Available online at www.sciencedirect.com
ScienceDirect
Current Opinion in Structural Biology

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)

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

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

0959-440X/$ – see front matter

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.

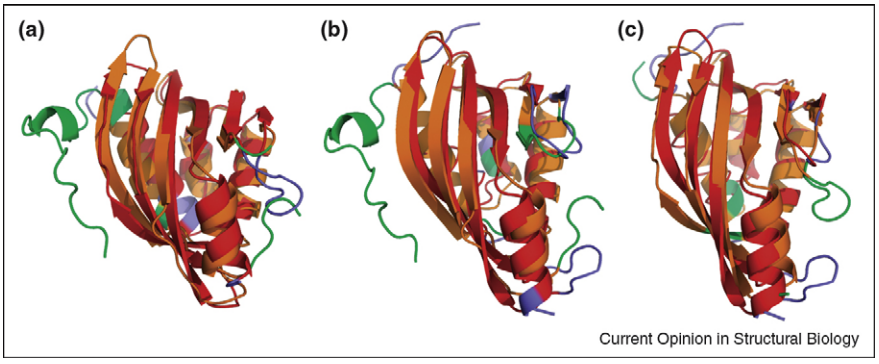
Hemoglobin and myoglobin, the very first structures solved by X-ray analysis, have closely related structures. Indeed, 135 (92%) of the 146 C^{α} atoms of the human β-hemoglobin chain (2hhb) can be structurally aligned with the sperm whale myoglobin chain (1 mbn) to a root mean square (rms) error of 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

Figure 1



Mutual structure alignments and superposition of three proteins of similar structure but distinct sequences. **(a–c)** Pairs of proteins. The first protein is in blue the second in green. Regions of structural similarity are highlighted in red (first protein) and orange (second protein), respectively. The three structures are, 3dm8-A, the A chain of a putative isomerase from Rhodopseudomonas palustris (Midwest Center of Structural Genomics, to be published), 1nww-B, the B chain of limonene-1,2-epoxide hydrolase from Rhodococcus erythropolis [62], and 3f9s-A, the A chain of a putative polyketide cyclase from Acidithiobacillus ferrooxidans (Joint Center for Structural Genomics, to be published). (a) 3dm8-A and 1nww-B, alignment length 116 (i.e. the number of structurally equivalent residues), rms-error 2.1 Å. (b) 3f9s-A and 1nnw-B, alignment length 110, rms-error 1.9 Å. (c) 3f9s-A and 3dm8-A, rms-error 2.4 Å. The percentage of identical residues in the respective alignments is in the order of <20%. Figures 1–6 were prepared using the programs TopMatch (structure alignment [29••]), TopDomain (domain decomposition), and PyMol (graphics).

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•]) and metamorphic proteins that have been observed in several alternative states (e.g. [14••,15••]).

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••,18•] designed two proteins of 88% sequence identity that fold into entirely different conformations. Getting back to natural sequences: Cordes and co-workers [19••] 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••]). 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••]. 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