Minireview

The role of protein structure in genomics

Francisco S. Domingues[a], Walter A. Koppensteiner[b], Manfred J. Sippl[a,b,*]

[a] *Center for Applied Molecular Engineering, Institute for Chemistry and Biochemistry, University of Salzburg, Jakob Haringer Strasse 3, A-5020 Salzburg, Austria*
[b] *ProCeryon Biosciences GmbH, Jakob Haringer Strasse 3, A-5020 Salzburg, Austria*

Received 5 May 2000

Edited by Gunnar von Heijne

Abstract The genome projects produce an enormous amount of sequence data that needs to be annotated in terms of molecular structure and biological function. These tasks have triggered additional initiatives like structural genomics. The intention is to determine as many protein structures as possible, in the most efficient way, and to exploit the solved structures for the assignment of biological function to hypothetical proteins. We discuss the impact of these developments on protein classification, gene function prediction, and protein structure prediction. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Structural genomics; Structure prediction; Structure classification; Functional genomics; Function prediction

1. Introduction

The genome projects generate biological sequence data on a breath-taking pace. The complete sequencing of the first genome (*Haemophilus influenzae*) [1], the first eukaryotic genome (*Saccharomyces cerevisiae*) [2], and the first genome of a multicellular organism (*Caenorhabditis elegans*) [3] are celebrated milestones, and now the human genome, the most precious of all, is on the finish line. In the wake of these achievements we are confronted with the information revolution in biology and medicine and we witness the creation of new paradigms like genomic medicine, genomic health care, and genomic and proteomic technologies [4]. No doubt we live in exciting times.

There is some way to go before we can take full advantage of the tremendous amount of information encoded in the genomes. Identifying the genes in a given genome, determining the biological role of their protein products, and understanding their interplay are challenging tasks that will keep biologists busy in the years to come.

In understanding the genomes molecular structure plays an important role. Lacking a three-dimensional structure our knowledge of the biological function of a protein and its molecular interactions is largely incomplete. One of the goals of structural genomics, therefore, is to determine as many structures as possible [5]. In spite of recent technical advances in experimental structure determination, discovering the three-dimensional structure of all protein products found in all the sequences genomes is a daunting task. The associated workload prevents to determine the structure of every protein by experimental techniques. But there is hope that in the long run computational techniques that are based on experimental data might provide a viable solution.

These initiatives in computational and structural genomics trigger a number of interesting questions and pose interesting problems. It is now well established that proteins often have similar folds, even if their sequences are seemingly unrelated [6]. Hence the number of sequences is much larger than the number of folds. Therefore the number of distinct folds could be small. If so, how many folds are there [7]?

Several classification schemes are available that organize our current knowledge of protein structures in a systematic way [8]. From these databases the number of distinct superfamilies seems to be in the order of one thousand. It is clear that such an estimate depends on the grain size used. Even if two protein structures are similar, they are never identical. Hence, a question particularly relevant for molecular modeling studies is whether fold space is discrete or continuous.

In the past protein structures were solved to understand their biochemical function in atomic detail. Hence, as a rule, the function was known before an attempt was made to determine the structure. The attitude of the structural genomics approach is radically different. Structures are solved to obtain information on the biochemical function and biological role of genes. But to what extent can we deduce function from structure? This is not as straightforward as one might hope. In fact this difficulty has triggered some doubts on the value of protein structure factories and the structural genomics initiative as a whole. Below we address these questions in some detail, concluding that the more structures are solved the better.

2. Protein structure classification

The Protein Data Bank [9] has been accumulating thousands of protein structures over the last years. In March 2000 the number of entries was 12 000. In order to make such a large amount of data understandable and usable, classification schemes have been implemented. Popular schemes are SCOP [10], CATH [11] and FSSP [12]. SCOP is a hierarchical classification based on human expertise where proteins are grouped according to structure and evolutionary relationships. CATH uses a semi-automated classification method that also follows a hierarchical scheme with clear structure similarity thresholds. FSSP is an automated classification method based on pairwise structure comparison. A

*Corresponding author. Fax: (43)-662-454889.
E-mail: sippl@came.sbg.ac.at

PII: S0014-5793(00)01678-1

major advantage of FSSP is the constant update with the latest PDB entries.

These classification databases provide an overall view of the protein structure space. Currently in SCOP release 1.48, November 1999, there are 520 folds and 771 superfamilies. In CATH, version 1.6, June 1999, there are 672 topologies and 1028 families. The total number of superfamilies is estimated to be approximately 1000 [13], but other estimates are in the range from 1000 to 6000 [8].

The question arises whether the classifications obtained are similar across the various schemes. Recently the grouping at fold and superfamily levels has been compared between SCOP, CATH, and the pairwise matches in FSSP [14]. Two thirds of all the pairs of proteins with the same fold in SCOP also matched in CATH and FSSP, but there is a difference in one third of the assignments. One reason for the observed differences might be that in SCOP human expertise on evolutionary relationships is used, while CATH and FSSP rely more on sequence similarity and geometric criteria. Another possibility is that the notion of discrete fold space is not adequate so that boundaries between folds are difficult to define.

3. Fold space, discrete or continuous

To address the question whether there are significant overlaps among distinct fold types, we deliberately searched the protein structure database using ProSUP [15] and found several examples. Fig. 1 demonstrates that three proteins considered to belong to distinct folds in SCOP show structure similarity. This is somewhat surprising considering that one of them (1tph1) is a TIM barrel and the other two (1tadC and 2pgd) have the frequently observed three layer α-β-α sandwich architecture. Also, CATH differs from SCOP by classifying 1tadC and 2pgd to belong to the same fold type. The example indicates that fold classifications are somewhat arbitrary. Fig. 1 can even be extended to include additional folds (Fig. 2), and one gets the impression that it is possible to move around in structure space by hopping from one fold type to the next. This is reminiscent of a continuous rather than a discrete fold space. Orengo et al. [11] already observed that the recurrence of structural motifs results in a continuum of fold types. This was observed in the case of the three layer α-β-α sandwich architectures and in the β sandwich architectures.

Then is fold space continuous or discrete? The question is somewhat academic as the answer is bound to lie between these two extremes. On the one hand the notion of discrete groups of folds forms the basis for useful classification schemes, but on the other hand the boundaries between these families are often difficult to define.

4. Structure and function

A question that comes up repeatedly in structure genomics is to what extent function can be deduced from structure. To answer that question one has to consider three categories. The fold topology which describes structural similarity, the homologous superfamily that implies evolutionary relationship, and functional similarity. Classification would be easy if proteins follow the principle that whenever two proteins have the same topology they belong to the same homologous superfamily and also have the same function. Of course, biology usually is not as easy as that. Functions can be associated with different folds (e.g. serine proteases), distinct homologous superfamilies can have the same fold [13] (e.g. α/β barrel or OB-fold), and homologous superfamilies can evolve into distinct functions [16]. There are even examples where proteins diverge to distinct structures [16].

Nevertheless, if structure and function coincide in a large number of cases, structural information would be helpful to assign function to new proteins. Hence the question is how often the functions of two proteins match when they are structurally similar and at the same time have no significant sequence similarity. This correlation of protein structure and function has been investigated by Koppensteiner et al. [17]. The result is that 66% of the proteins having a similar fold also have a similar function. In other words, observing structural similarity without sequence similarity implies functional relatedness in two out of three cases.

Currently for 17% of all protein sequences of complete genomes, functions can be assigned by sequence comparison [18]. Exploiting structural information to the largest possible extent could yield assignments for almost 50% [17]. The remaining question then is how we can obtain structural information for a protein to deduce its function. One possibility is to use structure prediction techniques like fold recognition and ab initio prediction [19–21]. Although these techniques are far from perfect it has been demonstrated that predicted structures can be used to assign functions. Xu et al. [22] predicted the function of two hypothetical *Methanococcus jannaschii* proteins. They suggested MJ0301 is a dihydropteroate synthase (DHPS) and MJ0757 a tymidylate synthase (TS). Both predictions have been verified experimentally. Similarly Fan et al. [23] identified the accessory subunit of mtDNA polymerase (polγ) to be structurally related to the anticodon-binding domain of class IIa aminoacyl-tRNA synthetases and assigned the functions of processivity clamp and primer recognition factor in mtDNA replication.

The second and most obvious possibility is to determine the structure by X-ray crystallography and NMR. In fact to determine as many possible structures as possible is the main goal of structural genomics initiatives. The solved structure can then be used to scan structure databases. If a related structure is found then frequently functional information can be deduced from the match. In the last 2 years several pilot studies have addressed the feasibility of this approach by solving the structures of several hypothetical proteins. A successful function assignment was reported by Hwang et al. [24] who identified the ORF MJ0226 of *M. jannaschii* as an novel nucleotide triphosphatase for the non-standard nucleotides XTP and ITP. This function was inferred from structural similarity to nucleotide binding proteins and has been confirmed experimentally. Colovos et al.[25] recognized the *ycaC* gene product of *Escherichia coli* as a hydrolase by structure comparison to *N*-carbamoylsarcosine amidohydrolase. In the case of of the yjgF gene product of *E. coli*, insights for the design of selective experiments were gained although a definite functional assignment was not possible[26]. Cases where the determination of structure resulted in a novel fold also can yield important structure information. In the work of Zarembinski et al. [27] the function of ORF MJ0577 of *M. jannaschii* was deduced from a bound ATP. Finally, by application of the Fuzzy Functional Form technique Fetrow et al. [28] dis-

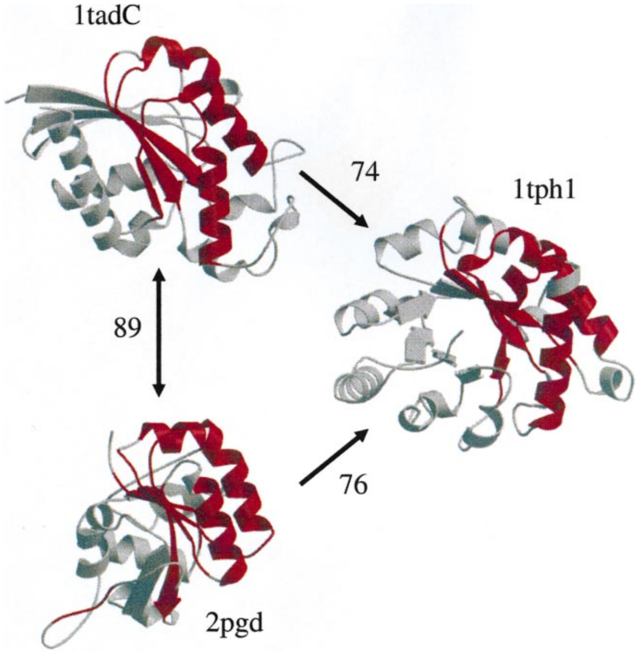

Fig. 1. Pairwise structure similarity across fold types. Structurally equivalent regions are according to ProSUP. Transducin-α (1tadC) has a three layer α-β-α sandwich architecture, it has 74 equivalent residues relative to triose phosphate isomerase (1tph1), marked red in both models. 6-Phosphogluconate dehydrogenase (2pgd), has also a α-β-α sandwich architecture. Marked red in the model are the 76 equivalent residues relative to 1tph1. In the superposition of 1tadC and 2pgd there are additional equivalent residues (not shown). The regions of similarity between these three folds overlap to a large extent.

covered an active site similar to those of thiol-disulfide oxidoreductases in serine/threonine protein phosphatase-1, although the folds are not related.

Another goal of structural genomics is to determine a set of protein folds that covers fold space. Here the question what are the chances that a structure determination results in a novel fold is important. From the structures submitted to PDB in 1998 only 200 domains had no sequence similarity to a known structure [17]. Only one quarter of these correspond to a novel fold. Taking into account the estimated number of transmembrane and non-globular proteins, the fraction of novel folds in complete genomes is estimated to be in the order of 16%, a surprisingly low number. Hence chances are high that a new structure resembles a known fold. Therefore to cover fold space one should use appropriate strategies to avoid redundant structure determination.

In terms of genome annotation novel folds are the most valuable since they provide structural data for whole protein families. A novel fold on average annotates 70 sequences in the current non-redundant sequence database of NCBI (Table 1).

5. Conclusion

Although function assignment from structure is not as straightforward as one might wish, it is clear that almost every new structure significantly increases our knowledge and improves our computational tools. First steps in the analysis of genomes are gene finding and annotation of putative genes. Gene and protein functions are discovered using search engines like BLAST [29] and FASTA [30] to scan sequence data-

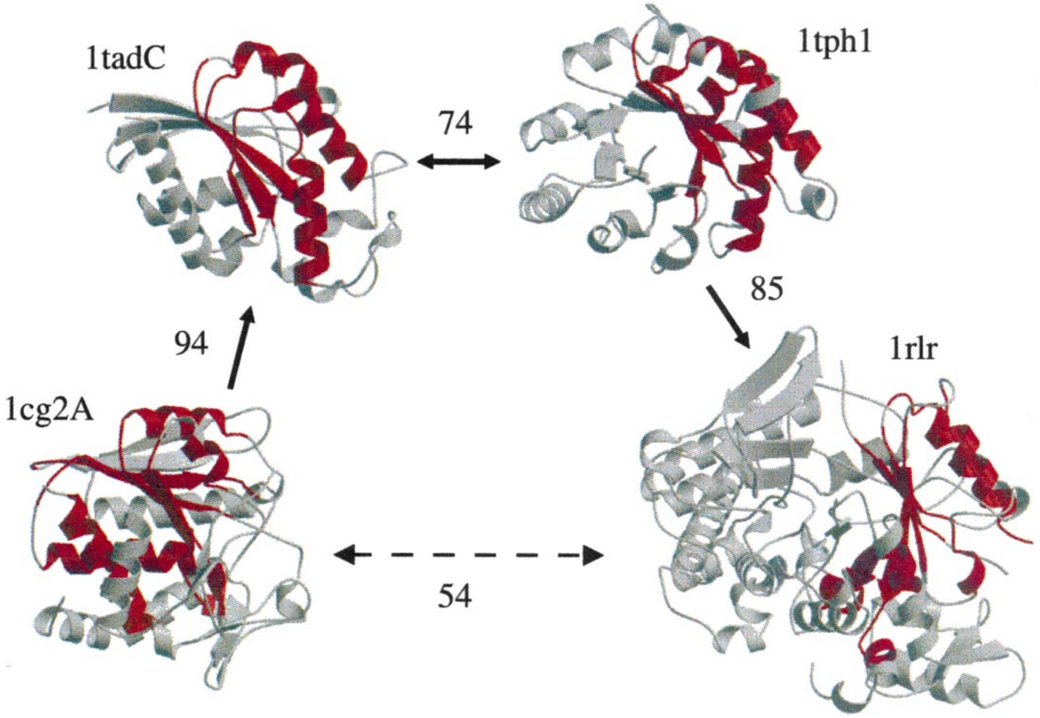

Fig. 2. Walking in fold space across fold types. Structurally equivalent regions are according to ProSUP. Carboxypeptidase G2 (1cg2A) shares structure similarity to transducin-α (1tadC), both have the same three layer α-β-α architecture. The 94 structurally equivalent residues are marked red in the 1cg2A model. Triose phosphate isomerase (1tph1), is a TIM barrel fold and shares 74 equivalent residues relative to 1tadC, marked red in both 1tph1 and 1tadC. Ribonucleotide reductase protein R1 (1rlr), has 85 equivalent residues relative to 1tph1, marked red in 1rlr. There is a lower degree of structure similarity between 1cg2A and 1rlr, where only to two helices and partially two strands match (not shown).

Table 1
From the analysis of Koppensteiner et al. [17] 30 novel folds were used to estimate the average number of sequences which can be annotated with a novel fold

PDB Id	Homologues		PDB Id	Homologues	
	Close	Distant		Close	Distant
pdb1a68.-	154	71	pdb1hp8.-	2	0
pdb1a74.A	4	0	pdb1hus.-	92	33
pdb1ahj.A	14	0	pdb1jdw.-	7	0
pdb1aiw.-	6	0	pdb1jsg.-	5	8
pdb1al0.1	8	0	pdb1kdx.A	18	3
pdb1ap8.-	35	4	pdb1noc.A	89	0
pdb1apj.-	4	1	pdb1rkd.-	13	248
pdb1aqt.-	20	76	pdb1skn.P	33	0
pdb1avq.A	5	8	pdb1toh.-	16	1
pdb1ay2.-	146	87	pdb1uch.-	21	11
pdb1baq.-	10	25	pdb1vgh.-	35	0
pdb1bbg.-	8	0	pdb1ygs.-	59	17
pdb1bgf.-	50	4	pdb2hgf.-	22	30
pdb1bnl.A	21	0	pdb2kin.A	245	77
pdb1fgj.A	5	0	pdb2kin.B	234	8

Homologous sequences for each fold were searched in the non-redundant sequence data base (NCBI server ftp://ncbi.nlm.nih.gov/blast/db/) using the iterative sequence search program PSI-Blast [40]. On average a novel fold can be linked to 46 close homologues (more than 30% sequence identity) and 23 distant homologues (less than 30% sequence identity). The number of homologues per fold varies considerable in the range of 2 (pdb1hp8) to 322 (pdb2kin.A).

bases for significant hits. Unfortunately, these databases contain numerous errors [31], and consequently the results are often misleading even if the sequence homology is significant. In contrast, if the structure of at least one member of a protein family is known, then the annotation is almost always reliable. Moreover, the structure often yields important information on key residues and binding sites for all family members that can form the basis of selective experiments.

In principle, the structure of proteins should be computable from their amino acid sequences. But this protein folding problem turned out to be a notoriously difficult. Nevertheless there are signs of progress [19–21]. Approaches to structure prediction come in three flavors. Comparative modeling starts with the structure of a protein having clear sequence homology to the target sequence. Fold recognition and threading methods scan fold libraries to identify structures compatible with the target sequence, and ab initio methods attempt to predict a structure from sequence information alone.

What these methods have in common is that practically all of them heavily rely on the knowledge pool of known structures in some way or another. Known structures are essential for comparative modeling and fold recognition. The secondary structure prediction techniques [32] are trained on libraries of known structures and many techniques that build structures from scratch either employ structural fragments derived from known folds [33,34] or employ knowledge based energy functions derived from the protein structure database [35–37]. Of course, the study and simulation of protein–ligand [38] and protein–protein interactions [39], and virtual screening depend on structures obtained from experiments. Hence, the large scale protein structure determination efforts will fuel genome annotation, functional assignments, and computational approaches with precious data, and the hope is justified that the expanding database of molecular structures will trigger major breakthroughs [5].

Acknowledgements: Support for this work was provided by the Fonds zur Förderung der wissenschaftlichen Forschung, Austria; Grant number P13710-MOB.

References

[1] Fleischmann, R.D., Adams, M.D., White, O., Clayton, R., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A. and Merrick, J.M. et al. (1995) Science 269, 496–512.

[2] Goffeau, A., Barrell, B.G., Bussey, H., Davies, R.W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J.D., Jacq, C., adn, E.J., Louis, M.J., Mewes, H.W., Murakami, Y., Philippsen, P., Tettelin, H. and Oliver, S.G. (1996) Science 274, 563–567.

[3] The *C. elegans*Sequencing Consortium, (1998) Science 282, 2012–2018.

[4] Sander, C. (2000) Science 287, 1977–1978.

[5] Burley, S.K., Almo, S.C., Bonanno, J.B., Capel, M., Chance, M.R., Gaasterland, T., Lin, D., Sali, A., Studier, F.W. and Swaminathan, S. (1999) Nat. Gen. 23, 151–157.

[6] Pastore, A. and Lesk, A.M. (1990) Proteins Struct. Funct. Genet. 8, 133–155.

[7] Chothia, C. (1992) Nature 357, 543–544.

[8] Swindells, M., Orengo, C., Jones, D., Hutchinson, E. and Thornton, J. (1998) Bioessays 20, 884–891.

[9] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I. and Bourne, P. (2000) Nucleic Acids Res. 28, 235–242.

[10] Murzin, A., Brenner, S., Hubbard, T. and Chothia, C. (1995) J. Mol. Biol. 247, 536–540.

[11] Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M. and Thornton, J. (1997) Structure 5, 1093–1108.

[12] Holm, L. and Sander, C. (1996) Science 273, 595–602.

[13] Brenner, S., Chothia, C. and Hubbard, T. (1997) Curr. Opin. Struct. Biol. 7, 369–376.

[14] Hadley, C. and Jones, D. (1999) Struct. Fold Des. 7, 1099–1112.

[15] Feng, Z.-K. and Sippl, M. (1996) Fold Des. 1, 123–132.

[16] Murzin, A. (1998) Curr. Opin. Struct. Biol. 8, 380–387.

[17] Koppensteiner, W.A., Lackner, P., Wiederstein, M. and Sippl, M.J. (2000) J. Mol. Biol. 296, 1139–1152.

[18] Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. and Sander, C. (1999) Bioinformatics 15, 391–412.

[19] Moult, J., Hubbard, T., Fidelis, K. and Pedersen, J.T. (1999) Proteins Struct. Funct. Genet. 3 (Suppl.), 2–6.

[20] Sippl, M.J. (1999) Structure 7, R81–R83.

[21] Koehl, P. and Levitt, M. (1999) Nat. Struct. Biol. 6, 108–111.

[22] Xu, H., Aurora, R., Rose, G.D. and White, R.H. (1999) Nat. Struct. Biol. 6, 750–754.

[23] Fan, L., Sanschagrin, P., Kaguni, L. and Kuhn, L. (1999) Proc. Natl. Acad. Sci. USA 96, 9527–9532.

[24] Hwang, K.Y., Chung, J.H., Kim, S.H., Han, Y.S. and Cho, Y. (1999) Nat. Struct. Biol. 6, 691–696.
[25] Colovos, C., Cascio, D. and Yeates, T. (1998) Structure 6, 1329–1337.
[26] Volz, K. (1999) Protein Sci. 8, 2428–2437.
[27] Zarembinski, T.I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R. and Kim, S.-H. (1998) Proc. Natl. Acad. Sci. USA 95, 15189–15193.
[28] Fetrow, J.S., Siew, N. and Skolnick, J. (1999) FASEB J. 13, 1866–1874.
[29] Altschul, S., Gish, W., Miller, W., Myers, E. and Lipman, D. (1990) J. Mol. Biol. 215, 403–410.
[30] Pearson, W.R. (1996) Methods Enzymol. 266, 227–258.
[31] Brenner, S. (1999) Trends Genet. 15, 132–133.
[32] Rost, B. and Sander, C. (1993) J. Mol. Biol. 232, 584–599.
[33] Sippl, M.H., Hendlich, M. and Lackner, P. (1992) Protein Sci. 1, 625–640.
[34] Simons, K.T., Bonneau, R., Ruczinski, I. and Baker, D. (1999) Proteins Struct. Funct. Genet. 3 (Suppl.), 171–176.
[35] Sippl, M. (1990) J. Mol. Biol. 213, 859–883.
[36] Sippl, M., Ortner, M., Jaritz, M., Lackner, P. and Flockner, H. (1996) Fold Des. 1, 289–298.
[37] Sippl, M.J. (1996) J. Mol. Biol. 260, 644–648.
[38] Gohlke, H., Hendlich, M. and Klebe, G. (2000) J. Mol. Biol. 295, 337–356.
[39] Robert, C.H. and Janin, J. (1998) J. Mol. Biol. 283, 1037–1047.
[40] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. (1997) Nucleic Acids Res. 25, 3389–3402.