Available online at www.sciencedirect.com

Fold space unlimited
Manfred J Sippl

You want to know how proteins do it? Take a walk in protein fold space. More often than not you will get a clue if not the answer. If you know what you are looking for and how to find it. In fact, there is more information than we can presently handle. Charting fold space and chasing its creatures has occupied us for the past decades. There is no end in sight.

Address
Division of Bioinformatics, University of Salzburg, Hellbrunnerstrasse 34, A-5020 Salzburg, Austria

Corresponding author: Sippl, Manfred J (sippl@came.sbg.ac.ac)

This review comes from a themed issue on Sequences and Topology
Edited by Anna Tramontano and Adam Godzik

Current Opinion in Structural Biology 2009, 19:312–320

0959-440X/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
DOI 10.1016/j.sbi.2009.03.010

Introduction
I received my first copy of the data base of protein structures [1] in 1977, dispatched by Olga Kennard in response to a letter of my thesis supervisor Hans Bernd Strack. The treasure, delivered by a postman, was wrapped in a big yellow envelope. Hands trembling, I uncovered the magnetic tape, 15 inches in diameter, containing all there was to know about protein structures: The atomic coordinates of 80 proteins.

Today, hands steady, I hit the return button to download the weekly release of PDB [2]. Some 150 files. A tiny epsilon compared to the 55 000 files that already sit on my hard disc with the volume doubling every four years. A hive of data that needs to be tamed and organized. A job for pioneers.

In 1994 Liisa Holm and Chris Sander started the FSSP/ Dali data base of aligned protein structures [3]. The following year Alexey Murzin, Steven Brenner, Tim Hubbard, and Cyrus Chothia released their first version of the SCOP data base [4] and Janet Thornton, Christine Orengo, David Jones, and co-workers published their first paper on the CATH data base [5]. These and many other specialized collections of protein structures and sequence families have become indispensable tools in protein structure research. They define the current state of the art in protein structure classification.

Anna Tramontano, the editor of this section, asked me to comment on two specific questions: Does it make sense trying to classify protein structures and what are the limits of current approaches? To address these points we have to call on a few protein structures. They talk, we prick up our ears. To set the stage we start with a brief summary of classic results.

A few classics of protein structure
Much of protein science rests on the hypothesis that the structure of a protein is determined by its amino acid sequence and the surrounding solvent [6]. In short: same sequence, same structure. Let us call this the law of protein folding.

Hemoglobin and myoglobin, the very first structures solved by X-ray analysis, have closely related structures. Indeed, 135 (92%) of the 146 Cα atoms of the human β-hemoglobin chain (2hhb) can be structurally aligned with the sperm whale myoglobin chain (1 mbn) to a root mean square (rms) error of 1.5 Å. Even so, there are only 25% identical amino acid pairs. Hence, right from the start it was clear that distinct sequences adopt very similar folds. Call this the first amendment of the law.

Back in 1973 Donald Wetlaufer [7] observed distinct structural regions in several globular proteins composed of a single polypeptide chain that he interpreted as independent folding units or domains. Moreover, a single structural domain may be composed of several pieces that are separated along the sequence. Hence the second amendment: A single protein chain may encode for more than one structural domain.

In 1993 Melanie Bennett, Senyon Choe and David Eisenberg, while solving the structure of monomeric and dimeric diphtheria toxins, observed a new mode of protein association called domain swapping or protein entanglement: Upon dimerization an unprecedented conformational rearrangement occurs: the entire R domain from each molecule of the dimer is exchanged for the R domain of the other [8]. Since then domain swapped jewels have become a commodity (e.g. [9,10,11]). Hence, the third amendment: Proteins may exchange domains where swapped and unswapped versions have only minor structural differences.

In 1984 Wolfgang Kabsch and Chris Sander found that pentapeptides of identical sequences may have completely
distinct structures. In 6 out of 25 cases they saw surprising structural adaptability: the same five residues are part of an α-helix in one protein and part of a β-strand in another [12].

Today we have a large collection of these chameleons, small and larger pieces of identical or similar sequences that fold into entirely different conformations (e.g. [13]/C15/C15) and metamorphic proteins that have been observed in several alternative states (e.g. [14]/C15/C15,15/C15). Changing gears, in 1997 Dalal, Balasubramanian, and Regan reported the design of a sequence with 50% identity to a β-sheet protein demonstrating that the artificial protein folds into a four-helix bundle [16]. Pressing still harder, in 2008 Bryan, Orban and co-workers [17/C15/C15,18] designed two proteins of 88% sequence identity that fold into entirely different conformations. Getting back to natural sequences: Cordes and co-workers [19/C15/C15,20/C15] found members of the Cro repressor family having sequence identities as high as 40% although half of their structures have switched from helices to strands (see also the commentary by Davidson [20/C15]). Hence, amendment number four: similar sequences may have distinct structures.

A few somersaults in fold space

Figure 1 shows three proteins whose structures, except for minor variations in loops and termini, are virtually identical. Alas, the sequence similarity among these proteins is low and there is no obvious relationship detectable at the sequence level. Amendment number one in action. What was a surprise a few years ago is now commonplace. Fold space abounds with such examples. In a sense structure similarity is orthogonal to sequence similarity. We need both sign posts to find our way through fold space. Hence, protein structure classification not only makes sense but also is vital.

Figure 2 compares the structures of 1n1c, the Tor-D chaperon from Shewanella massilia [21], and 1s9u, a proofreading chaperone from Salmonella typhimurium [22/C15]. Their sequence identity is 24% that is on the level of hemoglobin and myoglobin mentioned above. Thus, the two proteins should have similar folds. Not quite. A single chain of 1n1c has an open conformation, whereas the monomeric 1s9u chain folds into a compact globular domain. However, as shown in Figure 2, 1n1c forms two entangled domains, identical in terms of chemical composition and three-dimensional structure, a conformation that has been described as a case of extreme domain swapping [21], where each of the two hybrid domains has extensive structural similarity to the monomeric 1s9u domain.

The structures immediately suggest that a particular protein of this family exists as a certain mixture of entangled dimers and compact monomers. This is indeed quite plausible. With the exception of the switch region, the interactions within the monomer would be identical to those within the hybrid domain. If true, this would demonstrate that a globular domain does not necessarily correspond to a stable folding unit since the transition from the monomer to the dimer requires that the former easily disintegrates in two halves. This in turn would imply that the difference in energy between open and closed forms of the domain is small.

Arriving at definite answers to such speculations requires that we are able to find examples of proteins that adopt
Extreme domain swapping. (a) The chain structure of 1n1c, the TorD chaperon from Shewanella massilia [21]. The A and B chains of the dimer are in blue and yellow, respectively. (b) The domain structure of 1n1c. The arrow points to the switch region. Rewiring the chains at the switch region results in two identical monomers. (c) The upper domain of 1n1c (blue) and the regions of structural similarity (red) to the 1s9u-A monomer, a proofreading chaperone from Salmonella typhimurium [22]. (d) The structure of the 1s9u monomer with regions of structural similarity to 1n1c (c) in orange. The respective superposition yields an alignment length of 131 residues (65%), an rms-error of 2.7 Å, and a sequence identity 20%. The arrows in (c) and (d) point to a helix that is rotated in 1s9u by 90 degrees relative to the respective helix in 1n1c.

The asymmetric unit of 3cg0, a signal receiver domain from Desulfovibrio desulfuricans (New York SGX research Center for Structural Genomics, to be published). (a) The asymmetric unit contains four identical chains colored in blue, red, green, and yellow. (b) The double chain variant and (c) the single chain variant of the domain. The latter is oriented relative to the structure shown in (b). Except for the linker region at the right side of the domains, the two structures are virtually identical.
both conformations under comparable conditions. Figure 3 shows the asymmetric unit of the crystal structure of a signal receiver domain from *Desulfocibrio desulfuricans*. The structure was recently solved by the New York SGX Research Center for Structural Genomics and deposited with PDB in March 2008. The title contains the clause ‘an example of alternate folding’ (to be published). The asymmetric unit contains four identical chains, two of which form an entangled dimer whereas each of the remaining two chains forms a compact globular domain. The structures of all four domains are identical, except for a short loop that sends the chains of the distinct structural variants in opposite directions. Thus, we found what we were looking for: The compact globule and the swapped dimer within the same crystal. Pairs of monomers and entangled dimers occur with equal frequency so that, by definition, the difference in energy between the two states must be close to zero. Amendment numbers two and three in action. Yes, but with a slight twist.

To take this a step further let us turn to Figure 5. The related enzymes GDP-mannose dehydrogenase from *P. aeruginosa* [23], 1mv8, and UDP-glucose dehydrogenase from *S. pyogenes* [24], 1dlj, contain a central helical dimerization domain whose structure is strongly conserved, Figure 4d, but is formed in two different ways as shown Figure 4a,b. Again the difference between the two versions is confined to a short switch region, highlighted in red in Figure 4a,b. This looks familiar and UDP-glucose dehydrogenase has in fact been described as a domain swapped variant of the GPB-mannose dehydrogenase [24].

But note the difference. In our previous examples the respective domains are formed either from single chains or from two identical entangled chains whereas in the present case the domain cannot be constructed from a single chain alone. Both chains are required. Moreover, the domains have an exact twofold symmetry axis so that in both cases a rotation by 180 degrees reproduces exact copies of the molecules.

At this point it may seem futile to look for a domain of this type formed by a single chain. Alas, there it is. The second domain of 2pgd, ovine 6-phosphogluconate dehydrogenase [25], Figure 5c. The structural alignment with the bacterial 1dlj structure reveals extensive structural similarities although the respective sequences appear to be

---

**Figure 4**

A domain found in three related enzymes whose structure is formed in three distinct ways. (a) The chain structure of the central dimerization domain of 1mv8, GDP-mannose dehydrogenase from *P. aeruginosa* [23] (individual chains in blue and yellow), (b) the chain structure of the central dimerization domain of 1dlj, UDP-glucose dehydrogenase from *S. pyogenes* [24], (c) the single chain domain of 2pgd, ovine 6-phosphogluconate dehydrogenase [25]. The blue and yellow fragments highlight the correspondence with the chains shown in (b). The linker region joining these fragments is shown in green. The switch region is shown in red (a–c, f). (d) The 1dlj domain shown in (b) where the residues that are structurally equivalent to the 1mv8 domain are shown in raspberry red. (e) The 1dlj domain shown in (b) where the residues that are equivalent to the 2pgd domain are shown in raspberry red. (f) The monomeric 2pgd domain and a copy of this domain rotated by 180 degrees around the (pseudo) symmetry axis orthogonal to the paper plane. The sections colored raspberry red are structurally equivalent. The remaining colors are the same as in (c).
unrelated (sequence identity 10%). A comparison with 1mv8 yields the same picture.

The two halves of the 6-phosphogluconate dehydrogenase domain, corresponding to the two individual chains in the bacterial enzymes, are joined by a linker region of approximately 20 residues (highlighted in green in Figure 4c,f) that has no counterpart in the bacterial enzymes. The connectivity of the mammalian enzyme, that is, the conformation or topology of the switch region, clearly is of the UDP-glucose dehydrogenase type ((b) and (c) as opposed to (a) and (c) in Figure 5).

From the structural similarity of the 6-phosphogluconate dehydrogenase domain with the bacterial dimers it follows that the former must have an approximate twofold symmetry. A rotation of 180 degrees reveals that the domain is in fact highly symmetric, where 186 (72%) of the 258 Cα atoms can be superimposed to an rms-error as low as 1.4 Å (Figure 4f). However, the sequences of the two symmetric halves are entirely different. Less than 10% of the structurally equivalent pairs correspond to pairs of identical amino acids.

The complete structures of the three enzymes are shown in Figure 5. The bacterial dimers consist of five domains whereas a single chain of the mammalian enzyme has two structural domains that can be superimposed as a whole on the bacterial enzymes (Figure 5c). Hence, not only are the structures of the individual domains similar but also their mutual position and orientation in three dimensions. And all this although the structure of the symmetric domain is constructed in three different ways where the various variants have quite distinct sequences.

It is overwhelmingly tempting to speculate on the evolutionary events that might connect these molecules and on the implications regarding evolution of protein structure and biological function in general but we are running out of space here (but see, for example, Andreeva and Murzin [26]). Instead we take a final leap to amendment number four.
Figure 6 shows the NMR structures of 2k9a, human Binder of Arl2, a novel G protein binding domain [27**], and 2k0s, the homologous protein from zebra fish, determined by the Center for Eukaryotic Structural Genomics (to be published). The sequence identity is 72%. The locations of the helices along the sequences are similar but their spatial arrangement is entirely different. Hence, we have a case of two naturally occurring proteins of strong sequence homology folding into compact globular domains whose structures, except for their high helix content, are completely distinct.

Implications

The examples presented here have at least one common thread. They are to some extent surprising or, in the language of information theory, they are most informative. Such examples, however, are not rare. The repertoire of available structures contains a myriad of thrilling relationships. Taken as a whole they tell us a lot about the laws of protein folding and the peculiarities of fold space.

Domains and protein folds are mercurial objects that can be formed in many different ways. In the present context I have used the term domain to describe a compact structural unit where members of a particular type of domain can be identified by their mutual structural similarity, regardless of the connectivity of the individual chains.

When colors are used to visualize domains their identity seems rather obvious. But it is quite difficult to cast the concept of a domain in quantitative physical or mathematical terms. In 1973 Donald Wetlaufer [7] remarked: I do not feel that we have completely adequate criteria to define protein regions, but enough to make a start. His words prevail (see also Willy Taylor’s recent review [28*]).

I computed the domain decompositions and structural alignments of Figures 1–5 using TopDomain and TopMatch [29**], respectively, two programs that are implemented as public web services (see the reference to the COPS service below). A feature that distinguishes these tools from many other programs in current use is that they can be applied to PDB files and biological units as a whole. Moreover, TopMatch yields the complete spectrum of alignments. All potentially interesting structural relationships are reported, including permutations of sequences that change the connectivity but not the spatial arrangement of the permuted parts of a fold (e.g. Figure 4).

The similarities found among individual domains are not enough to capture the whole spectrum of relationships between protein structures. We also need to record the similarities among complete biological units (e.g. Figure 5). The largest structures contain dozens of chains and well over 10,000 amino acid residues. The size of these objects and the representation of structures as PDB files add significant technical and algorithmic challenges. For example, 1mv8 contains four identical but crystallographically independent chains, whereas 1dlj contains only a single chain. Hence, to uncover the whole bandwidth of structural relationships highlighted in Figures 4 and 5 one either has to generate the UDP-glucose dehydrogenase dimer from the crystallographic symmetry operations or download the respective biological unit from the PDB server.

The term ‘structure classification’ implicates several tasks. The first is a definition of the nature of objects (e.g. domains, biological units) amenable to classification along with properties that can be used to organize these objects (e.g. structural similarity). The second is to spot these objects (e.g. domain decomposition) and to
compute and record their mutual relationships (e.g. structure alignment). The latter requires suitable data structures for storage and retrieval of similarities and other relationships. But perhaps the most important aspect of any structure classification is that the complete repertoire of available structures is represented in a way that is accessible and comprehensible to consumers who are not necessarily experts in domain decomposition and structure comparison. This requires appropriate user interfaces for navigation in fold space and the instant visualization of structural similarities. The COPS web service (Classification Of Protein Structures) provides an example of current developments in this area [29**,30**] (accessible at http://cops.services.came.sbg.ac.at).

It is advisable to endow classifications with metric information. Metric relations are required for the quantification of structural similarities among domains or folds and for the analysis of global properties of fold space. The proper yardsticks are structure alignment tools. The major parameter returned by these programs is the alignment length. This parameter provides a quantitative and intuitive measure of the extent of similarity. Moreover, similarity defined in this way has the properties of a metric [31**]. Here is a final example illustrating the use and power of metric relationships.

A term frequently heard in the structural biology arena is the notion of a novel fold. A declaration like ‘this is a novel fold’ expresses an opinion rather than a factum. To be intelligible we need to be precise. The same applies to statements regarding the number of distinct folds. Whether or not two structures represent one and the same protein and that many proteins are composed of oligomers of identical chains. On the other extreme we have the opposite effect. Two structures almost always share some low percentage of similarity (e.g. a pair of helices). Hence, when $s < 20$, most of the domains are in the neighborhood of some other domain and consequently there are only very few folds that can be distinguished at low levels of similarity.

All this said a final word of caution is in order. I have assumed that all the structures presented here are correct. Grossly incorrect structures are comparatively rare (e.g. [33–36,37**]), but at the level of atomic detail errors abound (e.g. [38,39**]). Therefore, in a broader context, validation and quality control of structures is another topic in classification. In any case, none of the structures discussed here has been predicted by current theoretical approaches. So, structure solvers, please do not stop!

- **Figure 7.** Number of distinguishable folds (domains) as a function of relative similarity $s$. As shown in the insert, in the range $20 \leq s \leq 95$, the logarithm (log$_{10}$) of the number of distinguishable folds rises linearly as a function of relative similarity $s$. Hence, within this range the number of distinguishable folds, $n$, follows an exponential law $n = 10^\lambda s$, where $\lambda$ is the slope of log$_{10} n$ in the linear range.

The approximately 55 000 structures of the PDB repository split in roughly 200 000 compact structural domains or folds [30**]. Now imagine the domains to be points scattered in fold space. Pick any domain at random, draw a sphere of relative distance $d$ around it, and eliminate all neighbors that happen to lie within the boundary of that sphere. Repeat the recipe ($d$ constant) until you visited every domain. The number of repetitions (i.e. the number of spheres) obtained in this way provides a convenient estimate for the number of distinct folds at the level of relative similarity $s = 100 - d$. The number of distinct folds as a function of relative similarity is shown in Figure 7. At the level of $s = 80\%$, for example, we can distinguish approximately 20 000 folds. At the level of $s = 60\%$ the number drops below 10 000.

Taking the logarithm of the number of folds yields a linear function of $s$. In other words, the number of folds rises exponentially as a function of relative similarity. This implies that the domains are spread out evenly in fold space. From this perspective, fold space looks continuous rather than discrete [32]. The exponential law breaks down at high similarities. This reflects the facts that the PDB repository contains many duplicate entries of one and the same protein and that many proteins are composed of oligomers of identical chains. On the other hand we have the opposite effect. Two structures almost always share some low percentage of similarity (e.g. a pair of helices). Hence, when $s < 20$, most of the domains are in the neighborhood of some other domain and consequently there are only very few folds that can be distinguished at low levels of similarity.
Current literature
I close with a brief summary of current activities where I connect to Willy Taylor’s review of 2007 [28]. For recent updates on major structure classifications see [40–42], for collections of less common relationships like circular permutations consult [13, 43], for classification of protein oligomers and complexes see [44] and for discussions on the structure of fold space [45, 46]. Domain decomposition is covered in [47, 48, 49], genetic mechanisms of domain shuffling are discussed in [50, 51], and biological units and multiple crystal forms are addressed in [51, 52]. There is a huge body of literature on structure alignments. A list of 50 servers may be found in [53] and recent comparative analyses on structure alignment programs are found in [54, 55]. The use of structures in the prediction of protein function is addressed in [56–58]. Methods for validation and error correction are discussed in [59, 60, 61].

Acknowledgements
My sincere thanks to Hans Brandstetter and Simon Ginzinger who suggested several improvements on an earlier version of the manuscript, to Markus Gruber who provided the examples shown in Figures 1, 3 and 6 and prepared Figure 7, and to Markus Wiederstein who assisted in preparing the manuscript.

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- published in 2007
- published in 2008


