ANTHEPROT: An interactive graphics software for analyzing protein structures from sequences

C. Geourjon, G. Deléage and B. Roux

Laboratoire de Physico Chimie Biologique, Université Claude Bernard Lyon I, Villeurbanne, France

ANTHEPROT is a fully interactive program devoted to the analysis of protein structures using a graphics workstation. It presents four options: The first option can predict secondary structures using five methods, and hydrophobicity, solvent accessibility, flexibility and antigenicity profiles using eighteen scales. The user may introduce his own scales. The results displayed on the screen can be easily analyzed. The second option is for representing results concerning up to eight proteins by one method. To compare these proteins, it is possible to align the profiles or the predicted secondary structure according to various motifs. The secondary structure deduced from crystallographic data may also be introduced. The third option is designed to compare the primary structure of two proteins and to visualize on the screen regions that exhibit similarity. Six different comparison matrices may be used, but the user can also introduce his own matrices. The last option is for studying the proteolytic peptides resulting from a chemical or enzymatic digestion of a given protein. It is possible to analyze the protein cleavage using eleven chemical reagents or enzymes. The results are displayed on the screen as RP-HPLC chromatogram.

Keywords: protein structure, secondary structure prediction, hydrophobicity, homology, chemical and proteolytic digestion, reverse phase high pressure liquid chromatography (HPLC)

INTRODUCTION

In the absence of crystallographic data, structural features of proteins can be deduced from the analysis of protein sequences. Among such analytic techniques, the prediction of functionally important residues (FIR) is a promising tool for the near future. In addition, with the increasing number of protein sequences known from DNA cloning and sequencing, the need for a theoretical treatment of protein sequences has never been greater. (See Fasman[1] for a review.) In this context we have developed during the last two years a package named ANTHEPROT (theoretical analysis of proteins), which was originally designed for microcomputers.[2–4] Because interactive graphics are essential to protein sequence analysis, we report here a fully interactive graphics software for the prediction of protein structures from their sequences that is interfaced with the molecular modeling graphics software MAD. (The latter software was developed by R. Lahana at the Research Center of the Pierre Fabre Médicaments Inc.)

Address reprint requests to Dr. Deléage at the Laboratoire de Physico Chimie Biologique, Université Claude Bernard Lyon I, 43 Bd du 11 Novembre 1918, 69622 Villeurbanne Cédex, France.
Received 4 December 1990; revised 23 December 1990; accepted 4 January 1991

SYSTEM AND METHODS

Our CPU was an IBM 6150 workstation connected to an IBM 5080 graphics station. The software was developed in IBM VS/FORTRAN for general calculating purposes and in graPHIG'S for graphics subroutines with UNIX as the operating system. The software consists of more than 30 000 lines that are organized into 100 subroutines. This very modular form allows easy portability.

RESULTS

The ANTHEPROT program offers four main options. Each option yields a synthetic display that allows the user to obtain a global view of the information. If one asks for additional information, a more detailed view is generated. This view is specific to the option selected and gives the parameters used, homology or structural scores, percentages of secondary structures, scales or matrix used and so forth.

Options

The Ant option allows one to study a given protein by up to eight different methods. A complete list of the twenty-five methods available is given in Figure 1. Results from the methods are sorted by order of selection from top to bottom, the profiles being above the predicted structures, which are themselves above the secondary structures derived from X-ray data. Color Plate 1 shows the state of the screen for a synthetic view. Once the results have been displayed,

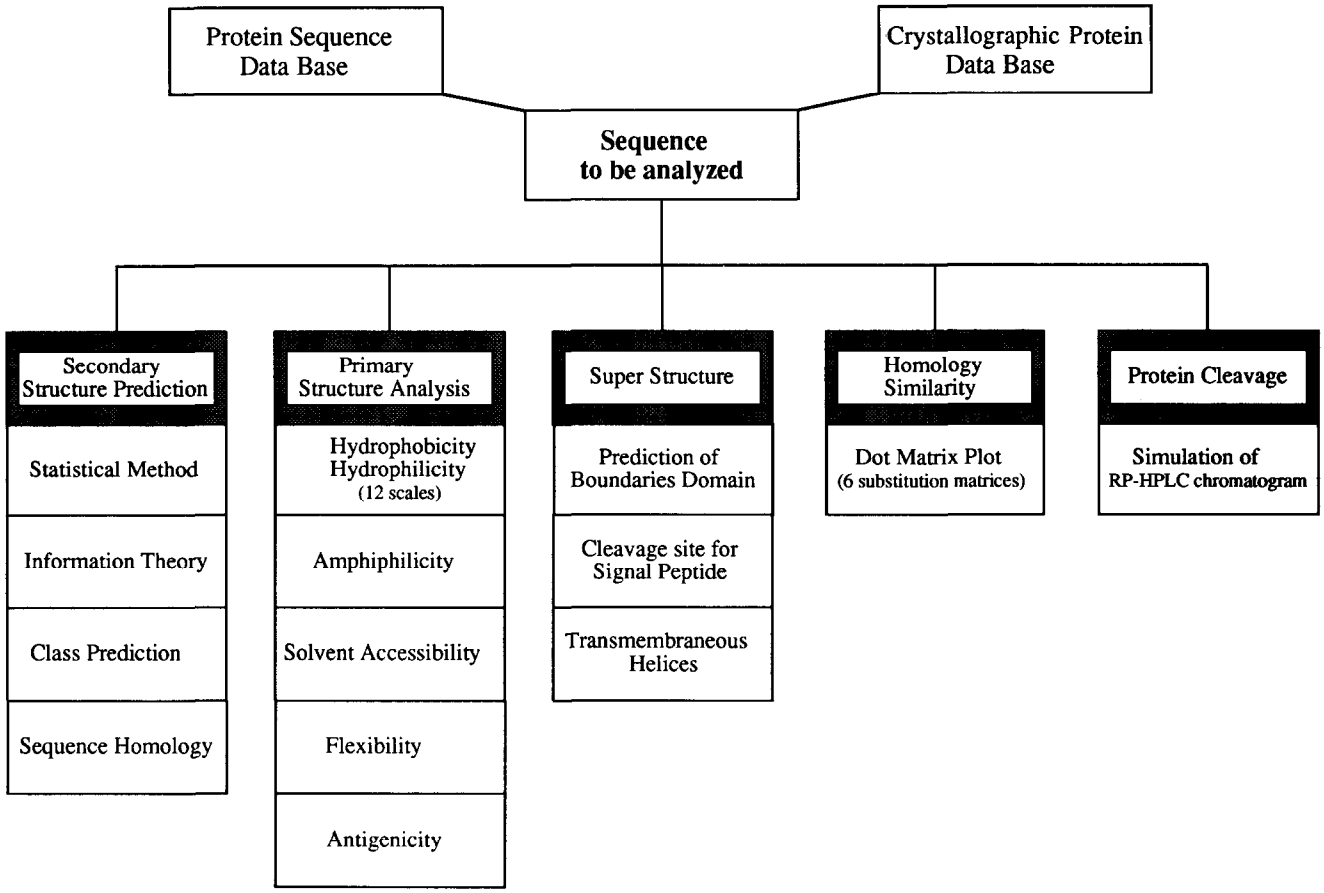

Figure 1. List of the different methods used in ANTHEPROT

it is possible to reduce or expand the scale, to translate the picture or to move a cursor onto the sequences. In addition, the state of amino acids indicated by the actual position of the cursor can be changed by the user. (The same can be done for longer strands of amino acids as well.) The percentages of secondary structures are actualized in real time in the right window.

The PMu option allows one to represent results from up to eight proteins by a single method. All of the tools previously described are also available in this option. In addition, the possibility of aligning secondary structures (or profiles) is offered. The position of such alignment is determined by the position of the cursors, which may be moved independently on each protein (only in this option). A typical example of the utility of such an alignment is given in Color Plate 2 for proteins that can bind adenine nucleotides. The alignment has been made at the level of the glycine-rich loops involved in ATP binding. The agreement between these structures is striking, indicating a good prediction accuracy in these functionally important regions.

The Day option is for dot matrix comparison of two proteins. In this two-dimensional representation of sequences, identical or homologous regions (depending on the matrix used) appear as diagonals. The extent of homology is coded by color as shown in Color Plate 3. Cursor movements are allowed with the possibility of following the diagonal (*xy* motion). All parameters can be changed: window width of comparison, matrix of substitution (chosen from among six) and threshold for homology. This interaction permits the user to adjust the parameters as a function of the results, ie., to reduce noise, to search for weak homology and so forth. Moreover a particular region limited within a square (100 × 100 amino acids) of this synthetic view may be enlarged. In this detailed view the actual score is given, the same tools as in synthetic view being available. The main advantage to working with subareas of 100 amino acids is that it takes less time to calculate the diagram, allowing real-time actualization of the screen.

The HPL option is for the prediction of proteolytic peptides obtained either by chemical attack or enzymatic digestion. The chromatographic behavior of the fragments is predicted in reverse phase HPLC (C18 column in water/trifluoracetic acid/acetonitrile solvent). The peaks are digitalized as Gaussian functions having an area proportional to peptide length; these functions constitute an approximation of the absorbance of the fragments at 220 nm, with standard deviations (peak widths) that depend on the resolution of the separation. For each elution time, the contribution of all peaks to the absorbance is calculated on an additive basis and the resulting diagram is generated. Information that is given as a function of the cursor's position include the peptide composition of a peak, the percent contribution of each peptide to the area (in case of overlapping peaks) and the sequence of the main component, as well as its predicted retention time. Selection of peaks by amino acid presence is also possible.

STRATEGY FOR APPLICATIONS

One may use ANTHEPROT to make secondary structure predictions based on several methods; it has been shown[4–6] that agreement between different methods can increase the accuracy of predictions to 70% (percentage of amino acids correctly predicted).

Another way to increase one's accuracy in predicting the structure of an unknown protein can be summarized in the following steps:

(1) Find homology with known protein structures using the Day module (dot matrix plots).
(2) Search for the parameters (decision constants, window width, matrix and homology thresholds) that give the best agreement between predicted and observed structures for this set or subset of known proteins. This should be done separately for each method using the Ant option.
(3) Apply the chosen methods to the unknown protein using these particular parameters (Ant menu).
(4) Compare the predicted structures for the unknown protein with observed ones from the known proteins. (Align the structures using the PMu module, taking into account the results of step 1.)
(5) Make iterative refinements by considering experimental data. (The HPL module can help in this endeavor.)
(6) Build a three-dimensional model with a molecular graphics software that can accept the predicted structure as an input file (such as MAD).

CONCLUSION AND PERSPECTIVES

The ANTHEPROT software allows the complete analysis of a protein sequence without the use of a keyboard, except for the one-time input of the sequence name. The graphics are predominant even in handling the program options. However improvements can still be realized concerning, for example, the management of the sequence database. We have done our best to increase the signal-to-noise ratio to allow easy extraction of information contained in a protein sequence and to permit a fairly good prediction of structure. In this context, ANTHEPROT will become a module of a molecular modeling package, Molecular Advanced Design (MAD).

ACKNOWLEDGEMENTS

Thanks are due to R. Lahana for his kind help in graPHIG'S language and for stimulating discussions. The authors also thank IBM France (Ecully) for allowing them access to the 5080 graphic station.

REFERENCES

1 Fasman, G.D. Developments of protein structure prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation* (G.D. Fasman, Ed.) Plenum Press, New York, 1989, Chap. 6, pp. 193–316
2 Deléage, G., Clerc, F.F., Roux, B. and Gautheron, D.C. ANTHEPROT: A package for protein sequence analysis using a microcomputer. *Cabios* 1988, **4**, 351–356
3 Deléage, G., Clerc, F.F. and Roux, B. ANTHEPROT: IBM PC and Apple Macintosh versions. *Cabios* 1989, **5**, 159–160
4 Deléage, G. and Roux, B. Use of class prediction to improve protein secondary structure prediction. Joint prediction with methods based on sequence homology. In *Prediction of Protein Structure and the Principles of Protein Conformation* (G.D. Fasman, Ed.) Plenum Press, New York, 1989, Chap. 13, pp. 587–597
5 Deléage, G. and Roux, B. An algorithm for protein secondary structure prediction based on class prediction. *Prot. Eng.* 1987, **1**, 289–294
6 Biou, V., Gibrat, J.F., Levin, J.M., Robson, B. and Garnier, J. Secondary structure prediction: combination of three different methods. *Prot. Eng.* 1988, **2**, 185–191