Comparison of protein structural profiles by interactive computer graphics

Stephen H Bryant* and Michael J E Sternberg

Laboratory of Molecular Biology, Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

This paper describes a novel computer graphics tool for predicting protein structures. The method is based on structural profiles, which are plots of hydrophobicity, parameters used for secondary structure prediction, or other residue-specific traits against sequence number. Similar structural profiles can indicate similar tertiary structures, in the absence of sequence homology. The profiles of reference proteins, with known structure, can be used for prediction. In the method presented here, structural profiles are compared by interactive computer graphics, using the program Multiplot. As a test, a structural profile comparison of several proteins known to have similar 3D structures is presented. Comparison of structural profiles detects similar folding of the two domains of rhodanese, which was not easily detected by sequence homology.

Keywords: proteins, protein structure prediction, hydrophobicity profiles, computer graphics

Received 7 July 1986
Accepted 28 July 1986

Computer graphics has often been used for the prediction of protein folding. Most straightforward has been the use of programs such as Frodo[1,2] to construct molecular models based on sequence homology with proteins whose 3D structure is known[3].

When no such homology can be detected less powerful methods of structure prediction must be used. One that has come into wide use that is also based on graphics is the display of hydrophobicity profiles[4]. These identify regions that will form regular secondary structure. Novotny and coworkers[5,6] have proposed that hydrophobicity plots and other structural profiles can also be used for the prediction of tertiary structure. They showed that similar folding of immunoglobulin domains can be detected by display of their structural profiles.

Comparison of structural profiles may become a powerful extension to the technique of predicting protein structure by homology. To explore this possibility the authors have examined the structural profiles of a series of proteins which are known to fold in a similar manner but which display decreasing sequence homology. These are compared using the interactive graphics program Multiplot, written for the Evans and Sutherland PS300.

*Current address: Department of Chemistry, Brookhaven National Laboratory, Upton, Long Island, NY 11973, USA

METHODS

Structural profiles

The structural profiles used are hydrophobicity and secondary structure propensities. Hydrophobicity is taken from the OMH scale of Sweet and Eisenberg[7], which is derived from evolutionary residue interchange frequencies within structural families. The scale of Bryant and Amzel[8], derived from nearest neighbour contact frequencies in folded protein, gave very similar results. Secondary structure propensities are taken from the scales of Chou and Fasman[9], which are directly based on the frequency with which each residue type appears in helix, sheet and turn structures. Secondary structure probabilities calculated according to the method of Garnier *et al.*[10] gave similar results. Where indicated, structural profiles were smoothed according to the '4(3RSR)2H twice' method of Tukey[11]. This is based on repeated medians taken over a window of three residues, and gives profiles smooth enough for comparison over long sequences.

Structural profiles are displayed as 2D plots, e.g. hydrophobicity against sequence number. Profiles may also be examined as 3D plots, e.g. helix and sheet propensities against sequence number. A preliminary example will be presented later.

Proteins examined

The sets of profiles examined were those of human immunoglobulin constant domains, from the structures of FAB New[12] and Fc[13], rhodanese domains[14], the nucleotide binding domains of dehydrogenases, from the structures of alcohol dehydrogenase[15], lactate dehydrogenase[16], and glyceraldehyde phosphate dehydrogenase[17], and α/β barrel structures, from triose phosphate isomerase (TIM)[18], and ribulose bisphosphate carboxylase (RuBisCo)[19,20].

Overview of Multiplot

Multiplot is a general purpose program for display of 2 and 3D plots. It requires an Evans and Sutherland PS300 with 1 Mbyte of main storage. The program is written in a combination of PS300 command language and FORTRAN 77, using the PSIO subroutine library provided by Evans and Sutherland. Multiplot is designed to be portable across host system software. A parallel interface to the PS300 is not required, though a version of the program using this interface, under VAX/VMS, is also available.

Up to four plots may be displayed at once, each containing a different type of profile, as shown for example in Colour Plate 1. The plots may be mapped interactively to any part of the PS300 screen. Profiles of up to six proteins may be included in each plot, with a selection of colours and labelling of data points, for example with residue type names. Profiles may be moved relative to one another, so the sequences to be compared need not have been aligned previously. The profiles in Colour Plate 1, for example, were aligned by eye, using Multiplot. One may zoom in on a particular region of sequence, and pan across the whole sequence, so that comparisons at any level of detail are possible. Similarity of structural profiles may be judged by interactive superpositioning.

Detailed description of Multiplot

Multiplot accepts data to be plotted in free format, as a list of x, y, and z coordinates and α-numeric data point labels. Other information necessary to construct the plots is requested in a series of prompts. Plot specifications given in this manner may be stored, and used again later to regenerate a display.

All interactions with the display are managed by the PS300 workstation. While introducing some limitation on possible interactions, this design allows the 3D and interactive graphics features of Multiplot to be most easily adapted to existing plotting and data management packages. Control of the graphics display is primarily via valuator dials, whose action is indicated by mnemonic labels. Labelled function keys are used for switching between different interaction and display modes.

Display modes fall into two groups, those acting at the level of a plot, and those acting at the level of an individual data item (i.e. a profile) within a plot. Plot-level display modes are 2D, 3D and invisible. Data-level display modes are simply visible and invisible.

Interaction modes also fall into two groups. One plot-level interaction mode controls the plot screen window, a rectangular box defined by six clipping planes. In 2D plots data and margin windows are automatically defined within the plot window, such that proper clipping is maintained. A second plot-level interaction mode controls the orientation, position and depth-cueing of 3D plots. A single data-level interaction mode controls translation and scaling along the plot x, y and z axes. The effect of data-level interactions does not depend on the orientation of a 3D plot, and extends to tick marks and scale labels in such a way that proper labelling of the plot is maintained.

The scope of each interaction may be specified via function keys. Plots and the data items they contain may be altered singly or in combination. This feature is important in the examination of structural profiles, which share sequence number as a common axis. The different types of profile displayed in each plot may be aligned against reference profiles simultaneously, under control of a single dial.

Interactions with the graphics display remain of acceptable quality when up to several thousands of data points are included. Mode switching is effectively instantaneous, as are rotations of 3D plots, and scaling or translation of a data item displayed in any or all of the plots. The complexity of the display is limited primarily by the speed of the PS300 processor, which may not refresh the calligraphic display frequently enough if too many vectors are present. In practice this is not a problem unless one attempts to display more than a few hundred α-numeric labels, which place the greatest load on the graphics processor.

RESULTS

Immunoglobulins

Shown in Colour Plate 1 is a structural profile comparison of several human immunoglobulin constant domains. All the profiles have been smoothed. These sequences show significant homology, and were predicted to show similar folds before their 3D structures were known. It was first pointed out by Novotny[5,6] that immunoglobulin sequences are also quite similar at the level of their structural profiles. The profiles shown in Colour Plate 1 have been aligned by eye, using Multiplot. This was easily done by comparing them pairwise, and momentarily switching off those not under examination. The alignment shown follows closely the established sequence alignment[21]. The profiles agree reasonably well, and clearly can be made to agree quite well throughout their length if allowance is made for small insertions and deletions in bend regions. The CH2 profiles (Fc, residues 1–110) agree less well than do the profiles of CL (Fab light chain), CH1 (Fab heavy chain) and CH3 (Fc, residues 110–220), reflecting the different manner of domain–domain association in CH2.

At this level of homology similar folding of an entire 110-residue domain can be recognized using structural profiles. The profile most powerful in aligning the sequences is hydrophobicity. The profiles based on frequency of residue occurrence in the various secondary structure types are for the most part correlated or anti-correlated with it. β-sheet profiles for the most part reflect hydrophobicity, and turn profiles its inverse.

Rhodanese

Colour Plate 2 shows a structural profile comparison of two regions of rhodanese, again using smoothed profiles. They are part of two domains with similar tertiary structure, which are considered to be the evolutionary product of an ancient gene duplication. Though it was subsequently shown that there is a low but significant sequence homology between these domains[22], this was not generally recognized prior to determination of the 3D structure.

The rhodanese structural profiles have again been aligned by eye, using Multiplot. This alignment could

be chosen easily, as no other regions of the Rhodanese structural profiles can be superimposed as well. It follows closely the correct alignment. Given at best a low level of sequence homology, similar folding of a 70-residue region resulting from ancient gene duplication was detected by the use of structural profiles.

Dehydrogenases

Alcohol dehydrogenase, lactate dehydrogenase and glyceraldehyde phosphate dehydrogenase display a similar nucleotide binding fold in one domain, but are otherwise dissimilar[23,24]. There is no overall sequence homology.

Smoothed profiles for the entire dehydrogenase nucleotide binding domain showed no recognizable similarity (data not shown). This is certainly due in part to a large number of insertions and deletions, which hinder recognition of similar features in the profiles. Colour Plate 3 shows unsmoothed structural profiles for a small region of this domain. These have been aligned according to the structural superposition based on 3D structure[23,24]. There is clearly some agreement, much of it due to conserved key residues. It is doubtful, however, that similar folding of this region could be identified by structural profiles alone.

TIM and RuBisCo

The 3D structure of RuBisCo is not yet known, but crystallographic work is in progress, and it has been reported that a portion of RuBisCo folds into an α/β barrel as seen in TIM[25,26]. The structural profiles of RuBisCo and TIM were compared to identify the regions of structural similarity. There is no known sequence homology.

Shown in Colour Plates 4(a) and (b) are smoothed structural profiles for two RuBisCo sequences. These have been aligned with one another by eye, using Multiplot. The alignment shown follows closely that proposed by Janson *et al.*[19] on the basis of hydrophobicity correlation[7]. Two possible alignments of TIM with RuBisCo are shown. There is some agreement at the level of matching peaks and valleys in the hydrophobicity trace. These do not seem to specify a unique alignment, however, and may only reflect the regular passage of the polypeptide chain in and out of the protein core. The authors have carried out a more detailed comparison of these sequences, using unsmoothed structural profiles, but again several alignments seemed possible. In this case also a structural similarity known to be present cannot be recognized.

DISCUSSION

The comparisons of structural profiles reported could not have been carried out without the interactive plotting capabilities of Multiplot. Sequences can be scanned quickly for similarities in structural profiles, which are easily seen by eye.

The authors have established that comparison of structural profiles can detect structural similarity when there is very low sequence homology, as in rhodanese. The method thus offers some promise as a prediction technique. On the other hand, the known structural similarities of dehydrogenase nucleotide binding domains, and of TIM and RuBisCo could not be detected. There is clearly some limit beyond which structural similarity cannot be recognized, possibly due to differences in loop size.

The method can be extended by use of other types of structural profile. A preliminary example is shown in Colour Plate 5, which shows a 3D structural profile alignment of rhodanese. The smoothed profiles are based on the side chain similarity scale of Bryant and Amzel[8] (upper), reflecting hydrophobicity (y axis) and charge (z axis), and a principal components analysis of Chou and Fasman type parameters (Bryant, unpublished) (lower), reflecting the presence of regular secondary structure (y axis), and the distinction between sheet and helical structures (z axis). Though it cannot be seen in 2D, the latter profile often shows a characteristic spiral pattern for β–α–β folding units, a feature that might be easily recognized in otherwise dissimilar profiles.

Visual comparison of structural profiles may complement methods of structure prediction based on templates[27] and patterns of key residues[28]. Recognition of similar features at the level of structural profiles may suggest structural homologies that can be explored further by modelling techniques.

ACKNOWLEDGEMENTS

The authors wish to thank Prof. T Blundell for valuable discussions. This work has been supported by a grant from the AFRC.

REFERENCES

1 **Jones, T A** *J. Appl. Cryst.* Vol 11 (1978) pp 268–272
2 **Jones, T A** *Methods in Enzymology* Vol 115 (1985) pp 157–171
3 **Blundell, T and Sternberg, M J E** *Trends. Biotech.* Vol 3 (1985) pp 228–235
4 **Kyte, J and Doolittle, R F** *J. Mol. Biol.* Vol 157 (1982) pp 105–132
5 **Novotny, J and Auffray, C** *Nucleic Acid Res.* Vol 12 (1984) pp 243–255
6 **Novotny, J et al.** *Proc. Natl. Acad. Sci. USA* Vol 83 (1986) pp 742–746
7 **Sweet, R and Eisenberg, D** *J. Mol. Biol.* Vol 171 (1983) pp 479–488
8 **Bryant, S and Amzel, L** *Int. J. Pept. Prot. Res.* (in press)
9 **Chou, P Y and Fasman, G D** *Adv. Enzym.* Vol 47 (1978) pp 45–148
10 **Garnier, J. et al.** *J. Mol. Biol.* Vol 120 (1978) pp 97–120
11 **Tukey, J** *Exploratory data analysis* Addison-Wesley, Reading, MA, USA, Chapters 7 and 16 (1977)
12 **Saul, F et al.** *J. Biol. Chem.* Vol 253 (1978) pp 585–597
13 **Diesenhofer, J** *Biochem.* Vol 20 (1981) pp 2362–2370
14 **Ploegman, J et al.** *Nature* Vol 273 (1978) pp 124–129
15 **Eklund, H et al.** *J. Mol. Biol.* Vol 102 (1976) pp 27–59
16 **White, J L et al.** *J. Mol. Biol.* Vol 102 (1976) pp 759–779
17 **Buehner, M et al.** *J. Mol. Biol.* Vol 90 (1974) pp 25–49

18 **Banner, D W et al.** *Nature* Vol 255 (1975) pp 609–614
19 **Janson, C et al.** *J. Biol. Chem.* Vol 259 (1984) pp 11594–11596
20 **Shinozaki, K and Sugiura, M** *Gene* Vol 20 (1982) pp 91–102
21 **Kabat, E A** *Structural concepts in immunology and immunochemistry* Holt, Rienhart and Winston, New York, NY, USA (1976) pp 281–285
22 **Keim, P et al.** *J. Mol. Biol.* Vol 151 (1981) pp 179–197
23 **Rossmann, M G et al.** *Nature* Vol 250 (1974) pp 194–199
24 **Ohlsson, I et al.** *J. Mol. Biol.* Vol 89 (1974) pp 339–354
25 **Brädén, C et al.** *Phil. Trans. R. Soc. London* Vol B313 (1986) pp 359-365
26 **Chapman, M S et al.** *Phil. Trans. R. Soc. London* Vol B313 (1986) pp 367–378
27 **Taylor, W R** *J. Mol. Biol.* Vol 188 (1986) pp 233–258
28 **Sternberg, M J E and Taylor, W R** *FEBS Lett.* Vol 175 (1974) pp 387–391