Interactive and graphic coupling between multiple alignments, secondary structure predictions and motif/pattern scanning into proteins

C. Geourjon and G. Deléage

Abstract

A computer module that includes multiple alignments, secondary structure prediction, and site and pattern search has been developed and integrated into our ANTHEPROM software for protein sequence analysis. All the programs can be invoked from within any routine, thus yielding multiple pathways to obtain final results. All the results are graphically displayed. The main feature of this module is that all methods are connected in an interactive graphic manner. This module has been designed to display easily the potential sites with conserved predicted structures.

Introduction

The most successful approach to the prediction of the structure and function of a protein is to identify similarities between the sequence of the molecule and that of other well-characterized proteins. When a strong sequence homology (identify > 30%) exists with a protein of known three-dimensional structure, then model-building techniques may be applied with some success (Blundell et al., 1987). Even though an incomplete similarity exists with a protein of known three-dimensional structure, the observation of conserved motifs or templates can provide important clues to identify the functional domains of the protein, and give guidance for the design.

We describe in this paper a computer module integrated into the ANTHEPROM package for the IBM RISC 6000 station (Geourjon et al., 1991) that addresses these questions. A large collection of computer-readable motifs is available, such as PROSITE (Bairoch, 1990). This database (release 8.10, March 1992) contains 605 known motifs encoded as a list of allowed residue types at a specific position along the sequence. The PATTERN program we have developed searches for the presence of these motifs, patterns and signatures described in PROSITE in one or more protein sequence(s). This pattern-matching program can also be used to search within a protein sequence database such as the SEQUENCE.PDB file, for a motif extracted from the PROSITE database.

In addition to motifs/signature/pattern searching programs, this module also includes multiple alignment methods and secondary structure prediction algorithms. All these methods can be graphically and interactively called from any of the other ones. The connection between all these methods provides a very powerful tool to identify conserved potential sites that have a particular secondary structure.

System and methods

Our CPU was an IBM RISC 6000 (model 320H) with a 3 D+ graphic card. The software was developed with IBM VS/FORTRAN (F77 compatible) and C language for general calculation purposes, and with PHIG’s for graphics primitives subroutines using ALX 3.1 (Unix) as the operating system. Thus, the program should work on any workstation having a F77 FORTRAN compatible and a PHIG’S+ compatible library (Silicon graphics, Hewlett Packard, etc.).

The program can be obtained for non-commercial use but collaborations are also possible at the source code level to incorporate new methods or algorithms into ANTHEPROM.

Algorithm and implementation

All the programs have been integrated into our ANTHEPROM software package (Deléage et al., 1988, 1989; Geourjon et al., 1991) in order to take advantage of its graphical interface with a mouse and menu-driven environment. Moreover, basic features (on-line help, text editor with many functions, interactive views for input—output key in) make a coherent and easy to use suite of programs. Table I lists the methods that constitute this new module and have been added to ANTHEPROM.

Database analysis

To make comparisons with protein sequences of known structures, one needs to get reliable databases containing both sequences and X-ray-deduced secondary structures. A possible source for such databases can be offered by the dictionary of secondary structure of proteins called DSSP (Kabsch and Sander, 1983), deposited at the EMBL server. We have written a subroutine that is able to convert all the individual DSSP files into one database containing only sequences and secondary structures. Alternatively, if the DSSP files are lacking, we have written a utility program that directly uses the PDB files to generate a SEQUENCE.PDB file containing all primary
sequences of PDB proteins. This program first reads the whole PDB database, then calculates the similarity between all pairs of sequences, and finally discards identical sequences. Some rules have been followed to resolve special cases. Primary sequences must contain at least 20 amino acids. Unknown or modified amino acids make the sequence to be discarded from the database. In the case of multimeric proteins, the sequence of each monomer is generated as an independent file with numbered suffixes. An auxiliary file called DATABASE.PDB, which contains the secondary structure appended to the sequences, is also generated. This file, which contains 250 proteins fragments exhibiting <50% identity between their sequences, is useful to check, compare and optimize secondary structure predictive methods. The SEQUENCE.PDB file, comprises 323 protein sequences having <-98% identity. These files can be used to compare a given protein sequence with that of all proteins with known structures.

A small additional motif database has been established (SIGNAL.DAT), which contains a list of 277 signal peptide sequences (Watson, 1984).

Multiple alignments and similarity search

In this chapter, we have integrated two of the most commonly used methods for making multiple alignments. The first program is Clustal (Higgins and Sharp, 1988) for which we have written a graphical interface to set all the parameters (gap penalties, window width, filtering level). The second is ALIGN from Feng and Doolittle (1987), in which we have incorporated the SCORE and PREALIGN routines from the same authors so as to yield alignments independently of the original order of sequences. In these two programs, I/O formats have been changed so as to accept our sequences data formats and to give standard formatted result files.

Figure 1 gives an example of the graphic display of the result file from a multiple alignment with the help of the built-in multi-editor. A consensus is calculated and asterisks (*) are used for identical residues that are aligned in all sequences. However, the user can interactively change the set of symbols to display conserved residues (e.g. plot a minus symbol in the consensus line for D or E aligned in all sequences). The editor allows the interactive search for a string in the currently edited file. This is particularly useful in the case of multiple alignments. For example, one can locate protein segments that share five amino acids in common just by searching the "*****" string of characters. In addition, aligned sequences are color coded in order to magnify the similarities. "Page down/up", "next/previous line", "next/previous 10 lines", "top/bottom of file", "next/previous string occurrence" are capabilities that are also supported by this graphic editor. The special button 'Site' displays the result of the PATTERN analysis as colored areas on the alignment (see the PATTERN section).

The second set of programs, PAIRWISE, is devoted to the screening of a protein sequence database with a given protein sequence in order to find the most similar regions. The algorithm used is based on that of Needleman and Wunsch (1970) as implemented in SCORE by Feng and Doolittle (1987); it makes all binary comparisons between the query protein sequence and the sequences of the database. Alternatively, the program PAIRWISE can use a database and make binary comparisons between all proteins sequences. This is particularly useful to build secondary structure databases containing proteins with <50% identities. It also allows the identification of regions of a protein that share sequence similarity with proteins
Fig. 2. Flow chart for the PATTERN program. The PATTERN program can be divided into four main steps. The input step is to read the site description file, to set the mismatch number ('O' is default value), the similarity threshold (100% is the default value) and the frequency above which a mismatch in the motif will be allowed. The syntax analyzer step is not time consuming and allows the use of the PROSITE.DAT file of the EMBL as it stands. The motif search step is a classical pattern-matching algorithm using a positional substitution matrix filled with 0 or 1 (bit matrix). The output step is in fact the input of other routines that are used to display the results graphically.

belonging to the Brookhaven Data Bank. Although the original algorithm was designed to look for the best global homology between sequences, it has been modified so as to be also able to find the best local similarity between proteins. An extensive searching process can find a peptide on one side (N side of sequence 1) similar to a peptide on the other side of the sequence (C side of sequence 2). In this last case, however, the comparison time is rather long. The searching procedure (local or global) is chosen as an external parameter just by choosing a moving window onto the query sequence. All peptides with a length that corresponds to the window width are tentatively aligned with the protein sequences of the database. The alignment method utilizes Dayhoff's PAM mutation matrix, thus allowing one to match either identical or conserved residues. However, the unity matrix can also be used so as to locate only identities.

Once the calculations have been carried out, the result file is loaded with the general-purpose graphical editor. The similarity scores are given for all pairs of sequences. The alignment between any pair of sequences can be immediately displayed after clicking onto the given pair. The alignment is color coded and the secondary structure (if any) is displayed. The potential sites can be displayed directly on the alignment.

PATTERN

Site database. The PROSITE database consists of two main files. The first one (PROSITE.DAT) offers basic information about protein patterns (i.e., title and description of patterns, signatures). The PATTERN program can directly use the PROSITE.DAT file distributed by the EMBL file server (Fuchs, 1990) as input motifs since it recognizes three types of headers: DE for the title, PA for the pattern and PD, which is the reference number in the documentation file. However, we also distribute a shortened version PROSITE.LST that can also be used if the user has no facilities to access EMBL. This site file only contains the lines beginning with the DE, PA and PD headers. Personal motifs and databases can also be used provided the PROSITE.DAT file syntax is obeyed.
The second file (PROSITE.DOC) contains textual documentation, including remarks and references about each motif. A numbering system gives a cross-reference system between the two files and is used to append the documentation to the result file for each motif found. At present, the SWISSPROT, NBRF and PDB modified formats are recognized on input. To increase the flexibility of the search, the program allows one to set either a mismatch number, or a similarity threshold before the search process is launched. In addition, it is also possible to select a pattern found into a protein and to scan the PDB sequences for its presence. The results are displayed into a special window.

**Algorithm.** The general flowchart of the PATTERN program is given in Figure 2. The PATTERN program has a syntax analyzer that allows the recognition of motifs and their conversion into a 'positional substitution matrix'. Since a pattern may represent different motifs to be searched for, the motifs are first scanned for the presence of commas and then all possible motifs are generated. For example, the pattern 'RT(2,3)T' means that each combination of R or T may be found two or three times, leading to 12 motifs. Thereafter, each position of a generated motif is converted to bits (0 or 1) of a 'positional substitution matrix'. In other words, if the motif is 11 amino acids long, there will be 11 different substitution matrices specifically filled with 0 or 1 depending on the amino acids that can be found at each given location. Finally, the protein sequence is coded and transformed into the motif through the matrix, thus leading to a substitution score that is compared with a threshold score.

This type of motif coding/comparison allows the user either to set a similarity threshold or a mismatch number before the identification process occurs. However, some motifs (such as N-glycosylation) are not very specific and should not be searched for with a mismatch unless the user wants to deal with rather long listings. The probability of finding a motif by chance is evaluated. This is achieved by calculation of a theoretical frequency for each site based on the occurrence frequency for each type of residue (Sternberg, 1991a,b). Thus, the theoretical frequency for each motif is calculated from the average protein's amino acid composition before entering the searching engine. Then the user is prompted to choose a frequency threshold to take into account a possibility of mismatch. In fact, the theoretical frequency for all motifs must be lower than this threshold to allow the mismatch occurrence. This is particularly useful to allow two or more mismatches in long, unfrequent motifs and to avoid long listings.

**Output.** The result file is directly loaded with the graphic editor as for multiple alignments. For each site found, the theoretical frequency is given as well as the fitting sequence, and the actual similarity score if a similarity level has been set. The corresponding documentation drawn from the PROSITE.DOC file is incorporated into the result file using the cross-reference system. The program automatically accommodates for the lack of the PROSITE.DOC file by just omitting this addressing reference step. A particular site can be selected from the edited file by mouse clicking and can be searched onto a sequence database such as the DATABASE.PDB file. Thus, it is possible to display the results on a single screen divided into two side-by-side windows: one window contains the sites found into a particular protein sequence, and the second one contains protein sequences of known structure, which exhibits the selected pattern. Although rapid enough, the search is faster if no similarity level and no mismatch options have been chosen (it takes typically 5 s to scan the 605 sites onto a 450 amino acid long protein without inclusion of the DOC file). The same algorithm has been implemented on IBM PC computers typically running for 70 s for the same scan on a 80286 16 MHz processor equipped with a 80287 coprocessor.

**Secondary structure prediction**

The secondary structure prediction can be used with the following strategy. The first step is to use the Ant menu (Geourjon et al., 1991) to predict the structure of a protein by different methods so as to see the agreement level between these methods. Once a predicted consensus structure has been defined, the second step is to use the Pmu menu (Geourjon et al., 1991) to compare the secondary structure of different proteins (generally related together) by a given predictive scheme. Alternatively, if the secondary structure of a protein of the same family is known (see the database tools), the parameters used for the prediction should be chosen so as to optimize the prediction. The results obtained with this Pmu menu are given in Figure 3; the predicted secondary structures have been aligned taking into account the alignment of primary sequences.

![Fig. 4. The module with a dynamic driving crossroad from the graphic display of results. The diagram shows the interactive graphic connection of multiple alignment, secondary structure, PATTERN, primary sequence comparisons and physico-chemical properties.](image-url)
using Clustal (Higgins and Sharp, 1988). The sequences are given in the rightmost window in the aligned format. The potential sites found with our PATTERN algorithm have also been displayed as colored squares. These square areas that can be activated by clicking the mouse in order either to describe briefly the currently selected site or display the whole related documentation in a special window.

Discussion
During recent years many methods that deal with either secondary structure prediction, or site/signature detection, sequence comparisons and multiple alignments have been developed. For example, many protein pattern recognition programs are available (Taylor, 1986; Cockwell and Gilles, 1989; Sibbald and Argos, 1990; Sibbald et al., 1991; Sternberg, 1991a,b; Fuchs, 1991). Most of these programs scan a protein sequence database for a motif. Our program PATTERN can also search for a given motif or a set of motifs, provided they obey the syntax used in the PROSITE definition of sites. This search can be performed either on a single sequence or a sequence database within a few seconds. Due to its syntax analyzer, the PATTERN program does not need any reformatted version of the PROSITE-DAT file for site description. That means that future releases of PROSITE will be usable through our PATTERN program without any changes. Moreover, mismatch or similarity level can be used as extended searching criteria. For example, the PATTERN program allows one to scan the PDB protein sequences for a particular site that has been previously found in a given sequence. However, few programs or software have incorporated these site/signature searches with secondary structure prediction and multiple alignments. Even where these programs do exist, no programs are yet available that combine all these methods in an interactive graphic manner. This means that the user can display all combinations of these methods just by using the mouse directly on the graphs. Thus the graphic display has to be considered as a ‘decision crossroad’ that allows the different programs to communicate (Figure 4). This also means that the pathways for getting a given screen can be multiple since all these methods are dynamically connected together. All these tools are very useful to identify the most conserved potential sites that may constitute structurally identical regions. All these features are prerequisite for three-dimensional modeling. It has to be mentioned that all the data plotted onto a single screen, such as Figure 3, could hardly be summarized into less than a hundred pages of text and that getting this screen can save hours of tedious manual comparisons of text files. This module is useful for optimizing the retrieval of all the information concerning a pattern or site. The next step will be to generate the atomic coordinates of each protein fragment in order to compare all the actual structures to obtain the knowledge-based three-dimensional modeling of a protein.

Acknowledgements
C.G. is a recipient of a fellowship from the Région Rhône-Alpes. Thanks are due to B Roux (Laboratoire de Physico-Chimie Biologique, UCL Lyon 1) for supporting this work and L Baggetto for improvements to the English.

References

Received on July 1, 1992; accepted on August 8, 1992