Volume 224, number 1, 149–155 FEB 05299 November 1987

Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences

C. Gaboriaud, V. Bissery, T. Benchetrit[+] and J.P. Mornon

Groupe Cristallographie et Simulations Interactives des Macromolécules Biologiques, Laboratoire de Minéralogie-Cristallographie, CNRS UA09, Universités P6 et P7, T16, 4 place Jussieu, 75252 Paris Cedex 05 and *[+] Département de Chimie Organique, UA498 CNRS, U266 INSERM, UER des Sciences Pharmaceutiques et Biologiques, 4 avenue de l'Observatoire, 75006 Paris, France*

Received 24 September 1987

A new method for comparing and aligning protein sequences is described. This method, hydrophobic cluster analysis (HCA), relies upon a two-dimensional (2D) representation of the sequences. Hydrophobic clusters are determined in this 2D pattern and then used for the sequence comparisons. The method does not require powerful computer resources and can deal with distantly related proteins, even if no 3D data are available. This is illustrated in the present report by a comparison of human haemoglobin with leghaemoglobin, a comparison of the two domains of liver rhodanese (thiosulphate sulphurtransferase) and a comparison of plastocyanin and azurin.

Protein sequence comparison; Conformation homology; Protein structure prediction

1. INTRODUCTION

With the increasing number of known amino acid sequences, the need for efficient tools to analyse and compare them becomes greater. An important challenge is to detect whether or not two amino acid sequences are likely to fold into closely similar three-dimensional structures. Indeed, if the sequence of a protein with an unknown structure can be related to one with a known architecture, the main chain folding for the latter provides a valid starting point for the structural modelling of the former (e.g. [1–3]).

Correspondence address: J.P. Mornon, Groupe Cristallographie et Simulations Interactives des Macromolécules Biologiques, Laboratoire de Minéralogie-Cristallographie, CNRS UA09, Universités P6 et P7, T16, 4 place Jussieu, 75252 Paris Cedex 05, France

Many classical methods of sequence comparison (e.g. [4]) focus on the detection of the maximum percentage of amino acid identity. Occasionally a marked similarity in amino acid sequences is found (e.g. >50% identity) and direct structural inferences can be made [5]. However, if the similarity is less marked, which does not necessarily imply any important difference in the overall three-dimensional (3D) folding, these methods are of no practical use. For example Argos [6] observed only 18% identity among 11215 properly aligned amino acids in superimposable 3D structures [6]. On the other hand, for distantly related proteins, the alignments obtained by classical methods can be inconsistent with those derived from the superimposition of the 3D structures (e.g. in [7]). This is mainly due to the omission in these methods of the structural requirement that insertions and deletions are expected in regions exposed to solvent in the tertiary structure and especially in regions that are not part of the secondary structure [7,8].

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/87/$3.50

To take into account these structural requirements, several methods have been proposed. They are still not of widespread use, however, because of their inherent drawbacks. Some [7,8] require 3D structural information for at least one of the compared sequences, which is often a severe limitation. Others require several long sets of programs because of the sophisticated use of structural criteria, secondary structure predictions and amino acid similarity [6,9].

The method proposed here, hydrophobic cluster analysis (HCA) does not share these drawbacks. It uses a representation of protein sequences on an α-helical 2D pattern, first proposed by Lim for his a priori prediction of the rough overall tertiary structure of several proteins [10]. The postulated premise is that the nascent polypeptide is a fluctuating α-helix when produced by the ribosome, before the native folding of the protein. Independently of this postulated mechanism and without assuming its real existence, the results show that this representation allows a more straightforward and efficient visual comparison of the structural features in protein sequences than the usual method of linear writing.

2. METHODS

2.1. *HCA plot*

Fig.1 illustrates the main principles of this 2D sequence representation, using human α_1-antitrypsin.

The G246–S283 region of this sequence is written on a classical α-helix (3.6 amino acids per turn) smoothed on a cylinder. After five turns, residues i and $i+18$ (e.g. G246 and E264) have similar positions parallel to the axis of the cylinder (see fig.1a).

To make this 3D representation easier to handle, the cylinder was then cut parallel to its axis (e.g. along G246–E264 line) and unrolled (see fig.1b).

As some adjacent amino acids (e.g. F253 and L254) are widely separated by the unfolding of the cylinder, the representation was duplicated, making the sequence easier to follow and giving a better impression of the environment of each amino acid (see fig.1c).

Sets of adjacent hydrophobic residues in the pattern were then encircled and termed hydrophobic clusters. The six residues adjacent to the amino

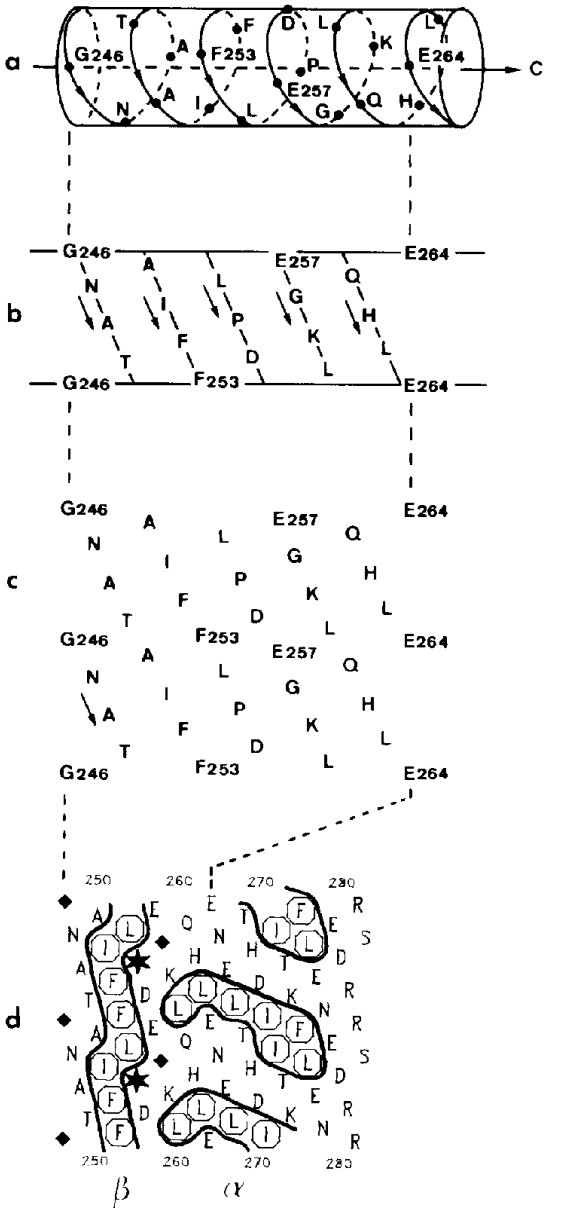

Fig.1. Principles of protein sequence analysis through HCA plots. The example is the G246...E264...S283 part of the sequence of human α_1-antitrypsin [19]: GNATAIFFLPDEGKLQHLE... See section 2 for full details.

acid i were designated $i-4$, $i-3$, $i-1$, $i+1$, $i+3$, $i+4$.

In order to define hydrophobic clusters, I, L, F, W, M, Y, V were considered as hydrophobic amino acids, whereas P was primarily considered

as a breaker of these clusters, and A and C as mimetic, i.e. hydrophobic only in a hydrophobic environment. This classification, albeit rather crude, is consistent with the OMH scale [11] and the Nozaki and Tanford scale [12].

To make the visual inspection of the pattern easier, the hydrophobic residues were encircled, and different symbols used for prolines (★) and glycines (◆) which are often present in loops, and cysteines (Ⓒ) which may be involved in disulfide bonds (fig.1d). To simplify the use of the method, a program, HCA plot, which plots the sequences on this pattern, was written in Fortran 77, using colour codes for amino acids.

2.2. *Sequence comparison*

HCA plots have been used for the comparison of a large number (>80) of protein sequences and based on this experience a few guidelines can be given to facilitate sequence comparison.

(i) A similarity of the hydrophobic clusters patterns is sought for this purpose. The overall distribution of the clusters along the sequences should be compared, as in the hydrophobic profiles method [13], but also taking into account their precise size, shape and orientation.

Hydrophobic clusters having the closest features are used as anchors for the structural alignment. The segmentation induced by prolines, glycines and hydrophilic areas is also a good criterion for the structural alignment of the protein sequences.

(ii) A precise sequence alignment from comparison of HCA plots can then be deduced, starting from the core of each aligned hydrophobic cluster and proceeding towards the boundaries of each structural segment. Thus insertions or deletions, if required, occur mainly in loops indicated by prolines, glycines and hydrophilic areas.

(iii) A numerical score of the matching of two hydrophobic clusters can be computed as follows:

$$\text{HCA homology score} = \frac{2CR \times 100}{RC_1 + RC_2}\%$$

where RC_1 (RC_2) is the number of hydrophobic residues in cluster 1 (cluster 2). CR is the number of hydrophobic residues in cluster 1 that are in correspondence with hydrophobic residues in cluster 2.

3. RESULTS

3.1. *Examples of sequence comparisons through HCA*

To illustrate the use of HCA, three alignments were chosen, described in recent papers as difficult to achieve correctly using classical methods [6,7,11,13]. These examples belong to different structural classes, as defined by Levitt and Chothia [14]: all α, α/β, all β.

3.1.1. Comparison of human haemoglobin (α-chain) and lupin leghaemoglobin (all α)

Although the different origins of these two proteins and their low sequence identity (⩽15%) make them very distantly related, their overall folding is similar [15].

With HCA plot patterns, it is very easy to see their structural homology and to propose a sequence alignment (see fig.2). The hydrophobic clusters C1 to C7 cover 65% of the proteins and their overall HCA homology is 80%. The alignment requires very few deletions. However, it is difficult to assign a precise alignment of the two clusters C4 because their sizes are too different.

With the exception of cluster C4, the superimposition of the 3D structures [15] of these two proteins confirms the proposed sequence alignment. It was also checked that the framed parts of the sequences (see fig.2) which anchor the structural alignment, belong to the conserved hydrophobic core of these two structures.

3.1.2. Comparison of the two domains of bovine liver rhodanese (α/β)

Despite very low sequence identity (⩽7%), these two domains have a similar structural organization [16]. Fig.3 shows the correspondences found on the HCA plot patterns. The hydrophobic residues that we postulate to be in structural correspondence are encircled in bold. Hydrophobic clusters C1 to C6 cover half of each domain and their overall HCA homology is 80%. Cluster C6 is the best conserved in the two domains (HCA homology score: 92%). Between clusters C2 and C3 the location of hydrophobic residues is quite different so that no straightforward alignment can be made.

On the whole, the superimposition of the 3D

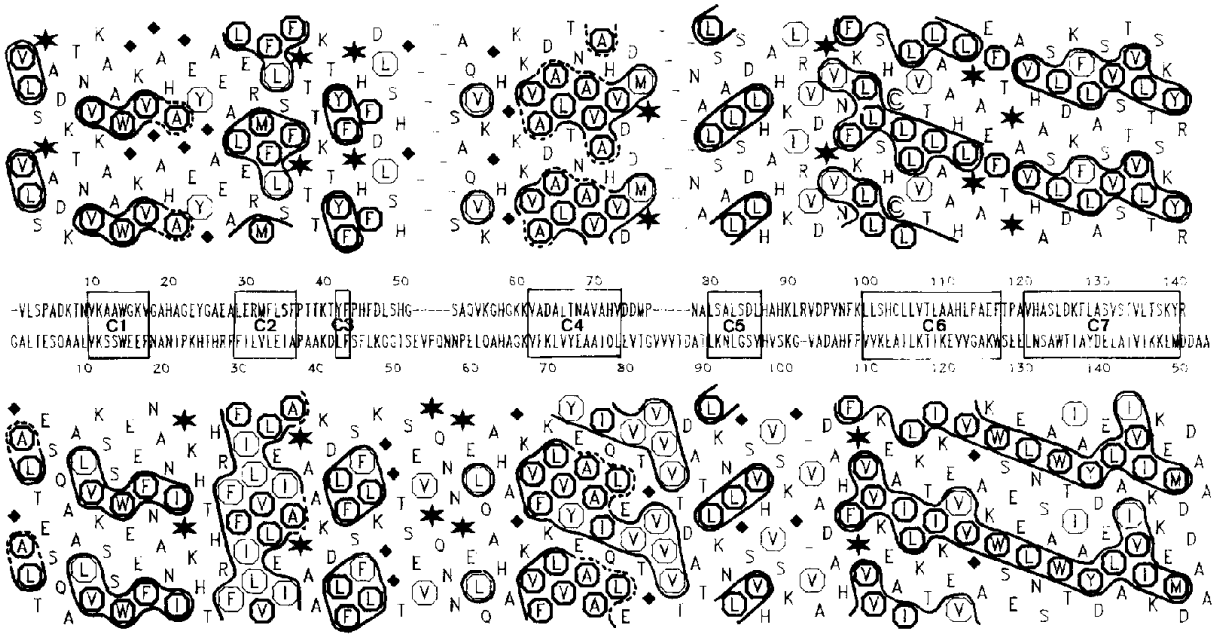

Fig.2. HCA plots and deduced sequence alignment of human haemoglobin α-chain (top) and lupin leghaemoglobin (bottom). The sequences of the conserved core of clusters C1 to C7 are framed. Dashed lines in cluster delineation indicate possible extensions suggested by the comparison between the two sequences. Deletions are indicated by $-$.

structures of these two domains confirms the proposed correspondences, except for cluster C1, which belongs to the loop junction between the two domains induced by the 139–149 proline rich region.

3.1.3. Comparison of azurin and plastocyanin (all β)

These two proteins share common folding features, as deduced from structural studies [17], although they belong to quite different functional protein families. Their sequence identity is at most 25%.

In fig.4, HCA plots suggest that the second half of these proteins (clusters C6 to C9) is structurally well conserved (HCA homology score: 80%). The major difference between the two proteins obviously comes from the presence of the long cluster A in azurin, which does not have any counterpart in plastocyanin. Regarding the N-terminal region, the distribution of the hydrophobic residues is roughly the same, but the precise shapes of the clusters differ (compact clusters or 'zig-zag' patterns), suggesting stronger divergence of the two foldings.

All these deduced features have been confirmed by visual examination of the two structures, as have the structural correspondences of hydrophobic residues, encircled in bold in fig.4. The C-terminal regions are highly homologous (rms (root mean square): 1.5 Å), the β-strands of the N-terminal regions are similarly arranged but with a greater rms, which is above 2 Å. The long A cluster of azurin is the only α-helix of these two proteins. The zig-zag patterns are associated with β-structures, as often observed.

3.2. *Criteria of validation of the method*

As illustrated in the previous section, this method has been tested on the comparison of different structural classes of globular proteins: all α, α/β and all β. In all β-proteins, the clusters are shorter and so the comparisons are not always as easy as for all α-proteins. Nonetheless, this method works well and is, for another example, particularly successful in the analysis of the serine protease family (not shown).

Since in practical cases the protein structures are unknown, we have noted some empirical criteria that can help in deciding whether or not two pro-

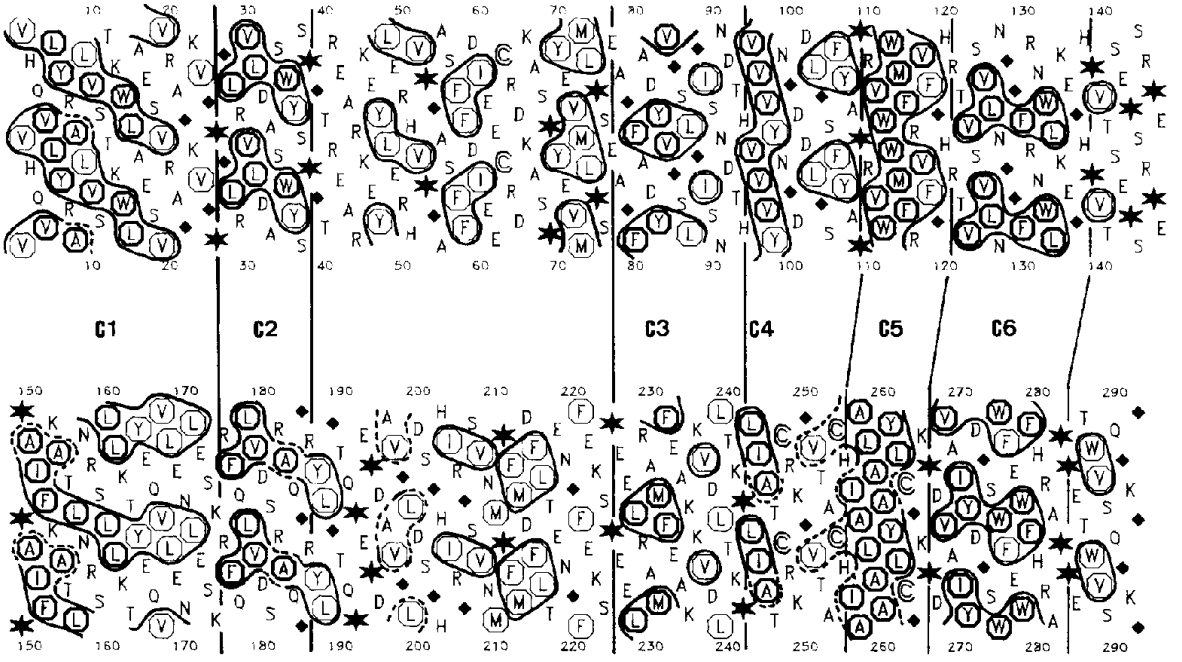

Fig.3. HCA plots of the two domains of bovine liver rhodanese showing their structural homology. Vertical lines indicate the proposed correspondences between the two domains.

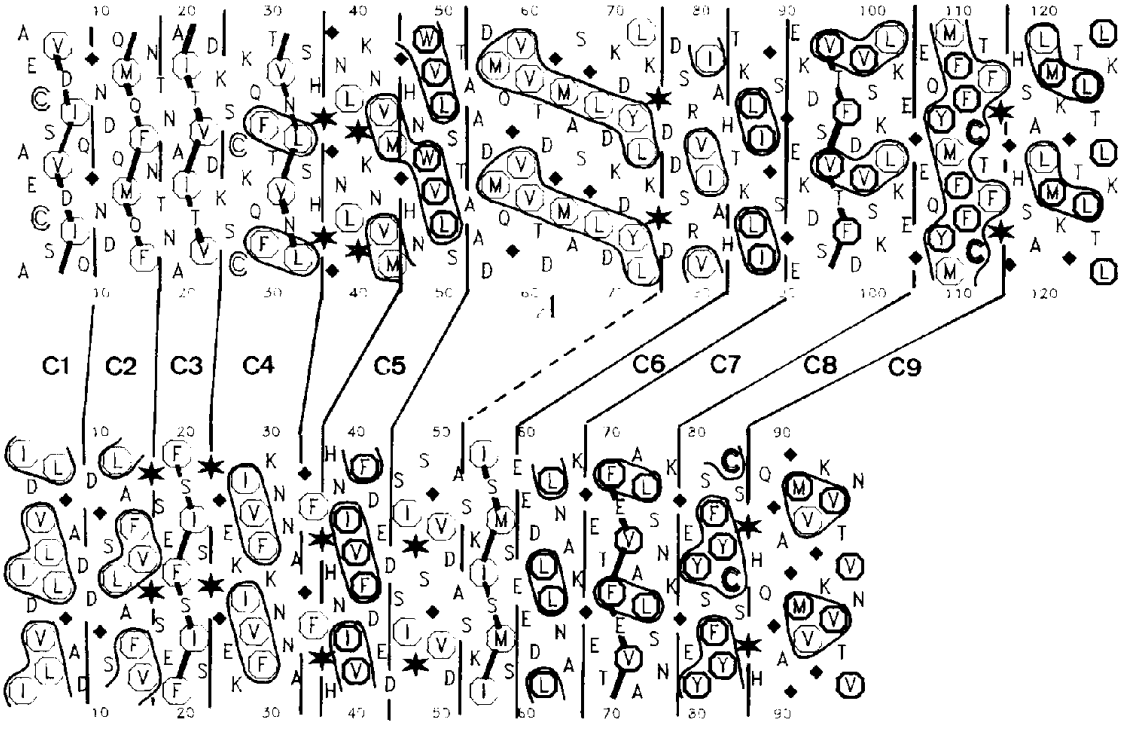

Fig.4. HCA plots of azurin (top) and plastocyanin (bottom). Vertical lines indicate the proposed correspondences between the two sequences. Proposed homologous hydrophobic residues are encircled in bold. 'Zig-zag' pattern lines are drawn.

teins are likely to have closely homologous folds (not shown).

The first criterion for assuming a common folding is the continuous presence of conserved clusters all along the sequences, or at least in 50-residue lengths. If two independent sequences are compared, several of these clusters are characterised by at least 80% HCA homology, especially when the clusters are short. The persistence of consensus hydrophobic clusters in different members of a given family is a stronger criterion, even if the HCA score sometimes decreases to 65%. This criterion is never met in proteins having very different folding topologies. In closely homologous proteins, the clusters as previously described correspond to regions superimposed with an rms less than 1 Å, or less than 2 Å in the worst cases.

The second criterion is the overall HCA homology score, calculated for the whole set of hydrophobic clusters, as given in the previous section. Different ranges of scores (s) have been observed in the following sensitivity tests: (i) proteins with close peptide chain folding (rms < 2 Å), $s > 75\%$, e.g. trypsin and kallikrein; (ii) proteins with similar secondary structure topologies, $50 < s < 65\%$, e.g. lactate and glyceraldehyde-3-phosphate dehydrogenases; (iii) proteins with completely different folds, $s < 50\%$, e.g. trypsin and concanavalin A.

These orders of magnitude can also be applied to the detection of local differences in the degree of structural homology, as exemplified in the previous section.

HCA was also used with proteins of unknown 3D structure. Structural homologies have been found between thermolysin and the neutral endopeptidase 24.11 [18], between steroid and thyroid receptors, and between uteroglobin and prostatic binding protein. These results will be published in detail elsewhere.

4. DISCUSSION

The results presented here are representative of the simplicity of handling HCA plots and show how strikingly the conserved structural features can be detected using this method, even in distantly related proteins. With closely related proteins, the alignment can be made at a first glance.

The efficiency of the method relies upon the 2D representation, which allows rapid perception of the long-range environment of each residue. Indeed, the α-helical pattern gives rise to efficient condensation of the information. This pattern is well adapted to the recognition of amphiphilic α-helices. It has also been observed that several defined types of short clusters are strongly associated with β-strands. Indeed, regular secondary structures in proteins have often similar lengths, clearly shorter than the protein itself. Since extended protein structures progress by 3.6 Å per residue, and α-helices progress only by 1.5 Å per residue, in almost all cases the hydrophobic clusters associated with α-helices contain many more residues than those associated with β-strands (regular or irregular). Thus, plausible rules of secondary prediction through cluster shape analysis are under study. Typical examples are presented in fig.5. Another useful aspect of this representation is its ability to incorporate general structural features together with specific residue patterns, as for example the conservation of disulfide bridges or glycosylation sites, which can be easily seen.

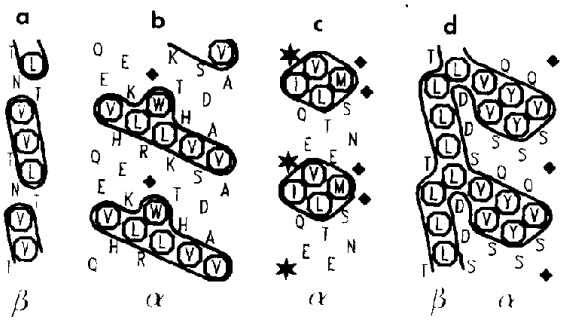

Fig.5. Typical shapes of clusters associated with definite secondary structures. The frequency of observation (O.F.) is defined as the ratio of the number of clusters associated with the indicated secondary structure to the total number of clusters displaying the typical shape in a sample of 28 independent proteins of the Protein Data Bank [20]. (a) β-Strand, O.F., 21/22. Cluster shown, segment (58–63) of dihydrofolate reductase. (b) α-Helix, O.F., 25/32. Cluster shown, segment (181–200) of triose phosphate isomerase. (c) α-Helix, O.F., 9/10. Cluster shown, segment (249–261) of alcohol dehydrogenase. (d) β-Strand, O.F., 17/20. Cluster presented here in combination with an α-helix, segment (214–238) of penicillopepsin. Another example is shown in fig.1d.

As the method has been designed for molecular modelling purposes the following characteristics are emphasised.

(i) The minimum set of insertions or deletions required to achieve global alignment of all the clusters is needed. The method is designed in such a way that these insertions will predominantly appear in surface areas, which is extremely important for molecular modelling purposes.

(ii) HCA plots seem to be an efficient tool for evaluation of whether two proteins are likely to share a close homologous folding, and where the closest features occur. In particular, we have noticed that, through the conserved hydrophobic clusters, much of the conserved hydrophobic cores can be predicted. A detailed statistical analysis is in progress to confirm this.

(iii) This representation allows easy observation of structurally conserved features in a particular protein family, and thus enables simultaneous manual alignment of several sequences of the same family.

(iv) At this stage of development the method cannot be used for databank screening, because of its manual handling, but it can be used easily as a further screening step. The process of automation is under study.

By writing to the author, the researcher's name will be added to a list for future HCA plot program distribution.

ACKNOWLEDGEMENTS

The authors are indebted to Ann Beaumont and David Marsh for language correction and to Marcel Bardet for technical support. Thanks are also due to Thierry Rabilloud for help in preparing the manuscript. This work was partly supported by the CNRS and the INSERM, and by a grant from the Fondation pour la Recherche Médicale. V.B. was supported by a fellowship from the Fondation pour la Recherche Médicale.

REFERENCES

[1] Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) Nature 326, 347–352.
[2] Greer, J. (1981) J. Mol. Biol. 153, 1027–1042.
[3] Jones, T.A. and Thirup, S. (1986) EMBO J. 5, 819–822.
[4] Bishop, M.J. and Rawlings, C.J. (1987) Nucleic Acid and Protein Sequence Analysis: A Practical Approach. In: The Practical Approach Series, IRL Press, Oxford.
[5] Chothia, C. and Lesk, A.M. (1986) EMBO J. 5, 823–826.
[6] Argos, P. (1987) J. Mol. Biol. 193, 385–396.
[7] Lesk, A.M., Levitt, M. and Chothia, C. (1986) Prot. Engineering 1, 77–78.
[8] Taylor, W.R. (1986) J. Mol. Biol. 188, 233, 258.
[9] Ptitsyn, O.B., Finkelstein, A.V., Kirpichnikov, M.P. and Skryabin, K.G. (1982) FEBS Lett. 147, 11–15.
[10] Lim, V.I. (1978) FEBS Lett. 89, 10–14.
[11] Sweet, R.M. and Eisenberg, D. (1983) J. Mol. Biol. 171, 479–488.
[12] Nozaki, Y. and Tandford, C. (1971) J. Biol. Chem. 246, 2211–2217.
[13] Bryant, S.H. and Sternberg, M.J.E. (1987) J. Mol. Graphics 5, 4–7.
[14] Levitt, M. and Chothia, C. (1976) Nature 261, 552–558.
[15] Lesk, A.M. and Chothia, C. (1980) J. Mol. Biol. 136, 225–270.
[16] Keim, P., Heinrikson, R.L. and Fitch, W.M. (1981) J. Mol. Biol. 151, 179–197.
[17] Adman, E.T. (1984) in: Metalloproteins: Metal Proteins with Redox Roles (Hanison, P.M. ed.) pp.1–42, Verlag Chemie, Basel.
[18] Benchetrit, T., Mornon, J.P., Crine, P. and Roques, B.P. (1987) Biochemistry, in press.
[19] Travis, J. and Salvesen, G.S. (1983) Annu. Rev. Biochem. 52, 655–709.
[20] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.