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

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 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.

The 147 globular proteins corresponding to 196 domains released in 1998 have no clear sequence similarity to previously known structures. However, 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 of the order of one-third of these relationships.

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

Introduction

Several bacterial and eucaryotic genomes (TIGR†) 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

*Corresponding author

E-mail address of the corresponding author:
sippl@came.sbg.ac.at

† http://www.tigr.org

Abbreviations used: FMN, flavin mononucleotide; PDB, Protein Data Bank; rms, root-mean-square; SCOP, structural classification of proteins.
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 represent 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. Here, 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 that are unrelated in sequence have a related function.

Superimposition of the Cα 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 Cβ 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.

Here, 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 Conclusion 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. When
redundancies of greater than 95% sequence identity are removed, 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. This 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, seven 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 $C^i$ atom positions, is needed for a proper definition of structurally equivalent residues. Often the $C^i$ 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 $\beta$-sheet but less so with $\alpha$-helices.

Figure 2 shows the superimposition of the $\epsilon$-subunit of F$_0$F$_1$-ATP synthase from *Escherichia coli* (1aqt) and the C-terminal subdomain of $\beta$-galactosidase from *E. coli* (1bgl). Based on $C^a$ superimposition, the similarity is extensive (Figure 2(a)). Considering side-chain orientation, only one sheet of the $\beta$-sandwich can be superimposed (Figure 2(b)) and the number of equivalent residues drops from 61 ($C^a$) to 42 ($C^a/C^p$), corresponding to a difference of 31%. From the distinct function and different side-chain orientations it seems likely that the ATP synthase and $\beta$-galactosidase domains are unrelated, although the topology of the $C^a$ 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 eight pairs at the substructure level. Figure 3 depicts an example of a small protein that superimposes with a compact substructure of a larger protein. Ragweed pollen allergen (1bbg) is a small (40 amino acid residues) z + b fold and superimposes with the C-terminal end of profilin (1pne) with 31 equivalent residues (considering side-chain orientation). The respective alignment has only three small 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. The problem is that many structurally equivalent pairs are exposed in one protein but buried in the other structure. Therefore, methods that take the physicochemical environment into account may fail in such cases (F. S. Domingues, P.L. & M.J.S., unpublished results).

Partial fold similarity was observed in eight cases. An example is the similarity of oncogene product p14TCL1 (1jsg) and avidin (1avd). Both consist of an eight-stranded b-barrel but only five strands are structurally equivalent. Surprisingly, the proteins lambda-exonuclease (1avq) and kinesin (2kin) are partially similar to functionally related proteins (Kovall & Matthews, 1997; Sack et al., 1997).

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

---

**Structural and functional similarities**

The results of our classification are summarized in Table 1. Of the 196 domains, 75% 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.

For comparison, Table 1 includes data based on Cα trace superimposition. Since this is less restrictive, more proteins having extensive similarity to a previously known structure are found. The effect is most pronounced for substructures, which increase from eight to 14 instances.

The SCOP data base was used to analyze functional relationships between the 196 domains and previously known structures (Table 2). However,
only 120 of these domains are classified in the used release of SCOP (August 1998). For the remaining 76 domains, evidence regarding functional relationships was obtained from the literature. Of the 196 domains, 97 (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 contain 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 protein pairs that are both structurally and functionally similar, corresponding to 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 that facilitates this distinction would be advantageous. The nine cases where an analogous protein has more equivalent residues in common with the target than a homologous protein are reduced to five when side-chain orientation is taken into account. Although 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 \( r \) (1a62).

An example is G:T/U mismatch-specific DNA glycosylase from *E. coli* (1mug) (Barrett et al., 1998), which is in the same SCOP superfamily as human uracil-DNA glycosylase (1akz). The \( C^\alpha \) 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 greater 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 \( \beta \)-subunit of protein farnesyltransferase (1ft1.B) has the fold of an \( \alpha \)-\( \alpha \) toroid consisting of six \( \alpha \)-helical hairpins arranged in a closed circular array. This type of fold tolerates variations in size. The functionally unrelated protein glucosylamylase (1gai) has the same fold with an identical number of \( \alpha \)-helical hairpins, while the functionally related
<table>
<thead>
<tr>
<th>Target</th>
<th>Domain</th>
<th>Length</th>
<th>Template</th>
<th>Eq (%)</th>
<th>Level</th>
<th>Homologue</th>
<th>Target name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a0b.-</td>
<td>125</td>
<td>1cgo.-</td>
<td>75</td>
<td>(60.0) Fold</td>
<td>Superfamily</td>
<td>C-terminal HPt domain of ArcB</td>
<td></td>
</tr>
<tr>
<td>1a0i.-</td>
<td>126</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td>ATP-dependent DNA ligase</td>
<td></td>
</tr>
<tr>
<td>1a17.-</td>
<td>93</td>
<td>1d7y.</td>
<td>36</td>
<td>(33.6) Fold</td>
<td>Protein phosphatase 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a2x.A-</td>
<td>128</td>
<td>1clp.A</td>
<td>75</td>
<td>(57.7) Fold</td>
<td>Pryrrolidone carboxyl peptidase</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5c.-</td>
<td>136</td>
<td>1bpm-1</td>
<td>83</td>
<td>(28.1) Fold</td>
<td>Tomchick et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5c.-</td>
<td>53</td>
<td>2rbg.w</td>
<td>61</td>
<td>(12.6) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a74.-</td>
<td>162</td>
<td>2ak7.A</td>
<td>39</td>
<td>(23.9) Fold</td>
<td>Homing endonuclease I-Ppol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a7j.-</td>
<td>150</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a22.-</td>
<td>237</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a3c.-</td>
<td>181</td>
<td>1hgx.A</td>
<td>136</td>
<td>(75.1) Fold</td>
<td>Tomchick et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5r.-</td>
<td>103</td>
<td>1ubi.</td>
<td>63</td>
<td>(61.2) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5t.-</td>
<td>167</td>
<td>2rbg.w</td>
<td>61</td>
<td>(12.6) Fold</td>
<td>Homing endonuclease I-Ppol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a62.-</td>
<td>150</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a68.-</td>
<td>93</td>
<td>1ubi.</td>
<td>63</td>
<td>(61.2) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a6q.-</td>
<td>295</td>
<td>1pm-1</td>
<td>83</td>
<td>(28.1) Fold</td>
<td>Tomchick et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a6q.-</td>
<td>53</td>
<td>2rbg.w</td>
<td>61</td>
<td>(12.6) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a74.-</td>
<td>162</td>
<td>2ak7.A</td>
<td>39</td>
<td>(23.9) Fold</td>
<td>Homing endonuclease I-Ppol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a7j.-</td>
<td>150</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a22.-</td>
<td>237</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a3c.-</td>
<td>181</td>
<td>1hgx.A</td>
<td>136</td>
<td>(75.1) Fold</td>
<td>Tomchick et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5r.-</td>
<td>103</td>
<td>1ubi.</td>
<td>63</td>
<td>(61.2) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a5t.-</td>
<td>167</td>
<td>2rbg.w</td>
<td>61</td>
<td>(12.6) Fold</td>
<td>Homing endonuclease I-Ppol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a62.-</td>
<td>150</td>
<td>1ckm.A</td>
<td>52</td>
<td>(47.7) Fold</td>
<td>Superfamily</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a68.-</td>
<td>93</td>
<td>1ubi.</td>
<td>63</td>
<td>(61.2) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a6q.-</td>
<td>295</td>
<td>1pm-1</td>
<td>83</td>
<td>(28.1) Fold</td>
<td>Tomchick et al. (1998)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1a6q.-</td>
<td>53</td>
<td>2rbg.w</td>
<td>61</td>
<td>(12.6) Fold</td>
<td>Bayer et al. (1998)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Structure-based Characterization of Novel Proteins

Phosphoserine aminotransferase

Transcription initiation factor 5A

Cystathionine β-lyase

N-ethylmaleimide-sensitive factor

Deoxyhypusine synthase

DNA-binding protein Dps

Co-chaperonin Gp31

Superfamily D-Glucose 6-phosphotransferase

3-Methyladenine DNA glycosylase II

G:T/U-specific DNA glycosylase

G-T/U-specific DNA glycosylase

HSⅤ (ClpQ) protease

Inhibitory receptor p58-cl42

Endo-β-1,4-glucanase

Inducible nitric oxide synthase

Nitrophorin 1

Perfringolysin O

Ferric iron binding protein

Growth hormone I

HESⅤ (ClpQ) protease

Inhibitory receptor p58-cl42

Endo-β-1,4-glucanase

Inducible nitric oxide synthase

Nitrophorin 1

Perfringolysin O
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 fewer 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 (Liepinsh et al., 1997). 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 (Liepinsh et al., 1998). 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 that most likely diverged by extensive permutations in the sequence is human deubiquitinating enzyme UCH-L3 (1uch). Both proteins are in the same SCOP superfamily of cysteine proteinases. Its catalytic triad superimposes quite well with that of other members of this superfamily. 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 that accommodates the catalytic cysteine residue is located between strands 2 and 3 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

Here, we investigated 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 sequences 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 that 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) obtained a much higher percentage, 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 FastA E-values in this analysis, the number of homologous proteins appear to be significantly smaller.

The question of 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 23 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

---

**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).
information that can be obtained from prediction methods that employ fold databases.

Twenty of the 23 fold recognition targets are similar to previously known structures and 48% of these are remote homologues (Murzin, 1999), a number that 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

Table 3. Highly similar analogous structure pairs

<table>
<thead>
<tr>
<th>Target name</th>
<th>PDB-Id</th>
<th>Length</th>
<th>Template name</th>
<th>PDB-ID</th>
<th>Equiv</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMN-binding protein (Liepinsk et al., 1997)</td>
<td>1axj</td>
<td>122</td>
<td>Hepatitis A virus 3C proteinase</td>
<td>1hav.A</td>
<td>67</td>
</tr>
<tr>
<td>Deoxyhypusine synthase (Liao et al., 1998)</td>
<td>1dhs</td>
<td>344</td>
<td>Phthalate dioxygenase reductase</td>
<td>2pia</td>
<td>41</td>
</tr>
<tr>
<td>Protein farnesyltransferase (Park et al., 1997)</td>
<td>1ft1.B</td>
<td>437</td>
<td>Pyruvate decarboxylase</td>
<td>1pvd</td>
<td>89</td>
</tr>
<tr>
<td>Transposase (Schumacher et al., 1997)</td>
<td>2ezk</td>
<td>99</td>
<td>Glucoamylase</td>
<td>1gai</td>
<td>152</td>
</tr>
<tr>
<td>KH domain of FMR1 (Musco et al., 1997)</td>
<td>2fmr</td>
<td>65</td>
<td>5-Epi-aristolochene synthase</td>
<td>5eas</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Trp repressor</td>
<td>1jhg.A</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>TC3 transposase</td>
<td>1tc3.C</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Signal transduction protein II</td>
<td>2pii</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Vigilin</td>
<td>1vig</td>
<td>44</td>
</tr>
</tbody>
</table>

Proteins that 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 (≈10%), while in the other cases the analogous have a substantially higher number of equivalent residues (see the text for a discussion).
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 that 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).

![Figure 7](image_url)

**Figure 7.** Current limits of structural and functional inference for proteomes inferred from structure data bases. Blue sector, the fraction of structures that can be inferred from pairwise sequence comparison (17%); gray sector, corresponds to the estimated percentage of transmembrane and non-globular proteins (20%). White sector, 16% of proteomes are most probably novel folds. Red sector, the fraction of proteins that are both structurally and functionally similar to an already determined protein structure (31%). Yellow sector, proteins with structural similarity to a known structure (16%).

### 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, 1996, 1998) and all sequences with significant similarity (E-value <0.01) to an 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 by the experimentalists (Table 2).

#### 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 P.L., W.A.K., M.J.S. & F.S. Domingues (unpublished). ProSup measures the extent of similarity by the number of structurally equivalent residues. Two residues are considered as equivalent if their \(C^\alpha\) atoms are closer than 5 Å after superimposition. The rms deviation of equivalent residues is held approximately constant by ProSup in the range of 2 to 3 Å. Side-chain orientation is represented by the \(C^\beta\) atom positions. When side-chain orientation is incorporated into structure superimposition, both \(C^\alpha\) and \(C^\beta\) atoms have to be closer than 5 Å after optimal superimposition of the \(C^\alpha\) trace. One feature of ProSup is the ability to generate a list of alternative alignments. For the present analysis, only the alignments with the highest number of equivalent residues are considered 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 (January 27, May 15 and September 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.) was this one 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 wissenschaftlichen Forschung.

References


*Edited by R. Huber*

(Received 9 September 1999; received in revised form 28 December 1999; accepted 28 December 1999)