CABIOS

Vol.4 no 3. 1988
Pages 351 – 356

ANTHEPROT: a package for protein sequence analysis using a microcomputer

Gilbert Deléage*, François F.Clerc, Bernard Roux and Danièle C.Gautheron

Abstract

A simple microcomputer package is described to make the theoretical analysis of protein sequences. Several methods designed to compare two sequences, to model proteolytic reactions and to predict the secondary structure, the hydrophobic/hydrophilic regions and the potential antigenic sites of proteins have been included in an Apple II microcomputer software. The package comprises 21 programs as well as the secondary structure database of Kabsch and Sander (1983).

Introduction

A precise knowledge of protein structure is required to elucidate the mechanisms of protein folding and the relationships between the structure and biological activities of proteins. In the absence of crystallographic data, many structural features can be deduced from the analysis of protein sequences. One of the most promising tools in the near future is the prediction of antigenic sites for the engineering of synthetic vaccines. In addition, with the increasing number of protein sequences known from DNA cloning and sequencing, the need in the theoretical treatment of protein sequences has never been greater. In this context, many different methods have been described but the possibility of their use is hindered either by the need of mainframes or minicomputers, or by the absence of compatibility at the language or machine levels. Being aware of this problem, we have developed software that uses and/or combines several of the most frequently used methods. The main advantages of this package are its evolutive potential since future methods will be rapidly implemented and that it does not need any special knowledge about computers.

System and methods

Minimal hardware requirements

These are: an Apple IIe (with Extended 80 columns text card) or Apple IIgs or Apple IIc plus monitor, one floppy disk drive with controller in slot 6, and any printer with interface in slot 1 (for text printout).

Laboratoire de Biologie et Technologie des Membranes du CNRS (UM 24), Université Claude Bernard Lyon I, 43, Bd du 11 Novembre 1918, 69622 Villeurbanne Cedex, France

**To whom reprint requests should be sent.*

Advisable material

Advisable materials are: a second floppy disk with the same controller (slot 6), a Hewlett Packard plotter (7470A or 7475), and an Imagewritter I or II with Super Serial Card (Apple) or special graphic interface available from Micro Informatique Diffusion (96, Bld Richard-Lenoir, 75011 Paris, France) or a Dot Matrix Printer (Apple) for graphic printing output.

Language and files

All the programs are written in Applesoft Basic and run in compiled form (128 kbyte Microsoft/Applesoft TASC compiler) for maximal speed and memory. All sequence data are stored on disk as Apple DOS 3.3 text files to allow corrections via a text file editor.

An Apple Macintosh and an IBM PC version are being developed and will be available in the near future.

Algorithms and programs

The options can be selected from a main menu (Table I). Their connections are given in Figure 1.

Table I. The options available with the ANTHEPROT package

Option number	Program name	Brief description
1	NOMPROT	Entry of a new sequence on disk
2	EDITSEQ	Edition of a sequence (1 or 3 letter code) and amino acid composition calculation
3	MOLWEIGHT	Molecular weight calculation
4	SPECVOL	Specific volume calculation
5	PEPMAP	Proteolytic attack modelling
6	CHROSIM	Chromatogram simulation
7	SEQ	Search for a given sequence in proteins
8	COMPARE	Search for homology between two sequences
9	CHOUFASMAN	Secondary structure prediction
10	DIRINFO	Secondary structure prediction
11	HOMOL	Secondary structure prediction
12	DOUBLEPRED	Secondary structure prediction
13	COPRED	Combination of secondary structure prediction methods
14	KYDOO	Hydropathy profile calculation
15	HYDROP	Antigenic potential calculation
16	ANTIGON	Antigenic potential calculation
17	FLEX	Local flexibility prediction
18	SURPRO	Surface profile calculation (antigenicity)
19	MHELIX	Intramembrane helices prediction
20	DOMAIN	Structural domain boundaries prediction
21	TRACE	Plotter driver

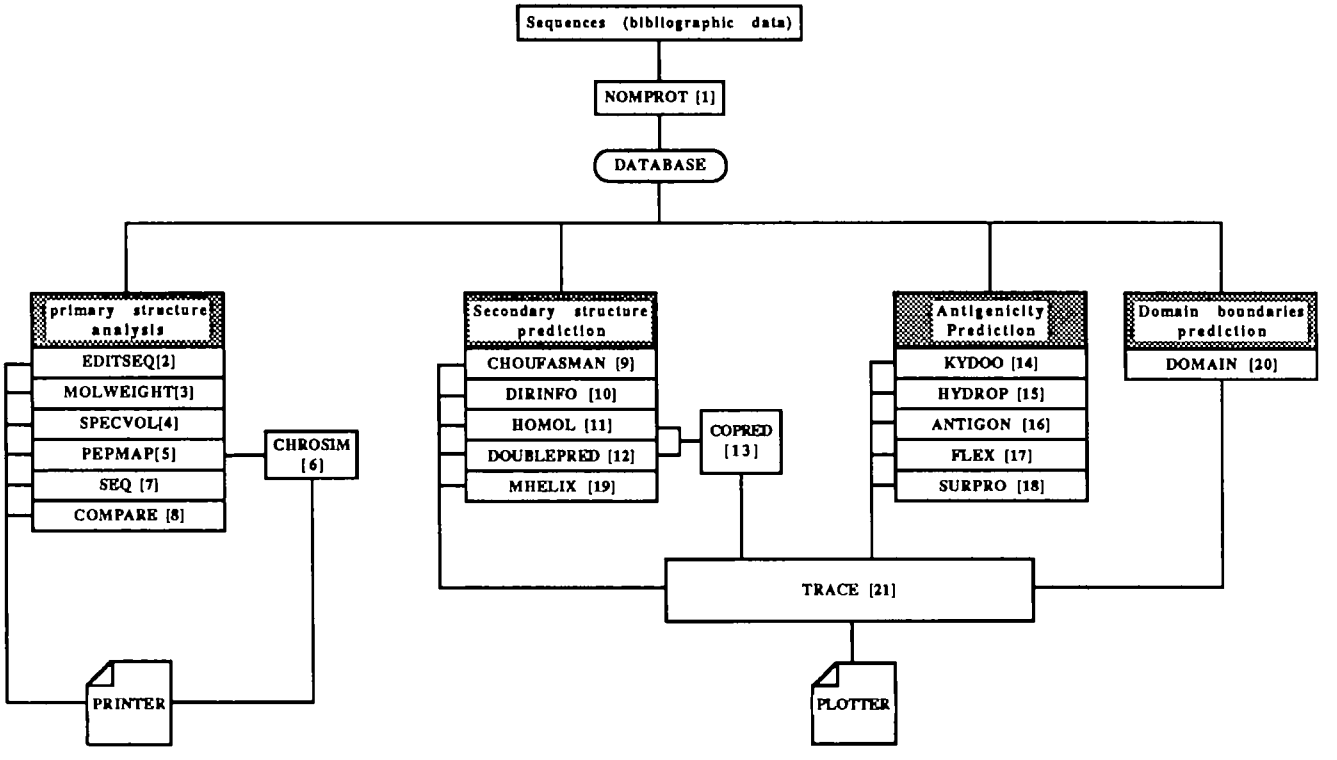

Fig. 1. Organization scheme of the ANTHEPROT package. The four treatment classes are shown in filled boxes. The connections between algorithms are represented by lines. The numbers in brackets refer to the option numbers given in Table I. It should be pointed out that all the algorithms connected to TRACE also allow direct printer outputs. The latter are not represented in the scheme to avoid overloading.

The first four options allow the input (NOMPROT) and the output of sequences with the one- or three-letter code for amino acids, the amino acids composition (with their percentage) by EDITSEQ and the calculation of molecular weight (MOLWEIGHT) and specific volume (SPECVOL) of proteins (Creighton, 1984).

The prediction of proteolytic peptides obtained either by chemical attack or enzymatic digestion can be achieved from the sequence of proteins with PEPMAP. Two types of calculations can be made from the peptide list.

(i) The electrophoretic behaviour in the Laemmli (1970) system is modelled from the molecular weight of peptides, and their distance of migration through a linear relationship which has been found particularly useful for small peptides.

(ii) The chromatographic behaviour in reverse-phase HPLC (C18 column in water/trifluoracetic acid/acetonitrile solvent system) is predicted according to Sasagawa *et al.* (1982). An example of such an analysis is given in Figure 2. The simulation has been achieved by the CHROSIM program using a file created by PEPMAP program. Peak height is proportional to the length of the peptides which is an approximation of their absorbance at 220 nm. In the CHROSIM program, an increment of time is chosen by the user on the basis of the expected width of peaks. For each peptide, a parameter $B = L/26 * R$ is calculated where L is the length of the peptide and R is its molar ratio. The peaks are built up as a series of values (B, $4B$, $16B$, $4B$, B) and centred on the coordinate of the calculated retention time of the corresponding peptide. The contributions of each peptide on each coordinate are added and the resulting values plotted on the graphic display. To enable an easy inter-

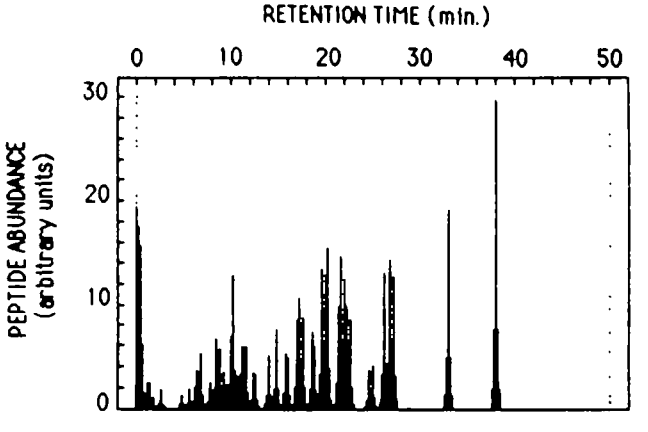

Fig. 2. Simulation of a reverse phase chromatographic separation: tryptic digest of bovine mitochondrial F1 ATPase α subunit. The chromatogram has been simulated using a linear gradient of acetonitrile of 1%/min (Sasagawa *et al.*, 1982)

```
PROTEIN  : ALPHA F1 BOV

E-E-T-G-T-A-E-V-S-S I-L-E-E-R-I-L-G-A-D T-S-V-D-L-E-E-T-G-R V-L-S-I-G-D-G-I-A-R
V-E-G-L-E-E-V-Q-A-E E-N-V-E-F-S-S-G-L-E G-N-S-L-N-L-E-E-P-D-N V-G-V-V-V-F-G-N-D-E
L-I-I-E-E-G-D-I-V-E-R T-G-A-I-V-D-V-P-V-G E-E-L-L-G-R-V-V-D-A L-G-N-A-I-D-G-E-G-P
I-G-S-E-A-E-E-E-V-G L-E-A-P-G-I-I-P-E-I S-V-E-E-P-N-Q-T-G-I E-A-V-D-S-L-V-P-I-G
R-G-Q-E-E-L-I-I-G-D R-Q-T-G-E-T-S-I-A-I D-T-I-I-N-Q-I-E-F-N D-G-T-D-E-E-I-I-L-Y
C-I-Y-V-A-I-G-Q-E-E S-T-V-A-Q-L-V-E-E-L T-D-A-
AMINO ACID NUMBER : 223

D-A-N-E-Y-T-I-V-V-S A-T-A-S-D-A-A-P-L-Q Y-L-A-P-Y-S-G-C-S-M G-E-Y-F-E-D-N-G-E-N
A-L-I-I-Y-D-D-L-S-E Q-A-V-A-Y-E-Q-N-S-L L-L-E-E-P-P-G-E-A Y-P-G-D-V-F-Y-L-N-S
E-L-L-E-E-A-A-I-N-N D-A-F-G-G-G-S-L-T-A L-P-V-I-E-T-Q-A-G-D V-S-A-Y-I-P-T-N-V-I
S-I-T-D-G-Q-I-F-L-E T-E-L-F-Y-I-G-I-E-P A-I-N-V-G-L-S-V-S-E V-G-S-A-A-Q-T-E-A-N
E-Q-V-A-G-T-N-E-L-E L-A-Q-Y-E-E-V-A-A-F A-Q-F-G-S-D-L-D-A-A T-Q-Q-L-L-S-E-
AMINO ACID NUMBER : 429

G-V-E-L-T-E-L-L-E-Q G-Q-Y-S-P-N-A-I-E-E Q-V-A-V-I-Y-A-G-V-E G-Y-L-D-E-L-E-P-S-K
I-T-I-F-E-N-A-F-L-S E-V-I-S-Q-N-Q-A-L-L G-E-I-E-T-D-G-I-S E-E-S-D-A-I-L-I-E-I
V-T-N-F-L-A-G-F-E-
AMINO ACID NUMBER : 509



*** AMINO ACID COMPOSITION ***

E                       6.48330059
A                       9.4302554
L                       9.4302554
B                       .982318271
N                       1.96463654
Q                       4.51866405
W                       0
V                       8.05500983
F                       2.75049116
K                       6.09037328
I                       7.66208252
D                       5.69744598
T                       5.10805501
S                       6.48330059
R                       6.28683694
C                       .392927308
M                       2.55402751
Y                       3.14341847
P                       3.33988212
G                       9.62671906
*** PROTEIN CLASS PREDICTION ***


*** DISTANCE BETWEEN THE PROTEIN AND THE CENTER OF EACH CLASS ***

        ALPHA =3.72941374
        BETA  =4.16989735
        ALPHA+BETA =3.7127906
        ALPHA/BETA =2.56932091
        IRREGULAR =8.61533307


*** CLASS (SUB-CLASS) ***


THE PROTEIN ALPHA F1 BOV IS ALPHA/BETA -ALPHA


*** WINDOW WIDTH = 9 ***
```

*** PREDICTION OF SECONDARY STRUCTURE ***

H = HELIX ,E = SHEET ,C = COIL, T = TURN

AA	PA	PB	PT	FT	PC	STRUCTURE
1	0	0	0	.636	.19	C
2	0	0	0	.892	.22	C
3	.766	.699	.844	.276	.8	C
4	.783	.758	.907	.247	.88	C
5	.884	.868	.902	.267	.9	C
6	.974	.836	.88	.107	.86	H
7	.98	.905	.853	.356	.83	H
8	.913	1.099	.768	.603	.82	E
9	.895	1.003	.843	.152	.86	E
10	.926	.976	.852	.094	.84	E
11	.977	1.081	.713	.053	.8	E
12	1.032	1.039	.717	.24	.78	E
13	1.112	.877	.807	.186	.76	H
14	1.106	.853	.836	.054	.76	H
15	1.081	.923	.763	.13	.77	H
16	1.021	.994	.723	.118	.8	H
17	.99	.917	.813	.147	.83	H
18	.866	.814	.993	1.096	.91	T
19	.939	.848	.859	.455	.88	H
20	.855	.827	.893	1.052	.95	C
21	.824	.913	.855	.271	.94	C
22	.832	.905	.911	.722	.92	C
23	.925	1.015	.772	.157	.85	E
24	.925	.853	.852	.181	.89	H
25	.98	.893	.847	.223	.84	H
26	1.053	.773	.928	.332	.81	H
27	1.041	.813	.905	.977	.81	H
28	.922	.874	.892	.384	.88	H
29	.856	.867	1.012	.212	.89	C
30	.927	.976	.85	.128	.84	E
31	.872	1.146	.771	.109	.83	E
32	.86	1.074	.806	.168	.86	E
33	.775	.99	.948	.628	.92	E
34	.792	1.098	.811	.994	.9	E
35	.736	.914	1.025	1.194	.96	T
36	.788	.811	1.015	.094	.97	C
37	.761	.885	1.022	.103	.95	C
38	.874	1.039	.768	.171	.88	E
39	.947	.889	.85	.096	.86	H
40	.955	.958	.802	.475	.84	E
41	.919	1.098	.748	.388	.82	E
42	.902	.947	.852	.364	.87	E
43	.841	.899	.992	.23	.89	C
44	.917	.968	.87	.655	.85	E
45	.941	.922	.877	.159	.84	H
46	.92	.844	.966	.166	.87	C
47	.988	.983	.824	.136	.8	H
48	1.091	.858	.807	.277	.78	H
49	1.158	.822	.784	.152	.75	H
50	1.198	.776	.833	.025	.73	H
51	1.217	.843	.768	.082	.7	H
52	1.166	.933	.696	.163	.73	H
53	1.082	1.052	.702	.256	.74	H
54	1.065	.894	.864	.304	.77	H
55	.977	.984	.84	1.558	.8	E
56	.86	.922	.975	2.218	.88	T

Fig. 3. Text printout for the program DOUBLEPRED. The amino acid sequence of the α subunit of the bovine mitochondrial F1 ATPase is given using the one letter code (IUPAC). The prediction of the class of the protein is then supplied. The amino acid number, the helix, sheet and turn parameters calculated over a nine-residue length segment are given in the AA, PA, PB, PT, PC columns respectively. Ft indicates the turn frequency (for details see Deléage and Roux, 1987; Chou and Fasman, 1974, 1978).

pretation, the programs provide a list of peptide characteristics in sequence position order for PEPMAP and in elution order for CHROSIM.

The SEQ program is a powerful option which permits to search successively for different sequences (continuous or discontinuous) up to 200 amino acids long in as many as 100 proteins of 1000 amino acids.

The sequence homology between two proteins can be estimated using COMPARE with three different substitution matrices: strict homology, the mutation data matrix (Dayhoff

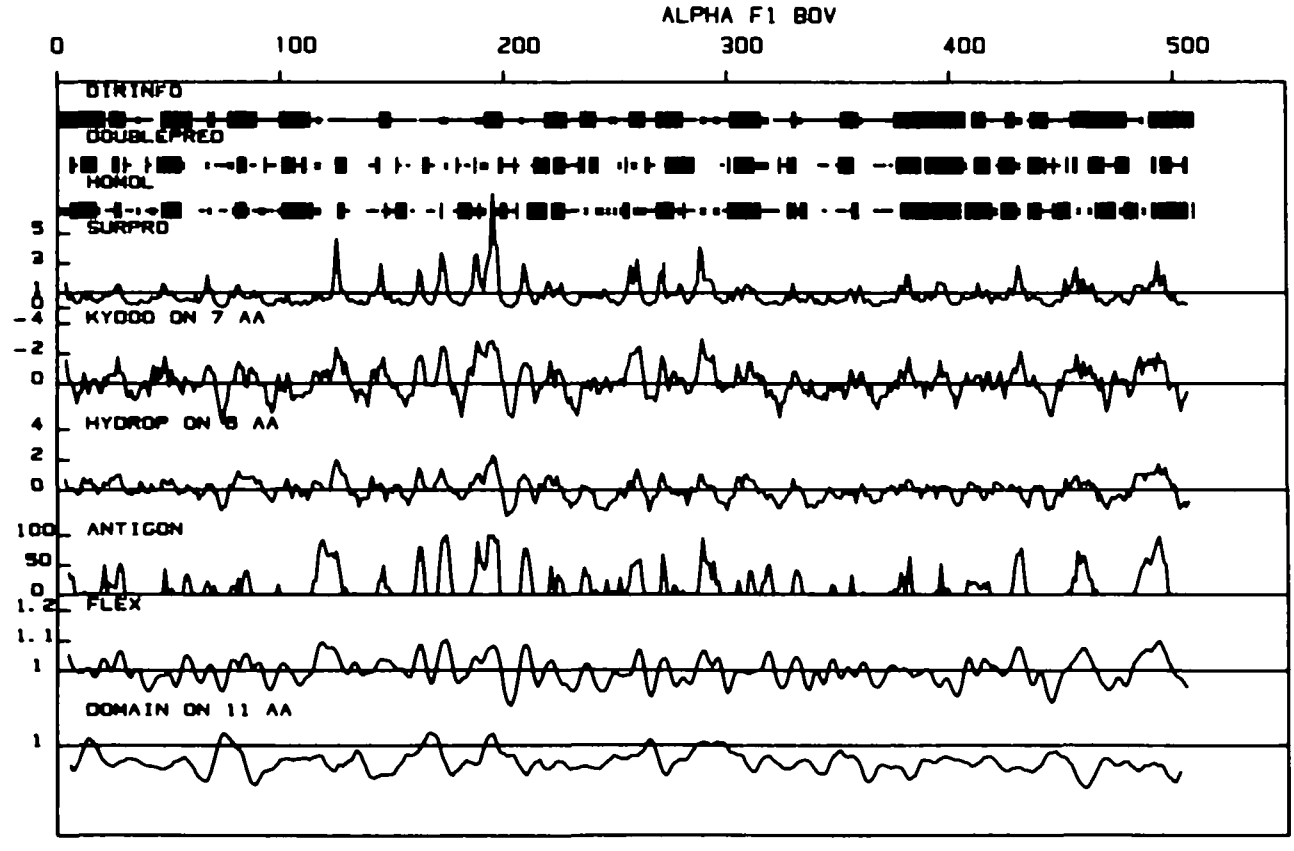

Fig. 4. Graphic display of a sequence analyzed by ANTHEPROT. This pattern is obtained with the bovine mitochondrial F1 ATPase α subunit, a 509 residues protein. The residue position is plotted in the horizontal axis The panels DIRINFO, DOUBLEPRED and HOMOL are the secondary structure predicted according to Garnier *et al.* (1978), Deléage and Roux (1987) and Levin *et al.* (1986) respectively. Three conformational states are visualized: helix (▮▮), sheet (------), turn (::::) whereas the blank regions are for the coil strands. The following panels are the profiles generated by the different programs which are mostly useful for determining the potential antigenic sites of a protein. The window widths appear in the panel comments when there are optional such as for KYDOO, HYDROP and DOMAIN otherwise the constant moving window is that described in the original method. For each graph, the scale is given as well as the mean value obtained for a random sequence. The time for drawing this figure is about 20 min at a pen velocity of 1 cm/s

et al., 1983) and the secondary structure similarity matrix (Levin *et al.*, 1986). The output is a classical dot-matrix plot (Staden, 1982).

The secondary structure of proteins can be predicted according to the methods of Chou and Fasman (1974, 1978) by CHOUFASMAN, Garnier *et al.* (1978) by DIRINFO, Levin *et al.* (1986) by HOMOL and Deléage and Roux (1987) by DOUBLEPRED. A typical example of printed text output is listed in Figure 3 for the analysis carried out with DOUBLEPRED on the α subunit of bovine mitochondrial F1 ATPase (Walker *et al.*, 1985). The amino acid composition is calculated to allow the protein class prediction to begin. The predicted class of a protein is determined by the lowest distance between the amino acid composition of the unknown protein and that of all proteins of a given class (Nakashima *et al.*, 1986). The secondary structure prediction is done taking into account the class prediction.

The hydrophobicity and hydrophilicity of proteins can be estimated according to Kyte and Doolittle (1982) by KYDOO. Several methods were designed for the prediction of potential antigenic sites. The first one (HYDROP) is based on the search for maximal hydrophilicity along the sequence (Hopp and Woods, 1981). The combination of methods using theoretical (Janin, 1979) and experimental (Parker *et al.*, 1986) hydrophilicity scales with that of prediction of polypeptide chain flexibility (Karplus and Schulz, 1985) is performed by ANTIGON (Parker *et al.*, 1986). In addition, the flexibility profile (FLEX) and the accessibility profile (SURPRO) of Boger *et al.* (1986) can also be used as separate criteria for antigenic sites prediction.

The algorithm of Rao and Argos (1986) to predict intramembrane helices is also available (MHELIX). The possible way of determining the boundaries of domains according to Vonderviszt and Simon (1986) is implemented in DOMAIN. All the algorithms of computerized methods have been keyed in as described by the authors. The Chou and Fasman (1978) method has been included as previously described (Deléage *et al.*, 1987). A typical graphic output delivered by the TRACE

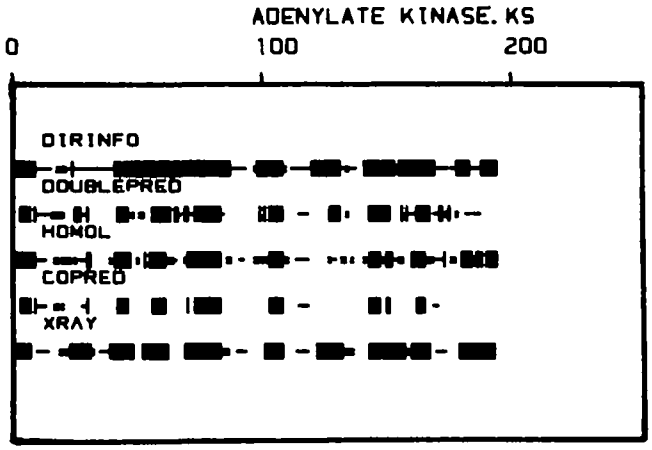

Fig. 5. Secondary structure of adenylate kinase. The secondary structure of adenylate kinase was predicted according to Garnier *et al.* (1978) (DIRINFO), Deléage and Roux (1987) (DOUBLEPRED), and to Levin *et al.* (1986) (HOMOL). The panel COPRED is for the predicted structure when the last two methods predict the same conformational state (helix, sheet and turn) for a given amino acid. For better clarity, the coil residues commonly predicted are not visualized. The X-ray panel indicates the observed secondary structure taken from Kabsch and Sander (1983). The accuracy of the combined prediction (COPRED) on adenylate kinase was 80% on a three state basis (helix, sheet and coil) when calculated on predicted residues.

option from the data saved on disk is shown in Figure 4. The graph has been delivered on a Hewlett Packard (HP 7470 A) plotter using files saved on disk by each program. Pen velocity (from 1 to 20 cm/s) and scale number (550 in this example) are chosen at the beginning of the program TRACE as well as the values to be plotted.

Implementation

Sequences up to 1000 amino acids can be analysed through the compiled form of all the programs. An automatic mode allows the study of up to 50 proteins successively via a particular program, except in the COMPARE and PEPMAP programs. Many options are available such as the display of the results (screen, printer or disk), decision constants equal to zero or not for the method of Garnier *et al.* (1978), thresholds for printing in the computerized version of the Chou and Fasman method and windows of different widths for hydrophobicity, hydrophilicity and domain boundaries profiles. The results of calculations can be saved on disk for a further exploitation, this option being particularly useful for anyone who has not the suited configuration for graphic printing.

The execution time of COMPARE program is proportional to the product of sequence lengths with probe length, the comparison of two 100 residue-long sequences with a probe length of 7 needs about 5 min on Apple IIgs. In the HOMOL program, each heptapeptide of the protein being studied must be compared with all heptapeptides of each protein of the database. Since the secondary structure database (Kabsch and Sander, 1983) comprises 60 proteins for a total number of amino acids higher than 10 000, this cannot be achieved in less than one night for a protein of 250 amino acids length (on Apple IIe). Nevertheless, the results can be stored on disk allowing a further listing or drawing of the predicted structure. The other programs are fast enough to allow the analysis of a protein of 500 amino acids in less than 5 min.

Together with the disks, the authors provide a short manual for the utilization of the programs and if desired an Applesoft BASIC version for further implementation.

This package is available on request provided you send two 5 1/4″ diskettes and set up your system configuration for graphic hard copy.

Discussion

The possibility of proteolysis modelling and chromatographic/electrophoretic simulation provides useful information for the design and the interpretation of such a study. The chromatographic simulation does not provide an exact location of the peaks but their distribution is correlated with experimental data, and the mean percent deviation of retention times is 9.2% (Sasagawa *et al.*, 1982).

The availability of a sequence comparison program will be useful in protein structure modelling (looking for homology with elucidated-structure proteins) and in phylogenic analysis.

Although a package for proteins secondary structure analysis has been recently described (Gribskov *et al.*, 1986) the main advantage of our software is that it permits the analysis by methods recently developed. Four secondary structure prediction methods are available, among which are two of the most currently reliable techniques. In addition, the combination of these two methods, which yields a significant improvement in secondary structure prediction accuracy (data not shown), is provided in the program COPRED. The power of such an approach is shown in Figure 5 for adenylate kinase for which most of the amino acids found in a given state by COPRED are effectively observed in this conformational state.

Another utility of the combination of methods based on different principles lies on the elimination of ambiguities. A typical example is given by the prediction of domains boundaries in proteins. Vonderviszt and Simon (1986) stress that their algorithm generates inexplicable false positives. Looking at the predicted secondary structure of a protein may help in eliminating the false minima observed in their output.

Besides, a powerful way for predicting antigenic potential sites is the consideration of several criteria. Parker *et al.* (1986) have pioneered such an approach and shown its reliability. Our package provides a greater diversity of recent and improved methods which can be associated on a single panel for an easy detection of regions with high antigenic potential.

Finally, this package gives the user the possibility of a complete analysis of protein sequence on a microcomputer with an

'artist' plot of the results which allows the fast retrieval of the information contained in a protein sequence.

Acknowledgements

Thanks are due to P Falson for computing assistance and to C Van Herrewège for the artwork. F F Clerc is a recipient of a BDI grant from the Centre National de la Recherche Scientifique.

References

Boger,J., Emini,E.A. and Schmidt,A. (1986) Surface probability profile. An heuristic approach to the selection of synthetic peptide antigens. *Reports on the Sixth International Congress in Immunology (Toronto)*, p. 250.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13**, 222–244.

Chou,P.Y. and Fasman,G.D. (1978) Prediction of secondary structure of proteins from amino acid sequence. *Adv. Enzymol. Relat. Subj. Biochem.*, **47**, 45–148.

Creighton,T.E. (ed.) (1984) *Proteins: Structure and Molecular Properties*. W.H.Freeman, p 7

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homology in protein sequence. *Methods Enzymol.*, **91**, 524–545.

Deléage,G. and Roux,B. (1987) An algorithm for protein secondary structure prediction based on class prediction. *Prot. Eng.*, **1(4)**, 289–294

Deléage,G., Tinland,B. and Roux,B. (1987) A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins. *Anal. Biochem.*, **163**, 292–297.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Gribskov,M., Burgess,R.R. and Devereux,J. (1986) PEPPLOT: a protein secondary structure analysis program for the UWGCG sequence analysis software package. *Nucleic Acids Res.*, **14**, 327–334

Hopp,T P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad Sci. USA*, **78**, 3824–3828.

Janin,J. (1979) Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Karplus,P.A. and Schulz,G.E. (1985) Prediction of chain flexibility in proteins. *Naturwissenschaften*, **72**, 212–213.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.*, **157**, 105–132

Laemmli,U.K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, **227**, 680–685.

Levin,J.M., Robson,B. and Garnier,J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett*, **205**, 303–308.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) The folding type of a protein is relevant to the amino acid composition. *J. Biochem., Tokyo*, **99**, 153–162

Parker,J.M.R., Guo,D. and Hodges,R.S. (1986) New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry*, **25**, 5425–5432

Rao,J.K.M. and Argos,P. (1986) A conformational preference parameter to predict helices in integral membrane proteins *Biochim. Biophys. Acta*, **869**, 197–214.

Sasagawa,T., Okuyama,T. and Teller,D.C. (1982) Prediction of peptide retention times in reverse-phase high performance liquid chromatography during linear gradient elution. *J. Chromatogr.*, **240**, 329–340.

Staden,R. (1982) An interactive graphics program for comparing and aligning nucleic and amino acid sequences *Nucleic Acids Res.*, **10**, 2951–2961.

Vonderviszt,F. and Simon,I. (1986) A possible way for prediction of domain boundaries in globular proteins from amino acid sequence. *Biochem. Biophys. Res. Commun.*, **139**, 11–17.

Walker,J.E., Fearnley,I.M., Gay,N.J., Gibson,B W , Northrop,F.D., Powell,S.J., Runswick,M.J., Saraste,M. and Tybulewicz,V.L.J. (1985) Primary structure and subunit stoichiometry of F1-ATPase from bovine mitochondria. *J. Mol. Biol.*, **184**, 677–701.

Received on November 24, 1987; accepted March 8, 1988