Characterization of Novel Proteins Based on Known Protein Structures.

Walter A. Koppensteiner, Peter Lackner,
Markus Wiederstein, and Manfred J. Sippl

Center for Applied Molecular Engineering
Institute for Chemistry and Biochemistry
University of Salzburg
Jakob Haringer Straße 3
A-5020 Salzburg, Austria

Tel: +43 662 8044 ext 5797
Fax: +43 662 454889
e-mail: sippl@came.sbg.ac.at

Running title: Structure-based characterization of novel proteins

Submitted to J. Mol. Biol.

Corresponding author: Manfred J. Sippl

Summary

The genome sciences face the challenge to characterize structure and function of a vast number of novel genes. Sequence search techniques are used to infer functional and structural information from similarities to experimentally characterized genes or proteins. The persistent goal is to refine these techniques and to develop alternative and complementary methods to increase the range of reliable inference.

Here we focus on the structural and functional assignments that can be inferred from the known three dimensional structures of proteins. The study uses all structures in the Brookhaven Protein Data Bank that were known by the end of 1997. The protein structures released in 1998 were then characterized in terms of functional and structural similarity to the previously known structures, yielding an estimate of the maximum amount of information on novel protein sequences that can be obtained from inference techniques.

147 globular proteins corresponding to 196 domains released in 1998 have no clear sequence similarity to previously known structures. 75% of the domains have extensive structure similarity to previously known folds, and most importantly, in two out of three cases similarity in structure coincides with related function. In view of this analysis, full utilization of existing structure data bases would provide information for many new targets even if the relationship is not accessible from sequence information alone. Currently, the most sophisticated techniques detect in the order of one third of these relationships.

Keywords: structural genomics; functional genomics; structure prediction; side chain orientation

Introduction

Several bacterial and eucaryotic genomes (TIGR, 1999) have been released and the completion of the human genome project (Rowen *et al.*, 1997) is on the way. The challenge is to assign biological function to the novel sequences of these genomes. A full characterization of a protein contains its molecular and cellular function, its three dimensional structure and its interaction with other molecules. Frequently, the function and biological role of a hypothetical protein is inherited from a characterized protein using sequence comparison methods (Altschul *et al.*, 1990; Pearson, 1996).

The basis of sequence comparison is the conservation of structure and function among related proteins (Sander & Schneider, 1991). The limit of reliable inference using sequence similarity is the so-called twilight zone where similarity becomes indistinguishable from random matches. On the other hand, proteins with insignificant sequence similarity can have similar tertiary structures (Pastore & Lesk, 1990). In fact, nature seems to be able to realize the enormously diverse biological functions by a limited number of folds (Chothia, 1992; Orengo *et al.*, 1994). Consequently, methods are desirable that can detect relationships beyond the twilight zone. Profile-based techniques like PSI-Blast (Altschul *et al.*, 1997) and Hidden-Markov models (Karplus *et al.*, 1998), and structure-based techniques like fold recognition (Bowie *et al.*, 1991; Sippl & Weitckus, 1992; Jones *et al.*, 1992; Bryant, 1996; Domingues *et al.*, 1999a; Jones, 1999) have made progress in this direction (Jones, 1997; Koehl & Levitt, 1999). The success of these methods is limited by the information on structure and function contained in data bases. An analysis of this information content is the focus of this paper.

The conservation of structure among distantly related proteins is the basis of structure-based functional annotations of uncharacterized proteins (Martin *et al.*, 1998; Russell *et al.*, 1998; Hegyi & Gerstein, 1999; Orengo *et al.*, 1999). The distinction of analogues and remote homologues has been perceived as a critical factor for the application of structures for function assignment (Flores *et al.*, 1993; Russell & Barton, 1994; Matsuo & Bryant, 1999). Analogous proteins are considered as a product of convergent evolution to a similar three dimensional structure while remote homologous originate from a common ancestor. There is agreement that a clear distinction is difficult to obtain because functional relatedness is often hard to prove (Holm & Sander, 1997; Murzin, 1998).

Here we explore to what extent information on structure and function contained in current data bases can be used to characterize novel genes. The limits of inference of structure and function from data bases can be explored from a set of experimentally determined structures. We present an analysis of structures released by PDB (Bernstein *et al.*, 1977) in 1998. We derive a data set of proteins which do not have sequence similarity to previously determined structures, i.e. proteins which were made public prior to 1998 and we investigate structural and functional relationships to the previously known proteins. We address the following questions: (1) How many new proteins have structural similarity to previously known proteins? (2) How many new proteins are functionally related to previously known proteins? (3) To what extent do structural and functional similarity coincide?

The extent of structural similarity required to model a target from a template depends on the intended purpose and the desired accuracy of the derived model. The level of detail can range from an all-atom model to a rough arrangement of secondary structure elements. Here we assume that a suitable template needs to share at least all secondary structure elements that build up the hydrophobic core with the target. If this condition is met then we call the structural relationship to be on the *fold* level. We consider two additional levels of similarity which are relevant for structure prediction called *substructure* and *partial fold* level: when a structure superimposes with a compact substructure of a larger protein, and when two folds have partial similarity where at least major parts of secondary structure elements are superimposable. All targets not assignable to one of these levels are classified as novel folds. We regarded structures A and B to have the same fold when all core secondary structure elements of A are superimposable with B, so that B can be an adequate template for A. This, however, does not always imply that A is also a suitable template for B. Here, the structures released in 1998 always represent the targets, and the previously determined structures the pool of possible templates.

The identification of functional relatedness among proteins requires a detailed and often elaborate analysis of the respective protein structures, especially when the similarity is weak. In the present analysis, we rely on the expert knowledge contained in the SCOP data base (Murzin *et al.*, 1995) and the reports of crystallographers and NMR spectroscopists. SCOP classifies protein domains in hierarchical levels called class, fold, superfamily, and family. This hierarchy reflects structural and functional similarities. Since SCOP is updated infrequently new structures are often not found in the data base, so that we had to consult the reports on the respective structure determinations. Using these sources we determine the functional relationships of domains released in 1998 to previously known structures. We find that two thirds of structurally similar pairs which are unrelated in sequence have a related function.

Superimposition of the C^{α} traces is the standard approach to determine the extent of structural similarity of two proteins (Taylor & Orengo, 1989; Holm & Sander, 1993). Here, we incorporate side chain orientation, represented by the C^{β} positions, into structure comparison. Side chain orientation is relevant for the selection of suitable templates since any attempt to model a structure will run into enormous difficulties when side chain orientation is not conserved. Also, side chain orientation is more conserved in remote homologues than in analogues.

In this paper we provide a detailed description of the data sets used, investigate the role of side chain orientation in distinguishing significant and insignificant structural similarity, and analyze how much structural information can be derived from previously known structures. Finally, we investigate the correlation of structural and functional relationships and provide evidence that side chain orientation facilitates the distinction of analogous and remote homologous proteins. In the conclusions we compare our results with a recent assessment of structure prediction methods and discuss the relevance of these findings for the structural genomics initiative and the annotation of whole genomes.

Results and Discussion

The data sets

In 1998 the PDB released 1792 new entries containing 3358 protein chains. PDB often contains multiple instances of the same protein. Removing redundancies of greater than 95% sequence identity only 664 new chains remain. This set represents the total amount of structural information released in 1998. From this set 490 sequences (74%) have significant sequence similarity to previously known structures. In principle, these structures could have been derived from the known data with a reasonable degree of accuracy.

We restrict the analysis to globular proteins and remove all non-globular chains, virus capsid and transmembrane proteins from the remaining 174 chains without significant sequence similarity to known structures. The remaining data set contains 147 protein chains (22% of 664) consisting of 196 domains. There are 107 single domain proteins in this data set, 32 contain two domains, 7 contain three domains, and one contains four domains. The domains have no clear relationship to previously determined proteins and are used for the subsequent structural and functional analysis. The data sets used in this analysis are shown in Figure 1.

Significant *versus* insignificant similarity

The analysis presented here relies on the detection of structural relationships and, consequently, on the distinction of significant and insignificant similarity of protein structures. In particular we find that side chain orientation, as represented by the C^β atom positions, is needed for a proper definition of structurally equivalent residues. Often the C^α traces of two proteins show extensive similarity while side chains point in opposite directions. This is frequently observed in structures with a high content of β-sheet but less so with α-helices.

Figure 2 shows the superimposition of the ϵ-subunit of F1F0-ATP synthase from *E. coli* (1aqt) and C-terminal domain of β-galactosidase from *E. coli* (1bgl). Based on C^α superimposition the similarity is extensive (Figure 2(a)). Considering side chain orientation only one sheet of the β-sandwich can be superimposed (Figure 2(b)) and the number of equivalent residues drops from 61 (C^α) to 42 (C^α/C^β), corresponding to a difference of 31%. From the distinct function and different side chain orientations it seems likely that the ATP synthase and β-galactosidase domains are unrelated, although the topology of the C^α traces appear similar. Therefore, we classify these domains as different folds. An attempt to derive a structural model for one of these domains using the other domain as a template will be very difficult, since the orientation of many side chains has to be reversed.

On average 55% of the residues of structurally similar domains are equivalent. The respective number for unrelated pairs, including similarities of substructures and other partial similarities, is 36%. The averages differ considerably, but the corresponding distributions overlap, so that the number of equivalent residues is insufficient to separate cases of significant similarity from others. Nevertheless, many pairs having more than 50% equivalent residues, relative to the target size, share the same fold. This covers 66% of the significant similarities with an error rate of 4%. Extreme cases are 1ips.A and 1cau.A, two proteins having the same fold although they share only

23% equivalent residues. Another such case is 1kdx.A and 1vnc which have different folds but 69% equivalent residues

In this analysis we observed 8 pairs at the substructure level. Figure 3 depicts an example of a small protein which superimposes with a compact substructure of a larger protein. Ragweed pollen allergen (1bbg) is a small (40 aa) α+β fold and superimposes with the C-terminal end of profilin (1pne) with 31 equivalent residues (with C^β filter). The three strands and the helix superimpose with the alignment having three short gaps. In principle the larger protein could serve as a scaffold for modeling the smaller protein. Then, given the sequence of the smaller protein, the question is whether the structural similarity to the larger protein could be detected by prediction methods. From a previous study it is clear that this is a difficult problem (Domingues *et al.*, 1999b). In particular, many structurally equivalent pairs are exposed in one protein but buried in the other structure. Methods that take the physicochemical environment into account are likely to fail in such cases.

Partial fold similarity was observed in 8 cases. An example is the similarity of oncogene product $p14^{TCL1}$ (1jsg) and avidin (1avd). Both consist of a 8-stranded β-barrel but only 5 strands are structurally equivalent. Surprisingly, the proteins λ-exonuclease (1avq) and kinesin (2kin) are partially similar to functionally related proteins (Table 1).

Structural and functional similarities

The results of our classification are summarized in Table 2. 75% of the 196 domains have similarity on the fold level to previously known proteins. Only 49 domains of the structures released in 1998 have no clear structural similarity to a previously determined structure. The latter corresponds to 7.4% of the 664 non-redundant sequences with 92.6% being structurally similar to a previously known structure.

The SCOP data base was used to analyze functional relationships between the 196 domains and previously known structures (Table 1). However, only 120 of these domains are classified in the latest release of SCOP (August 1998). For the remaining 76 domains evidence regarding functional relationships was obtained from the literature. 97 of the 196 domains (49.5%) are functionally related to previously known proteins.

Structural similarity does not always coincide with functional relationship. On the other hand, in most cases functionally related pairs also have extensive structural similarity. Figure 4 depicts the concordance of structural and functional relationships of the domains. The functionally related domains contains three groups: for 45% the most similar structural template is functionally related, for 3% the functionally related template does not coincide with the "best" template (see below) and 1.5% are functionally related (same SCOP superfamily) but have weak structural similarity. From the functionally unrelated domains 27% have structural similarity to known folds while 23.5% are novel folds, substructures or partially similar folds. There are 94 proteins pairs which are both structurally and functionally similar corresponding to a fraction of 64% of the structurally similar proteins.

Homologous *versus* analogous proteins

Typically, a protein has structural similarity to several proteins and expert knowledge is essential to distinguish homologous from analogous proteins. A measure which distinguish homologous from analogous pairs would be advantageous. The 9 cases where and analogous protein has more equivalent residues in common with the target than a homologous protein are reduced to 5 when side chain orientation is taken into account. Also not perfect, side chain orientation helps to distinguish analogous from homologous pairs. We also encountered some instances where the consideration of side chain orientation reveals a closer homologue (SCOP family instead of superfamily), e.g. for the RNA-binding domain of the transcriptional terminator protein ρ (1a62).

An example is G:T/U mismatch-specific DNA glycosylase from *E. coli* (1mug) which is in the same SCOP superfamily as human uracil-DNA glycosylase (1akz). The C^α trace of 1mug shares higher similarity with cutinase from *Fusarium solani* (1cex, 97 equivalent residues, different folds in SCOP) than with 1akz (89 residues). When side chain orientation is taken into account the ranking is reversed: 1mug shares 77 equivalent residues with 1akz and 73 with 1cex. A similar situation is observed for the proteins interleukin-1 receptor (1itb.B), the T-domain of the brachyury transcription factor (1xbr.A), and robustoxin (1qdp).

For five proteins an analogue has higher similarity than the homologue also when side chain orientation is taken into account (Table 3). We describe two examples where side chain orientation fails to correctly discriminate homologous from analogous pairs.

The β-subunit of protein farnesyltransferase (1ft1.B) has the fold of an α-α toroid consisting of six α-helical hairpins arranged in a closed circular array. This type of fold tolerates variations in size. The functionally unrelated protein glucoamylase (1gai) has the same fold with an identical number of α-helical hairpins while the functionally related protein 5-epi-aristolochene (5eas) consists of only four α-helical hairpins. This is an extreme example where a functionally unrelated protein has extensive structure similarity whereas a functionally related protein is much smaller and, therefore, has less equivalent residues (152 with 1gai *versus* 105 with 5eas).

The second example concerns the FMN-binding protein from *Desulfovibrio vulgaris* (1axj) which forms a small β-barrel and binds the cofactor flavin mononucleotide (FMN). This protein has high similarity (67 equivalent residues) to the C-terminal domain of hepatitis A virus 3C proteinase (1hav.A) (Figure 5) which is an analogous relationship. The structural relationship of 1axj with the N-terminal domain of phthalate dioxygenase reductase from *Pseudomonas cepacia* (2pia) is a typical example of partial fold similarity. 1axj shares only 41 equivalent residues with 2pia which binds FMN in the same region as 1axj (Figure 6(a)). Both proteins are in the same SCOP Ferredoxin reductase-like superfamily. The sequence of FMN binding protein is circularly permuted relative to 2pia so that the N-terminal β-hairpin of 2pia corresponds to the C-terminal hairpin in 1axj (Figure 6(b)). The superimposition algorithm preserves sequence order so that these substructures cannot be superimposed.

Homologous proteins having weak structural similarity

There are a few examples of proteins having related function but rather dissimilar structures. Such similarities can have several origins. An example which most likely diverged by extensive permutations in the sequence is human deubiquitinating enzyme UCH-L3 (1uch). This protein belongs to the SCOP superfamily of cysteine proteinases. Its catalytic triad superimposes quite well with that of other members of this superfamily (Johnston *et al.*, 1997), however, the sequential order of the catalytic residues are different in deubiquitinating enzyme and cysteine proteinases of the papain family. The structural similarity is confined to a five-stranded β-sheet while the flanking helices are not superimposable. The main difference in topology is that the helix which accommodates the catalytic cysteine residue is located between strands two and three in deubiquitinating enzyme and at the N-terminus in papain-like cysteine proteinases. This is an example where the functional relatedness of two proteins, which most likely evolved from a common ancestor, is most difficult to detect by automated structure based methods.

Conclusion

In this paper we investigate the use of protein structure in the structural and functional characterization of protein sequences. We used the 3358 protein chains released in 1998 by the PDB to simulate a situation where a set of sequences has to be annotated. A large part of these entries is redundant in the sense that the sequence have a high percentage of identities to previously known structures. Removing the sequence redundancies and all non-globular structures, 147 proteins corresponding to 196 domains remain. This set of 196 domains is most interesting since it contains the information which is not accessible by sequence based methods. The goal of this analysis was to document the information that could be gained from these proteins if proper search tools are available.

The analysis reveals that 75% of these 196 domains have extensive similarity to previously determined structures. These 75% correspond to the maximum amount of structural information that could be retrieved from a data base if a perfect technique is available. Most likely this amount will increase with the growing number of available protein structures. Two thirds of structurally similar proteins are also functionally related. With respect to the 196 domains this corresponds to 50% of functional coincidences.

In a recent study (Orengo *et al.*, 1999) obtain a much higher percentage of 83% of functional coincidences. Our analysis differs in several aspects. Their cutoff for significant sequence similarity is 30% sequence identity. However, as was discussed by Brenner *et al.* (1998) the use of extreme value statistics reveals significant similarities way below this threshold. Therefore, their estimate of 83% contains cases that are considered to be homologous by these techniques. Since we used FastaA E-values in this analysis the number of homologous proteins appear to be significantly smaller.

The question to what extent current techniques are able to use the available information is most interesting. A reasonable estimate requires a suitable benchmark. Fortunately, the recent CASP experiment (Koehl & Levitt, 1999; Sippl, 1999) provides an invaluable estimate of the current state of the art in structure prediction. In particular, the results obtained in the fold recognition category are relevant in view of the current analysis. These techniques predict structures using fold data bases. In the last CASP experiment there were 21 targets in the fold recognition category, i.e. sequences having no significant sequence similarity to proteins of known structure. Hence, the CASP experiment yields an estimate for the amount of structural information that can be obtained from prediction methods which employ fold data bases.

19 of the 21 fold recognition targets are similar to previously known structures and 42% of these are remote homologues (Murzin, 1999), a number which is similar to the 50% homologues obtained here. Therefore, the prediction success observed in CASP3 can be used to estimate how many of the structural coincidences observed for the 1998 proteins could be obtained from current structure prediction methods. The success rate is in the order of 30 to 40% (Murzin, 1999). Hence, there is an enormous potential to increase the amount of information that can be used for the structural and functional characterization of novel proteins by using and improving these techniques.

These results are relevant in light of the structural genomics initiatives whose goal is to determine at least one member of all biologically relevant protein families. On the one hand, the structure data bases created in these projects will enable template based prediction methods to compute approximate structures for virtually every protein of interest, at least in principle. On the other hand the prediction methods will be an asset for structure determination, since they are instrumental in identifying proteins which most likely have a novel fold.

Large scale annotations of whole proteomes indicate that currently sufficiently accurate models can be built for 8 to 17% of the sequences based on sequence comparison and comparative modeling (Andrade *et al.*, 1999). This limit can be further extended when all remote sequence-structure relationships are exploited. To estimate this range we assume that the fold types in whole proteomes are distributed like the domains in PDB. Then the fraction of transmembrane and other non-globular proteins is approximately 20% (Frishman & Mewes, 1999). The estimated number of proteins unrelated in sequence but having significant similarity to a known structure is $(100 - 17 - 20) \times 0.75 = 47\%$. Two thirds of these will be functionally related according to the results obtained here (Figure 7).

Materials and Methods

Preparation of dataset

The data sets for the presented analysis were extracted from the structures released by the PDB in 1998 where we regarded the time stamp of the file at the PDB server as the release date. (This is the date where a structure becomes available for public use. In some cases it might happen that there is a long delay between publication and release.) These sequences were filtered with a threshold of 95% sequence identity to remove identical and highly similar sequences (mutants) to derive a set of unique sequences. This is necessary because PDB files may contain several instances of the same sequence and because experimentalist often deposit slightly different structures of the same protein, like ligated and non-ligated forms. These unique sequences were compared with the sequences in the PDB at the end of 1997 using FastA (Pearson, 1998; Pearson, 1996) and all sequences which has significant similarity (E-value $<$ 0.01) to an already existing protein structure were removed from the set. From the resulting set all chains with non-compact structure, transmembrane proteins and virus capsid proteins were also eliminated. All structures of the final data set were split into domains according to information from SCOP (Murzin *et al.*, 1995), CATH (Orengo *et al.*, 1997) or publications of the experimentalists (Table 1).

Structure comparison

All structure comparisons were performed with the program ProSup which implements rigid-body superimposition of two proteins (Feng & Sippl, 1996). Parameters were set as described by Lackner *et al.* (1999). ProSup measures the extent of similarity by the number of structurally equivalent residues. Two residues are considered as equivalent if their C^α atoms are closer than 5Å after superimposition. The rms deviation of equivalent residues is held approximately constant by ProSup in the range of 2 to 3Å. Side chain orientation is represented by the C^β atom positions. When side chain orientation is incorporated into structure superimposition both C^α and C^β atoms have to be closer than 5Å after optimal superimposition of the C^α trace. One feature of ProSup is to generate a list of alternative alignments. For the present analysis only the alignments with the highest number of equivalent residues are regarded for evaluation.

Each domain of the data set was compared with the structures of the PDB at the time of its release. To save computing time and to avoid too many structure libraries we chose the following strategy: three snapshots (Jan 27, May 15, and Sep 9) of the PDB with less than 40% sequence identity were extracted and the domains of the data set were compared with the entries of the current snapshot at the time of release. The results of the ProSup data base searches (a list of protein chains sorted by the number of equivalent residues) were inspected by eye. The selected "best" template was normally the chain with highest number of equivalent residues. Only in cases, where another chain gave a more reasonable alignment (shorter gaps, more compact in three dimensions, etc.) this one was considered as the "best" template.

Acknowledgments

We are grateful to Alessandro Monge and Francisco S. Domingues for valuable suggestions to improve the manuscript. This work was supported by grants P11601-GEN, P11205-MOB and P13710-MOB of the Austrian *Fonds zur Förderung der wissenschaflichen Forschung*.

References

Altschul, S. F., Gisch, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acid Res.* **25**, 3389–3402.

Andrade, M. A., Brown, N. P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C., & Sander, C. (1999). Automated genome sequence analysis and annotation. *Bioinformatics,* **15**, 391–412.

Bayer, P., Arndt, A., Metzger, S., Mahajan, R., Melchior, F., Jaenicke, R., & Becker, J. (1998). Structure determination of the small ubiquitin-related modifier SUMO-1. *J. Mol. Biol.* **280**, 275–286.

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

Bowie, J. U., Lüthy, R., & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–170.

Brenner, S. E., Chothia, C., & Hubbard, T. J. (1998). Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci. USA,* **95**, 6073–6078.

Bryant, S. H. (1996). Evaluation of threading specificity and accuracy. *Proteins: Struct. Funct. Genet.* **26**, 172–185.

Chen, X., Vinkemeier, U., Zhao, Y., Jeruzalmi, D., Darnell Jr, J. E., & Kuriyan, J. (1998). Crystal structure of a tyrosine phosphorylated STAT-1 dimer bound to DNA. *Cell,* **93**, 827–839.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature,* **357**, 543–544.

Clausen, T., Huber, R., Laber, B., Pohlenz, H. D., & Messerschmidt, A. (1996). Crystal structure of the pyridoxal-5'-phosphate dependent cystathionine β-lyase from *Escherichia coli* at 1.83 Å. *J. Mol. Biol.* **262**, 202–224.

Domingues, F. S., Koppensteiner, W. A., Jaritz, M., Prlic, A., Weichenberger, C., Wiederstein, M., Flöckner, H., Lackner, P., & Sippl, M. J. (1999a). Sustained performance of knowledge-based potentials in fold recognition. *Proteins: Struct. Funct. Genet.* , in press.

Domingues, F. S., Lackner, P., & Sippl, M. J. (1999b). Structure based evaluation of sequence comparison and fold recognition alignment accuracy. *Submitted,* .

Fan, Q. R., Mosyak, L., Winter, C. C., Wagtmann, N., Long, E. O., & Wiley, D. C. (1997). Structure of the inhibitory receptor for human natural killer cells resembles haematopoietic receptors. *Nature,* **389**, 96–100.

Feng, Z.-K. & Sippl, M. J. (1996). Optimum superimposition of protein structures: ambiguities and implications. *Fold. Des.* **1**, 123–132.

Finnin, M. S., Cicero, M. P., Davies, C., Porter, S. J., White, S. W., & Kreuzer, K. N. (1997). The activation domain of the MotA transcription factor from bacteriophage T4. *EMBO. J,* **16**, 1992–2003.

Flores, T. P., Orengo, C. A., Moss, D. S., & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.* **2**, 1811–1826.

Frishman, D. & Mewes, H.-W. (1999). Genome-based structural biology. *Prog. Biophys. Mol. Biol.* **72**, 1–17.

Gallagher, D. T., Gilliland, G. L., Xiao, G., Zondlo, J., Fisher, K. E., Chinchilla, D., & Eisenstein, E. (1998). Structure and control of pyridoxal phosphate dependent allosteric threonine deaminase. *Structure,* **6**, 465–475.

Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R., & Hogle, J. M. (1998). The crystal structure of Dps, a ferritin homolog that binds and protects DNA. *Nature Struct. Biol.* **5**, 294–303.

Guo, F., Gopaul, D. N., & van Duyne, G. D. (1997). Structure of Cre recombinase complexed with DNA in a site-specific recombination synapse. *Nature,* **389**, 40–46.

Hansen, J. L., Long, A. M., & Schultz, S. C. (1997). Structure of the RNA-dependent RNA polymerase of poliovirus. *Structure,* **15**, 1109–1122.

Hegyi, H. & 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.

Hester, G., Stark, W., Moser, M., Kallen, J., Markovic-Housley, Z., & Jansonius, J. N. (1999). Crystal structure of phosphoserine aminotransferase from *Escherichia coli* at 2.3 Å resolution: comparison of the unligated enzyme and a complex with α-methyl-L-glutamate. *J. Mol. Biol.* **286**, 829–850.

Hofmann, B., Schomburg, D., & Hecht, H. J. (1993). Crystal structure of a thiol proteinase from *Staphylococcus aureus* V-8 in the E-64 inhibitor complex. *Acta Cryst. A,* **49**, 102.

Holm, L. & Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123–138.

Holm, L. & Sander, C. (1997). New structure — novel fold? *Structure,* **5**, 165–171.

Hunt, J. F., van der Vies, S. M., Henry, L., & Deisenhofer, J. (1997). Structural adaptations in the specialized bacteriophage T4 co-chaperonin Gp31 expand the size of the Anfinsen cage. *Cell,* **90**, 361–370.

Igarashi, N., Moriyama, H., Fujiwara, T., Fukumori, Y., & Tanaka, N. (1997). The 2.8 Å structure of hydroxylamine oxidoreductase from a nitrifying chemoautotrophic bacterium, Nitrosomonas europaea. *Nature Struct. Biol.* **4**, 276–284.

Johnston, S. C., Larsen, C. N., Cook, W. J., Wilkinson, K. D., & Hill, C. P. (1997). Crystal structure of a deubiquitinating enzyme (human UCH-L3) at 1.8 Å resolution. *Embo J.* **16**, 3787–3796.

Jones, D. T. (1997). Progress in protein structure prediction. *Curr. Opin. Struct. Biol.* **7**, 377–387.

Jones, D. T. (1999). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Kakuta, Y., Pedersen, L. G., Carter, C. W., Negishi, M., & Pedersen, L. (1997). Crystal structure of estrogen sulphotransferase. *Nature Struct. Biol.* **4**, 904–908.

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden Markov models for detecting remote protein homologies. *Bioinformatics,* **14**, 846–856.

Kim, H. Y. & Cho, Y. (1997). Structural similarity between the pocket region of retinoblastoma tumour suppressor and the cyclin-box. *Nature Struct. Biol.* **4**, 390–395.

Koehl, P. & Levitt, M. (1999). A brighter future for protein structure prediction. *Nature Struct. Biol.* **6**, 108–111.

Kraulis, P. J. (1991). Molscript: A program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.

Krell, T., Horsburgh, M. J., Cooper, A., Kelly, S. M., & Coggins, J. R. (1996). Localization of the active site of type II dehydroquinases. Identification of a common arginine-containing motif in the two classes of dehydroquinases. *J. Biol. Chem.* **271**, 24492–24497.

Lackner, P., Koppensteiner, W. A., Domingues, F. S., & Sippl, M. J. (1999). Recurrent problems in protein structure comparison. *In preparation,* .

Lenzen, C. U., Steinmann, D., Whiteheart, S. W., & Weis, W. I. (1998). Crystal structure of the hexamerization domain of N-ethylmaleimide-sensitive fusion protein. *Cell,* **94**, 525–536.

Liao, D. I., Wolff, E. C., Park, M. H., & Davies, D. R. (1998). Crystal structure of the NAD complex of human deoxyhypusine synthase: an enzyme with a ball-and-chain mechanism for blocking the active site. *Structure,* **6**, 23–32.

Liepinsh, E., Kitamura, M., Murakami, T., Nakaya, T., & Otting, G. (1997). Pathway of chymotrypsin evolution suggested by the structure of the FMN-binding protein from *Desulfovibrio vulgaris*. *Nature Struct. Biol.* **4**, 975–979.

Liepinsh, E., Kitamura, M., Murakami, T., Nakaya, T., & Otting, G. (1998). Common ancestor of serine proteases and flavin-binding domains. *Nature Struct. Biol.* **5**, 102–103.

Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B. Taroni, C., & Thornton, J. M. (1998). Protein folds and functions. *Structure,* **15**, 875–884.

Matsuo, Y. & Bryant, S. H. (1999). Identification of homologous core structures. *Proteins: Struct. Funct. Genet.* **35**, 70–79.

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380–387.

Murzin, A. G. (1999). Structure classification-based assessment of CASP3 predictions for the fold recognition targets. *Proteins: Struct. Funct. Genet.* , in press.

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

Musco, G., Kharrat, A., Stier, G., Fraternali, F., Gibson, T. J., Nilges, M., & Pastore, A. (1997). The solution structure of the first KH domain of FMR1, the protein responsible for the fragile X syndrome. *Nature Struct. Biol.* **4**, 712–716.

Nogales, E., Wolf, S. G., & Downing, K. H. (1998). Structure of the $\alpha\beta$ tubulin dimer by electron crystallography. *Nature,* **391**, 199–203.

Orengo, C. A., Jones, D. T., & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631–634.

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

Orengo, C. A., Pearl, F. M., Bray, J. E., Todd, A. E., Martin, A. C., Lo Conte, L., & Thornton, J. M. (1999). The CATH Database provides insights into protein structure/function relationships. *Nucl. Acid Res.* **27**, 275–279.

Pappa, H., Murray-Rust, J., Dekker, L. V., Parker, P. J., & McDonald, N. Q. (1998). Crystal structure of the C2 domain from protein kinase C-δ. *Structure,* **6**, 885–894.

Park, H. W., Boduluri, S. R., Moomaw, J. F., Casey, P. J., & Beese, L. S. (1997). Crystal structure of protein farnesyltransferase at 2.25 Angstrom resolution. *Science,* **275**, 1800–1804.

Pastore, A. & Lesk, A. M. (1990). Comparison of the structures of globins and phycocyanins: evidence for evolutionary relationship. *Proteins: Struct. Funct. Genet.* **8**, 133–155.

Pearson, W. R. (1996). Effective protein sequence comparison. *Meth. Enzymol.* **266**, 227–258.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *J. Mol. Biol.* **276**, 71–84.

Peat, T., Newman, J., Waldo, G. S., Berendzen, J., & Terwilliger, T. C. (1998). Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution. *Structure,* **6**, 1207–1214.

Pickersgill, R., Smith, D., Worboys, K., & Jenkins, J. (1998). Crystal structure of polygalacturonase from *Erwinia carotovora ssp. carotovora*. *J. Biol. Chem.* **273**, 24660–24664.

Ranganathan, R., Lu, K. P., Hunter, T., & Noel, J. P. (1997). Structural and functional analysis of the mitotic rotamase Pin1 suggests substrate recognition is phosphorylation dependent. *Cell,* **89**, 875–886.

Rowen, L., Mahairas, G., & Hood, L. (1997). Sequencing the human genome. *Science,* **278**, 605–607.

Russell, R. B. & 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., Sasieni, P. D., & Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903–918.

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Schumacher, S., Clubb, R. T., Cai, M., Mizuuchi, K., Clore, G. M., & Gronenborn, A. M. (1997). Solution structure of the Mu end DNA-binding Iβ subdomain of phage Mu transposase: modular DNA recognition by two tethered domains. *EMBO J.* **16**, 7532–7541.

Sippl, M. J. (1999). Who solved the protein folding problem? *Structure,* **7**, R81–R83.

Sippl, M. J. & Weitckus, S. (1992). Detection of native like models for amino acid sequences of unknown three dimensional structure in a data base of known protein conformations. *Proteins: Struct. Funct. Genet.* **13**, 258–271.

Sulzenbacher, G., Shareck, F., Morosoli, R., Dupont, C., & Davies, G. J. (1997). The *Streptomyces lividans* family 12 endoglucanase: construction of the catalytic core, expression, and X-ray structure at 1.75 Å resolution. *Biochemistry,* **36**, 16032–16039.

Taylor, W. R. & Orengo, C. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.

TIGR (1999). The Institute for Genomic Research. http://www.tigr.org.

Tomchick, D. R., Turner, R. J., Switzer, R. L., & Smith, J. L. (1998). Adaptation of an enzyme to regulatory function: structure of *Bacillus subtilis* PyrR, a *pyr* RNA-binding attenuation protein and uracil phosphoribosyltransferase. *Structure,* **6**, 337–350.

Van Asselt, E. J., Perrakis, A., Kalk, K. H., Lamzin, V. S., & Dijkstra, B. W. (1998). Accelerated X-ray structure elucidation of a 36 kDa muramidase/transglycosylase using wARP. *Acta Cryst. D,* **54**, 58–73.

Yao, N., Hesson, T., Cable, M., Hong, Z., Kwong, A. D., Le, H. V., & Weber, P. C. (1997). Structure of the hepatitis C virus RNA helicase domain. *Nature Struct. Biol.* **4**, 463–467.

Zhang, Z., Huang, L., Shulmeister, V. M., Chi, Y. I., Kim, K. K., Hung, L. W., Crofts, A. R., Berry, E. A., & Kim, S. (1998). Electron transfer by domain movement in cytochrome bc_1. *Nature,* **392**, 677–684.

Table legends

Table 1: Structural and functional relationships of all data set domains. The columns indicate the target name (pdb-id.chain-id.model-id) with eventual domain boundaries, the length of the domain, the name of the "best" template, the absolute number of equivalent residues and relative to the target domain length, the assigned level of structural similarity (fold, partial, substructure, novel; see text for explanation) and the functional relationship of target and template. The *Homologue* column specifies level and source of functional relationship. This could be SCOP *superfamily* or *family* or information obtained from the structure publications (*yes*). In the latter case the footnotes refer to the subsequently listed references. A 'yes' in parenthesis means that a homologue exists but an analogue with higher similarity is the best template. 1:(Tomchick *et al.*, 1998) 2:(Bayer *et al.*, 1998) 3:(Kim & Cho, 1997) 4:(Kakuta *et al.*, 1997) 5:(Liepinsh *et al.*, 1997; Liepinsh *et al.*, 1998) 6:(Zhang *et al.*, 1998) 7:(Pappa *et al.*, 1998) 8:(Chen *et al.*, 1998) 9:(Pickersgill *et al.*, 1998) 10:(Finnin *et al.*, 1997) 11:(Hester *et al.*, 1999) 12:(Peat *et al.*, 1998) 13:(Clausen *et al.*, 1996) 14:(Guo *et al.*, 1997) 15:(Hofmann *et al.*, 1993) 16:(Lenzen *et al.*, 1998) 17:(Liao *et al.*, 1998) 18:(Grant *et al.*, 1998) 19:(Igarashi *et al.*, 1997) 20:(Park *et al.*, 1997) 21:(Hunt *et al.*, 1997) 22:(Yao *et al.*, 1997) 23:(Van Asselt *et al.*, 1998) 24:(Fan *et al.*, 1997) 25:(Sulzenbacher *et al.*, 1997) 26:(Ranganathan *et al.*, 1997) 27:(Hansen *et al.*, 1997) 28:(Krell *et al.*, 1996) 29:(Gallagher *et al.*, 1998) 30:(Nogales *et al.*, 1998) 31:(Schumacher *et al.*, 1997) 32:(Musco *et al.*, 1997)

Table 2: Refer to the text for explanations of the similarity levels. Columns 2 to 5 indicates the absolute and relative numbers of instances of a particular similarity level found in the data set, both for C^α superimposition (columns 2 and 3) and C^α/C^β superimposition (columns 4 and 5).

Table 3: Proteins which have an analogous structure with higher similarity than a potential homologue. The analogue is always specified before the homologue. In

three cases (1dhs, 2ezk, and 2fmr) the difference in equivalent residues is only marginal ($\approx 10\%$) while in the other cases the analogous have a substantially higher number of equivalent residues (see text for a discussion). The footnotes refer to the subsequently listed references. 1: (Liepinsh *et al.*, 1997; Liepinsh *et al.*, 1998) 2: (Liao *et al.*, 1998) 3: (Park *et al.*, 1997) 4: (Schumacher *et al.*, 1997) 5: (Musco *et al.*, 1997)

Figure legends

Figure 1: Preparation of data sets.

Figure 2: Comparison of the superimposition of the ϵ-subunit of F1F0-ATP synthase from *E. coli* (1aqt, left) and the C-terminal domain of β-galactosidase from *E. coli* (1bgl, right). Structurally equivalent residues are dark gray. (a) C^{α} superimposition: strands of both sheets of the sandwich are equivalent. (b) C^{α}/C^{β} superimposition: only the strands of one sheet are equivalent because side chain orientations are different. All ribbon representations were produced with the program Molscript (Kraulis, 1991).

Figure 3: The small protein ragweed pollen allergen (1bbg, 40 aa) superimposed with a substructure of profilin (1pne, 140 aa). All secondary structure elements of 1bbg superimposes with the C-terminal part of 1pne (residues 84 to 129, 31 equivalent residues). Stretches of structurally equivalent residues are dark gray.

Figure 4: Coincidence of structural and functional similarities of the data set domains. Red sector: domains which have extensive structural and functional relationship to a known structure; orange sector: domains where a template with functional relationship exists but an analogous proteins has highest structural similarity; yellow sector: domains having a structural template with different function; magenta sector: domains which are classified in the same SCOP superfamily without extensive structural similarity; blue sectors: domains with a novel fold including

substructures (light blue) and partial fold similarity (very light blue).

Figure 5: Superimposition of FMN-binding protein from *Desulfovibrio vulgaris* (1axj) with the C-terminal domain of hepatitis A virus 3C proteinase (1hav.A). All strands of the core have structurally equivalent counterparts (67 equivalent residues, dark gray).

Figure 6: (a) Superimposition of FMN-binding protein from *Desulfovibrio vulgaris* (1axj) with the N-terminal domain of phthalate dioxygenase reductase form *Pseudomonas cepacia* (2pia). Both structures are ligated with the cofactor FMN in a similar orientation. Four of the six strands are shared by both structures (41 equivalent residues, colored red). The green colored hairpins are in the same structural position but not structurally equivalent because the sequences are circularly permuted. (b) Alignment of the structure superimposition. Helices and strands are depicted by blue boxes and gray arrows, respectively.

Figure 7: Current limits of structural and functional inference for proteomes inferred from structure data bases. The blue sector indicates the fraction of structures which can be inferred from pairwise sequence comparison (17%) while the gray sector corresponds to the estimated percentage of transmembrane and non-globular proteins (20%). 16% of proteomes are most probably novel folds (white sector). The red sector shows the fraction of proteins which are both structurally and functionally similar to an already determined protein structure (31%). The yellow sector represents the proteins with structural similarity to a known structure (16%).

Abbreviations

FMN	flavin mononucleotide
PDB	Protein Data Bank
rms	root mean square
SCOP	structural classification of proteins

Table 1: Structural and functional relationships of the data set domains

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1a0b.-.-		125	1cgo.-.-	75	60.0	fold	
1a0i.-.-	2-240	239	1ckm.A.-	126	52.7	fold	superfamily
1a0i.-.-	241-349	109	1ah9.-.1	52	47.7	fold	superfamily
1a17.-.-		166	1sly.-.-	93	56.0	fold	
1a2z.A.-		220	1ecp.A.-	128	58.2	fold	
1a3c.-.-		181	1hgx.A.-	136	75.1	fold	yes[1]
1a5r.-.1		103	1ubi.-.-	63	61.2	fold	yes[2]
1a5t.-.-	1-167	167	2reb.-.-	104	62.3	fold	superfamily
1a5t.-.-	168-330	163	1wer.-.-	71	43.6	novel	
1a62.-.-		130	1mjc.-.-	49	37.7	fold	family
1a68.-.-		95	1tfr.-.-	39	41.1	novel	
1a6q.-.-	2-296	295	1pma.1.-	83	28.1	partial	
1a6q.-.-	297-368	72	1bgw.-.-	53	73.6	fold	
1a74.A.-		163	1kit.-.-	39	23.9	novel	
1a7j.-.-		290	2ak3.A.-	105	36.2	fold	superfamily
1aa2.-.-		108	1aoa.-.-	96	88.9	fold	family
1acc.-.-	14-258	245	1msp.A.-	62	25.3	fold	
1acc.-.-	259-485	227	1rho.A.-	62	27.3	fold	
1acc.-.-	486-604	119	1gua.B.-	49	41.2	fold	
1acc.-.-	605-735	131	1rho.A.-	56	42.7	fold	
1ad1.A.-		266	1fdy.D.-	161	60.5	fold	
1ad6.-.-		185	1vin.-.-	88	47.6	fold	yes[3]
1af7.-.-	11-90	90	1pnk.B.-	54	60.0	fold	
1af7.-.-	91-284	194	1vid.-.-	107	55.2	fold	superfamily
1af9.-.-	875-1109	235	1sac.A.-	118	50.2	fold	superfamily

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1af9.-.-	1110-1315	206	1tie.-.-	128	62.1	fold	superfamily
1ah1.-.1		129	1tcr.A.-	89	69.0	fold	family
1ahj.A.-		207	1rom.-.-	55	26.6	novel	
1ahj.B.-	1-112	112	1cpc.A.-	48	42.9	novel	
1ahj.B.-	113-212	100	1pse.-.1	49	49.0	fold	
1ahk.-.-		129	1fie.A.-	50	38.8	fold	family
1aiw.-.1		62	1fgp.-.1	31	50.0	partial	
1aj6.-.-		219	1yer.-.-	119	54.3	fold	superfamily
1al0.1.-		152	2cae.-.-	51	33.6	novel	
1am2.-.-		199	1at0.-.-	124	62.3	fold	superfamily
1amx.-.-		180	1anu.-.-	71	39.4	fold	
1ao6.A.-	1-190	190	1ar1.A.-	62	32.6	novel	
1ao6.A.-	191-382	192	1rom.-.-	69	35.9	novel	
1ao6.A.-	383-582	200	1ddt.-.-	63	31.5	novel	
1aos.A.-	1-112	112	1fur.A.-	60	53.6	fold	superfamily
1aos.A.-	113-362	250	1fur.A.-	232	92.8	fold	superfamily
1aos.A.-	363-462	100	2abk.-.-	55	55.0	fold	
1ap0.-.1		73	1sap.-.-	42	57.5	fold	superfamily
1ap8.-.1		213	1vao.A.-	71	33.3	partial	
1apj.-.1		74	1uby.-.-	28	37.8	novel	
1aqt.-.-		138	1xsm.-.-	52	37.7	novel	
1aqu.A.-		297	1gky.-.-	97	32.7	fold	yes[4]
1auk.-.-		489	1alk.A.-	168	34.4	fold	superfamily
1auo.A.-		218	1tht.A.-	147	67.4	fold	superfamily
1auv.A.-		311	1gsa.-.-	195	62.7	fold	superfamily
1auz.-.1		116	1wab.-.-	70	60.3	fold	
1avq.A.-		228	1cfr.-.-	64	28.1	partial	superfamily
1aw5.-.-		326	1ad1.A.-	165	50.6	fold	

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1aw8.B.-		91	1cxs.A.-	53	58.2	fold	superfamily
1axj.-.1		122	1hav.A.-	67	54.9	fold	(yes)[5]
1ay2.-.-		158	1bcp.A.-	47	29.7	novel	
1azo.-.-		232	1pvu.A.-	67	28.9	fold	superfamily
1bak.-.1		119	1btk.B.-	82	68.9	fold	family
1baq.-.1		139	1ivh.A.-	60	43.2	substructure	
1bbg.-.-		40	1pne.-.-	31	77.5	substructure	
1bcc.A.-	4-234	231	1bia.-.-	76	32.9	fold	
1bcc.A.-	235-445	211	1azs.B.-	69	32.7	fold	
1bcc.D.-		241	1cxc.-.-	69	28.6	fold	yes[6]
1bd9.A.-		187	1rlw.-.-	58	31.0	fold	
1bdy.A.-		123	1rlw.-.-	85	69.1	fold	yes[7]
1bf5.A.-	136-317	182	1dlc.-.-	103	56.6	fold	
1bf5.A.-	318-490	173	1a02.N.-	87	50.3	fold	yes[8]
1bf5.A.-	491-710	220	1sha.A.-	69	31.4	fold	yes[8]
1bg8.A.-		89	1ao6.A.-	56	62.9	fold	
1bgf.-.-		124	2brd.-.-	63	50.8	novel	
1bhe.-.-		376	1rmg.-.-	274	72.9	fold	yes[9]
1bja.A.-		95	1smt.A.-	72	75.8	fold	yes[10]
1bjn.A.-	1-258	258	1cl1.A.-	137	53.1	fold	yes[11]
1bjn.A.-	259-362	104	2dkb.-.-	84	80.8	fold	yes[11]
1bkb.-.-	4-75	72	1vie.-.-	46	63.9	fold	
1bkb.-.-	76-139	64	1ah9.-.1	55	85.9	fold	yes[12]
1bnl.A.-		178	1ixx.A.-	56	31.5	novel	
1bqv.-.1		110	1rlr.-.-	51	46.4	fold	
1brz.-.1		54	2sn3.-.-	39	72.2	fold	superfamily
1c25.-.-		161	1rhs.-.-	84	52.2	fold	
1c3d.-.-		294	1sqc.-.-	188	63.9	fold	

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1cl1.A.-	1-256	256	1ajs.A.-	155	60.5	fold	yes[13]
1cl1.A.-	257-395	139	2dkb.-.-	80	57.6	fold	yes[13]
1crx.A.-	20-130	111	1a0p.-.-	65	58.6	fold	yes[14]
1crx.A.-	131-341	211	1ae9.A.-	121	57.3	fold	yes[14]
1cv8.-.-	2-79	78	1ppn.-.-	52	66.7	fold	yes[15]
1cv8.-.-	80-174	95	1the.A.-	65	68.4	fold	yes[15]
1d2n.A.-	505-676	172	1a5t.-.-	120	69.8	fold	yes[16]
1d2n.A.-	677-750	74	1maz.-.-	48	64.9	substructure	
1dfx.-.-		125	1f13.A.-	59	47.2	fold	
1dhs.-.-		344	2mas.A.-	98	28.5	fold	(yes)[17]
1dps.A.-		167	1aew.-.-	128	76.6	fold	yes[18]
1e2a.A.-		105	1aj3.-.1	74	70.5	fold	
1fgj.A.-		546	1occ.A.-	92	16.8	novel	(yes)[19]
1fsz.-.-	23-231	209	1hdc.D.-	119	56.9	fold	
1fsz.-.-	232-356	125	1com.K.-	67	53.6	fold	
1ft1.A.-		377	1lrv.-.-	80	21.2	fold	
1ft1.B.-		437	1gai.-.-	152	34.8	fold	(yes)[20]
1ftr.A.-	1-17, 163-296	151	1cg2.A.-	65	43.0	fold	
1ftr.A.-	18-162	145	1cg2.A.-	70	48.3	fold	
1g31.A.-		111	1aon.O.-	61	55.0	fold	yes[21]
1gc1.G.-		321	1ord.A.-	47	14.6	novel	
1hei.A.-	187-323	136	1pjr.-.-	88	64.7	fold	yes[22]
1hei.A.-	324-485	161	1pjr.-.-	70	43.5	fold	yes[22]
1hei.A.-	486-629	144	1ah7.-.-	50	34.7	novel	
1hkb.A.-	16-476	461	1gla.G.-	175	38.0	fold	superfamily
1hkb.A.-	477-914	438	1gla.G.-	174	39.7	fold	superfamily
1hp8.-.-		68	1sly.-.-	47	69.1	substructure	

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1hus.-.-		155	1rlr.-.-	52	33.5	novel	
1if1.A.-		113	1pue.E.-	67	59.3	fold	superfamily
1ihp.-.-		438	1rpa.-.-	202	46.1	fold	superfamily
1ips.A.-		331	1cau.A.-	76	23.0	fold	
1itb.B.-	1-100	100	2ncm.-.1	71	71.0	fold	family
1itb.B.-	101-204	104	2ncm.-.1	73	70.2	fold	family
1itb.B.-	205-315	111	1tlk.-.-	91	82.0	fold	family
1jdw.-.-		423	1esc.-.-	57	13.5	novel	
1jfr.A.-		262	1bro.A.-	160	61.1	fold	superfamily
1jsg.-.-		114	1avd.A.-	46	40.4	partial	
1kdx.A.1		81	1vnc.-.-	52	64.2	substructure	
1kwa.A.-		88	1pdr.-.-	76	86.4	fold	family
1ltm.-.-	42-264	223	154l.-.-	74	33.2	fold	yes[23]
1ltm.-.-	265-361	97	1csh.-.-	44	45.4	novel	
1lxd.A.-		100	1gua..-.	65	65.0	fold	family
1mb1.-.-		130	1apm.E.-	46	35.4	fold	
1moq.-.-	243-450	208	1fsz.-.-	96	46.2	fold	
1moq.-.-	451-608	158	1bmt.A.-	82	51.9	fold	
1mpg.A.-	1-99	99	1ytb.A.-	56	56.6	fold	superfamily
1mpg.A.-	100-282	183	2abk.-.-	130	71.0	fold	superfamily
1mro.A.-	102-276	175	1bdp.-.-	57	32.6	novel	
1mro.A.-	277-549	273	1rom.-.-	76	27.8	novel	
1mro.B.-	1-43, 181-442	305	1aj8.A.-	69	22.6	novel	
1mro.B.-	44-180	137	4rnp.A.-	64	46.7	substructure	
1mro.C.-		247	1vhi.A.-	59	23.9	novel	
1mrp.-.-	1-101, 226-277	153	1dmb.-.-	123	80.4	fold	family

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1mrp.-.-	102-235, 278-309	166	1sbp.-.-	110	66.3	fold	family
1mug.A.-		168	1akz.-.-	77	45.8	fold	superfamily
1ned.A.-		183	1pma.1.-	156	85.2	fold	family
1nkr.-.-	6-101	96	1zxq.-.-	73	76.0	fold	yes[24]
1nkr.-.-	102-200	99	1ac6.A.-	76	76.8	fold	yes[24]
1nlr.-.-		234	1xyo.A.-	133	56.8	fold	yes[25]
1noc.A.-		388	2min.B.-	54	13.9	novel	
1np1.A.-		184	1aqb.-.-	124	67.4	fold	family
1pfo.-.-	30-388	359	1auk.-.-	55	15.3	novel	
1pfo.-.-	389-500	112	1tsh.A.-	52	46.4	fold	
1phm.-.-	200-354	155	1ahs.A.-	65	41.9	fold	
1phm.-.-	45-199	155	1cwp.A.-	65	41.9	fold	
1pin.A.-		163	1fkb.-.-	41	25.2	fold	yes[26]
1poi.A.-		317	1dea.B.-	75	23.7	fold	
1poi.B.-		260	1qor.A.-	83	31.9	fold	
1prx.A.-		224	1gp1.A.-	111	49.6	fold	superfamily
1ps1.A.-		337	5eas.-.-	215	63.8	fold	superfamily
1qdp.-.1		42	1agg.-.1	30	71.4	fold	family
1rdr.-.-	67-367	301	1mml.-.-	65	21.6	fold	yes[27]
1rdr.-.-	12-37, 368-461	120	2lbd.-.-	50	41.7	novel	
1rkd.-.-		309	1nah.-.-	100	32.4	partial	
1rmg.-.-		422	1tyu.-.-	216	51.2	fold	
1ryp.2.-		233	1pma.1.-	184	79.0	fold	family
1sfp.-.-		114	2bpa.2.-	70	61.4	fold	
1shk.A.-		173	1ukz.-.-	108	62.4	fold	yes[28]
1skn.P.-		92	1cnt.1.-	50	54.3	novel	

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
1tdj.-.-	5-335	331	2tys.B.-	247	74.6	fold	yes[29]
1tdj.-.-	336-497	162	1psd.A.-	63	38.9	fold	
1tmk.A.-		216	1kin.A.-	132	61.1	fold	family
1toh.-.-		343	1jet.A.-	60	17.5	novel	
1tub.A.-	1-205	205	1fsz.-.-	127	62.0	fold	yes[30]
1tub.A.-	206-381	176	1fsz.-.-	109	61.9	fold	yes[30]
1tyf.A.-		193	1nzy.A.-	131	67.9	fold	superfamily
1uag.-.-	1-93	93	1mio.B.-	75	80.6	fold	
1uag.-.-	94-298	205	1rkd.-.-	105	51.2	fold	
1uag.-.-	299-437	139	1iso.-.-	74	53.2	fold	
1uch.-.-		230	1the.A.-	55	23.9	novel	superfamily
1uea.B.-	1-107	107	1tii.D.-	54	50.5	fold	
1uea.B.-	108-181	74	3grs.-.-	36	48.6	novel	
1uox.-.-		295	1gtq.A.-	84	28.5	fold	
1vde.A.-	1-180, 416-454	219	1at0.-.-	117	53.4	fold	superfamily
1vde.A.-	181-298	118	1af5.-.-	66	55.9	fold	superfamily
1vde.A.-	299-415	117	1af5.-.-	69	59.0	fold	superfamily
1vgh.-.1		55	2mbr.-.-	32	58.2	novel	
1vub.A.-		101	1vie.-.-	38	37.6	fold	
1wja.A.-		47	2cgp.A.-	35	74.5	fold	
1xat.-.-	3-163	161	1tdt.A.-	89	55.3	fold	superfamily
1xat.-.-	164-210	47	1dkx.A.-	42	89.4	fold	
1xbr.A.-		184	1svc.P.-	71	38.6	fold	superfamily
1ygs.-.-		234	1a25.A.-	52	22.2	partial	
1yub.-.-	1-181	181	1vpt.-.-	103	56.9	fold	family
1yub.-.-	182-245	64	1ddf.-.-	42	65.6	fold	
2ezk.-.-		99	1jhg.A.-	49	49.5	fold	(yes)[31]

Target	Domain	Length	Template	Eq	Eq%	Level	Homologue
2fmr.-.1		65	2pii.-.-	49	75.4	fold	(yes)[32]
2hfh.-.1		109	1hst.A.-	51	46.8	fold	superfamily
2hgf.-.-		97	1gup.A.-	46	47.4	substructure	
2kin.A.-		238	1vom.-.-	90	37.8	partial	family
2kin.B.-		100	1aa6.-.-	43	43.0	substructure	
2pth.-.-		193	1ecp.A.-	110	57.0	fold	
2sak.-.-		121	2qil.A.-	51	42.1	fold	
4rnp.A.-	7-294	294	2dtr.-.-	61	20.7	novel	
4rnp.A.-	295-883	589	1bdp.-.-	204	34.6	fold	superfamily

Table 2: Classification by C^{α} and C^{α}/C^{β} superimposition

Level of similarity	C^{α} superimposition		C^{α}/C^{β} superimposition	
Fold	153	78.1%	147	75.0%
Substructure	14	7.1%	8	4.1%
Partial fold	9	4.6%	8	4.1%
Novel	20	10.2%	33	16.8%

Table 3: Highly similar analogous structure pairs

Target	PDB-Id	Length	Template	PDB-Id	Equiv
FMN-binding protein[1]	1axj	122	Hepatitis A virus 3C proteinase	1hav.A	67
			Phthalate dioxygenase reductase	2pia	41
Deoxyhypusine synthase[2]	1dhs	344	Purine nucleoside hydrolase	2mas	98
			Pyruvate decarboxylase	1pvd	89
Protein farnesyltransferase[3]	1ft1.B	437	Glucoamylase	1gai	152
			5-epi-Aristolochene synthase	5eas	105
Transposase[4]	2ezk	99	Trp repressor	1jhg.A	49
			TC3 transposase	1tc3.C	43
KH domain of FMR1[5]	2fmr	65	Signal transduction protein PII	2pii	49
			Vigilin	1vig	44

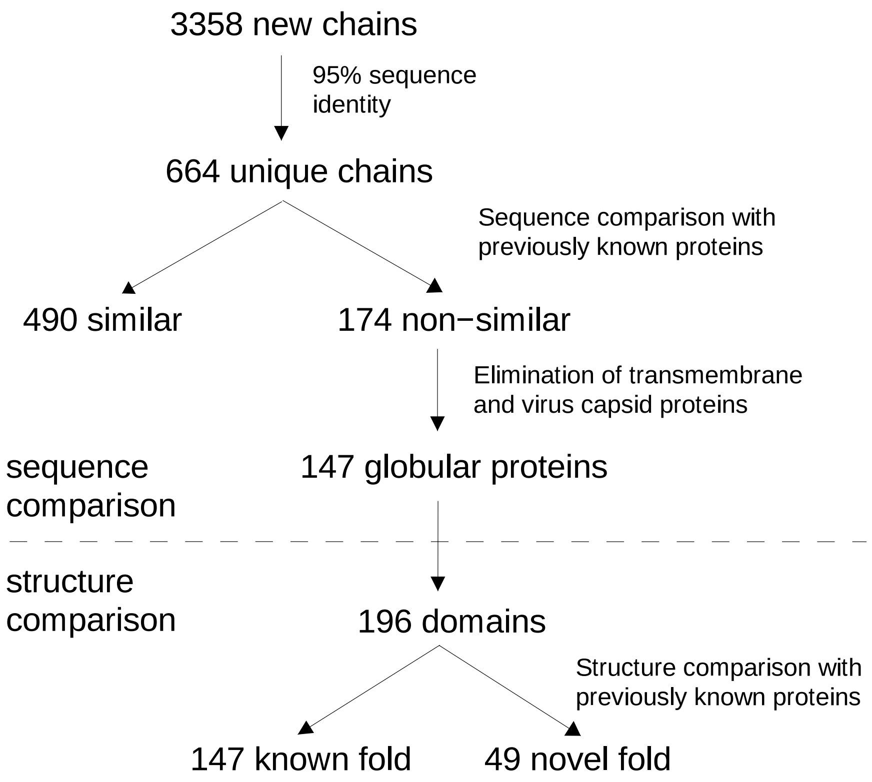
3358 new chains
95% sequence identity
664 unique chains
Sequence comparison with previously known proteins
490 similar
174 non−similar
Elimination of transmembrane and virus capsid proteins
sequence comparison
147 globular proteins
structure comparison
196 domains
Structure comparison with previously known proteins
147 known fold
49 novel fold

Figure 1:

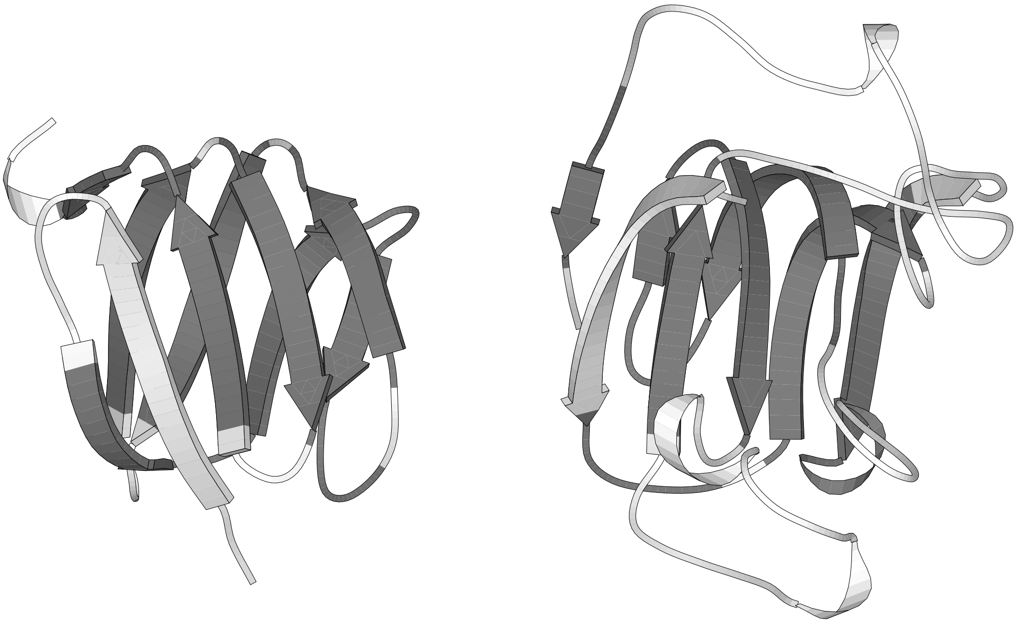

Figure 2: (a)