PREDITOP: A program for antigenicity prediction

J.L. Pellequer†, E. Westhof

PREDITOP: A program for antigenicity prediction

J.L. Pellequer*† and E. Westhof‡

*UPR Structure des Macromolécules Biologiques et Mécanismes de Reconnaissance, *Laboratoire d'Immunochimie and ‡Equipe de Modélisation et de Simulation des Acides Nucléiques, Institut de Biologie Moléculaire et Cellulaire du CNRS, France*

A program (PREDITOP) for predicting the location of antigenic regions (or epitopes) on proteins is described. This program and the associated ones are written in Turbo Pascal and run on IBM-PC compatibles. The program contains 22 normalized scales, corresponding to hydrophilicity, accessibility, flexibility, or secondary structure propensities. New scales are easily implemented. An hydrophobic moment procedure has also been implemented in order to determine amphiphilic helices. The program generates a result file where the values represent a particular physicochemical aspect of the studied protein. PREDITOP can display one or several result files by simple graphical superimposition. Curve combinations can be done by the ADDITIO or MULTIPLI routines which create a new result file by adding or multiplying previously calculated files representing several propensities. The program is useful and efficient for identifying potential antigenic regions in a protein with the aim of raising antibodies against synthesized peptides which cross-react with the native protein.

Keywords: antigenicity of proteins, prediction of epitopes, hydrophilicity, segmental flexibility, accessibility, hydrophobic moment, secondary structure

INTRODUCTION

Antigenicity reflects the ability of a molecule to be recognized by an antibody. The region of the antibody that binds to the antigen is made up of six complementary-determining regions, and is called a *paratope*. The antigenic region recognized by the paratope is named the *epitope*. The most common antigens are proteins, and their epitopes are of two types: continuous and discontinuous. A continuous epitope is made of consecutive amino acids in the protein sequence. A discontinuous epitope is a region where the recognized amino acids are brought together in three-dimensional space, but are distant in the sequence. It is generally assumed that most antigenic regions in proteins are constituted of discontinuous epitopes.[1–3]

There are two ways of delineating protein epitopes. The first approach is X-ray crystallography of antibody-antigen complexes. Up to now, five complexes have been solved, three were Fab-Lysozyme complexes[4–6] and two were Fab-influenza neuraminidase complexes.[7,8] An idiotypic complex (Fab-Fab) has also been recently solved.[9]

The second approach is to study the reaction of antibody with protein fragments or synthetic peptides. The antibodies may have been raised either against the native protein or against a protein fragment. In both cases, the antibody should cross-react with the native protein and with the protein fragment.

The localization of the antigenic regions in a protein is of particular interest for the development of synthetic vaccines. One expects that a peptide mimicking a protein region will be able to induce an antibody response leading to recognition of the parent protein. A second application of epitope localization is the use of specific antibodies for screening genomic expression banks and localizing the protein *in situ*.

This report describes a package to help to localize the continuous epitopes of a protein from its sequence. The aim of the antigenicity prediction method is to identify one or several regions to be synthesized for the production of antipeptide antibodies cross-reactive with the parent protein. In this case, only continuous epitopes are considered. It should be emphasized that our aim was not to predict all antigenic sites of a protein, but to localize a small number of sites with a high degree of confidence in order to suggest which peptides should be chemically synthesized. The first step of this study was to develop a program which could predict and plot antigenic regions. The classical calculation procedure (i.e., the window assignment) was first developed by Hopp and Woods.[10]

PROGRAM DESCRIPTION

The PREDITOP software calculates and plots antigenicity prediction profiles based on propensity scales. The program contains four main calculation procedures and a complete graphic representation of the calculated curves. The results are expressed as a graph, where peaks should correspond to

Color Plates for this article are on pages 191–192.

Address reprint requests to Dr. Westhof at UPR Structure des Macromolécules, Biologiques et Mécanismes de Reconnaissance, Institut de Biologie Moléculaire et Cellulaire du CNRS, 15 rue René Descartes, 67084 Strasbourg Cedex, France.

Received 24 August 1992; revised 23 October 1992; accepted 27 October 1992

the antigenic parts of the protein and valleys to the non-antigenic parts, which should generally correspond to the interior of the protein. The x-axis represents the amino acid sequence and the y-axis represents the relative propensity of the studied sequence in the chosen scale.

The program is written in Turbo Pascal V5.5. It runs on the IBM® PC/XT/AT or compatible computers with MS-DOS® 3.XX. The program requires 256 Kb of RAM and is compatible with the graphic card CGA (320×200, 4 colors), EGA (640×350, 16 colors), VGA (640×480, 16 colors), or HCG (720×348, monochrome). The program detects automatically the graphic card. Better results are obtained with the VGA video card. The calculations are fast, e.g., 200 amino acids are computed per second on a 386-20 MHz. The program has been written with a user-friendly approach by implementing systematic default values. The program does not require, but is compatible with, an arithmetical coprocessor 387-25 Mhz.

INPUT/OUTPUT

The program needs three input files which are the propensity scale, the sequence and the journal file. As shown in Figure 1a, two of the three files are obtained by specific subroutines named NORMALIS and LECTURE. The third one, the journal file, is a faked input file, which is required at the beginning as a blank file and which is also an output file.

The propensity scales correspond to a broad variety of physicochemical parameters which have been correlated with the location of the continuous epitopes in a few well-characterized proteins, such as hydrophilicity,[10,11] accessibility,[12-14] and flexibility of short segments of polypetide chains.[15,16] A scale is composed of 20 values, each assigned to an amino acid. Those values sort the amino acids according to one parameter, as cited above. For instance, the most hydrophilic amino acid has the greater value in a hydrophilicity scale. The package contains 22 scales, all normalized so that the results can be superimposed and mathematically manipulated.

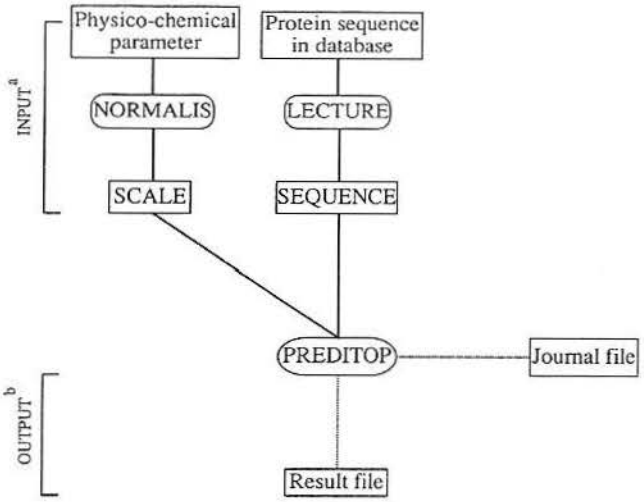


Figure 1. PREDITOP input/output flowchart. The rectangular boxes are data files and the oval boxes are programs. a) The upper part is the input and contains two ASCII files: the scale and the sequence. The first line with rectangular boxes corresponds to the database. The second one corresponds to the computed data files. Connections are made with solid lines. b) The lower part is the output and contains two files: the result file and the updated journal file. Connections are made with dashed lines.

The normalization was obtained in the following way. First, the mean of the scale to be normalized is set to zero by subtracting its original mean. Then, the values are normalized between +3 and −3, which is an arbitrary choice, given by the limits of the first hydrophilicity scale[10] used for antigenicity prediction. Thus, the new normalized value is equal to

$$\frac{(\text{old value}) \times 3}{\text{maximum old value}}$$

This procedure will thus compress or expand a published scale between the ±3 boundaries. It was assumed that the possible distortion in the standard deviation of the original scale is a minor drawback in comparison to the advantages gained afterwards by the processing of the calculated curves. The hydrophilicity scale of Parker *et al.*[11] is represented below (in the order required by PREDITOP)

Arg : 0.87	Asp : 2.46	Glu : 1.86	Lys : 1.28
Ser : 1.50	Asn : 1.64	Gln : 1.37	Gly : 1.28
Pro : 0.30	Thr : 1.15	Ala : 0.03	His : 0.30
Cys : 0.11	Met : −1.41	Val : −1.27	Ile : −2.45
Leu : −2.78	Tyr : −0.78	Phe : −2.78	Trp : −3.00

We remark that no values are available for the undetermined amino acids B (ASN or ASP) and Z (GLN or GLU). Thus, the program checks the protein sequence in order to search for such undefined amino acids and asks in which amino acid type they should be changed.

The protein sequences could be obtained by database searching using the UWGCG package, for example.[17] Then a minor operation has to be done for transcribing this sequence into a good format for PREDITOP. We mark the beginning of the sequence by \\ and the end of the sequence by a point, using any text editor. The LECTURE program transcribes the original sequence file in an ASCII format, where the first line corresponds to the total number of amino acids and each following line corresponds to one amino acid in a one letter code.

The journal file which is, at the beginning, an empty ASCII file, will contain names of all result files at the end of the run as well as the methods used to construct them (Figure 1a).

The output files are the result file of the calculations and an update of the journal file (Figure 1b). The files are both in ASCII form, and may be created by a single DOS command or with any text editor. The use of ASCII files is slower than binary files, but allows a quick visualization of the file contents. The result file is composed of the following: the first line contains the total number of amino acids of the protein, and each following line contains a real value corresponding to the calculated value of the center of the window. The two last lines correspond to the mean and the standard deviation of the values.

COMPUTATION

Calculations are based on a window assignment. A window is a section of a protein sequence composed of consecutive amino acids. A value is assigned to each amino acid of the window according to the selected scale. An arithmetical mean is then made within this window, and the mean is assigned to the center of the window in the case of the classical method.[10] Then, the window is shifted by one residue and the procedure continues. Each window center is saved in an output file. However, in the classical calculation, the window is smoothed by a Gaussian function in order to stress the center of the window as follows:

$$F(x) = \frac{1}{\sigma\sqrt{2\Pi}} \exp\left(\frac{-(x_i - x_j)^2}{2\sigma^2}\right)$$

where x_i is the point to be smoothed with i varying from 1 to window length and x_j is the mean of the window.[18] A sigma value (σ) of 2 is well suited for a nice smoothing. Each value within the window is multiplied by the corresponding value of the Gaussian function.

The four main calculation procedures are (Figure 2):

(1) A standard calculation based on an arithmetical mean using a window of generally seven amino acid length, but any length is accepted[10]
(2) Calculation of flexibility according to the Karplus and Schulz algorithm based on a triple scale[19]
(3) Calculation of surface accessibility according to the formula of Emini *et al.*[20] based on multiplication instead of addition within the window
(4) Amphiphilic helix determination based on the hydrophobic moment calculation according to the Eisenberg *et al.* formula.[21]

Figure 2 lists the authors who have developed the propensity scales (classical scales) or who have developed the calculation methods (other scales).

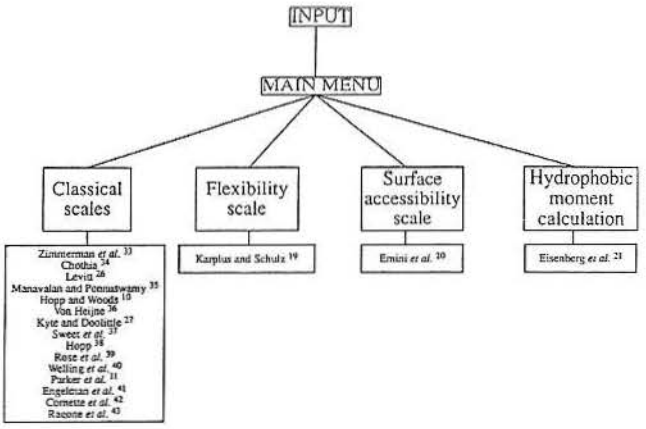


Figure 2. Four calculation procedures available from the main menu. The first procedure is the standard one, and may use all the scales listed below the box. The scales may have different physicochemical origins, such as hydrophilicity, accessibility, flexibility and secondary structure. The last three procedures are different calculation methods. For this reason, only the name of the authors who developed those methods are given.

The calculation based on a flexibility scale[19] is similar to classical calculation, except that the center is the first amino acid of the six amino acids window length, and there are three scales for describing flexibility instead of a single one. The three scales correspond to different degrees of rigidity in the neighboring residues. According to the rigidity of a neighbor, a value of one of the three scales is assigned to the current amino acid.

The calculation based on surface accessibility scale[20] is quite different because it is based on a product instead of an addition within the window. The accessibility profile is obtained using the formula

$$Sn = \left(\prod_{i=1}^{6} \delta_{n+4-i}\right)(0.37)^{-6}$$

where Sn is the surface probability, δn is the fractional surface probability value,[22] and i varies from 1 to 6.

The hydrophobic moment calculation is based on the Eisenberg *et al.* formula:[21]

$$\mu = \sqrt{\left(\sum_{i=1}^{n} H_i \cos\delta_i\right)^2 + \left(\sum_{i=1}^{n} H_i \sin\delta_i\right)^2}$$

where H_i is the hydrophobicity of the amino acid and δ_i is the successive angle between an amino acid and the next (97° for a right helix). The length n of the window is generally assumed to be 11 residues.[23] This hydrophobic moment is a widely used method for determining amphipathic helices in a protein. We adopt the following scheme to calculate the hydrophobic moment. Since the angle step in proteins is not always the ideal value of 100°, we chose to use three curves where the angles vary from 90° to 104° (i.e., 90°, 97°, and 104°). The three curves are then added by the ADDITIO subroutine. The ADDITIO routine has two rules for adding the curves. When the two values to be added have the same sign, then the added value is the sum of the two values. Otherwise, the output value is set to zero. At the end, the curve is normalized between +3 and −3. The effect of the last rule is to eliminate the uncertain helical assignments. We can see the effect of the ADDITIO routine in Color Plate 1. When such helices were predicted with the ADDITIO routine, the positions of 35 amphiphilic helices were correctly predicted out of a total of 55 predicted helices (Table 1).

The ADDITIO routine can also be used for adding the curves resulting from different propensity scales. Owing to the uncertainties in the scales, only a weight of 1 is implemented.

Another routine, MULTIPLI, allows for simple mathematical treatment of the curves. This routine uses a more drastic method, since it consists in multiplying the result curves instead of adding them. The calculation rules are identical, which means that a zero value is obtained when the two values to be multiplied are of opposite signs. Color Plate 2 shows the expected great difference obtained after the use of those two routines. A systematic comparison of the addition and multiplication routines has been performed (Pellequer *et al.*[24]). In summary, the addition of curves predicts more peaks, with the disadvantage that some wrongly predicted peaks are retained, while the multiplication of curves leads to less peaks and to fewer false peaks.

Table 1. Results obtained for the prediction of amphiphilic helices*

Protein codes	Number of peaks		
	Total	Predicted	Correctly predicted
CHO	2	3	2
CYT	4	4	3
IFB	5	6	4
LEG	7	5	5
LYS	4	3	2
MHR	4	4	4
MYO	8	6	5
RAS	5	5	3
REN	4	12	2
SCO	1	3	1
TMV	5	4	4
TOTAL	49	55	35

* Abbreviations are: CHO for cholera toxin, CYT for cytochrome c, IFB for β-interferon, LEG for leghemoglobin, LYS for lysozyme, MHR for myohemerythrin, MYO for myoglobin, RAS for h-RAS p21 oncogene, REN for renin, SCO for scorpion neurotoxin, and TMV for tobacco mosaïc virus protein.

GRAPHICS

The PREDITOP graphic routine basically plots a result file which contains real values normalized between -3 and $+3$. The mean of the values is set to zero.

Three other items, if they are known, can be also displayed (Figure 3). They are the location of the epitopes, i.e., the first and last amino acids of an epitope, the secondary structure, i.e., the first and last amino acids adopting any of the three regular secondary structures (helix, sheet, and turn), and finally, the sequence in the one-letter code format. Any number of those files can be displayed. Color Plate 3 is a graphical representation example. A complete menu makes provision for the graphic representation parameters. It contains x and y scaling, x and y translation on the screen. When using a color screen, the curves are colored. When using a monochrome screen, the curves are dashed. One can also choose the histogram form for representing the calculated curves.

There are two ways of using the graphical representation. First, we can superimpose several result files which have the same length in residues. This allows the examination of several physicochemical parameters. In other words, we can identify the best peaks if it is assumed that antigenic sites are composed of hydrophilic, accessible and flexible amino acids. We can see the advantage of superimposition on Color Plate 4 where several relevant items of information are plotted on the same graph. Evidently, one could study many other sequential aspects of a protein. For example, we can construct a scale where ARG and LYS are at value $+3$ and ASP and GLU at value -3 and all others values at zero; such a scale allows a rapid localization of the charged amino acids in a protein. Thus, one can construct scales for any particular purpose and visualize them through a graphical representation.

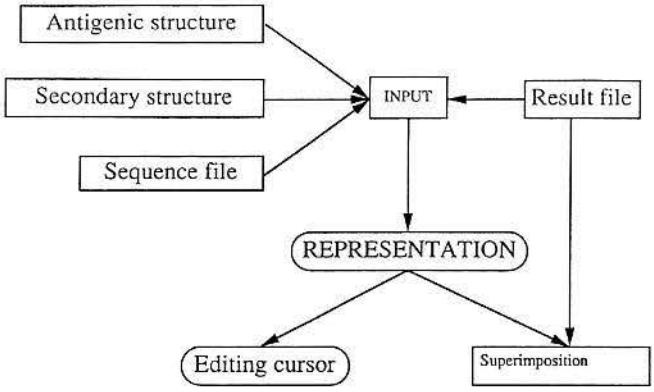


Figure 3. Graphical procedure. Rectangular boxes correspond to data files, and oval boxes correspond to independent graphic routines. The arrows show how the graphical procedure may be used. On the upper left, there are three accessory files when sufficient data are available but these are not a requirement for the graphical representation. On the upper right, the file which contains the data to be ploted is represented. The superimposition and the editing cursor are discussed in the text.

Second, an editing cursor makes it possible to move along the sequence by pressing the left and right arrow keys. This function displays the amino acid name and its position in the sequence, and also displays the propensity value of this amino acid according to the scale used. A standard deviation around the mean gives an estimation of the validity of the peak displayed. We generally use a threshold of 0.7 times σ. This limit corresponds to 25% of the total predicted amino acids and may be changed by the user (Color Plate 3). The curves can be printed by a simple hardcopy.

CONCLUSION

The PREDITOP program and its associated routines make it possible to visualize graphically the behavior of a physicochemical parameter along the sequence of a protein (Figure 4). The program was mainly used to predict antigenicity by studying several parameters simultaneously. The package is composed of five programs.

The first program, NORMALIS, is the starting point, since it normalizes the propensity scale if it is not one of the 22 built-in scales. Then, the sequence of interest should be entered either manually, by typing the sequence in a text editor, or by transferring it from a database. The program LECTURE transforms the basic sequence to an ASCII format usable by PREDITOP. At this level, we can introduce the protein code. When LECTURE runs, we must assign a three-letter name to the studied protein. This name is called protein code, and will be a part of all the filenames used in the package. For example, in the case of myoglobin, MYO is the protein code, and the filenames for the sequence file, result file, secondary structure file, and epitope file are: DATAMYO, MYO1.HOP (result file constructed with the hydrophilicity scale of Hopp and Woods[10]), MYOSTRU, and MYOEPIT, respectively.

As explained above, PREDITOP calculates a profile of one parameter with a protein sequence. At this level, two

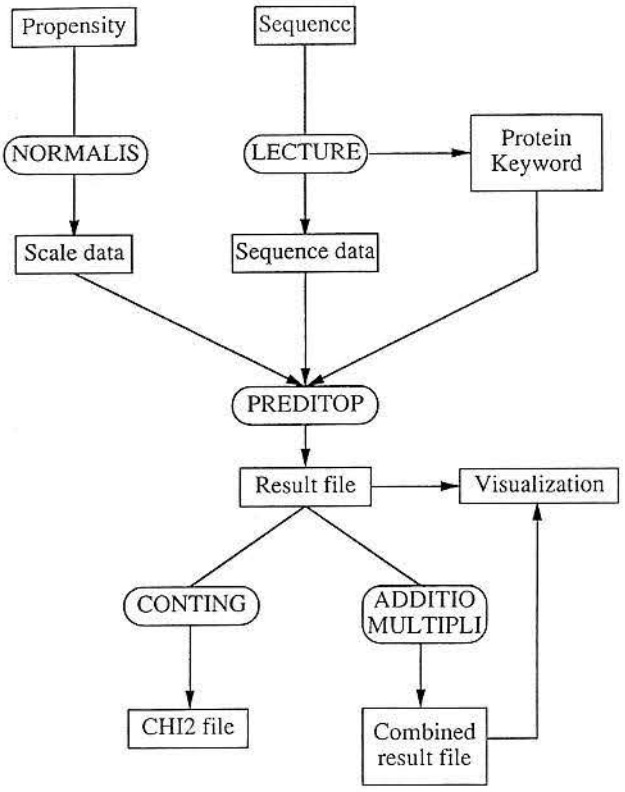


Figure 4. PREDITOP environment. The upper part is composed of three input lines. The first one is the database, the second line is the program and the third line is the data files. The protein keyword is the common code of the program names for input, for output and for representation. This keyword is determined by the LECTURE program. The lower part is composed of three parts. The first corresponds to the visualization routine, the second corresponds to the evaluation of the accuracy of the predicted curves when antigenic data are available (CONTING), the third corresponds to the combined curves obtained by the ADDITIO or the MULTIPLI programs. Evidently, those combined curves could be visualized, since all calculated curves are normalized between −3 and +3.

approaches are offered. First, one can estimate the validity of the prediction if sufficient information is available. Second, one can combine several approaches by adding result files residue by residue in a new result file. This operation is performed by the routines ADDITIO and MULTIPLI, which give normalized result files. In our case, we have selected well-known epitopes on several proteins and tested several scales.[25] The program CONTING determines two parameters: the numbers of well or wrongly predicted amino acids and a statistical expression of this number: x^2. The threshold is based on a value equal to 0.7 times the standard deviation. The accuracy of the prediction method has been measured with 22 scales applied to 11 proteins.[25] We conclude that none of the single prediction scales in current use gives a level of correct prediction higher than 60%. We have found[25] that the turn prediction scale of Levitt[26] and the accessibility prediction scale of Emini *et al.*[20] give the best result with approximately 60% of amino acids correctly predicted. Recently, in order to increase the percentage of correctly predicted amino acids and, especially, to minimize the number of incorrect predictions, we tested the MULTIPLI routine (Pelleque *et al.*[24]). The use of this routine leads to few predicted epitopes but with an accuracy in correctly predicted epitopes around 70%. An example of such a prediction, using the multiplication procedure, is given in Color Plate 5.

The PREDITOP program is a newly written program and not an updated version of the Hydrphil/Hydgraf programs which were written a few years ago.[18] PREDITOP links together calculation and graphical representation routines. The major changes reside in simplicity and convenient use. Considerable improvements have been made since the first prediction programs[10,27] in which no graphical representation was implemented and where the input (e.g., protein sequence) had to be typed by hand. Later versions of such programs[28] had graphic routines integated,[29,30] but the main disadvantage is that no superimposition could be obtained. PREDITOP may superimpose up to 10 different physicochemical parameters.

The UWGCG package[17] offers a program to calculate protein antigenicity by the means of the antigenic index[31] in which four approaches are used. The antigenic index is obtained by weighted addition of those parameters. In this case, we have no choice of the weights, scales, and calculation parameters such as window length, window center, or smoothing, and no superimposition is available.

Finally, one of the major advantages of PREDITOP is that it makes it possible to create a new scale and calculate a profile which could be superimposed with other classical parameters unlike the EPITPLOT program written in BASIC by Menéndez-Arias and Rodriguez.[32] In all other programs the scales are integrated into the program and cannot be changed by the user.

ACKNOWLEDGEMENTS

We thank G.D. de Marcillac for his help during the early stages of this work, and M.H.V. Van Regenmortel for advice and numerous fruitful discussions.

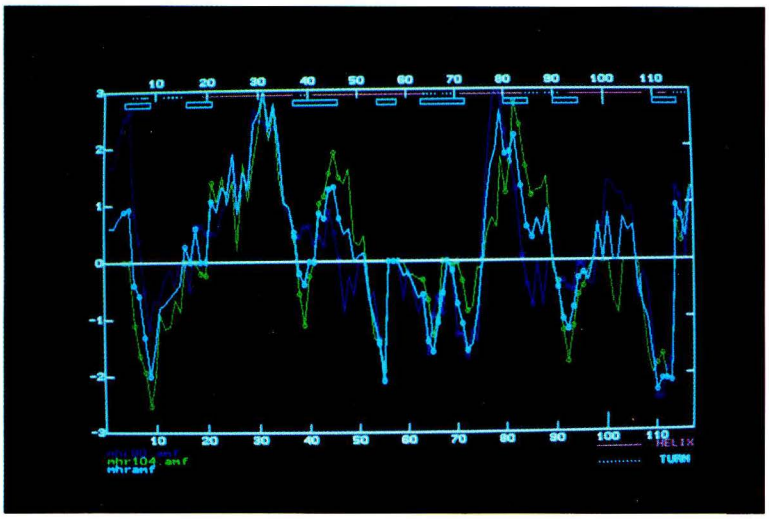
REFERENCES

1 Van Regenmortel, M.H.V. Antigenic cross-reactivity between proteins and peptides: new insights and applications. *Trends Biochem. Sci.* 1987, **12,** 237–240
2 Getzoff, E.D., Tainer, J.A., Lerner, R.A., and Geysen, H.M. The chemistry and mechanisms of antibody binding to protein antigens. *Adv. Immunol.* 1988, **43,** 1–98
3 Van Regenmortel, M.H.V. The concept and operational definition of protein epitopes. *Phil. Trans. R. Soc. Lond.* 1989, **B 323,** 451–466
4 Amit, A.G., Mariuzza, R.A., Phillips. S.E.V., and Poljak, R.J. Three-dimensional structure of an antigen-antibody complex at 2.8 Å resolution. *Science* 1986, **233,** 747–753
5 Sheriff, S., Silverton, E.W., Padlan, E.A., Cohen, G.H., Smith-Gill, S., Finzel, B.C., and Davies, D.R. Three-dimensional structure of an antibody-antigen complex. *Proc. Natl. Acad. Sci. USA* 1987, **84,** 8075–8079

6 Padlan, E.A., Silverton, E.W., Sheriff, S., and Cohen, G.H. Structure of an antibody-antigen complex: crystal structure of the HyHEL-10 Fab-lysozyme complex. *Proc. Nat. Acad. Sci. USA* 1989, **86,** 5938–5942
7 Colman, P.M., Laver, W.G., Varghese, J.N., Baker, A.T., Tulloch, P.A., Air, G.M., and Webster, R.G. Three-dimensional structure of a complex of antibody with influenza virus neuraminidase. *Nature* 1987, **326,** 358–363
8 Tulip, W.R., Varghese, J.N., Webster, R.G., Air, G.M., Laver, W.G., and Colman, P.M. Crystal structures of neuraminidase-antibody complexes. *Cold Spring Harbor Symp. Quant. Biol.* 1989, **54,** 257–263
9 Bentley, G.A., Boulot, G., Riottot, M.M., and Poljak, R.J. Three-dimensional structure of an idiotype-anti-idiotype complex. *Nature* 1990, **348,** 254–257
10 Hopp, T.P., and Woods, K.R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* 1981, **78,** 3824–3828
11 Parker, J.M.R., Guo, D., and Hodges, R.S. 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* 1986, **25,** 5425–5432
12 Barlow, D.J., Edwards, M.S., and Thornton, J.M. Continuous and discontinuous protein antigenic determinants. *Nature* 1986, **322,** 747–748
13 Thornton, J.M., Edwards, M.S., Taylor, W.R., and Barlow, D.J. Location of 'continuous' antigenic determinants in the protruding regions of proteins. *EMBO J.* 1986, **5,** 409–413
14 Novotny, J., Handschumacher, M., Haber, E., Bruccoleri, R.E., Carlson, W.B., Fanning, D.W., Smith, J.A., and Rose, G.D. Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains). *Proc. Natl. Acad. Sci. USA* 1986, **83,** 226–230
15 Westhof, E., Altschuh, D., Moras, D., Bloomer, A.C., Mondragon, A., Klug, A., and Van Regenmortel, M.H.V. Correlation between segmental mobility and the location of antigenic determinants in proteins. *Nature* 1984, **311,** 123–126
16 Tainer, J.A., Getzoff, E.D., Alexander, H., Houghten, R.A., Olson, A.J., Lerner, R.A., and Hendrickson, W.A. The reactivity of anti-peptide antibodies is a function of the atomic mobility of sites in a protein. *Nature* 1984, **312,** 127–134
17 Devereux, J., Haeberli, P., and Smithies, O. A comprehensive set of sequence analyses program for the VAX. *Nucl. Acid. Res.* 1984, **12,** 387–395
18 Van Regenmortel, M.H.V., and de Marcillac, G.D. An assessment of prediction methods for locating continuous epitopes in proteins. *Immunol. Lett.* 1988, **17,** 95–107
19 Karplus, P.A., and Schulz, G.E. Prediction of chain flexibility in proteins. A tool for the selection of peptide antigens. *Naturwissenschaften* 1985, **72,** S. 212
20 Emini, E.A., Hughes, J.V., Perlow, D.S., and Boger, J. Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide. *J. Virol.* 1985, **55,** 836–839
21 Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* 1984, **81,** 140–144
22 Janin, J., Wodak, S., Levitt, M., and Maigret, B. Conformation of amino acid side-chains in proteins. *J. Mol. Biol.* 1978, **125,** 357–386
23 Dohlman, J.G., De Loof, H., Prabhakaran, M., Koopman, W.J., and Segrest, J.P. Identification of peptide hormones of the amphipathic helix class using the helical hydrophobic moment algorithm. *Proteins* 1989, **6,** 61–69
24 Pellequer, J.L., Westhof, E., and Van Regenmortel, M.H.V. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett.*, in press
25 Pellequer, J.L., Westhof, E., and Van Regenmortel, M.H.V. Predicting the location of continuous epitopes in proteins from their primary structures. *Methods Enzymol.* 1991, **203,** 176–201
26 Levitt, M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 1978, **17,** 4277–4284
27 Kyte, J., and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157,** 105–132
28 Stern, P.S. Predicting antigenic sites on proteins. *Trends Biotech.* 1991, **9,** 163–169
29 Krystek, S.R.J., Reichert, L.E.J., and Andersen, T.T. Analysis of computer-generated hydropathy profiles for human glycoprotein and lactogenic hormones. *Endocrinology* 1985, **117,** 1110–1124
30 Mandler, J. Antigen: protein surface residue prediction. *Comput. Applic. Biosci.* 1988, **4,** 493
31 Jameson, B.A., and Wolf, H. Predicting antigenicity from protein primary structure. *Comput. Applic. Biosci.* 1988, **4,** 181–186
32 Menéndez-Arias, L., and Rodriguez, R. A BASIC microcomputer program for prediction of B and T cell epitopes in proteins. *Comput. Applic. Biosci.* 1990, **6,** 101–105
33 Zimmerman, J.M. The characterization of amino acid sequences in proteins by statistical methods. *J. Theoret. Biol.* 1968, **21,** 170–201
34 Chothia, C. The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.* 1976, **105,** 1–14
35 Manavalan, P., and Ponnuswamy, P.K. Hydrophobic character of amino acid residues in globular proteins. *Nature* 1978, **275,** 673–674
36 Von Heijne, G. On the hydrophobic nature of signal sequences. *Eur. J. Biochem.* 1981, **116,** 419–422
37 Sweet, R.M., and Eisenberg, D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.* 1983, **171,** 479–488
38 Hopp, T.P. Protein antigen conformation: folding patterns and predictive algorithms; selection of antigenic and immunogenic peptides. *Ann. Sclavo* 1984, **2,** 47–60
39 Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985, **229,** 834–838
40 Welling, G.W., Weijer, W.J., Van der Zee, R., and

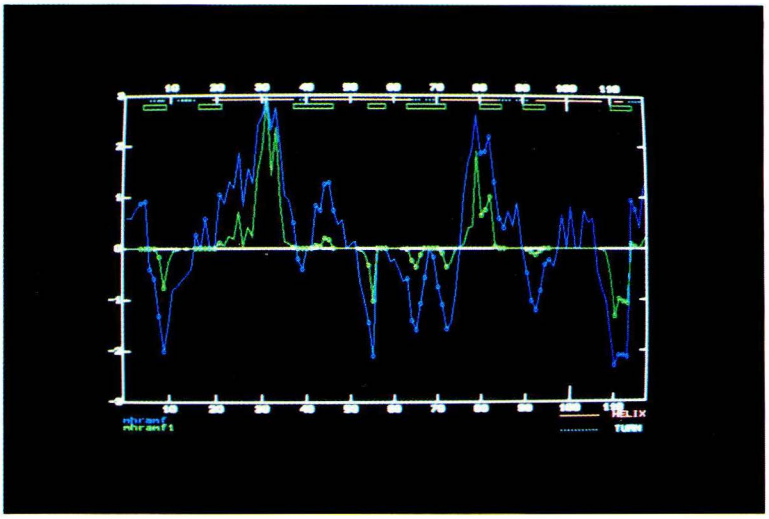
Welling-Wester, S. Prediction of sequential antigenic regions in proteins. *FEBS letter* 1985, **188,** 215–218
41 Engleman, D.M., Steitz, T.A., and Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Ann. Rev. Biophys. Biophys. Chem.* 1986, **15,** 321–353
42 Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A., and DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* 1987, **195,** 659–685
43 Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A.M., and Colonna, G. Flexibility plot of proteins. *Prot. Eng.* 1989, **2,** 497–504

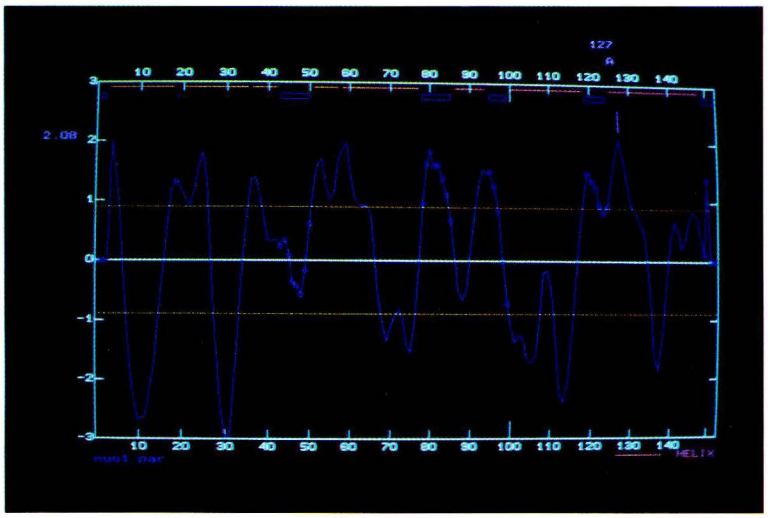
PREDITOP: A program for antigenicity prediction



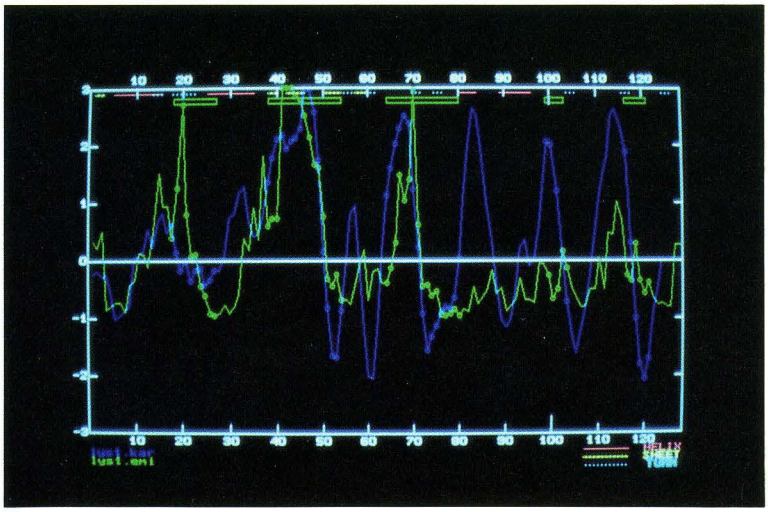
Color Plate 1. Graphical representation of amphiphilic helix prediction based on hydrophobic moment calculations of the myohemerythrin protein. Curves named MHR90.AMF and MHR104.AMF represent the calculation with a step angle of 90° and 104°, respectively. MHRAMF represents the addition of the three files MHR90, MHR97 (not shown), MHR104. We remark that the MHRAMF curve diminishes a wrong prediction by decreasing the height of the first peak of the curve which does not correspond to a helix (red line) but does not eliminate it. All other helices are well predicted.

Color Plate 2. Superimposition of the MHRAMF curve (in blue as in the Color Plate 1) and the MHRAMF1 (in green) curve which corresponds to the multiplication of the three helix result files (as in Color Plate 1). The major difference between the two curves is that the multiplication gives no wrong prediction but diminishes the number of total peaks, while the addition gives more peaks but with wrong predictions. The multiplication procedure gives peaks with smaller size than the addition procedure due to the weak number obtained by the multiplication of values that are below the unity.

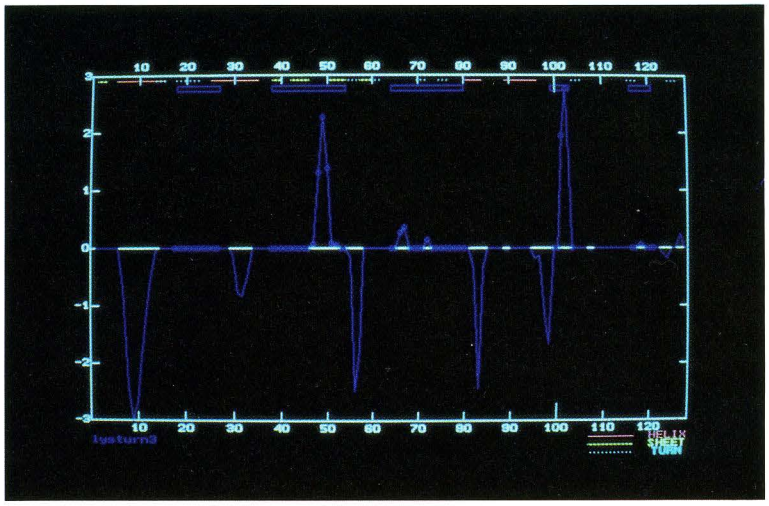




Color Plate 3. Hydrophilicity profile of myoglobin constructed with the scale of Parker *et al.*[11] This graph uses a scale normalized between +3 and −3. The two orange lines on each side of the mean correspond to ±0.7 × standard deviation. Such an interval includes 50% of the amino acids of the protein. Blue rectangles at the top of the curves correspond to the known protein epitopes; the circles drawn on the curves correspond to the same residues. The secondary structure pattern, if known, is shown above the rectangles. A red line corresponds to a helix (the only pattern on this curve). In the program a yellow dashed line corresponds to a sheet and a cyan dotted line to turns. The purple number and letter above the graph give the position of the cursor on the protein and the name of the pointed amino acid; the propensity value of this amino acid is given on the left of the graph.



Color Plate 4. Superimposition of two curves representing an accessibility (LYS1.EMI determined by Emini *et al.*[20]) and a flexibility (LYS1.KAR determined by Karplus and Schulz[19]) prediction of the lysozyme protein. This graph reveals that the two higher superimposed peaks fall in antigenic regions. This fact highlights the use of the superimposition.



Color Plate 5. Graphical representation of the new method for antigenicity prediction. The curve represents turn prediction of the lysozyme protein. The few visualized peaks are obtained by multiplying four files predicting turns in proteins (Pellequer *et al.*[24]). This curve represents a prediction with no errors in peak assignments.