25 Borsani, G., DeGrandi, A. Ballabio, A., Bulfone, A., Bernard, L., Banfi, S., Gattuso, C., Mariani, M., Dixon, M., Donnai, D. et al. (1999) Hum. Mol. Genet. **8**, 11–23

26 Iwabe, N., Kuma, K. and Miyata, T. (1996) Mol. Biol. Evol. **13**, 483–493

27 Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L. and Postlethwait, J. (1999) Genetics **151**, 1531–1545

Received 9 September 1999

Searching for the ideal forms of proteins

W. R. Taylor
Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, U.K.

Abstract

A modification of the Structure Alignment Program (SAP), combined with a novel automatic method for the definition of structural elements, correctly identified the core folds of a variety of small β/α proteins when compared with a series of ideal architectures. This approach opens the possibility of not just determining whether one structure is like another, but given a range of ideal forms, determining what the protein is. Preliminary studies have shown it to work equally well on the all α-class and the all-β class of protein, each of which have corresponding ideal forms. Given the speed of the algorithm, it will be possible to compare all of these against the Protein Structure Database and determine the extent to which the current ideal forms can account for the variety of protein structure. Analysis of the remainder should provide a base for the development of further forms.

Introduction

With the large number of protein structures now known, it is difficult to gain an overview of their variety of forms and even more difficult to comprehend how each structure relates to its neighbours. Systematic attempts have been made to instil order into this bewildering variety, most notably in the heirarch classifications captured in the SCOP [1], CATH [2] and DALI [3] structure databases. The order in these collections is based on the pairwise comparison of protein structures using either intuitive (SCOP), automatic (DALI) or combined (CATH) approaches. The classification of structures based on comparison is strongest when the proteins are most similar, so all these collections differ little in their allocation of similar structures. The more difficult task is when there is some similarity in the fold of more than two proteins with each having different features in common. In this situation, the decision to group proteins together can often be arbitrary or, more cautiously, not made. The latter solution leads to a large number of distinct entities and rather than producing a tall classification tree, results in a low bush.

This situation is similar to that of the naturalist of the 19th Century who classified groups based on numbers of legs, teeth, bones and other features, giving rise to strong relationships between similar animals (or plants using different features of course) but less as the similarity became more distant, and if the organisms shared no common features, then they could not be considered related. To move on from this ‘collecting’ phase requires the adoption of an underlying theory that can structure the different groups given only weak evidence and, in the absence of data, can provide a default relationship as a working model until proven otherwise. For the naturalist, the underlying theory was provided through the ideas of evolution and ultimately through the modern phylogenetic analysis of sequence data. It might seem that a definitive resolution of the protein classification problem could also be attained along similar lines. However, protein structures are more strongly conserved through evolution than the sequences that embody them, which implies that the difficult areas in protein structure classification cannot be resolved through sequence comparison.

Without recourse to an evolutionary history, an alternative approach is to search for unifying structural principles that can be represented as idealized protein—protein archetypes, or their underlying Platonic forms. An indication of what

Key words: protein architecture, protein classification, protein comparison, protein structure.

Abbreviations used: SAP, Structure Alignment Program; RMS, root mean square; PDB, Protein Structure Database; 3chy, chemotaxis-Y protein; 5 nul, short-chain flavodoxin; 2fcr, long-chain flavodoxin; 1 etu, ribosomal elongation factor Tu; 5p21, ras oncogene protein p21; 3adk, adenylate kinase; 1kev, alcohol dehydrogenase.

these may be like can be found through comparison, in which some folds are seen to occur more frequently than others [4,5]. However, this observation is a derivative of the data and depends on what we have seen already; this does not have the power to bridge across areas of sparse data. A more direct attempt (with strong affinity to the Platonic forms) was developed by Murzin and Finkelstein [6] for the all-α class of protein. These were represented by quasi-regular polyhedra and provided a good model for the cores of many all-α proteins. Similar 'stick' models were also easily constructed for the all-β class of protein [7,8], the alternating β/α class of protein [9] and, as a special case, the eight-fold β/α barrel [10] (for a review, see Finkelstein and Ptitsyn [11]).

Generally, these ideal forms have been taken as frameworks for structure prediction (e.g. Taylor [12], using the polyhedral constructions [6]) but the degree to which they can represent the observed protein structures has never been systematically investigated. In all discussions and analysis, there has been a considerable degree of intuition involved in their assignment and assessment. In the current study, a method is described that can make this analysis automatic, avoiding many of the assumptions that have caused problems in the past (including the definition of secondary structure).

Methods

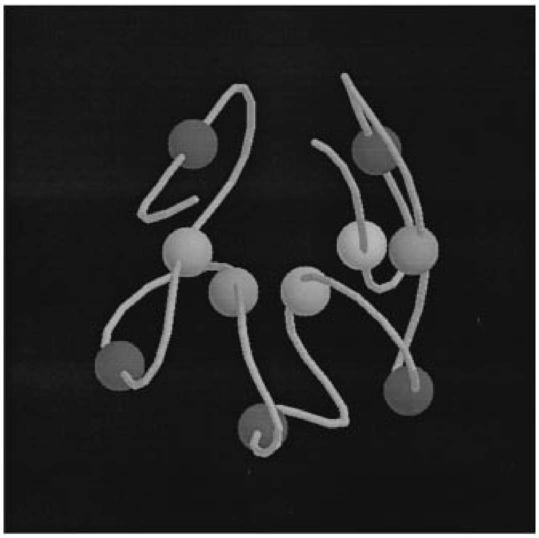
Sub-structure definition

The comparison of the ideal forms to protein structures can best be made by reducing the 'real' protein to a 'stick' representation. Commonly this is done by representing the linear secondary structures by their axes. This depends on having unambiguous definitions of secondary structure, which, despite automatic approaches, are often sensitive to structure quality. This area of ambiguity can be avoided by relying on a purely geometric definition of line segments. The axis of a secondary structure is typically taken as the line with minimum deviation (least-squares) from the α-carbons. This can be found as the principle axis of the equivalent inertial ellipsoid [13]. More generally, if the size of the three inertial axes are given by A, B and C (in descending order), then for a good linear structure the ratio $A/(B+C)$ will be large. This ratio can be calculated for all segment sizes at all residue positions and the optimal combination of segments found by dynamic-programming in a similar way to the definition of transmembrane segments [14].

Figure 1

Simplified representation of 3chy

The smoothed backbone trace of the chemotaxis-Y protein (3chy) is shown with the mid-points of the automatically defined line segments shown as spheres. These are shaded by their residue-density (see the text) with more dense segments (helices) shaded darker. The three-layer 2-5-3 structure can be clearly seen.



To make the calculation more equivalent over β-strands and α-helices, the protein structure was initially smoothed as described in [15]. No smoothing or inertial ratios, however, were calculated over chain breaks. This approach parses the protein structure into lines and each line can be characterized by the number of residues/length (referred to below as its residue-density). This measure is effectively equivalent to a definition of secondary structure but, as will be seen below, it is not necessary to make this explicit, thereby allowing more freedom for ambiguous structures (loops, 3_{10}-helices or distorted β-strands) to adopt different positions (Figure 1).

Stick-figure comparison

Angle and distance matching

The stick figures might be compared directly with each other using a structure comparison program such as Structure Alignment Program (SAP); however, the fold of the ideal forms were not specified initially since direct comparison would require testing every possible fold over the ideal form. Even for small proteins (of more than ten segments) the combinations become excessive. To avoid this, the stick figures were further reduced into a matrix of pairwise line interactions. As in other similar comparison methods, these were characterized by their distance and angle. The former was taken as the closest approach of the two line segments while the latter was the unsigned dihedral angle. These two measures are independent of line direction and so eliminate the difference between parallel and antiparallel interactions. Some interactions will be more important than others and this was quantified by the degree of overlap of their line segments.

Using these values for any given match of the real protein to the ideal form, a root mean square (RMS) deviation can be calculated over all pairs of segments for both the angles and distances weighted by overlap. When this is small, a good fit to the ideal form will have been found.

Finding the best segment assignment

In the SAP program, consecutive triples of points are taken in each structure and the similarity of the remaining points compared in the coordinate frame defined by each triple. This assessment is made on the basis of point separation and relative orientation and the best matching pairs found by dynamic-programming [16,17].

The current problem can be approached in a similar way, except that each triple was selected on the basis of local structural similarity and were not necessarily adjacent in the sequence. Similarly, the dynamic-programming algorithm cannot be used as it assumes that the equivalent points will be in linear order. Instead, the 'stable-marriage' algorithm [18] was used to reconcile the matrix of conflicting preferences into a one-to-one pairwise assignment.

This assignment was not taken as absolute and some limited recombination among the weaker pairs was allowed. In the results reported below, the 25 best-matching triples were used to each generate 25 minor variants. This process was repeated 24 times and in each subsequent calculation, some random noise was introduced into the score matrix before calculation of the variations. This latter device is similar to the introduction of noise in the SAP program [17]. As each calculation (including the line segment definition stage) takes less than one second (on a 600 MHz Pentium processor), computation time is not a limiting factor in this approach.

All alignments with a weighted distance RMS deviation of less than 5 Å and a weighted angle RMS deviation of less than 0.5 radians were considered for assessment in three dimensions.

Final evaluation

From the alignment of segments generated by the preceeding method, it is possible to construct an ideal stick figure with the same fold as the real protein. This reintroduces direction to the sticks and allows a direct comparison between the two structures. To make this comparison even more direct, the stick lengths of the real protein were set to the same length as their ideal counterparts (typically 10 Å). These equivalent stick figures were then passed to the SAP program for a full three-dimensional comparison (Figure 2).

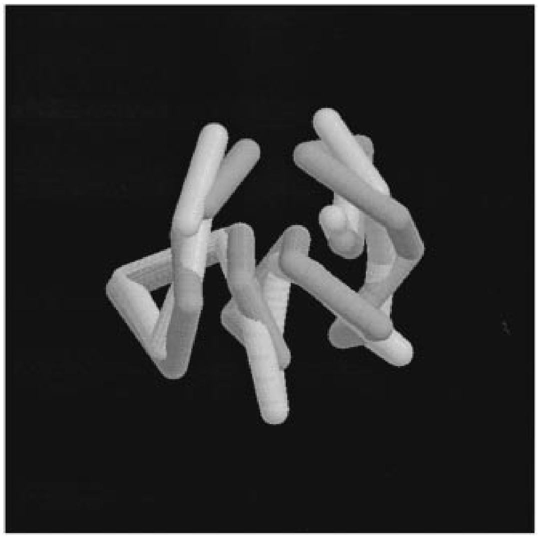
From an initial assessment of the results, based on the model with the best SAP score, some unusual behaviour was seen. Occasionally, an α-helix would be matched in the position normally assumed to be a β-strand. While valid from a geometric perspective, this behaviour was not desirable for real proteins and was discouraged by multiplying the SAP score by a factor reflecting the difference in secondary structure. However, as secondary structure was never explicitly defined, this was taken as a difference in their residue-densities (see above).

Similarly, some small line segments were chosen in preference to the larger segments. Again, while representing valid geometric solutions, it was preferable to see the larger structures matched-up. This was encouraged by similarly

Figure 2

Superposed stick figures of 3chy and its ideal form

The stick figure representation of 3chy (white) superposed on the corresponding stick-figure of the ideal form (grey) is shown in the same orientation as Figure 1. The structures match with a 3.4 Å RMS deviation over all 20 matched end-points (Table 1).



multiplying the SAP score by a factor derived from the length of the real secondary structure.

Some solutions were still found with transposed β-strands. While these were easy to recognize visually in the superposed stick figures, they were less easy to avoid by constraining the filter on the distance of RMS deviation. If this were made too strict, then equivalent α-helices would be missed that had a deviation just as great as two transposed strands. Instead, the SAP score was divided by the weighted RMS deviation, giving a stronger penalty against any errors in close pairs of β-strands.

Data

The method was tested on small members of the alternating β/α family of proteins. These exhibit a wide variety of different folds based on a core architecture of a central β-sheet packed on both sides by α-helices. A shorthand can be used to describe these proteins by summarizing the numbers of α-helices above the sheet, the number of β-strands in the sheet, and the number of α helices below—not unlike the system for classifying steam locomotives by their wheels (number of leading bogies, drive wheels and trailing bogies). In this system, Gordon (the large engine) is a 2-6-6 class, whereas Thomas is only a 0-6-0 class.

Real proteins

The test set of real proteins is described below with their Protein Structure Data bank identifier in parentheses and their packing class in brackets. Where it is uncertain how many helices pack against the sheet, a ' + ' is shown to indicate other uncounted helices. Similarly, where the number is ambiguous (for example, whether distorted strands or helices are counted) then the options are separated by slashes.

Three proteins were selected from the flavodoxin fold group, including the chemotaxis Y protein (3chy) [2-5-3], a simple (short-chain) flavodoxin (5nul) [2-5-2/3] and the long-chain variant (2fcr), which has an extended β-sheet [2-6/7-2/3]. Two proteins were taken from the small G-protein family, including the ribosomal elongation factor Tu (1etu) [2-6-3] and the *ras* oncogene protein p21 (5p21) [2-6-3]. Although these latter two have the same core fold, they have different arrangements on one edge of their β-sheets where the elongation factor has an inserted domain in its full structure. This gave a chance to test the algorithm with a structure that contains a chain break. Adenylate kinase (3adk) [2-5-3 +] was taken as a further example of a protein with a different core fold from all the others. Finally, a 'classic' Rossman fold domain was extracted from the structure of alcohol dehydrogenase (1kev 152:293) [3-6-2/3]. This is an interesting inclusion, as unlike the other structures (above), it begins with an α-helix and not a buried β-strand.

Ideal forms

The ideal form taken to represent these proteins was similar to that used previously for prediction [9] and consisted of a core β-sheet with a 20° twist between β-strands which were spaced at 5 Å at their mid-points. α-helices were placed above and below this sheet using a construction that preserved the local interactions with the sheet—as previously used in the construction of ideal frameworks for transmembrane helices [19], creating a realistic staggered packing between the helices. Each helix lay, on average, 10 Å above the sheet and each secondary structure was 10 Å in length.

From this general structure four instances were constructed with five and six β-strands and differing numbers of α-helices. Using the 'locomotive' classification scheme, these followed the progression: 2-5-2, 2-5-3, 2-6-3 and 3-6-3.

Results

Each of the protein structures described above was compared with each of the four ideal forms. Their goodness-of-fit was evaluated by the RMS deviation of the real stick figure from the ideal stick figure, as calculated by the SAP program, based on the aligned segment end-points (Table 1).

The overall level of the RMS deviations is roughly what would be expected for the comparison of the α-carbon coordinates between any unrelated pair of these proteins and indicates that the ideal form does not show any particular bias towards a particular fold. Each result was examined individually using Rasmol and all the matches were found to have the correct corresponding topology with no strand transpositions or misassigned secondary structures, as had sometimes been seen in the initial testing of the method. The RMS deviations generally got slightly larger as more elements were incorporated, but this was not always so: for example, both 3chy and 3adk have lower values with the 2-5-3 form relative to the 2-5-2 form. This does not, however, imply a better fit of the common elements but only a better than average fit for the additional element.

Table 1

RMS deviations from the ideal forms for a range of small *β*/*α* class proteins specified by the Protein Data Bank (PDB) codes

Each column gives the RMS deviation to the ideal form specified by its 'locomotive' class corresponding to the number of *α*-*β*-*α* segments in each layer (see text for details). The RMS values are unweighted over all the equivalent end-points of the secondary structures, the number of which is given in parentheses at the top of each column. A dash indicates that either no solution was found by the matching program, or it did not incorporate all the elements of the ideal form. Each match was examined and all were found to be a good topological match.

	2-5-2 (18)	3-5-2 (20)	3-6-2 (22)	3-6-3 (24)
3chy	3.305	3.260	—	—
5nul	4.002	4.471	—	—
2fcr	4.997	5.073	5.237	—
3adk	5.774	5.070	—	—
letu	5.418	5.484	5.821	—
5p21	4.917	5.227	5.428	6.773
1kev	2.800	2.891	3.264	—

In the flavodoxin fold group, 3chy attained the best fit with its full complement of structures and any matches with larger ideal forms simply reproduced this full fit. With 5nul the best fit was equivalent to 3chy and was attained by matching a small 3_{10}-helix in the place of a corresponding α-helix in 3chy. 2fcr produced an extended fit to the first six-strand form by matching each of the parts of the edge strand that is broken by the large insert. This is not unreasonable as these two 'halves' have a region of overlap where they hydrogen-bond to each other. The fit to adenylate kinase (3adk) reached only the 2-5-3 ideal form. Although there are other helices below the sheet, only three of these pack.

In the G-protein fold group, both proteins were correctly fitted to a six-strand sheet with an antiparallel hairpin on the edge of the sheet. Despite the similarity in the two proteins, 5p21 continued to have a reasonable fit with an additional helix (3-6-3 class). Graphical investigation revealed that this extra segment was an extended loop on the edge of the domain, in the place where an extra helix might lie. The extension was not made in letu, however, as this is the location of the missing domain.

The 'classic' Rossman fold has a 2-6-2 packing class but the addition of the N-terminal helix in (1kev) raises this to the 2-6-3 class which was correctly identified—also with the lowest RMS deviation seen across the proteins considered. There is a small C-terminal helix in 1kev but this does not interact with the sheet.

Conclusions

The ability of this method to find solutions up to, but not beyond, the core fold of the protein opens the possibility for its use as a classification tool. Given a series of ideal forms, it is necessary only to present these in order of size and select the largest solution. Unlike the visual analysis of topology cartoons, this approach is completely automatic and is focused on the well-packed core elements of the structure (which are not always obvious in topology cartoons). Finding solutions based on the core also means that two proteins can be compared even though they do not have the same overall fold. This can be done by looking back at their match to smaller ideal forms and if a common solution is found then this can be taken as a measure of relatedness. For example, even though Gordon and Thomas are engines with quite distinctive characters and functions, they have the common feature of six drive-wheels and a common fit to the 0-6-0 classification, giving them a stronger relationship compared with many smaller tank engines.

Looking ahead, the use of such analysis of protein structure will reveal the extent to which the ideal forms are able to account for the variety of protein structure. This is important for the prediction of structure from sequence. At the moment the most successful prediction schemes are based on comparison of a sequence with known structures. Given a complete range of ideal forms, this limitation could be overcome.

References

1 Murzin, A. G., Brenner, S. E., Hubbard, T. and Chothia C. (1995) J. Mol. Biol. **247**, 536–540

2 Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. and Thornton, J. M. (1997) Structure **5**, 1093–1108

3 Holm, L. and Sander C. (1997) Nucleic Acids Res. **25**, 231–234

4 Murzin, A. G. and Chothia C. (1992) Curr. Opin. Struct. Biol. **2**, 895–903

5 Orengo, C. A., Jones, D. T. and Thornton, J. M. (1994) Nature (London) **372**, 631–634

6 Murzin, A. G. and Finkelstein, A. V. (1988) J. Mol. Biol. **204**, 749–769

7 Cohen, F. E., Sternberg, M. J. E. and Taylor, W. R. (1980) Nature (London) **285**, 378–382

8 Finkelstein, A. V. and Reva, B. A. (1991) Nature (London) **351**, 497–499

9 Cohen, F. E., Sternberg, M. J. and Taylor, W. R. (1981) J. Mol. Biol. **147**, 253–272

10 Lesk, A. M., Branden, C. I. and Chothia, C. (1989) Protein Struct. Funct. Genet. **5**, 139–148

11 Finkelstein, A. V. and Ptitsyn, O. B. (1987) Prog. Biophys. Mol. Biol. **50**, 171–190

12 Taylor, W. R. (1993) Protein Eng. **6**, 593–604

13 Taylor, W. R., Thornton, J. M. and Turnell, G. (1983) J. Mol. Graphics **1**, 30–38

14 Jones, D. T., Taylor, W. R. and Thornton, J. M. (1994) Biochemistry **33**, 3038–3049

15 Taylor, W. R. (1999) Protein Eng. **12**, 203–216

16 Taylor, W. R. and Orengo, C. A. (1989) J. Mol. Biol. **208**, 1–22

17 Taylor, W. R. (1999) Protein Sci. **8**, 654–665

18 Sedgewick, R. (1990) Algorithms in C. Addison-Wesley

19 Taylor, W. R., Jones, D. T. and Green, N. M. (1994) Protein Struct. Funct. Genet. **18**, 281–294

Received 1 September 1999

Using the CATH domain database to assign structures and functions to the genome sequences

F. Pearl*, A. E. Todd*, J. E. Bray*, A. C. R. Martin*, A. A. Salamov†, M. Suwa†, M. B. Swindells*, J. M. Thornton*‡ and C. A. Orengo*[1]

*Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, U.K., †Helix Research Institute, 1532-3 Yana, Kisarazu-Shi, Chiba, 292, Japan, and ‡Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, U.K.

Abstract

The CATH database of protein structures contains ~ 18000 domains organized according to their (C)lass, (A)rchitecture, (T)opology and (H)omologous superfamily [1]. Relationships between evolutionary related structures (homologues) within the database have been used to test the sensitivity of various sequence search methods in order to identify relatives in Genbank and other sequence databases [2]. Subsequent application of the most sensitive and efficient algorithms, gapped blast and the profile based method, Position Specific Iterated Basic Local Alignment Tool (PSI-BLAST) [3], could be used to assign structural data to between 22 and 36% of microbial genomes in order to improve functional annotation and enhance understanding of biological mechanism. However, on a cautionary note, an analysis of functional conservation within fold groups and homologous superfamilies in the CATH database, revealed that whilst function was conserved in nearly 55% of enzyme families, function had diverged considerably, in some highly populated families. In these families, functional properties should be inherited far more cautiously and the probable effects of substitutions in key functional residues carefully assessed.

Introduction

There are nearly 20000 known domain structures in the Protein Databank [4] now held in the Research Collaboratory for Protein Structures at Rutgers University. These data still lag considerably behind the known sequences (~ 400000 currently in Genbank). However, with the advent of the structure genomic initiatives [5] we can expect the numbers to increase substantially and there are suggestions that we may know all the major folds in nature within the next 5 years. Once their structures have been determined, interest will focus on methods for assigning functional properties to these proteins.

In order to recognize and understand structural and functional relationships between proteins, we have clustered all the known structures into fold groups and evolutionary superfamilies using a combination of automatic and manual

Key words: protein structure classification.

Abbreviations used: CATH, class, architecture, topology, homologous; PSI-BLAST, Position Specific Iterated Basic Local Alignment Tool; DDD, DALI Domain Directory; DHS, Dictionary of Homologous Superfamilies; FMN, flavin mononucleotide.

[1] To whom correspondence should be addressed.