Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Vol. 159, No. 2, 1989
March 15, 1989

BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
Pages 687-693

PREDICT7, A PROGRAM FOR PROTEIN STRUCTURE PREDICTION

R.S. Cármenes*, J.P. Freije, M.M. Molina and J.M. Martín

Departamento de Biología Funcional,
Area de Bioquímica y Biología Molecular,
Universidad de Oviedo, 33071-Oviedo, Spain

Received December 22, 1988

We describe a program for protein sequence analysis which runs in IBM PC computers. Protein sequences are loaded from files in Mount-Conrad and Lipman-Pearson format. Seven features are analyzed: hydrophilicity, hydropathy, surface probability, side chain flexibility, antigenicity, secondary structure and N-glycosylation sites. Numeric results can be shown, printed or stored in files exportable to other programs. Graphics of up to four predictions can be displayed on the screen, printed out or plotted, with several definable options. This program has been designed to be fast, user-friendly and to be shared with the scientific community.

DNA sequencing is a much easier work than protein structure analysis, and as that technology is spreading and becoming popular, there is an increasing amount of DNA sequences encoding putative proteins. More than 8000 protein sequences have been identified so far, in contrast with only 300-400 proteins of known tertiary structure (1).

To demonstrate that a particular open reading frame really encodes a polypeptide of the predicted size is a long task needing a lot of laboratory work. Every time that a putative coding sequence is found, several key questions arise. What kind of protein is it? Which is its structure, location, function and antigenic properties? Even when we have already demonstrated that it encodes a polypeptide of known size, several of these questions remain to be answered. Predictive methods can give at this point invaluable help, provided their applicability and limitations are properly understood. Most predictive methods are based on statistical evaluations and none can give better than 60% accuracy (2). Predictions can not be taken as definitive conclusions but they give information helping to interpret other results or to design new experiments.

From the above considerations, it becomes obvious that, rather than using any single predictive analysis on its own, several of them should be considered simultaneously. Since there is a lack of programs allowing complex

0006-291X/89 $1.50

protein structure analysis, including several predictions and good graphical outputs, we have devised a project to develop this kind of programs. In this paper we present PREDICT7 which analyzes protein structural features using seven different algorithms. They include secondary structure analysis, hydrophilicity, side chain flexibility, surface probability, antigenicity and location of putative N-glycosylation sites.

Materials and Methods

(a) Hardware

PREDICT7 runs on IBM PC/XT/AT or PS/2 compatible machines with 512K memory. A graphic card (CGA, EGA, VGA or Hercules) is recommended to display graphics on the screen. No specific installation of the program is necessary. Hardcopies of the results can be obtained either through a printer using the PrtScr facility of PC-DOS, suitable for drafts, or using any HP7475A compatible plotter to give publication-quality graphics.

(b) Software

PREDICT7 was written using Turbo Pascal version 4.0 (Turbo Pascal is a trademark of Borland International, Inc.). The Pascal compiler is not necessary to run PREDICT7, which has been developed as a stand-alone program.

(c) Predictive algorithms

Several predictive algorithms have been implemented. Hydrophilicity and hydropathy values are calculated according to the scales of Hopp & Woods (3) and Kyte & Doolittle (4) respectively. Antigenicity is calculated using the scale of Welling et al. (5). Side chain flexibility is predicted following the Karplus & Schulz (6) method. The set of proteins from which data derive is the same as in the UWGCG package (7) in order to maintain consistency with data published by others using the same method but under different implementations. Emini et al. (8) surface probability is calculated with the single probabilities from Janin et al. (9). Secondary structure is calculated using the Garnier-Osguthorpe algorithm (10). N-glycosylation sites are searched as Asn-X-Thr or Asn-X-Ser sequences, where X is any residue (11).

Results and Discussion

(a) Description of the program

To use PREDICT7, no programing experience is needed. Each protein sequence must be entered in a text file using the one-letter standard amino acid code (both lower and upper case are accepted) in lines up to 80 characters long. Lines starting with ; or > will be ignored (reserved for comments), as will any blank space or non standard amino acid character. This simple and flexible format is compatible with that used in other popular sequence analysis programs such as the Mount-Conrad (12) or the Lipman-Pearson (13) packages. Both are good user-supported scientific software, the first mainly intended to analyze DNA sequences and translate them into amino acid sequences, and the second devised to protein data-bank retrieve and homology search. The program we describe in this paper complements the former two and protein sequence files obtained as an output from any of them are happily accepted as input files for analysis by PREDICT7.

Once the program is called, and a short description of the available predictions has been displayed, the user is requested to enter the sequence file-name. The sequence is then loaded and displayed on the screen together with its total length. Up to 1800 amino acid residues are accepted for analysis, which is far enough to deal with most protein sequences. Then the user is asked to enter the window size for hydrophilicity-hydropathy-antigenicity calculations and the Garnier's decision constant for α-helix and ß-sheet prediction. A default window size value of 6 is suggested, according to the recommendations of Hopp (14). To make the program easier to use and as an option, the sequence file-name and window size can be specified as parameters following the program name when calling it. In the example we shall examine later, this would be done by typing PREDICT7 IMP.AA 6.

Once the sequence has been loaded and the constants defined, calculations of the seven predictions will start, taking about 3 seconds for a typical 250 residue sequence in an IBM AT computer. After calculations have finished, the main menu is displayed. From it and its submenus, many options are available covering several aspects:

-Kind of output. Both numeric results and graphics can be obtained in various ways. They can be stored in a file for later use, displayed on the screen, printed out, or plotted using HP-compatible plotters.

-Graphical options. The four predictions to be shown, their order, and the region of the sequence (the total length by default) can be defined. The user can, as well, choose whether to write the scales, ticks, zero lines or axis labels. All are shown by default, but if a different set of definitions is frequently used, it can be saved in a file called PREDICT7.DAT.

-Standard serial port settings. The plotter is connected to the computer through a RS232 interface. Port number, bauds, data bits, stop bits, and parity checking bit have to be set to the appropriate values. This can be easily performed through a straight forward submenu and the new settings saved as before.

Single keystrokes allow shifting to submenus or options. The previous menu can be reached at any time by simply pressing the escape key (Esc). When data have been saved in a file, this is reminded to the user as exiting the program. Much care has been taken in designing the user interface in order to make the program as friendly as possible so that it can be used without reference to external instructions. Simple on-line help screens are available from any menu pressing the F1 key.

(b) Using the program

As an example, we have run the program using the sequence of one of the proteins we are currently investigating. This is the matrix protein of porcine transmissible gastroenteritis coronavirus (TGEV), a 262 residue

```
Program Predict7  (c) 1988 by R.S.Cármenes et al.

Structural predictions of protein from sequence in file imp.aa.
Done on thursday 6 december, 1988 at 10:00.

NG     N-glycosylation sites.
HydfHW hydrophilicity      (Hopp & Woods 1981).
HydpKD hydropathy          (Kyte & Doolittle 1982).
FlexKS flexibility         (Karplus & Schulz 1985).
SurfEJ surface probability (Emini 1985 & Janin 1978).
AgctW  antigenicity        (Welling et al. 1985).
SecGR  secondary structure (Garnier et al. 1978).

Window size for AgctW, HydfHW and HydKD = 6.
Garnier's decision constant for helices = -100.
Garnier's decision constant for sheets  = 0.
```

Amino ac.		NG	SecGR	HydfHW	HydpKD	FlexKS	Log of SurfEJ	AgctW
1	Met		Helix	-0.475	1.400	1.000	0.147	-0.007
2	Lys		Helix	-0.740	1.880	1.000	0.020	0.009
3	Leu		Helix	-0.917	2.317	1.000	-0.191	-0.041
4	Leu		Helix	-1.000	2.633	1.000	-0.275	0.036
5	Leu		Helix	-1.583	3.583	0.923	-0.844	0.020
6	Ile		Helix	-1.450	3.367	0.908	-1.094	-0.012
7	Leu		Helix	-1.400	3.433	0.906	-1.160	-0.027
8	Ala		Helix	-1.400	3.550	0.904	-1.243	-0.088
9	Cys		Helix	-1.183	3.100	0.904	-1.157	-0.020
10	Val		Helix	-1.050	2.883	0.904	-1.408	-0.052

Fig 1 Part of the printed numeric output. Protein sequence file, date of analysis, sequence length, and averaging window are always specified. Then the scores for hydrophilicity, hydropathy, antigenicity, flexibility and surface probability, and the predicted secondary structure assigned to each residue are shown. Ending the output are the calculated probabilities for every secondary configuration (helix, sheet, turns and coil).

integral membrane protein whose amino acid sequence has been deduced from its coding nucleotide sequence (15,16). A part of the numeric output is shown in Fig 1. Fig 2-A shows one of the graphics screens that can be obtained, and Fig 2-B graphics as plotted using different options.

The first overall characteristic observed is its global hydrophobicity. A clearly hydrophobic region can be observed at the N-terminus of the protein, corresponding to about the first 20 amino acid residues. This area also corresponds to an α-helix area in the Garnier's secondary structure prediction, and shows low surface probability. It has been experimentally determined that the first 17 residues of this protein correspond to its signal peptide (16).

Next to the signal sequence comes a less hydrophobic stretch, where 2 of the 3 potential glycosylation sites of the protein are located. This portion is, accordingly, believed to be located outside of the viral particles. Three more strongly hydrophobic regions are found, all three separated by short hydrophilic sequences where turns and high surface probabilities are predicted. This can be interpreted as three membrane spanning regions, linked by turns situated near the membrane surfaces.

The half carboxylic end of the protein is less hydrophobic, and alternated α-helix and ß-sheet regions are predicted separated by turns. This may be interpreted as an α/ß structure located either inside or outside the viral particle. The integral membrane protein from this and other

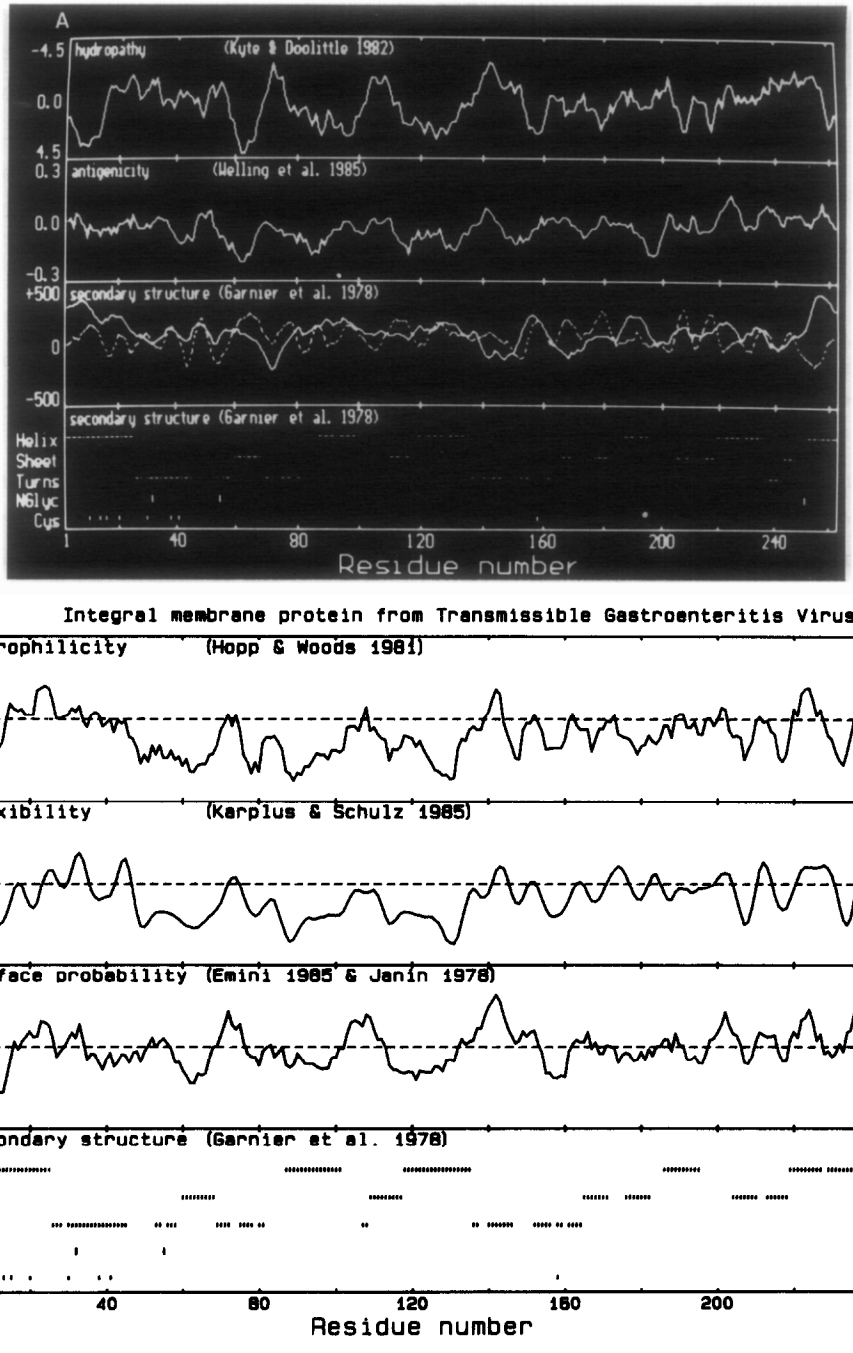

Fig 2 Graphical output as seen on the screen with a Hercules Graphic Card (A) or plotted (B). Some options have been changed in each to show several possible graphic plots.

coronaviruses are assumed to have this portion inside the virus (16,17,18). The last 25 residues correspond to the most hydrophilic part of the protein, and high scores are given for surface probability, side chain flexibility and

antigenicity. This stretch also contains the third potential glycosylation site. Preliminary experimental evidence found in our laboratory suggests that an antigenic site to a specific monoclonal antibody can be situated in this area. The presence of the N-glycosylation site and its antigenicity can support the hypothesis that this part of the protein could be situated outside of the viral particle, in contrast to the previously proposed structure.

(c) Conclusions

From the study of this example, several general considerations can be made. Firstly, several predictive methods should be used, in order to contrast them. Secondly, part of the information obtained is somewhat redundant with some of the predictions. This is the reason beneath the decision of not including every prediction in graphical outputs. We believe four of them can be representative enough, avoiding overcrowded figures. And thirdly, predicted structures should be carefully contrasted with the experimental available information as it comes: each can complement the other, helping to fit data in the puzzle. A review on this subject relative to integral membrane proteins has recently been published (19) and illustrates how both kinds of data should be considered.

There is a number of protein structure prediction programs available. Some are integrated in complex commercial DNA/protein analysis packages, while others are freely available to the scientific community. These later programs have shown its usefulness, but have three major shortages. They usually do not have any graphical capacity, which is essential for a proper understanding of the predicted structures, they can not simultaneously analyze more than one protein structure feature, and they are not very fast. Although some of the commercial programs do overcome these deficiencies they are not freely available to the academic community, are usually expensive and sometimes need mainframe computers to be implemented. From these considerations, we believe PREDICT7 can be useful to other investigators in this field.

(d) Availability of the program

PREDICT7 is available to anyone for non-commercial use upon request by sending a $5\frac{1}{4}$ inch formatted blank diskette to the authors. A copy of the program can also be obtained via EARN/BITNET by requesting it to CMSMD11@EOVUOV11.

Acknowledgments

This work was supported by a CEC-BAP grant number BAP-0219-E. M.M.M was supported by a *Fondo de Investigaciones Sanitarias* fellowship, and J.M.M. by a *Ministerio de Educación y Ciencia* fellowship. We would like to express our gratitude to F.Parra and C.L.Otín for the interest shown during the development of this work, and to J.Riera for his technical advice.

References

1.-B.Jasny (1988) Science **240**, 722-723.
2.-M.J.Rooman & S.J.Wodak (1988) Nature **335**, 45-49.
3.-T.P.Hopp & K.R.Woods (1981) Proc.Natl.Acad.Sci.USA **78**, 3824-3828.
4.-J.Kyte & R.K.Doolittle (1982) J.Mol.Biol. **157**, 105-132.
5.-G.W.Welling, W.J.Weijer, R.van der Zee & S.Welling-Wester (1985) FEBS Lett. **188**, 215-218.
6.-P.A.Karplus & G.E.Schulz (1985) Naturwissenschaften **72**, 212-213.
7.-J.Devereux, P.Haeberli & O.Smithies (1984) Nuc.Acid Res. **12**, 387-396.
8.-E.A.Emini, J.V.Hughes, D.S.Perlow & J.Boger (1985) J.Virol. **55**, 836-839.
9.-J.Janin, S.Wodak, M.Levitt & M.Maigret (1978) J.Mol.Biol. **125**, 357-386.
10.-J.Garnier, D.J.Osguthorpe & B.Robson (1978) J.Mol.Biol. **120**, 97-120.
11.-S.C.Hubbard & R.J.Ivatt (1981) Ann.Rev.Biochem. **50**, 555-583.
12.-D.W.Mount & B.Conrad (1986) Nuc.Acid Res. **14**, 443-454.
13.-W.R.Pearson & D.J.Lipman (1988) Proc.Natl.Acad.Sci.USA **85**, 2444-2448.
14.-T.H.Hopp (1986) J.Immunol.Meth. **88**, 1-18.
15.-P.Britton, R.S.Cármenes, K.W.Page & D.J.Garwes (1988) Mol.Microbiol. **2**, 497-505.
16.-H.Laude, D.Rasschaert & J.C.Huet (1987) J.Gen.Virol. **68**, 1687-1693.
17.-J.Armstrong, H.Niemann, S.Smeekens, P.J.M.Rottier & G.Warren (1984) Nature **308**, 751-752.
18.-P.J.M.Rottier, G.W.Welling, S.Welling-Wester, H.G.M.Niesters, J.A.Lenstra & B.van der Zeijst (1986) Biochem. **25**, 1335-1339.
19.-P.D.McCrea, D.M.Engelman & J.L.Popot (1988) Trends Biochem.Sci. **13**, 289-290.