is important to note that both Scop and Pfam were constructed using manual judgement to infer domain boundaries and evolutionary relationships between domain families. We therefore believe that the differences between the two databases are representative of the differences between structure and sequence approaches in general, and that idiosyncrasies have had relatively little effect on the results.4

Slightly more than half of the Pfam families correspond to a Scop entry, see Figure 1 and Table 1. The half that does not match Scop consists of families for which either no structure is known, or for which the sequence similarity to a protein of known fold is undetectable. The reciprocal correspondence of Scop to Pfam can be calculated at various clustering levels in Scop. We found this figure to range from 70% at the family level to 76% on the fold level. Pfam families that match Scop

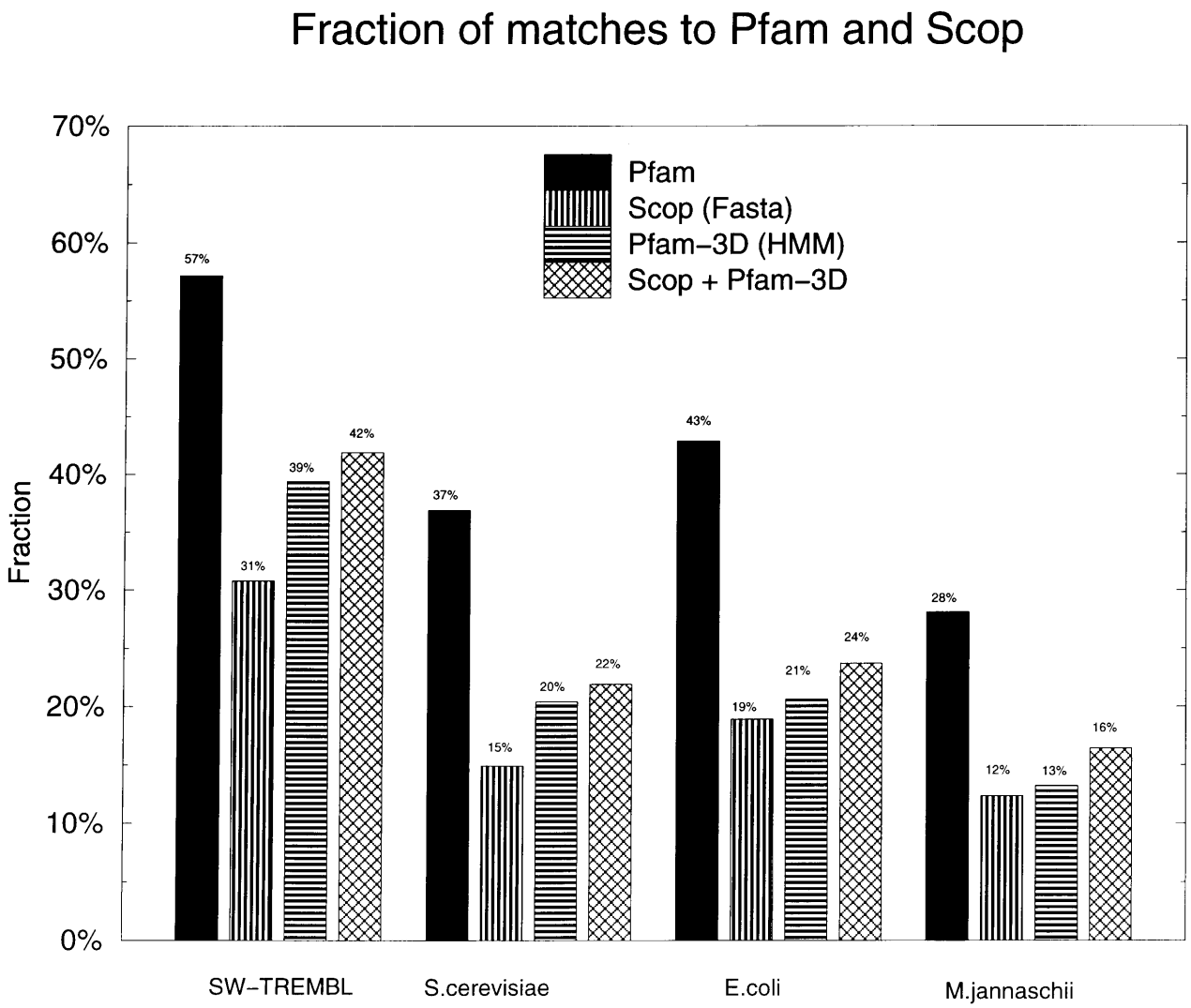

Fig. 5. The fraction of proteins in SWISSPROT and in three complete genomes that matches Pfam 3.3 and that can be assigned a fold. Within each dataset the leftmost filled bar shows the fraction of proteins that matches a Pfam 3.3 family HMM. The next three bars show the fraction that can be assigned a fold using FASTA searching against Scop, HMM-searching against Pfam-3D, and the union of these two methods.

(57% of all Pfam families) are biased towards large families, since these contain 69% of the sequences in Pfam. This bias may be caused by a greater interest in research on proteins of known structure as well as a greater incentive to solve the structure of large protein families.

Scop domains are on average 20% shorter than Pfam 3.3 domains

In both Scop and Pfam, the main reason for dividing a protein chain into several domains is that other proteins are similar over only a portion of the chain. The domains in both databases are thus normally independent and complete folding units, and the definitions should agree in most cases. The main difference is that they are based on structural information in Scop and on sequence similarity alone in Pfam. To quantify the difference between domain definitions in the two databases, we plotted the average length of corresponding Scop and Pfam families against each other, as shown in Figure 2. On average, the domain lengths in Scop are 80% of the corresponding Pfam 3.3 domains; the correlation coefficient was 0.45. We suspect that Scop domains are shorter than Pfam domains for two main reasons: (1) that only a fragment of the chain was used to determine the structure, and (2) structural similarity indicated a domain boundary that was not detectable in the sequences, thus leading to domain splitting only in Scop. Figure 3 shows that Scop splits Pfam domains four times more often than Pfam splits Scop domains.

Table 3. Pfam families with hits from more than once Scop family

Pfam AC	Pfam family	Scop families
Pfam family created on Scop superfamily level (see fig 3a)		
PF00027	Cyclic nucleotide-binding domain	Regulatory subunit of Protein kinase A Catabolite gene activator protein, N-terminal domain,
PF00030	BetaGamma crystallin	beta-gamma-Crystallin Ca-binding development proteins
PF00036	EF hand	Calmodulin-like Parvalbumin Calbindin D9K EF-hand modules in multidomain proteins S100 proteins
PF00037	4Fe-4S ferredoxins and related iron-sulfur cluster binding domains.	Short-chain ferredoxins 7-Fe ferredoxin Single 4Fe-4S cluster ferredoxin Arcaeal ferredoxins
PF00047	Immunoglobulin domain	V set domains (antibody variable domain-like) C1 set domains (antibody constant domain-like) I set domains
PF00061	lipocalin	Retinol binding protein-like Fatty acid binding protein-like
PF00069	Eukaryotic protein kinase domain	Serinethreonin kinases Tyrosine kinase
PF00085	Thioredoxin	Protein disulfide isomerase, N-terminal domain Thioltransferase
PF00087	Snake toxin	Snake venom toxins Dendroaspin
PF00089	Trypsin	Eukaryotic proteases Prokaryotic proteases
PF00132	Bacterial transferase hexapeptide (four repeats)	UDP N-acetylglucosamine acyltransferase Xenobiotic acetyltransferase Tetrahydrodipicolinate-N-succinlytransferase, THDP- succinlytransferase, DapD Carbonic anhydrase
PF00135	Carboxylesterases	Acetylcholinesterase-like Fungal lipases

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF00141	Peroxidase	Myeloperoxidase-like Cytochrome c peroxidase-like
PF00161	Ribosome inactivating protein	Plant cytotoxins Shiga toxin, A-chain
PF00168	C2 domain	Synaptotagmin-like (S variant) PLC-like (P variant)
PF00175	Oxidoreductase FADNAD-binding domain	Reductases NADPH-cytochrome p450 reductase Flavohemoglobin, C-terminal domain Phthalate dioxygenase reductase
PF00187	Chitin recognition protein	Agglutinin (lectin) domain Antimicrobial peptide 2, AC-AMP2
PF00206	Lyase	Argininosuccinate lyase Fumarase L-aspartase
PF00246	Zinc carboxypeptidase	Carboxypeptidase T Pancreatic carboxypeptidases
PF00296	Bacterial luciferase	Bacterial luciferase (alkanal monooxygenase) non-fluorescent flavoprotein (luxF, FP390)
PF00300	Phosphoglycerate mutase family	6-phosphofructo-2-kinasefructose-2,6-bisphosphatase, phosphatase domain Phosphoglycerate mutase
PF00328	Histidine acid phosphatase	Phytase (myo-inositol-hexakisphosphate-3-phosphohydrolase) Acid phosphatase
PF00388	Phosphatidylinositol-specific phospholipase C, X domain	Mammalian PLC Bacterial PLC
PF00432	Prenyltransferase and squalene oxidase repeats	Protein farnesyltransferase, beta-subunit Terpene syntases
PF00459	Inositol monophosphatase family	Inositol polyphosphate 1-phospatase Inositol monophosphatase
PF00537	long chain scorpion toxins	Long-chain scorpion toxins Short-chain scorpion toxins
PF00544	Pectate lyase	Pectate lyase Pectin lyase A

Pfam HMMs detect all members in 88% of the corresponding Scop families

To quantify the similarity between the family member definitions in Pfam and Scop, we counted the proportion of members in each Scop family that was detected by the Pfam HMM. Of the 920 families in Scop, 642 (70%) Scop families contain at least one member that is detected by a Pfam HMM. In 562 of these, all Scop members are detected by the Pfam HMM. Thus, in nearly all cases either all or no Scop members matched Pfam. In 80 (12%) of these Scop families, did a fraction of the Scop members match Pfam, i.e. in most of these cases the fraction comes close to 0 or 100%.

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF00561	alphabeta hydrolase fold	Haloperoxidase Haloalkane dehalogenase Bacterial lipase Hydroxynitrile lyase
PF00808	Histone-like transcription factors (CBFNF-Y) and archaeal histones.	Archaeal histone Nucleosome core histones
PF01023	S-100ICaBP type calcium binding domain	S100 proteins Calbindin D9K
PF01367	5'-3' exonuclease	T5 5'-exonuclease T4 RNase H
Pfam family divided into domains in Scop (see fig 3b)		
PF00009	Elongation factor Tu family	Elongation factors Elongation factor Tu (EF-Tu), the C-terminal domain
PF00012	Hsp70 protein	ActinHSP70 Heat shock protein 70kD (HSP70), C-terminal, substrate-binding fragment
PF00034	Cytochrome c	monodomain cytochrome c FAD-linked oxidases, C-terminal domain FAD-linked oxidases, N-terminal domain N-terminal (heme c) domain of cytochrome cd1-nitrite reductase Two-domain cytochrome c
PF00043	Glutathione S-transferases.	Glutathione S-transferases, C-terminal domain Glutathione S-transferases, N-terminal domain
PF00044	glyceraldehyde 3-phosphate dehydrogenases	Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) Glyceraldehyde-3-phosphate dehydrogenase-like, N-terminal domain
PF00056	lactatemalate dehydrogenase	Lactate & malate dehydrogenases, C-terminal domain Lactate & malate dehydrogenases, N-terminal domain
PF00070	Pyridine nucleotide-disulphide oxidoreductase class-I	FADNAD-linked reductases, N-terminal and central domains FADNAD-linked reductases, dimerisation (C-terminal) domain
PF00118	TCP-1cpn60 chaperonin family	Prokaryotic chaperone Eukaryotic chaperone GroEL, the ATPase domain The intermediate domain of GroEL

There can be several reasons that one or more sequences in the same Scop family are not detected by the Pfam HMM. (1) The protein missed is not closely related to all the other members of the family, i.e. according to Scop the proteins belong to the same family but the sequence identitity is too low for automatic methods to detect this. (2) The Pfam HMM is defined over a smaller region in Pfam than in Scop. (3) Errors in Scop or in Pfam. By studying a few of these cases it seems as if (1) is the most common explanation to why a sequence is missed.

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF00128	Alpha amylase	alpha-Amylases, beta-sheet domain alpha-Amylases, N-terminal domain
PF00136	DNA polymerase family B	DNA polymerase I (Klenow fragment) 3'-5' exonuclease domain of phage DNA polymerases
PF00152	tRNA synthetases class II	Class II aminoacyl-tRNA synthetases (aaRS), catalyic domain An anticodon-binding domain
PF00173	Heme-binding domain in cytochrome b5 and oxidoreductases	Cytochrome b5 FMN-linked oxidoreductases
PF00174	Oxidoreductase molybdopterin binding domain	Sulfite oxidase, the middle, catalytic domain E set domains
PF00186	Dihydrofolate reductase	Dihydrofolate reductases Thymidylate synthase
PF00204	DNA topoisomerase II (N-terminal region)	DNA gyrase B, N-terminal domain type II DNA topoisomerase a type II DNA topoisomerase domain
PF00205	Thiamine pyrophosphate enzymes	Pyruvate oxidase and decarboxylase Pyruvate oxidase and decarboxylase, middle domain
PF00224	Pyruvate kinase	Pyruvate kinase Pyruvate kinase, C-terminal domain Pyruvate kinase beta-barrel domain
PF00239	Site-specific recombinases	gamma,delta resolvase, large fragment Recombinase DNA-binding domain
PF00289	Carbamoyl-phosphate synthase (CPSase)	Biotin carboxylaseCarbamoyl phosphate synthetase Biotin carboxylaseCarbamoyl phosphate synthetase Biotin carboxylase subunit of acetyl-CoA carboxylase, C-terminal domain Carbamoyl phosphate synthetase, large subunit connection domain
PF00303	Thymidylate synthase	Dihydrofolate reductases Thymidylate synthase
PF00317	Ribonucleotide reductase	R1 subunit of ribonucleotide reductase, C-terminal domain R1 subunit of ribonucleotide reductase, N-terminal domain
PF00368	Hydroxymethylglutaryl-coenzyme A reductase	NAD-binding domain of HMG-CoA reductase Substrate-binding domain of HMG-CoA reductase

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF00369	Hydroxymethylglutaryl-coenzyme A reductase	NAD-binding domain of HMG-CoA reductase Substrate-binding domain of HMG-CoA reductase
PF00391	PEP-utilizing enzymes	Pyruvate phosphate dikinase, C-terminal domain N-terminal domain of enzyme I of the PEP:sugar phosphotransferase system Pyruvate phosphate dikinase, central domain
PF00403	Heavy-metal-associated domain	FADNAD-linked reductases, dimerisation (C-terminal) domain FADNAD-linked reductases, N-terminal and central domains Metal-binding domain
PF00406	Adenylate kinase	Nucleotide and nucleoside kinases Bacterial ADK, insert zinc finger domain
PF00408	Phosphoglucomutasephosphomannomutase	Phosphoglucomutase, first 3 domains Phosphoglucomutase, the C-terminal domain
PF00429	ENV polyprotein (coat polyprotein)	F-MuLV receptor-binding domain MMLV p15 fragment (residues 409-426)
PF00456	Transketolase	Transketolase Transketolase, C-terminal domain
PF00469	Negative factor, (F-Protein) or Nef.	Regulatory factor Nef HIV-1 Nef protein N-terminal fragment 1-25
PF00503	G-protein alpha subunit	G proteins Transducin (alpha subunit), insertion domain
PF00509	Haemagglutinin	Hemagglutinin, headpiece Influenza hemagglutinin (stalk)
PF00552	Integrase	Retroviral integrase DNA-binding domain of HIV-1 integrase
PF00555	delta endotoxin	delta-Endotoxin (insectocide), middle domain delta-Endotoxin, C-terminal domain
PF00579	tRNA synthetases class I (Trp and Tyr)	Class I aminoacyl-tRNA synthetases (RS), catalytic domain Tyrosyl-tRNA synthetase, middle domain
PF00587	tRNA synthetases class II (Gly, His, Pro and Ser)	Class II aminoacyl-tRNA synthetases (aaRS), catalyic domain An anticodon-binding domain of Class II aaRS
PF00607	gag gene protein p24 (core nucleocapsid protein).	HIV-1 capsid protein, N-terminal core domain HIV capsid C-terminal domain

A list of the largest protein families without a known structure

One of the goals with this study was to produce a list of large protein families for which no fold has been assigned. This list should be of interest for the ongoing projects to determine the structure of all new folds. The largest of these families are shown in Table 2. The complete list can be obtained from http://www.biokemi.su.se/~arne/pfam-scop/Scop-pfam.no-match. For many of the largest families it may have been

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF00665	retroviral pol related endonuclease	Retroviral integrase N-terminal Zn binding domain of HIV-1 integrase
PF00679	Elongation factor G C-terminus	Elongation factor G (EF-G), domain V Translational machinery components
PF00703	Glycosyl hydrolases family 2	beta-glycanases beta-Galactosidaseglucuronidase domain
PF00704	Glycosyl hydrolases family 18	type II chitinase Chitinase A, insertion domain
PF00707	Translation initiation factor IF-3	Translation initiation factor IF3 Translation initiation factor IF3, N-terminal domain
PF00713	Hirudin	Hirudin-like Non-folded peptides
PF00725	3-hydroxyacyl-CoA dehydrogenase	Acyl-CoA dehydrogenase 6-phosphogluconate dehydrogenase-like, N-terminal domain
PF00728	Glycosyl hydrolase family 20	Bacterial chitobiase, catalytic domain Bacterial chitobiase, Domain 2
PF00749	tRNA synthetases class I (E and Q)	Gln-tRNA synthetase (GlnRS), C-terminal (anticodon-binding) domain Anticodon-binding (C-terminal) domain of glutamyl-tRNA synthetase (GluRS)
PF00870	P53	p53-like transcription factors p53 tetramerization domain
PF00958	GMP synthase, C-terminal domain	GMP synthetase, the C-terminal, dimerisation domain N-type ATP pyrophosphatases
PF01077	Nitrite and sulphite reductase	Sulfite reductase hemoprotein (SiRHP), domains 2 and 4 Sulfite reductase, domains 1 and 3
PF01117	Aerolysin	(Pro)aerolysin, the pore-forming lobe Alpha-hemolysin
PF01179	Copper amine oxidase	Copper amine oxidase, domain 3 (catalytic) Copper amine oxidase, domains 1 and 2
PF01276	OrnLysArg decarboxylase	Ornithine decarboxylase major domain Ornithine decarboxylase C-terminal domain
PF01315	Aldehyde oxidase and xanthine dehydrogenase, C terminus	Aldehyde oxidoreductase, molybdemum cofactor-binding domain Aldehyde oxidoreductase, domain 3

difficult to obtain a structure because they are localized in the membrane, such as 7tm_1, or have a non-globular fold, such as filament. Out of the five largest families there are four transmembrane families and one non-globular family.

Table 2 could be used by structural biologists involved in structural genomics projects to indicate which large families still need to have their structure solved. As high-quality multiple sequence alignments of these families are already pro-

Table 3. *Continued*

Pfam AC	Pfam family	Scop families
PF01316	Arginine repressor	C-terminal domain of arginine repressor Arginine repressor (ArgR), N-terminal DNA-binding domain
PF01317	Biotin repressor	Biotin holoenzyme synthetase Biotin repressor (BirA)
PF01324	Diphtheria toxin	ADP-ribosylating toxins Diphtheria toxin, middle domain Diphtheria toxin, C-terminal domain
PF01330	Bacterial DNA recombination protein, RuvA	DNA helicase RuvA subunit, the middle domain DNA helicase RuvA subunit, C-terminal domain
PF01360	Monooxygenase	FAD-linked reductases, N-terminal domain p-Hydroxybenzoate hydroxylase
PF01397	Terpene synthase family	5-Epi-aristolochene synthase, C-terminal domain 5-Epi-aristolochene synthase, N-terminal domain

vided through Pfam, these proteins should be ideal targets for protein structure prediction attempts.

The level of heterogeneity in Scop and Pfam families

The definition of a protein family is sometimes rather arbitrary. It may therefore be interesting to examine what families in the two databases are equivalent and which differ. Families in Scop and in Pfam are defined manually with different objectives; in Scop a family is created based on one of two criteria, either having more than 30% sequence identity or a 'lower sequence identity but whose functions and structures are very similar' (Murzin *et al.*,1995), while families in Pfam are defined with a focus on creating good multiple sequence alignments and HMMs.

Most of the families in Pfam and Scop are equivalent to each other: in the 802 Pfam families that match a Scop family, 712 (89%) match only one Scop family. Conversely, of the 642 Scop families that match a Pfam family, 580 (90%) match only one Pfam family. This shows that in most cases there are no differences between the family definitions in Pfam and in Scop.

There are 90 occurences when multiple Scop families match one Pfam family, some of these are listed in Table 3. This is due to two reasons (Figure 3). (1) In 31 cases the Scop families that match the same Pfam family belong to the same superfamily, thus Pfam and Scop do correspond at the superfamily level. (2) In 59 cases the Scop families that match the same Pfam family correspond to different Scop-domains, thus a result of different domain definitions in Scop and Pfam.

Table 4 lists the 62 examples where multiple Pfam families correspond to one Scop family. In 44 of these cases, Scop was clustered at a higher level than in Pfam, see Figure 3. For instance, hsp70 and actin are in the same Scop family while they are separate families in Pfam. The reason for this is that hsp70 and actin are very difficult to align on a sequence basis, while after structural superposition they can be showed to have common ancestry (Flaherty *et al.*, 1991). In the 18 remaining cases, Pfam had split the family up into subdomains, presumably because it was not possible to produce a good multiple alignment of all members over the entire length. We noticed that a few of these cases are caused by errors in Scop 1.39. For instance, the unrelated domains dihydrofolate reductase and thymidylate synthase were present on the same Scop sequence.

We further examined the size distribution of the families, in three categories: families uniquely present in Scop, uniquely in Pfam, and families found in both databases (see Figure 4). This shows that the families only present in Scop are predominantly small, while the families uniquely in Pfam are predominantly large. The largest Scop family outside Pfam was the 'DNA-binding domain of HIV-1 integrase', and only 6 Scop families with more than 50 sw-34 entries were missing from Pfam 3.3. However, there are several Scop families missing from Pfam 3.3 that are larger than the smallest Pfam 3.3 families. The largest of these families will be included in the release of Pfam 4.0.

Complete genomes matched to structure and sequence families

A number of genomes have recently been completely sequenced. In this study we have compared what fraction of these genomes can be matched to a Pfam 3.3 family and to

Table 4. Scop families with members that match more than one Pfam family

Scop family	Pfam families	Pfam AC
Scop families split in Pfam (see fig 3c)		
Ferritin	Ferritins	PF00210
	Bacterioferritin	PF01334
Ribonucleotide reductase-like	Ribonucleotide reductases	PF00268
	Fatty acid desaturase	PF00487
Long-chain cytokines	Somatotropin hormone family	PF00103
	Interleukin-6G-CSFMGF family	PF00489
	Ciliary neurotrophic factor	PF01110
	LIF OSM family	PF01291
Short-chain cytokines	Interleukin 2	PF00715
	Interleukin 4	PF00727
	Granulocyte-macrophage colony-stimulating factor	PF01109
Interferonsinterleukin-10 (IL-10)	Interferon alphabeta domain	PF00143
	Interferon gamma	PF00714
	Interleukin 10	PF00726
Prokaryotic DNA-bending protein	Bacterial DNA-binding protein	PF00216
	H-NS histone family	PF00816
The C-terminal domain of alpha and beta subunits of F1 ATP synthase	ATP synthase ab C terminal	PF00306
	ATP synthase Alpha chain, C terminal	PF00422
Celluales catalytic domain	Glycosyl hydrolases family 9	PF00371
	Glycosyl hydrolases family 8	PF01270
E set domains	Oxidoreductase molybdopterin binding domain	PF00174
	hemocyanin family	PF00372
	Rel homology domain (RHD).	PF00554
	FilaminABP280 repeat.	PF00630
	Flavivirus glycoprotein	PF00869
Fibronectin type III	Fibronectin type III domain	PF00041
	Tissue factor	PF01108
Cellulose-binding domain family III	Cellulose binding domain	PF00942
	Cohesin domain	PF00963

a protein of known structure, see Figure 5. Pfam 3.3 matches 57% of the proteins in SW+TREMBL, while 37% of the proteins in *Saccharomyces cerevisiae*, 43% of the proteins in *Escherichia coli*, and 28% in *Methanococcus jannaschii* match Pfam 3.3. All assignments are available from http://www.biokemi.su.se/~arne/pfam-scop/ pfam_scop_foldassignments.{swtrembl,ec,sc,mj}.gz. As expected, the figures for the complete genomes are lower than for SW+TREMBL, since Pfam 3.3 is biased towards the largest families in SW+TREMBL.

To determine the proportion of the proteins in these genomes that can be assigned a fold, we used three different methods (see Materials and methods). First, a FASTA search against Scop sequences was carried out for each protein sequence. Second, the sequences were compared to the HMMs of 'Pfam-3D', the Pfam 3.3 families that significantly match

Table 4. *Continued*

Scop family	Pfam families	Pfam AC
p53-like transcription factors	P53	PF00870
	T-box	PF00907
Cold shock DNA-binding domain-like	'Cold-shock' DNA-binding domain	PF00313
	S1 RNA binding domain	PF00575
Plant virus proteins	Viral coat protein (S domain)	PF00729
	Tymovirus coat protein	PF00983
	Bromovirus coat protein	PF01318
Animal virus proteins	picornavirus capsid protein	PF00073
	Polyomavirus coat protein	PF00718
	Parvoviral coat protein	PF00740
	Hexon, adenovirus major coat protein	PF01065
beta-glycanases	Cellulase (glycosyl hydrolase family 5)	PF00150
	Glycosyl hydrolase family 10	PF00331
	Glycosyl hydrolases family 17	PF00332
	Glycosyl hydrolases family 2	PF00703
type II chitinase	Chitinases, family 2	PF00192
	Glycosyl hydrolases family 18	PF00704
Aldolase	Fructose-bisphosphate aldolase class-I	PF00274
	Dihydrodipicolinate synthetase family	PF00701
Tryptophan biosynthesis enzymes	Indole-3-glycerol phosphate synthases	PF00218
	Tryptophan synthase alpha chain	PF00290
	N-(5'phosphoribosyl)antranilate (PRA) isomerase	PF00697
Nucleotide and nucleoside kinases	Adenylate kinase	PF00406
	Guanylate kinase	PF00625
	Thymidine kinase from herpesvirus	PF00693
G proteins	ADP-ribosylation factor family	PF00025
	Ras family	PF00071
	G-protein alpha subunit	PF00503
Nitrogenase iron protein-like	4Fe-4S iron sulfur cluster binding proteins, NifHfrxC family	PF00142
	SRP54-type protein	PF00448
	Adenylosuccinate synthetase	PF00709
CoA-dependent acetyltransferases	2-oxo acid dehydrogenases acyltransferase (catalytic domain)	PF00198
	Chloramphenicol acetyltransferase	PF00302

a protein of known structure. Third, if either of the two previous methods matched a sequence to a known structure, it was counted as 'structure known', i.e. the union of the two previous methods.

As seen in Figure 5, 31% of SW+TREMBL matched a sequence of known 3D structure in PDB, and 39% matched Pfam-3D. The two methods combined could however assign 42% of SW+TREMBL to a known structure. Applied to proteins from completely sequenced genomes, the methods combined could assign a structure to 22% of the proteins in *S. cerevisiae* and 24% in *E. coli*, and to 16% in *M. jannaschii*. These fold assignment rates are 25% higher than has been reported in another study using FASTA only (Frishman and Mewes, 1997).

Table 4. *Continued*

Scop family	Pfam families	Pfam AC
Thioltransferase	Thioredoxin	PF00085
	Glutaredoxin	PF00462
ActinHSP70	Hsp70 protein	PF00012
	Actin	PF00022
RNA methylases	Ribosomal RNA adenine dimethylases	PF00398
	Poly A polymerase regulatory subunit	PF01358
Purine and uridine phosphorylases	phosphorylases family 2	PF00896
	Phosphorylase family	PF01048
L-arabinose binding protein-like	Periplasmic binding proteins and LacI family.	PF00532
	Receptor family ligand binding region	PF01094
Phosphate binding protein-like	Bacterial regulatory helix-turn-helix proteins, lysR family	PF00126
	Bacterial extracellular solute-binding proteins, family 5	PF00496
	Bacterial extracellular solute-binding proteins, family 3	PF00497
	Prokaryotic sulphate- and thiosulphate-binding protein	PF01100
Translational machinery components	Ribosomal protein S5	PF00333
	Elongation factor G C-terminus	PF00679
ADP-ribosylating toxins	Diphtheria toxin	PF01324
	Heat-labile enterotoxin alpha chain	PF01375
MHC antigen-recognition domain	Class I Histocompatibility antigen, domains alpha 1 and 2	PF00129
	Class II histocompatibility antigen, beta domain	PF00969
	Class II histocompatibility antigen, alpha domain	PF00993
Glucose 6-phosphate dehydrogenase-like	Glucose-6-phosphate dehydrogenase	PF00479
	Oxidoreductase family	PF01408
Tetrahydrobiopterin biosynthesis enzymes	GTP cyclohydrolase I	PF01227
	6-pyruvoyl tetrahydropterin synthase	PF01242
Class II aminoacyl-tRNA synthetases (aaRS), catalyic domain	tRNA synthetases class II	PF00152
	tRNA synthetases class II (Gly, His, Pro and Ser)	PF00587
DNase I-like	Deoxyribonuclease I (DNase I)	PF01181
	AP endonucleases family 1	PF01260
beta-LactamaseD-ala carboxypeptidase	Beta-lactamase	PF00144
	Penicillin binding protein transpeptidase domain	PF00905
DNA polymerase I (Klenow fragment)	DNA polymerase family B	PF00136
	DNA polymerase family A	PF00476

Relative to using the methods individually, combining the methods increased the fold assignment rate by 25–45% for the three genomes in this study compared to using only FASTA, and by 10–25% compared to only using Pfam. Since we used both FASTA and Pfam 3.3 HMM searching with conservative cutoffs, we do not believe that in combining the two methods we increased the amount of noise. Rather, the increase in fold assignment is due to combining the higher sensitivity of the Pfam 3.3 HMMs with the broader coverage of small families in Scop.

Table 4. *Continued*

Scop family	Pfam families	Pfam AC
Membrane all-alpha	Cytochrome C and Quinol oxidase polypeptide I	PF00115
	Photosynthetic reaction center protein	PF00124
	ATP synthase subunit C	PF00137
	Cytochrome c oxidase subunit III	PF00510
	Bacteriorhodopsin	PF01036
Gonadotropin, A and B chains	Cystine-knot domain	PF00007
	Glycoprotein hormones	PF00236
Short-chain scorpion toxins	Scorpion short toxins	PF00451
	long chain scorpion toxins	PF00537
Defensin	Mammalian defensins	PF00323
	Anenome neurotoxin	PF00706
	Beta defensins	PF00711
Transmembrane helical fragments	Neurotransmitter-gated ion-channel	PF00065
	Anion exchanger family	PF00955
	Glycophorin A	PF01102
Non-folded peptides	IQ calmodulin-binding motif	PF00612
	Hirudin	PF00713
Scop family divided into two domains in Pfam(see fig 3d)		
Nucleosome core histones	Core histone H2AH2BH3H4	PF00125
	Histone-like transcription factors (CBFNF-Y) and archaeal histones.	PF00808
Calbindin D9K	EF hand	PF00036
	S-100ICaBP type calcium binding domain	PF01023
S100 proteins	EF hand	PF00036
	S-100ICaBP type calcium binding domain	PF01023
Legume lectins	Legume lectins alpha domain	PF00138
	Legume lectins beta domain	PF00139
Pleckstrin-homology domain (PH domain)	PH (pleckstrin homology) domain	PF00169
	BTK motif	PF00779
Mammalian PLC	Phosphatidylinositol-specific phospholipase C, Y domain	PF00387
	Phosphatidylinositol-specific phospholipase C, X domain	PF00388
Inosine monophosphate dehydrogenase (IMPDH)	IMP dehydrogenase GMP reductase	PF00478
	CBS domain	PF00571

Conclusions

In summary, the majority of the protein families present in both Scop and Pfam correspond well to each other. Given the differences in goals, underlying data, and methodological approach between Pfam and Scop, we were surprised how similar the family definitions are. In the cases where families of the two databases correspond poorly, this is usually due to the different goals of the databases, but sometimes to more or less arbitrary differences in the family definitions. We have exploited the comparsion results to provide a list of the largest proteins families with unknown structure. By using

Table 4. *Continued*

Scop family	Pfam families	Pfam AC
FMN-linked oxidoreductases	Heme-binding domain in cytochrome b5 and oxidoreductases	PF00173
	FMN oxidoreductase	PF00724
	FMN-dependent dehydrogenase	PF01070
	Dihydroorotate dehydrogenase	PF01180
Caspase	ICE-like protease (caspase) p10 domain.	PF00655
	ICE-like protease (caspase) p20 domain.	PF00656
Tyrosine-dependent oxidoreductases	short chain dehydrogenase	PF00106
	Short chain dehydrogenasereductase C-terminus	PF00678
	3-beta hydroxysteriod dehydrogenaseisomerase family	PF01073
	NAD dependant epimerasedehydratase family	PF01370
FADNAD-linked reductases, N-terminal and central domains	Pyridine nucleotide-disulphide oxidoreductase class-I	PF00070
	Heavy-metal-associated domain	PF00403
	UDP-glucoseGDP-mannose dehydrogenase family	PF00984
Retroviral integrase	Integrase	PF00552
	retroviral pol related endonuclease	PF00665
Dihydrofolate reductases	Dihydrofolate reductase	PF00186
	Thymidylate synthase	PF00303
Substrate-binding domain of HMG-CoA reductase	Hydroxymethylglutaryl-coenzyme A reductase	PF00368
	Hydroxymethylglutaryl-coenzyme A reductase	PF00369
NAD-binding domain of HMG-CoA reductase	Hydroxymethylglutaryl-coenzyme A reductase	PF00368
	Hydroxymethylglutaryl-coenzyme A reductase	PF00369
FADNAD-linked reductases, dimerisation (C-terminal) domain	Pyridine nucleotide-disulphide oxidoreductase class-I	PF00070
	Heavy-metal-associated domain	PF00403
Thymidylate synthase	Dihydrofolate reductase	PF00186
	Thymidylate synthase	PF00303
Serinethreonin kinases	Eukaryotic protein kinase domain	PF00069
	Protein kinase C terminal domain	PF00433

Pfam-3D in combination with Scop, the fraction of proteins for which a fold can be assigned was increased significantly. For proteins in complete genomes, this fraction was markedly lower in *Methanococcus jannaschii* than in *Saccharomyces cerevisiae* and *Escherichia coli*.

Acknowledgments

We thank Gunnar von Heijne and Erik Wallin for valuable discussions. AE was supported through grants from the Swedish natural science research council, the Swedish research council for engineering sciences and the Magnus Bergvall foundation.

References

Abola,E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) Data Commission of the international union of crystallography. In Allen,F.H., Bergerhoff,G. and Sievers,R. (eds), *Databases-Information Content, Software Systems, Scientific Applications, Protein Data Bank*. Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, pp. 107–132.

Bairoch,A. and Apweiler,R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.*, **24,** 17–21.

Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Finn,R.D. and Sonnhammer,E.L. (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res.*, **27,** 260–262.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer Jr,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112,** 535–542.

Blattner,F.R., Plunkett,G.r., Bloch,C.A., Perna,N.T., Burland,V., Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F., Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J., Mau,B. and Shao,Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277,** 1453–1474.

Brenner,S., Hubbard,T., Murzin,A. and Chothia,C. (1995) Gene duplications in *H. influenzae*. *Nature*, **378,** 140.

Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M. and Venter,J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*, **273,** 1058–1073.

Casari,G., De Daruvar,A., Sander,C. and Schneider,R. (1996) Bioinformatics and the discovery of gene function. *Trends Genet.*, **12,** 244–245.

Clayton,R.A., White,O., Ketchum,K.A. and Venter,J.C. (1997) The first genome from the third domain of life. *Nature*, **387,** 459–462.

Downing,K.H. and Nogales,E. (1998) New insights into microtubule structure and function from the atomic model of tubulin. *Eur. Biophys. J.*, **27,** 431–436.

Eddy,S.R. (1997) HMMER – Hidden Markov model software. URL: http://genome.wustl.edu/eddy/hmmer.html.

Flaherty,K., McKay,D., Kabsch,W. and Holmes,K. (1991) Similarity of the three-dimensional structures of actin and the atpase fragment of a 70-kda heat shock cognate protein. *Proc. Natl Acad. Sci. USA*, **88**, 5041–5045.

Frishman,D. and Mewes,H.W. (1997) Protein structural classes in five complete genomes. *Nature Struct. Biol.*, **4,** 626–628.

George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1996) The PIR-International Protein Sequence Database. *Nucleic Acids Res.*, **24,** 17–21.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273,** 595–603.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25,** 231–234.

Krogh,A., Brown,M., Mian,I.S., Sjölander,K. and Haussler,D. (1994) Hidden Markov models in computational biology. *J. Mol. Biol.*, **235,** 1501–1531.

Linial,M., Linial,N., Tishby,N. and Yona,G. (1997) Global self-organization of all known protein sequences reveals inherent biological signatures. *J. Mol. Biol.*, **268,** 539–556.

Miller,M.H. and Scheraga,H.A. (1976) Calculation of the structures of collagen models role of interchain interactions in determining the triple-helical coiled-coil conformation. I. *J. Polym. Sci.*, **54,** 171.

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.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D., Swindells,M. and Thornton,J.M. (1997) CATH – a hierarchic classification of protein domain structures. *Structure*, **5,** 1093–1108.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence analysis. *Proc. Natl Acad. Sci. USA*, **85,** 2444–2448.

Sonnhammer,E.L. and Kahn,D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.*, **3,** 482–492.

Sonnhammer,E.L., Eddy,S.R. and Durbin,R. (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins, Structure Function and Genetics*, **28,** 405–420.

Sonnhammer,E.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998a) Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26,** 322–325.

Sonnhammer,E.L., von Heijne,G. and Krogh,A. (1998b) A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc. of Sixth Int. Conf. on Intelligent Systems for Molecular Biology.* AAAI Press, Menlo Park.

Tatusov,R.L., Mushegian,A.R., Bork,P., Brown,N.P., Hayes,W.S., Borodovsky,M. and Rudd,K.E. (1996) Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Curr. Biol.*, **6,** 279–291.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem.*, **18,** 268–285.

Wu,C.H., Zhao,S., Chen,H.L., Lo,C.J. and McLarty,J. (1996) Motif identification neural design for rapid and sensitive protein family search. *Comput. Applic. Biosci.*, **12,** 109–118.