BUTTERWORTH HEINEMANN

# ANTHEPROT 2.0: A three-dimensional module fully coupled with protein sequence analysis methods

**C. Geourjon and G. Deléage**

*Institut de Biologie et de Chimie des Protéines, UPR 412-CNRS, F-69367 Lyon Cedex 07, France*

*ANTHEPROT is a fully interactive graphics program devoted to the analysis of the sequences and structures of proteins. This program, originally developed to facilitate the protein sequence analysis coupled with multiple alignments and predicted secondary structures of proteins,[1,2] now comprises a powerful 3D module to display and handle macromolecular structures. All the methods that were previously integrated into ANTHEPROT are now directly coupled with a 3D window that provides the user all the classic features of a molecular modeling package. Indeed, it allows real-time rotation and translation of 3D structures with many kinds of models in depth-cueing mode (space filling, backbone, wire models, main chain, and ribbons), selections (atom type, residue type, segments, and chain), color-coding systems (amino acid properties, predicted or observed secondary structures, temperature B factor, and subunits), geometric calculations (Ramachandran plot, distances, and angles), and fitting molecules. Stereo views are possible as well as HPGL standard files. A module specifically devoted to the determination of 3D structures using nuclear magnetic resonance is also available. This major release of our program for IBM rs6000 workstations is available by anonymous ftp to ibcp.fr for academic institutions.*

*Keywords: amino acid sequence, homology modeling, protein structure, secondary structure prediction, NMR modeling*

## INTRODUCTION

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. In this context, we have developed a package called ANTHEPROT (ANalysis THE PROTeins), which was originally designed for microcomputers[3–5] and for protein sequence analysis.[1] In addition, with the progress in the determination of three-dimensional (3D) structures by nuclear magnetic resonance together with the increasing number of applications of homology-based modeling approaches, the need for powerful tools in these fields has increased. In this article we present our new developments, including a 3D module that is fully coupled with all of the 1D and 2D methods of protein sequence analysis.

## SYSTEMS AND METHODS

All programs were developed for IBM rs6000 (AIX 3.2.4) workstations with any of the 3D graphic cards with Z buffer and double buffer (3D+, GT4x, and GTO). The software was developed in Fortran and C for general calculating purposes and in graPHIG's for graphic primitives. The software consists of more than 70 000 lines of program contained in 600 subroutines. A migration to other platforms can be planned if there are enough requests for it (and financial support). On the other hand, the program has been successfully run (with reasonable performance) through X11 windows on Silicon graphics and SUN machines. A 3D interactive graphic viewer is also available within the ANTHEPROT version for PC-compatible computers.

## AVAILABILITY

The ANTHEPROT 2.0 release is available by anonymous ftp to ibcp.fr or by Gopher to merlot.welch.jhu.edu (or gopher.gdb.org). An installation procedure is provided to run the program with a minimum of effort. A W3 server demonstrating some of ANTHEPROT 2.0 program capa-

Color plates for this article are on pp. 199 and 200.

Address reprint requests to Dr. Deléage at the Institut de Biologie et de Chimie des Protéines, UPR 412-CNRS, 7 passage du Vercors, F-69367 Lyon Cedex 07, France. e-mail: deleage@ibcp.fr.
Received 15 March 1995; accepted 18 April 1995.



655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00035-5

bilities is available through the Uniform Resource Locator (http://www.ibcp.fr).

## RESULTS

### New features

The main options of the ANTHEPROT 2.0 package are given in Figure 1. The new options that make this second release of major significance are highlighted in boxes with heavy borders. Among them, the 3D module is a completely new part of the package that has been fully coupled to all existing modules. The new secondary structure prediction method (SOPM) that we have developed[6] has also been added. A nuclear magnetic resonance (NMR) module is also included so as to take benefit of all the 3D features.

### General features of the three-dimensional module

The ANTHEPROT 2.0 program consists of about 70 menus that are all mouse activated. The use of ANTHEPROT 2.0 requires very few keyboard inputs. All the available options in the previous release have been kept and improved in this new release. From 8 to up to 30 different macromolecular structures (depending on available memory) can be loaded with the help of a fully mouse-driven menu. These molecules can be grouped as a single object (global mode) after a fitting procedure, for example, or separated from the others just by clicking on the given structure to be activated. In this latter mode all future actions will concern only the selected molecule. This mode is a powerful tool with which to investigate interactions between molecules (e.g., docking of a substrate onto a protein) or protein–protein recognition or protein–DNA interactions.

Once the molecules have been displayed on screen, it is possible to reduce or expand the scale or to translate the picture and attain information about the cursor location simply by clicking on the given structure (atom type and number, amino acid type and number, and chain number are provided in the message window). All molecular movements are achieved with mouse button combinations (the left mouse button is for rotation around any of the three axes, the right mouse button is for scaling, and the middle mouse button is for translation). Moreover, any atoms in the molecule can be identified just by clicking on them.

Many criteria to attain various types of selections can be invoked (chain, segments, amino acids, atoms, etc.), and these selections are appended into a heap. This selection mode allows all logical operators (and, or, xor, not). Selections are put into a selection heap. Thus, these selections can be combined in an infinite number of ways and any change in visualization affects the selection heap. If additional selections are needed, the current heap must be cleared out, allowing another (new and empty) heap to be automatically created. A typical example of a selection heap is given in Color Plate 1 for the fructose repressor protein. This powerful selection mode can be used to show all $C_\alpha$ in gray with a superimposed ribbon in yellow, all glutamines in an atomic space filling (CPK) mode, and Tyr-19 and Tyr-28 as red and blue wire, respectively. The distance between oxygen atoms of Tyr-19 and Tyr-28 residues is given (4.24 Å) as a dotted green line.

All graphics can be saved in a file written in standard HPGL langage, which can be further plotted by an HPGL-compatible plotter (or printer).

### Coupling the three-dimensional module with methods of protein sequence analysis

The 3D module is fully coupled with all methods of protein sequence analysis. For example, sequences and structures can be loaded at the same time. For PDB files, a sequence file deduced from the ATOM information (with the .SEQ extension) is automatically created in order to allow its complete analysis. The window is then divided into the sequence editor area and the structure area (see Color Plate 2). Thereafter, the sequences can be aligned and the color coding of the homology imposed on the structure, allowing one to easily detect the location of conserved amino acids in the structure. A typical example is given in Color Plate 2, in which a stereo view of the LacI repressor headpiece is shown with some other aligned members of the same family.

### Site and motif detection in structure database

To help in the search for the known 3D motifs, ANTHEPROT 2.0 can list all entries from the PROSITE.DAT file.[7] The user may select a particular site and the program searches for matching sequences in the Protein DataBank. Once this search is completed, the program extracts the atomic coordinates of the PDB files and automatically superimposes the segments by searching for the best fit between all hits (data not shown).

### The nuclear magnetic resonance module

The determination of the 3D structure of proteins, using NMR, is achieved by restrained molecular modeling. This modeling leads to an iterative refinement of the structures that involves multiple views of the generated molecules. To help in this refinement step, we have developed an NMR module to analyze the resulting structures (Color Plate 3). For example, our program is able to load a table of restraints in the Xplor 3.1 format[8] and to display them directly on the structures as dotted lines. These dotted lines are clickable and the information dealing with the restraint is provided (atoms, imposed value, and actual distance). These restraints can be displayed on the basis of intraresidue, sequential, and medium or long-range restraint type. These restraints are color coded as a function of their actual value in the structure. The same strategy can be applied with residual violations (if any) that are color coded according to the intensity (red is for a violation greater than 0.5 Å, green is for a violation lower than 0.5 Å). There is also a neighbor option in which the atoms located at a distance range from selected atoms can be joined by dotted lines. This tool is useful when comparing the experimental restraints with the maximal set of possible nuclear Overhauser effects (NOEs) that could be observed. Obviously, all these tools are combined with the selection mode.

The fit of molecules is an important criterion in assessing the efficiency of the refinement procedure after NMR de-

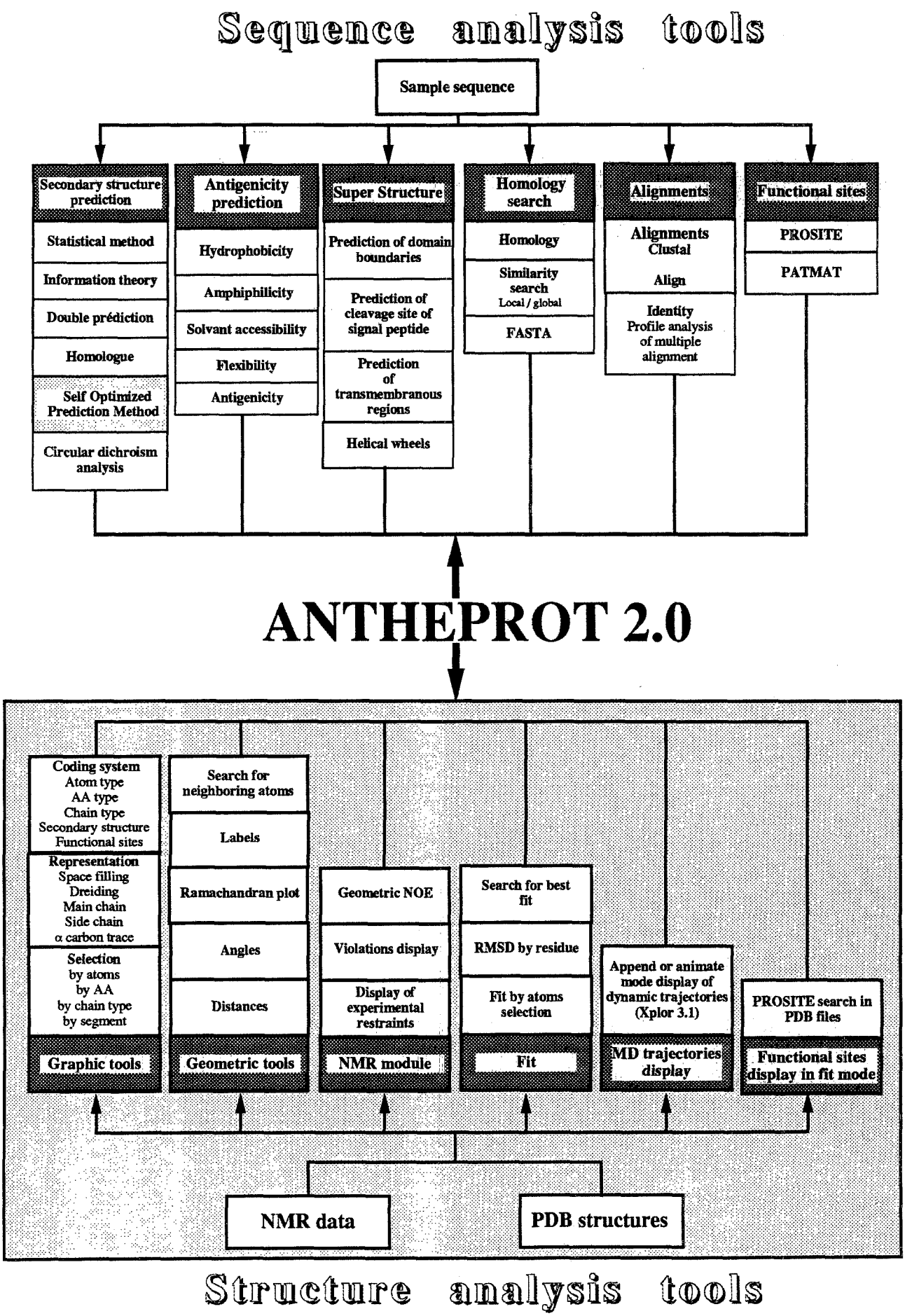


*Figure 1. Concept of coupling the 2D analysis with the new 3D tools (gray background) available in ANTHEPROT 2.0.*

termination of structures. In this context, we have included a powerful subroutine to fit molecules. This subroutine permits the user to select the atom types that must be superimposed (all atoms, main chain or α carbons). However, from NMR studies, some underdetermined regions may exist that hinder a correct fit of molecules. For that reason, the fit procedure allows molecules to be superimposed only on given selected regions (e.g., secondary structure) and an array containing the pairwise comparisons is generated. A typical example of fit onto helical regions is given in Color Plate 4 for the headpiece of the fructose repressor.[9] The corresponding Ramachandran plot is also displayed on the same screen. This allows the user to click interactively either on the dot in the Ramachandran plot,[10] or on the structure or the sequence window. Most interesting is the possibility of searching for the best fit between two molecules of different lengths (the smaller molecule is moved along the larger one).

## DISCUSSION

Until now, programs could be characterized as sequence analysis software (mostly developed for microcomputers or SUN workstations) or as molecular modeling programs (most working on powerful graphics workstations). Although several software packages for the analysis of sequences exist, they are mainly devoted to nucleic acid treatments; if they were initially designed to analyze protein sequences they are often devoid of interactive graphics options. On the other hand, the molecular modeling programs are mostly graphic but are rather poor in protein sequence analysis. Our major goal in releasing ANTHEPROT 2.0 is to couple many sequence analysis methods, techniques for secondary structure prediction, site and signature detection, and multiple alignment with most of the possibilities of a molecular modeling program with interactive (4D concept) graphics capabilities. The main feature of ANTHEPROT 2.0 is that it deals with application fields that range from the search for protein sequences to sequence databases to the visualization of dynamic trajectories generated by the Xplor program. Moreover, ANTHEPROT is a protein sequence analysis program that has migrated toward 3D capabilities. An interesting alternative approach is presented in the RasMol program, which now accepts as input a query in the Prosite dictionary syntax for potential site and signature detection.[11]

With the development of structure determination by NMR, the design of the functional domain is a prerequisite step before proceeding with the rather tedious NMR studies. ANTHEPROT 2.0 is a program that now makes realistic the "domain design" concept. This concept will become more and more useful before addressing NMR studies. The newly designed domain can be overexpressed, purified to homogeneity, and subjected to NMR methods in order to determine its structure for structure–function relationship studies.

## ACKNOWLEDGMENTS

C. Geourjon is a recipient of a fellowship from the Région Rhône-Alpes. Thanks are due to B. Kieffer for the source core to fit molecules.

## REFERENCES

1 Geourjon, C., Deléage, G., and Roux, B. ANTHEPROT: An interactive graphic software for analyzing protein structures from sequences. *J. Mol. Graphics* 1991, **9,** 188–190
2 Geourjon, C., and Deléage, G. Interactive and graphic coupling between multiple alignments, secondary structure prediction, and motif/pattern scanning into protein sequences. *Comput. Appl. Biosci.* 1993, **9,** 87–91
3 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
4 Deléage, G., Clerc, F.F., and Roux, B. ANTHEPROT: IBM PC and Apple Macintosh versions. *Cabios* 1989, **5,** 159–160
5 Deléage, G., and Geourjon, C. An interactive graphic program for calculating the secondary structure content of proteins from circular dichroism spectrum. *Comput. Appl. Biosci.* 1993, **9,** 197–199
6 Geourjon, C., and Deléage, G. SOPM: A self-optimised method for protein secondary structure prediction. *Protein Eng.* 1994, **7,** 157–164
7 Bairoch, A. The PROSITE dictionary of sites and patterns in proteins: Its current status. *Nucleic Acids Res.* 1993, **21,** 3097–3103
8 Brunger, A. (1991) X-Plor, version 3.1, a system for X-ray crystallography and NMR, Yale University Press.
9 Pennin, F., Geourjon, C., Montserret, R., Yang, Y., Fillon, A., Bonod, C., Cortay, J.C., Nègre, D., Cózzone, A.J.C. and Deléage, G. (submitted) Determination of the tertiary structure of DNA binding domain from E. Coli FruR protein by NMR.
10 Ramachandran, G.N., and Sasisiekharan, V. Conformation of polypeptides and proteins. *Adv. Protein Chem.* 1968, **23,** 283–437
11 Saqui, M.A.S., and Sayle, R. PdbMotif—a tool for the automatic identification and display of motifs in protein structures. *Cabios* 1994, **5,** 545–546
12 Higgins, D.G., and Sharp, P.M. CLUSTAL: A package for performing multiple sequence alignments on a microcomputer. *Gene* 1988, **73,** 237–244