HCABAND : A computer program for the 2D-helical representation of protein sequences

Bernard Henrissat, E. Raimbaud, V. Tran, J.P. Mornon

CABIOS

Vol.6. no.1. 1990
Pages 3–5

HCABAND: A computer program for the 2D-helical representation of protein sequences

B.Henrissat*, E.Raimbaud[1], V.Tran[1] and J.P.Mornon[2]

Abstract

A BASIC program, HCABAND, is described which is used to convert the linear amino acid sequence of proteins into a word processor-readable file to generate the 2D-helical representation required for hydrophobic cluster analysis. The user can specify the width of the plot and can use the word processor macro-commands to facilitate visual inspection of the plots. The plots can be easily stored on diskettes, mixed with text and printed. HCABAND features are generally applicable and can be implemented on virtually any microcomputer.

Introduction

With the advent of DNA sequencing, the amino acid sequences of a large number of proteins are now available. This has led to an increasing number of methods for sequence comparison and for secondary structure prediction (Schulz, 1988; Taylor, 1988; Thornton, 1988; Hodgman, 1989). One of the major goals of protein comparison is to determine whether two sequences are related and, consequently, whether they have a similar fold. Most classical methods are based on the maximization of alignment scores with various comparison matrices. However, these methods are not reliable when the amino acid identity is low (Lesk *et al.*, 1986; Barton and Sternberg, 1987). Hydrophobic cluster analysis (HCA) is a method to compare amino acid sequences (Gaboriaud *et al.*, 1987) which is derived from the theory of Lim (1974) and the representation of Schiffer and Edmunson (1967). HCA has proven to be able to detect topological similarities between proteins of low amino acid identity (Benchetrit *et al.*, 1988; Henrissat *et al.*, 1988, 1989). The method involves the drawing of the sequences on a theoretical α-helix where the hydrophobic residues form clusters. The shape, size, orientation and relative position of the clusters are then compared and sequence similarity, where it exists, may be readily revealed. One of the main advantages of HCA is that it allows distant information to become visible more readily than with methods which are based solely on singlet amino acid property/identity. This communication describes a simple computer program which converts the amino acid sequence of a protein into the 2D-helical plot required for HCA.

Centre de Recherches sur les Macromolécules Végétales, CNRS, BP 53X, F-38041 Grenoble, [1]Institut National de la Recherche Agronomique, BP 527, F-44026 Nantes and [2]Laboratoire de Minéralogie-Cristallographie, CNRS UA09, Universités P6 et P7, T16, 4 place Jussieu, F-75252 Paris cedex 05, France
**To whom reprint requests should be sent*

System and methods

An IBM PC AT or compatible is required. The program was written with Microsoft GWBASIC V3.11 on a Tandon Target microcomputer but will run on any IBM PC or close compatible under MS DOS 2.0 or higher and GWBASIC or BASICA. It can be adapted to other computer systems with few modifications. HCA files can be printed under DOS command but are preferably processed through a word processor. Visual impression is enhanced if a line spacing <0.5 can be defined. We routinely use Microsoft Word V3.20 and V4.00 which give satisfactory results. Other word processors that can read ASCII files should also work. HCA plots were printed on dot matrix (EPSON-FX) or laser (HP Laserjet; NEC LC 890 Postscript) printers through the commands of the word processor.

Algorithm

With the method developed by Mornon and co-workers (Gaboriaud *et al.*, 1987), the linear protein sequence is written on a helix smoothed on a cylinder. For such a graphical representation the two important parameters are NT and NR in the following relationship: after NT turns, residues i and (i + NR) have similar positions parallel to the axis of the cylinder. In their plot, NT = 5 and NR = 18 and this describes a classical α-helix with 3.6 amino acids per turn.

Satisfactory plots can also be obtained for other pairs of NT and NR values, e.g. (4, 15) or (3, 11). The resulting helices have respectively 3.75 and 3.67 residues per turn which is close to the 'standard' value of 3.6. For reasons of convenience (Henrissat *et al.*, 1988), we have selected the pair with the lowest value of NT, e.g. NT = 3 and NR = 11. It must be noted that these values can be changed easily since they appear as parameters in the program.

The main feature of the algorithm is to convert an initial sequence (one line) into a two-dimensional array with the second dimension being the NR value (NR lines) (see Figure 1a). In order to position each symbol in the right line and column, we have created an intermediate table ID (ID[i], i varying from 1 to NR) which contains the position of the first symbol of the

line i with the following formula:

$$ID\ [i] = (j \bmod NR) + 1$$

with $j = (i - 1) \cdot X$ and $X = 4$. For values of NT = 4 and NR = 15, the value of X is 4. The value of X is 11 when NT = 5 and NR = 18.

The next operation is the filling of the two-dimensional array with blanks. For the first line, symbol 1 is located at the first rank (ID [1] = 1), symbol 2 (1 + 11) at rank (1 + 11), symbol (i + 22) at rank (i + 22) etc. For the second line symbol 2 is located at rank 5 (ID [2] = 5), symbol (2 + 11) at rank (5 + 11), symbol (2 + 22) at rank (5 + 22) etc. The use of the ID table allows the filling of the 11 lines (Figure 1A). For printing, one can duplicate the band and reduce the line spacing (Figure 1B).

The algorithm is general and it must be noted that since all the operations are done in the text mode, the program can be easily adapted to virtually any computer. A variety of word processors can be used and high quality plots can be obtained with laser printing.

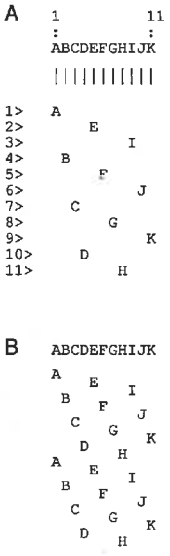

Fig. 1. Construction of the plot with NT = 3 and NR = 11. (**A**) Filling of the two-dimensional array. (**B**) Duplication of the band and reduction of the line spacing from 1 to 0.5.

Implementation

Step 1. generation of HCA files

The sequence files are ASCII files containing only the protein sequence in one-letter code (up to 5000 amino acids long). They are identified by a .SEK suffix. They can be extracted from a databank or, alternatively, they can be created manually with a word processor. It is recommended that control characters are avoided. Since protein sequences are frequently too long to fit a single band in the plot, blocks have been defined so that the program can automatically calculate a line break. Blocks have a width of NR characters and therfore correspond to 11 amino acids. The optimal 'number of blocks' is therefore a function of page size and orientation, character size and margins. The program starts by listing the protein sequence files available on the diskette drive. After the user has typed the protein sequence file name and the 'number of blocks' (i.e. how many times 11 amino acids are desired per page width), the program generates a new file with the same name but with a .HCA suffix.

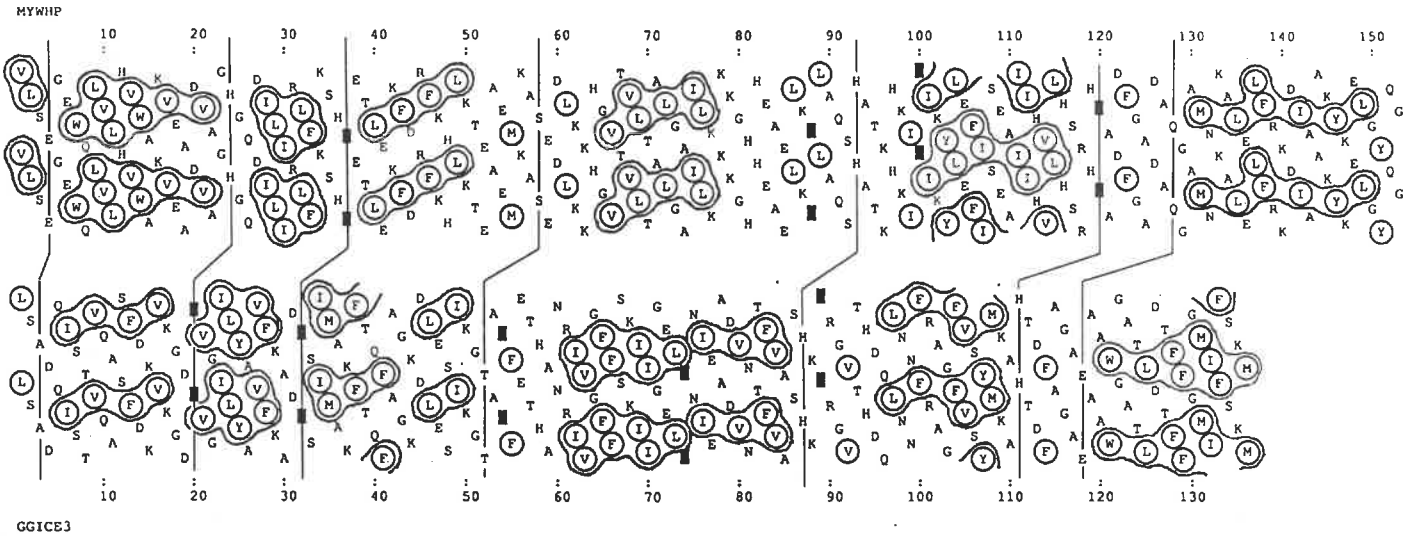

Fig. 2. HCA plots of sperm whale metmyoglobin (top) and larval *Chironomous thummi* globin (bottom). Prolines are represented by ■. Hydrophobic amino acids have been circled and their clusters have been drawn. Vertical lines indicate the proposed correspondences between the two sequences.

Step 2. word processing HCA files into HCA plots

Through the word processor's commands, the new files (with a .HCA suffix) can be loaded from the diskette and merged in order to avoid repetitive processing. A character font with fixed spacing is necessary. For a better visual appearance and to save space, line spacing should be reduced to ~0.25, 0.3, 0.4 and 0.5 line for character size of 6, 8, 10 or 12 respectively. Throughout the plot one can replace any amino acid with special properties by a dedicated ASCII character. For example, prolines are of special interest in hydrophobic cluster analysis since they are cluster breakers (Gaboriaud *et al.*, 1987). We found it convenient to replace all the prolines (P) in the plot by (■) which can be spotted more easily. This can be done automatically by the word processor. Similarly one can replace glycines (G) by (◙), etc. To avoid interruption of clusters one can duplicate each band (Gaboriaud *et al.*, 1987). Numbering of the amino acids can be combined with text if needed. The HCA plot can then be saved on disk and printed. As an example, the HCA plots of sperm whale metmyoglobin and *Chironomous thummi* globin are presented in Figure 2. The 3D structures of these two proteins are similar although they display an amino acid conservation of ~22% (Bashford *et al.*, 1987). The similarity between the HCA plots is striking and allows straightforward alignment of their sequences.

Other HCA plot programs

HCABAND is an elementary program for use with the most simple and common microcomputers. HCA plot programs using graphical displays are currently being developed in the laboratory of J.P.M. For example, MACHCA operates from any Macintosh microcomputer and allows the automatic contouring of clusters. An IBM PC version of MACHCA is scheduled to take advantage of the graphical capabilities of these microcomputers. GKSHCA is a more sophisticated program to be used with any display or plotter compatible with the graphics library GKS; it produces compact coloured HCA plots with automatic clustering and several numerical predictions. A fully interactive software, MANSEK, derived from GKSHCA is under present development and will be described elsewhere.

MACHCA can be obtained from J.P.Mornon on request.

Acknowledgements

The authors wish to thank S.Mortimer for his help with presentation in English.

References

Barton,G.J. and Sternberg,M.J.E. (1987) Evaluation and improvements in the automatic alignment of protein sequences. *Protein Eng.*, **1**, 89–94.

Bashford,D., Chothia,C. and Lesk,A.M. (1987) Determinants of a protein fold. Unique features of the globin amino acid sequences. *J. Mol. Biol.*, **196**, 199–216.

Benchetrit,T., Bissery,V., Mornon,J.P., Devault,A., Crine,P. and Roques,B.P. (1988) Primary structure homologies between two zinc metallopeptidases, the neutral endopeptidase 24.11 ('enkephalinase') and thermolysin, through clustering analysis. *Biochemistry*, **27**, 592–596.

Gaboriaud,C., Bissery,V., Benchetrit,T. and Mornon,J.P. (1987) Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. *FEBS Lett.*, **224**, 149–155.

Henrissat,B., Popineau,Y. and Kader,Y. (1988) Hydrophobic cluster analysis of plant protein sequences. A domain homology between storage and lipid transfer proteins. *Biochem. J.*, **255**, 901–905.

Henrissat,B., Claeyssens,M., Tomme,P., Lemesle,L. and Mornon,J.P. (1989) Cellulase families revealed by hydrophobic cluster analysis. *Gene*, **81**, 83–95.

Hodgman,T.C. (1989) The elucidation of protein function by sequence motif analysis. *Comput. Applic. Biosci.*, **5**, 1–13.

Lesk,A.M., Levitt,M. and Chothia,C. (1986) Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Protein Eng.*, **1**, 77–78.

Lim,V.I. (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.*, **88**, 857–872.

Schiffer,M. and Edmunson,A.B. (1967) Use of helical wheels to represent the structures of proteins and to identify segments with helical potential. *Biophys. J.*, **7**, 121–135.

Schulz,G.E. (1988) A critical evaluation of methods for prediction of protein secondary structure. *Ann. Rev. Biophys. Biophys. Chem.*, **17**, 1–21.

Taylor,W.R. (1988) Pattern matching methods in protein sequence comparison and structure prediction. *Protein Eng.*, **2**, 77–86.

Thornton,J.M. (1988) The shape of things to come. *Nature*, **335**, 10–11.

Received on May 9, 1989; accepted on August 28, 1989