ELSEVIER

Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction

Rachel Kolodny, Donald Petrey and Barry Honig

The identification of geometric relationships between protein structures offers a powerful approach to predicting the structure and function of proteins. Methods to detect such relationships range from human pattern recognition to a variety of mathematical algorithms. A number of schemes for the classification of protein structure have found widespread use and these implicitly assume the organization of protein structure space into discrete categories. Recently, an alternative view has emerged in which protein fold space is seen as continuous and multidimensional. Significant relationships have been observed between proteins that belong to what have been termed different 'folds'. There has been progress in the use of these relationships in the prediction of protein structure and function.

Addresses
Howard Hughes Medical Institute, Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Columbia University, 1130 St Nicholas Avenue, Room 815, New York, NY 10032, USA

Corresponding author: Honig, Barry (bh6@columbia.edu)

Current Opinion in Structural Biology 2006, **16**:393–398

This review comes from a themed issue on
Sequences and topology
Edited by Nick V Grishin and Sarah A Teichmann

Available online 4th May 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.04.007

Introduction

Proteins have complex three-dimensional shapes that, by eye, often bear striking similarity to one another over their entire lengths or over shorter regions. In parallel to what can be deduced from pure sequence relationships, structural similarities also suggest the possibility of evolutionary relationships between proteins. Indeed, because it is widely accepted that structure is better conserved than sequence (at least given our current ability to detect sequence relationships), the identification of structural relationships between proteins can provide important structural and functional information not available from sequence analysis alone. However, detecting geometric relationships between proteins is a far more uncertain process than the identification of pure sequence relationships, as the latter can be clearly defined in statistical terms. In contrast, there is considerable ambiguity in how to describe a geometric relationship between two proteins, resulting in the large number of approaches to this problem described in the literature.

One effective but qualitative approach is based on manual pattern recognition. Richardson's [1] classical review of structural motifs in proteins was a striking example that has evolved over the years into manually curated structure classification schemes, as epitomized by the SCOP [2] and CATH [3] databases. Implicit in SCOP and CATH is a hierarchical view whereby 'structure space' is divided into isolated, non-overlapping 'islands' that are denoted by categories such as folds. It is perhaps surprising that the concept of a fold has entered the vocabulary of structural biology in the complete absence of a clear quantitative measure of how such an entity should be described. Implicit in the hierarchical view is that protein structure space is discrete, in the sense that if a particular protein belongs to one category it does not belong to some other category.

Does the use of inherently rigid classification schemes limit our recognition of important relationships that exist between proteins that have been segregated into different categories? In principle, one could consider overlapping classifications, whereby each object is assigned to multiple classes; unfortunately, there are no overlapping classifications of protein structure space. Indeed, there is growing evidence that protein structure space is continuous, in the sense that there are meaningful structural relationships between proteins that are classified very differently. In this review, we discuss these alternative perspectives, and argue that both hierarchical and continuous views have ranges of validity. We suggest that the development of computational tools and algorithms that recognize both descriptions of structure space can enhance our ability to predict protein structure and function.

Protein structure alignment

Structural alignment programs define scoring functions that measure the geometric similarity between proteins and use various algorithms to search for two substructures such that these functions are optimal. Most existing similarity measures can be classified into two main types depending on what they compare: the distances between corresponding pairs of atoms in the two structures (e.g. [4–6]); and the relative positions of the corresponding atoms of two proteins that have been superimposed (e.g. [7,8,9•,10,11]). It had been expected that the structural

alignment problem, under either of these formulations, is NP-hard [12]; however, Kolodny and Linial [13•] recently reported a polynomial time algorithm that guarantees finding an (approximate) optimal solution for a whole class of scoring functions of the second type. Their main conclusion is that any efficient solution to the structural alignment problem must search the 'superposition space' of the two structures being compared or, equivalently, optimize a scoring function of the second type.

Several recent studies have introduced new structure similarity measures that are quite different from those used in traditional approaches. Rogen and Fain [14] suggest describing the shape of a protein backbone by a vector of 30 values inspired by mathematical knot theory and define the similarity between two structures as the (Euclidean) distance between their corresponding vectors. Calculating the similarity of two structures under this measure is instantaneous. More importantly, it is a pseudo-metric and hence satisfies the triangle inequality, which is paramount to automatic clustering, or visualization, of protein structure space. Note that any similarity measure between two proteins that is defined on substructures of these two proteins cannot satisfy the triangle inequality [14]. Erdmann [15] suggests another knot-theory-inspired similarity measure and provides algorithms to calculate it. Ye and Godzik [16], and Shatsky *et al.* [17] suggest flexible structural alignment algorithms, whereby one of the two proteins being compared is bent at several hinge points; the similarity is measured on corresponding rigid parts. This approach is especially important given the large conformational changes proteins can undergo. Friedberg and Godzik [18••] suggest a similarity measure for protein folds, which is a normalized count of the number of fragment pairwise alignments between proteins populating those folds.

The availability of so many structural alignment programs makes it difficult to establish common standards as to how structural similarity should be described. Some groups have carried out comparisons of different programs, using receiver operating characteristic (ROC) curves to evaluate how well the similarities found by a structural alignment method imitate a gold standard classification [19]. Then, using CATH [20,21] or SCOP [22] as the gold standard, they compare the ROC curves of different methods. One problem with this approach is that a program is penalized for detecting cross-fold similarities, even though clearly many such similarities exist. Also, the structural alignment program SSAP influenced the creation of CATH, making it, in effect, the gold standard structural alignment program [21]. To address this issue, Kolodny *et al.* [23••] recently evaluated structural alignment programs by directly comparing properties such as alignment length, RMS distance and number of gaps for more than four million protein pairs. The direct comparison of alignments from different programs also allows the creation of a 'best-of-all' method, which returns, for every pair of structures, the best alignment found by several programs; this 'joint effort' outperforms all the individual methods that it uses.

The nature of fold space

SCOP [2] and CATH [3] describe fold space in very similar ways. In SCOP's manual classification, the first two levels, 'class' and 'fold', are defined based purely on structure; the next level, 'superfamily', takes into account both structure and function, and the level below accounts for sequence as well, thus grouping proteins with clear evolutionary relationships. CATH combines manual classification with the automatic structural alignment program SSAP [6]: the topmost level, 'class', is based on secondary structure composition; the second level, 'architecture', is classified manually; the third level, 'topology (fold family)', depends on the shape and connectivity of the secondary structures, and is classified using SSAP; and the last level, 'homology', uses sequence information. Ultimately, the presence or absence of a structural relationship between two proteins is determined by the category to which they are assigned.

FSSP [24] is a database that does not use a classification scheme. All-on-all alignments are available and a continuous measure of structural similarity is provided. Isolated examples of relationships between proteins that would be treated as unrelated based on hierarchical protein classification schemes have been observed for some time [25]. However, that this might be a more general feature of protein structure space has only recently been widely recognized. This issue was discussed extensively by Yang and Honig [8], who carried out an all-on-all alignment of proteins in the PDB using the PrISM program. As emphasized in that work, there is no unambiguous way of clustering proteins into discrete groups, as a significant number of overlaps and ambiguities will inevitably exist.

Recent applications of structure alignment that do not incorporate categorizations from the hierarchical databases and rely only on objective measures of similarity have provided further examples of cross-fold similarities. A detailed analysis using the structure alignment program CE [5] was among the first studies to describe the interrelationships between protein substructures [26]. Others [23••,27] have noticed a similar phenomenon in CATH. In one study [27], structure alignment was used to calculate a property of proteins, termed 'gregariousness', which reflects how often a given protein has some substructure in common with other proteins that are classified as belonging to different folds. It was found that, for some classes of proteins, there were a significant number of cross-fold similarities between substructures and that, for these classes, a continuous view of fold space may be more appropriate. Kihara and Skolnick [28] have shown that small proteins (up to 100 residues) can be similar to

other proteins with very different secondary structure compositions ('class' in the CATH hierarchy). For 24% of the small proteins in their database, there is some other protein classified as having a different class with an RMSD of less than 3.5 Å that overlaps 60% of its structure.

Newer measures of similarity, such as that developed by Rogen and Fain [14], and Hou *et al.* [29], have the useful property that they enable visualization of protein fold space. That is, proteins can be represented as points in two-dimensional or three-dimensional space, with the distance between them reflecting their structural similarity. The pictures (Figure 2 of both studies) produced in this way are compelling illustrations of the overlap of different regions of fold space. A further example of this phenomenon was described by Krishna and Grishin [30••] in their discussion of 'structural drift'. They found that two proteins that are likely to be evolutionarily related based on sequence and functional analysis belong to multiple fold categories, depending on how their domains are defined (Figure 1 in [30••]). In other words, such proteins are hybrids of two overlapping subdomains that would be classified as different folds according to traditional classification schemes.

Figure 1 illustrates the complexities associated both with classification and with ambiguities in structure alignment programs. The figure shows a structural alignment of the DNA mismatch repair protein PMS2 (PDB code 1ea6) and TM1457 (PDB code 1s12), a protein of unknown function from *Thermotoga maritima*. TM1457 was described as having a new fold by the group that determined the structure [31], based on their analysis of a search for similar structures using the DALI server [4]. Moreover, TM1457 was a target (T201) at the last Comparative Assessment of Protein Structure Prediction (CASP) conference and was classified as a 'new fold' target (based on a search for similar structures in the PDB using the program MAMMOTH) [32]. Despite being classified as having a new fold, TM1457 is structurally similar to PMS2 (as identified by the SKA program [33], which is a modified version of the PrISM structure alignment tool [8]). Both proteins have an identical arrangement of secondary structure elements (SSEs) (although, as shown in Figure 1a, PMS2 has several additional SSEs). Standard measures, such as the DALI Z-score, suggest that these two proteins are unrelated. However, as each SSE in TM1457 has a structurally equivalent SSE in PMS2, the possibility exists that TM1457 is a related protein, but represents a 'substructure' of PMS2.

That these two proteins might actually be related is suggested by analysis of their molecular surfaces. In

Figure 1

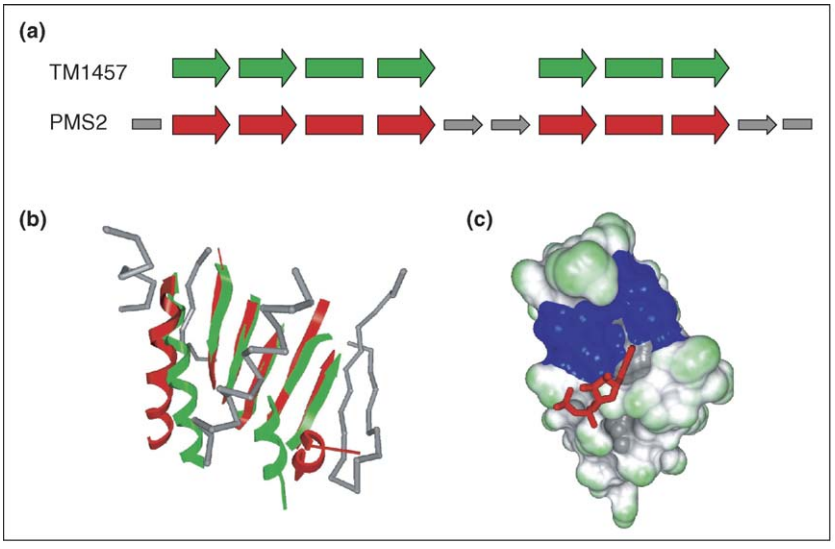

Illustration of a structural relationship between TM1457, a hypothetical protein from *T. maritima*, and the DNA mismatch repair protein PMS2. The alignment covers 78 residues, representing 81% of the smaller protein (TM1457), and has a Cα RMSD of 3.7 Å for the aligned residues. **(a)** Alignment of the SSEs comprising TM1457 (green) and PMS2 (red). Arrows represent β strands and rectangles represent α helices. SSEs present in PMS2 but not in TM1457 are in grey. **(b)** Structure alignment of TM1457 and PMS2 using the same color scheme as in (a). **(c)** The molecular surface of TM1457, showing a cleft that may be a ligand-binding site. An ADP molecule, shown in red, is taken from the structure of PMS2, with the transformation relating the two structures having been applied to its coordinates. The color scheme is based on two properties of the protein: the curvature of the surface, with convex regions colored green and concave regions colored grey; and sequence conservation, with residues in blue strongly conserved between sequence homologs of TM1457.

particular, when the two proteins are structurally aligned, a cleft on the surface of TM1457, identified by Shin *et al.* [31], aligns nicely with a cleft on the surface of PMS2 that is known to bind ADP (Figure 1c). Moreover, in an analysis of a multiple sequence alignment of TM1457 homologs, Shin *et al.* noticed two conserved regions that line the cleft identified as a putative binding site (mapped to the molecular surface in blue in Figure 1c). Although this function prediction has not been tested experimentally, we stress that the classification of TM1457 as a new fold obscures the intriguing possibility that this protein might be a nucleotide-binding protein.

Does the description of fold space matter? Applications

The discrete and the continuous views of fold space have different advantages. The hierarchical classifications of proteins into evolutionarily related sequence families and superfamilies can be carried out in a relatively unambiguous fashion, and have the advantage that they are annotated and validated by experts in the field. Also, the sequence neighbors of every protein are well defined. The organization of this information into well-maintained databases is clearly extremely valuable. The separation of proteins into folds/topologies, however, is more ambiguous and here, in particular, a continuous view may be more appropriate.

As has been recently discussed [34•], the way fold space is described is particularly important for protein structure/function prediction. Indeed, pairs of proteins that are classified as belonging to different folds can, in fact, be quite similar over large regions of their structure and share a common function (Figure 1 in [34•]). Furthermore, it has become apparent that relationships between remotely related proteins can be used in the definition of fragments that can be assembled in the modeling of proteins of unknown structure. As is evident from recent CASP results, 'fragment-based' methods have proved to be particularly successful in structure prediction [35]. Such methods exploit local sequence similarities between template structures and a query sequence whose structure is to be predicted, and assemble final, compact structures from the fragment templates that have been identified. These templates may involve as few as approximately ten residues [36•,37•], partial domains [38•,39•] or larger substructures [40], and do not, in general, involve proteins that are categorized as being in the same fold.

Fragment-based methods are also finding increasing application in function prediction [41]. Friedberg and Godzik [18••,42] compared a new fragment-based similarity measure of protein structures and a protein function similarity measure that is based on gene ontology (GO) descriptors for function annotation. They show a significant correlation between these two measures. This is of particular interest because their fragment-based similarity measure finds many cases of SCOP cross-fold similarities. Proteins belonging to different SCOP folds are not normally expected to have a similar function. Hou *et al.* [43••] created a three-dimensional map of protein structure space (denoted SSM) using the similarities calculated by DALI and studied the pairwise distances within this map. They test how different distance measures predict function, as defined using GO annotation, and show that SSM distance performs better than the raw DALI score, the DALI Z-score and the sequence-based BLAST E-value. Note that the SSM is a metric with the property that a set of ''modest but consistent'' similarity scores between a group of proteins will place them in the same region of three-dimensional Euclidean space. It is possible that the reliance on a set of similarities, as opposed to a simple pairwise similarity measure, may be the source of the improved function annotation.

Conclusions

The increasing number of protein structures in the PDB and the availability of many fast programs that compare protein structures reveal many unsuspected similarities in protein structure space. Traditional discrete hierarchical classification schemes group proteins with clear evolutionary relationships. At the structural level, these classifications constitute an abstraction that groups structures into topologies and folds based on similarities that have been detected based, in part, on visual inspection. Basing a classification scheme on such an abstraction can effectively reveal common patterns, but it can also obscure meaningful geometric relationships between proteins that have been placed in different categories. As we have seen, such relationships can be used in the prediction of structure and function. Consequently, it might be preferable if the term 'fold' was reserved for general descriptions of a given protein (i.e. ''the protein has a fold consisting of ...'') and was not to be used to imply the existence of a unique relationship between SSEs.

In our opinion, instead of describing a protein as belonging to an existing or new fold, it would be more informative to report the value of a quantitative measure of structural similarity to one or more existing proteins. The similarity can then be quantified by the alignment's properties (e.g. RMSD and length); given the alignment, these quantities are well defined and easily calculated. Unfortunately, an alignment that is found with one program will not necessarily be found with another. Nor is there evidence that some programs are significantly more effective than others for all structure pairs. A natural response to this inherent ambiguity is to use a variety of programs for a given application.

We suggest that the use of structural alignment methods offers the promise of identifying structural and functional relationships between proteins that have not been detected so far. It is difficult at this stage to do this in

an entirely automatic fashion, but effective tools are available that facilitate the discovery of functional insights, such as the one suggested in Figure 1. Structural genomics initiatives around the world produce large quantities of structural data that can be used as a basis for the discovery of new sequence/structure relationships between proteins. It is our opinion that we should be approaching the new data with the understanding that we may emerge from the next few years with very different views of protein sequence/structure/function space than we have today. Designing data structures and algorithms that recognize that our views may be in the process of changing may thus be of considerable value.

Acknowledgements

We are grateful to Michael Levitt, Chris Tang, Mickey Kosloff and Burkhard Rost for many helpful discussions on the topics covered in this review. This work was supported in part by the Northeast Structural Genomics Consortium (NESG – GM074958). The thinking reflected in this review has evolved in part as a result of facing the challenges of NESG target selection.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Richardson JS: **The anatomy and taxonomy of protein structure**. *Adv Protein Chem* 1981, **34**:167-339.

2. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data**. *Nucleic Acids Res* 2004, **32**:D226-D229.

3. Pearl F, Todd A, Sillitoe I, Dibley M, Redfern O, Lewis T, Bennett C, Marsden R, Grant A, Lee D *et al.*: **The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis**. *Nucleic Acids Res* 2005, **33**:D247-D251.

4. Holm L, Sander C: **Protein structure comparison by alignment of distance matrices**. *J Mol Biol* 1993, **233**:123-138.

5. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Eng* 1998, **11**:739-747.

6. Taylor WR, Orengo CA: **Protein structure alignment**. *J Mol Biol* 1989, **208**:1-22.

7. Subbiah S, Laurents DV, Levitt M: **Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core**. *Curr Biol* 1993, **3**:141-148.

8. Yang A-S, Honig B: **An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance**. *J Mol Biol* 2000, **301**:665-678.

9. • Krissinel E, Henrick K: **Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions**. *Acta Crystallogr D Biol Crystallogr* 2004, **60**:2256-2268.
The authors present a new publicly available method for structural alignment. This fast method performed well in a recent evaluation by Kolodny *et al.* [23••].

10. Kleywegt GJ: **Use of non-crystallographic symmetry in protein structure refinement**. *Acta Crystallogr D Biol Crystallogr* 1996, **52**:842-857.

11. Madej T, Gibrat JF, Bryant SH: **Threading a database of protein cores**. *Proteins* 1995, **23**:356-369.

12. Eidhammer I, Jonassen I, Taylor WR: **Structure comparison and structure patterns**. *J Comput Biol* 2000, **7**:685-716.

13. • Kolodny R, Linial N: **Approximate protein structural alignment in polynomial time**. *Proc Natl Acad Sci USA* 2004, **101**:12201-12206.
A theoretical study showing that protein structural alignment can be solved in (high) polynomial time by exhaustive enumeration of all rotations and translations of one structure with respect to the other.

14. Rogen P, Fain B: **Automatic classification of protein structure by using Gauss integrals**. *Proc Natl Acad Sci USA* 2003, **100**:119-124.

15. Erdmann M: **Protein similarity from knot theory and geometric convolution**. In *Proceedings of the Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB), March 2004*; *San Diego, USA: ACM Press; 2004:* 195-204.

16. Ye Y, Godzik A: **Flexible structure alignment by chaining aligned fragment pairs allowing twists**. *Bioinformatics* 2003, **19**:II246-II255.

17. Shatsky M, Nussinov R, Wolfson HJ: **FlexProt: alignment of flexible protein structures without a predefinition of hinge regions**. *J Comput Biol* 2004, **11**:83-106.

18. •• Friedberg I, Godzik A: **Connecting the protein structure universe by using sparse recurring fragments**. *Structure* 2005, **13**:1213-1224.
The authors study the structure of protein fold space, using a new interfold similarity measure based on the frequency of fragments shared between folds. A web-based application that accompanies this work enables interactive probing of their perspective of fold space. They also show that their similarity measure correlates well with a function similarity measure.

19. Gribskov M, Robinson NL: **The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching**. *Comput Chem* 1996, **20**:25-343.

20. Novotny M, Madsen D, Kleywegt GJ: **Evaluation of protein fold comparison servers**. *Proteins* 2004, **54**:260-270.

21. Sierk ML, Pearson WR: **Sensitivity and selectivity in protein structure comparison**. *Protein Sci* 2004, **13**:773-785.

22. Leplae R, Hubbard TJ: **MaxBench: evaluation of sequence and structure comparison methods**. *Bioinformatics* 2002, **18**:494-495.

23. •• Kolodny R, Koehl P, Levitt M: **Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures**. *J Mol Biol* 2005, **346**:1173-1188.
This study describes a large-scale comparison of six publicly available protein structural alignment methods. The authors argue that, instead of comparing the methods using ROC curves and a classification 'gold standard', one can directly compare different alignments. This strategy also enables the definition of a method, denoted 'best-of-all', that runs several structural alignment algorithms and returns their best alignment.

24. Holm L, Sander C: **The FSSP database: fold classification based on structure-structure alignment of proteins**. *Nucleic Acids Res* 1996, **24**:206-209.

25. Holm L, Sander C: **Globin fold in a bacterial toxin**. *Nature* 1993, **361**:309.

26. Shindyalov IN, Bourne PE: **An alternative view of protein fold space**. *Proteins* 2000, **38**:247-260.

27. Harrison A, Pearl F, Mott R, Thornton J, Orengo C: **Quantifying the similarities within fold space**. *J Mol Biol* 2002, **323**:909-926.

28. Kihara D, Skolnick J: **The PDB is a covering set of small protein structures**. *J Mol Biol* 2003, **334**:793-802.

29. Hou J, Sims GE, Zhang C, Kim SH: **A global representation of the protein fold space**. *Proc Natl Acad Sci USA* 2003, **100**:2386-2390.

30. •• Krishna SS, Grishin NV: **Structural drift: a possible path to protein fold change**. *Bioinformatics* 2005, **21**:1308-1310.

The authors describe a mechanism of evolution of protein structures and an example of it, coined 'structural drift'. Structural drift is characterized by proteins that are a hybrid of two overlapping subdomains, which are similar to proteins with very different structures.

31. Shin DH, Lou Y, Jancarik J, Yokota H, Kim R, Kim S-H: **Crystal structure of TM1457 from *Thermotoga maritima***. *J Struct Biol* 2005, **152**:113-117.

32. Tress M, Tai CH, Wang G, Ezkurdia I, López G, Valencia A, Lee B, Dunbrack RL Jr: **Domain definition and target classification for CASP6**. *Proteins* 2005, **61**:8-18.

33. Petrey D, Honig B: **GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences**. *Methods Enzymol* 2003, **374**:492-509.

34. • Petrey D, Honig B: **Protein structure prediction: inroads to biology**. *Mol Cell* 2005, **20**:811-819.
This review of the overall protein structure prediction process contains a discussion of the importance of the description of fold space in template-based modeling.

35. Moult J, Fidelis K, Tramontano A, Rost B, Hubbard T: **Critical assessment of methods of protein structure prediction (CASP) - round VI**. *Proteins* 2005, **61**:3-7.

36. • Bradley P, Malmström L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KMS, Baker D: **Free modeling with Rosetta in CASP6**. *Proteins* 2005, **61**:128-134.
The approaches to protein structure prediction described in [36•–39•,40] are all variations on a theme involving the identification of local similarities between a protein whose structure is to be predicted and multiple template structures that generally have different folds/topologies. The fragments are eventually assembled into a final compact fold. The success of these methods at CASP6 demonstrates the importance of understanding the local relationships between protein structures that the hierarchical classification schemes obscure.

37. • Jones DT, Bryson K, Coleman A, McGuffin LJ, Sadowski MI, Sodhi JS, Ward JJ: **Prediction of novel and analogous folds using fragment assembly and fold recognition**. *Proteins* 2005, **61**:143-151.
See annotation to [36•].

38. • Zhang Y, Arakaki AK, Skolnick J: **TASSER: an automated method for the prediction of protein tertiary structures in CASP6**. *Proteins* 2005, **61**:91-98.
See annotation to [36•].

39. • Zhang Y, Skolnick J: **The protein structure prediction problem could be solved using the current PDB library**. *Proc Natl Acad Sci USA* 2005, **102**:1029-1034.
See annotation to [36•].

40. Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM: **A ''FRankenstein's monster'' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation**. *Proteins* 2003, **53**:369-379.

41. Szustakowski JD, Kasif S, Weng Z: **Less is more: towards an optimal universal description of protein folds**. *Bioinformatics* 2005, **21**:II66-II71.

42. Friedberg I, Godzik A: **Fragnostic: walking through protein structure space**. *Nucleic Acids Res* 2005, **33**:W249-W251.

43. •• Hou J, Jun SR, Zhang C, Kim SH: **Global mapping of the protein structure space and application in structure-based inference of protein function**. *Proc Natl Acad Sci USA* 2005, **102**:3651-3656.
The authors extend their previous work and construct a three-dimensional map of 'protein structure space' using multidimensional scaling (MDS). After showing that proteins with similar structures and functions co-localize in the map, they use the structure space map (SSM) distance for function prediction. For this task, SSM distance outperforms structural and sequence similarity measures.