A COMPUTER PROGRAM FOR PREDICTING PROTEIN ANTIGENIC DETERMINANTS

THOMAS P. HOPP* and KENNETH R. WOODS
The Lindsley F. Kimball Research Institute, The New York Blood Center, New York, NY 10021, U.S.A.; and
The Department of Biochemistry, Cornell University Medical College, New York, NY 10021, U.S.A.

(Received 20 October 1982; accepted 16 November 1982)

Abstract—A computerized method for predicting the locations of protein antigenic determinants is presented, which requires only the amino acid sequence of a protein, and no other information. This procedure has been used to predict the major antigenic determinant of the hepatitis B surface antigen, as well as antigenic sites on a series of test proteins of known antigenic structure (Hopp & Woods (1981) Proc. natt. Acad. Sci. U.S.A. 78, 3824-3828.) The method is suitable for use in smaller personal computers, and it is written in the BASIC language, in order to make it available to investigators with limited computer experience and/or resources. A means of locating multiple antigenic sites on a homologous series of proteins is demonstrated using the influenza hemagglutinin as an example.

INTRODUCTION

In an earlier report from this laboratory (Hopp & Woods, 1981) we described a method for deducing antigenic portions of proteins using no other information besides the amino acid sequence. The method is based on calculated estimates of local hydrophilicity along the polypeptide chain, and the assumption that hydrophilic regions are predominantly surface-oriented and therefore potentially antigenic. This approach has been applied to the sequence of the hepatitis B surface antigen (HBsAg) in order to identify and synthesize a short peptide bearing a major antigenic determinant of that virus (Hopp, 1981). This peptide was found to bind specifically to natural antibody to HBsAg, and was also shown to be capable of inducing anti-HBsAg antibody responses when coupled to carriers and used to immunize mice (Prince et al., 1982). Because the prediction method is highly successful in locating antigenic sites in the sequences of the twelve test antigens from which it was developed (Hopp & Woods, 1981) it is probable that antigenic sites on many other proteins can be correctly identified by use of this procedure. This paper describes a computerized version of the prediction method, and demonstrates its use on several viral surface antigens.

METHOD

The computer program for the prediction procedure is listed in Fig. 1. It is written in Hewlett-Packard BASIC and is suitable for use in the HP-85 computer. Only minor changes are required to convert the program into Apple BASIC or other similar languages. A printout of hydrophilicity values for myoglobin is shown in Fig. 2.

The following points are important to the proper execution of the program.

1. Amino acids are entered as the one letter codes defined by the IUPAC-IUB Commission on Biochemical Nomenclature and found in The Atlas of Protein Sequence and Structure (Dayhoff, 1976).

2. Up to 500 amino acids can be entered; larger proteins must be divided into two or more parts for analysis.

3. No provision is made for entering B or Z codes corresponding to unassigned amide states for the acid residues (Asx or Gx). The difference in hydrophilicity values for the acids versus amides is so great that antigenic determinant predictions are severely affected by changing the amide assignment of a particular residue, and therefore, it is not rec-
Fig. 1. Antigenic determinant prediction program, suitable for use in the Hewlett-Packard HP-85 computer. The program is interactive, with prompts for entry of the necessary information: 'name of protein' (up to 10 characters); 'number of amino acids' (up to 500); and the sequence, entered as single letters, each followed by the 'end of line' key. The standard one letter amino acid codes are: D, aspartic acid; N, asparagine; T, threonine; S, serine; E, glutamic acid; Q, glutamine; P, proline; G, glycine; A, alanine; C, cysteine; V, valine; M, methionine; I, isoleucine; L, leucine; Y, tyrosine; F, phenylalanine; W, tryptophan; K, lysine; H, histidine; R, arginine.

It is recommended that predictions be made for any protein for which a complete amide assignment has not yet been reported.

(4) The six residue hydrophilicity average is listed opposite the first amino acid of the group, for example, in myoglobin the first average value, 0.500, is shown opposite valine 1; it actually represents the average of the hydrophilicity values for the six residues from valine 1 to glutamic acid 6. Similarly, the value of 0.183 at leucine 2 actually represents the average for residues 2 through 7. No values are listed opposite the five C-terminal residues. However, their hydrophilicity values have been averaged along with that of glutamic acid 148, as the last six residue average; this average is listed opposite glutamic acid 148. Therefore, even though there are five fewer averages than the total number of residues, all residues have been included in at least one average value.

(5) Following the hydrophilicity printout, a list of up to 10 peaks is presented. These are ranked according to peak height, with the highest point listed first (in this case, position 58). Peaks 2, 3, etc. represent the second highest peak, third highest peak, and so forth, in order of decreasing peak height. As detailed in the previous paper (Hopp & Woods, 1981) only the three highest peaks are strongly correlated with antigenicity. Smaller peaks should not be considered as potential antigenic sites.
Fig. 2. Printout for sperm whale myoglobin. This output is obtained after entering the amino acid sequence of myoglobin. Amino acids are printed out as the three letter codes in order to facilitate reading of the sequence.

program lists both as peaks, so long as they have at least three residues intervening between them. The N-terminal-most peak is listed first and the C-terminal-most peak is ranked next, even though they should have an equal ranking. Where two identical values occur closer then three residues apart (as in myoglobin positions 77 and 78) only the first position is listed, in order to prevent multiple rankings for what is in essence the same peak.

APPLICATION OF THE METHOD

A suitable approach for applying this method is to synthesize the predicted antigenic determinant by the Merrifield peptide synthesis procedure, and then test the peptide for antigenic and immunogenic activity. It is prudent to synthesize more than just the six residues yielding the highest hydrophilicity average for several reasons: (1) other investigators have
found that additional amino acids flanking the antigenic sequence often enhance the antigenic reactivity of a given sequence, probably by imparting a more native conformation to the sequence in question (Atassi & Saplin, 1968; Crumpton, 1974); (2) the predicted point sometimes lies immediately to one side of the natural antigenic determinant, so that several additional residues are required to ensure a good overlap of the synthetic peptide on the antigenic site (Hopp & Woods, 1981). Therefore, it is appropriate to synthesize peptides of no fewer than twelve residues, including three residues on either side of the six amino acids that yielded the highest average hydrophilicity value.

**Prediction and synthesis of an HBsAg antigenic determinant**

A plot of the computer generated prediction profile for HBsAg is shown in Fig. 3. The surface antigen is an unusually hydrophobic protein, containing a large proportion of apolar and aromatic amino acids, and this is reflected in the hydrophilicity profile which for the most part lies below the zero line. There are, however, a number of more hydrophilic sequences which presumably have some degree of exposure to the aqueous environment on the surface of the virus. The largest of these peaks (peak 1) corresponds to amino acids 141 to 146 of the protein. This region was synthesized in a peptide containing residues 138–149, and found to bind up to 9%, of antibodies directed against HBsAg (Hopp, 1981) and furthermore to elicit anti-HBsAg responses in animals (Prince et al., 1982). More recently Bhatnagar et al., (1981) have shown that a similar peptide (residues 139–147) is capable of inhibiting up to 80%, of the reaction of HBsAg with its antibody, and have identified it as the major, or a determinant of HBsAg. These results suggest that the present prediction method may facilitate the search for the antigenic sites on proteins, and short-cut the laborious procedures of chemical modification or cross-reaction studies on homologous proteins that have been necessary in the past.

**Application to influenza hemagglutinin**

The unreliability of lower peaks in predicting antigenic sites has led to an alternative method for obtaining multiple predictions for a given protein. Hydrophilicity analysis is applied to a homologous series of protein antigens, and the most prominent peaks from the whole group are used to predict antigenic sites on different parts of the molecule. Figure 4 illustrates the use of hydrophilicity scans of the hemagglutinin of influenza virus to deduce likely antigenic sites.

The antigenic structure of the hemagglutinin is known to vary with substitutions of amino acids on its surface. This variability is reflected in the hydrophilicity plots for the five antigenically distinct hemagglutinins shown. The highest peak of hydrophilicity can be seen to reside at a different position in each of the hemagglutinins, and the second and third highest peaks occupy different positions, as well. Because many of the antigenic residues of the hemagglutinin have already been identified (Wiley et al., 1981) it is possible to develop a strategy that can be followed to quickly synthesize peptides corresponding to a number of predicted antigenically active sites. Thus, if the antigenic structure of the hemagglutinin were not known, the following procedure would be carried out.

1. A twelve amino acid peptide, corresponding to a sequence centered on the highest predicted peak would be synthesized for each hemagglutinin, and tested for antigenicity and/or immunogenicity.

2. The second highest peaks, and then the third highest peaks would be synthesized, with one important restriction: peptides that overlap substantially with previously synthesized regions are omitted (regardless of whether the
Hydrophilicity analysis of influenza hemagglutinins (HA1 subunit). The letters A to D refer to the antigenic sites described by Wiley et al. (1981); vertical lines in the bottom panel represent individual antigenic residues. The hemagglutinins are: WSN, A/WSN/33 (Hatt et al., 1981); PR, A/Puerto Rico/8/34 (Winter et al., 1981); JAP, A/Japan/305/1980; MEM, A/Memphis/102/72 (Ward & Dopheide 1980); X31, A/Aichi/2/68 (Ward & Dopheide, 1981). Peaks are numbered according to their height in the hydrophilicity profile.

Using this conservative strategy, and assuming that antigenic activity can be demonstrated on the appropriate peptides, it would be possible to explore a major portion of the antigenic structure of the hemagglutinin with very few incorrect prediction assignments. As outlined in Table 1, a series of peptides would be synthesized and tested in the following order:

1. WSN peptide 1, incorporating two residues of antigenic site D;
2. PR peptide 1, incorporating one additional residue assigned to site D;
3. JAP peptide 1, no known antigenic residues incorporated;
4. MEM peptide 1, incorporating five residues of site A;
5. X31 peptide 1, an unconfirmed prediction (actually coincident with the protease activation site, see below);
6. WSN peptide 2, an unconfirmed prediction;
7. PR peptide 2, not done—covers the same area as WSN peptide 1. For similar reasons, MEM-2, X31-2, WSN-3 among others would not be done;
8. JAP peptide 2, incorporating four residues of site B;
9. JAP peptide 3, incorporating two additional residues of site B.

By this approach, after synthesizing only eight peptides, three of the four major antigenic sites of the molecule (Sites A, B and D) would have been covered, at least partially, and a total of 14 out of the 25 known antigenic residues ($56\%$) would have been incorporated into test peptides.

Although the prediction success rate is lower in this example than in the test system used to develop the method, several factors that improve the outlook should be considered: (1)
delineation of the antigenic sites of the influenza hemagglutinin is by no means complete, leaving open the possibility that in some cases the predictions were locating as yet unrecognized antigenic sites; (2) Muller et al. (1982) have recently synthesized a peptide that substantially overlaps the 'wrong' predicted peptide at site JAP-1, and shown that it can produce anti-hemagglutinin responses upon immunization. This suggests that this region may be previously undetected antigenic site, or may represent an area where an unnatural response may be generated, producing antibodies that bind the hemagglutinin in an altogether new site; (3) the 'unconfirmed prediction' at site X31-1 actually covers the location of the proteolytic cleavage which is necessary to activate the virus for membrane penetration. Despite the lack of known antigenicity at this site, it is suggested that the prediction method may be of value in locating other types of surface oriented activity on protein molecules; (4) in several instances, the 'unconfirmed predictions' were actually found to be located on highly exposed portions of the three-dimensional structure (Wilson et al., 1981) of the hemagglutinin, suggesting that these areas might be capable of inducing anti-hemagglutinin antibody responses despite their lack of correlation with natural antigenic determinants.

DISCUSSION

Prediction of protein antigenic determinants is likely to become important in advancing our understanding of protein immunoochemistry and as an aid in the development of synthetic vaccines, monoclonal antibodies and immunologic reagents. The laboriousness of alternative procedures has been a major impediment to this field in the past. The present method should eliminate much of the preliminary work necessary to an intelligent attack on the immunoochemistry of protein antigens of viruses and other pathogens. This is demonstrated by its high success rate with the model systems from which it was developed, by the prediction and synthesis of the major antigenic determinant of HBsAg, and now by the demonstration of its potential for rapidly identifying multiple antigenic sites on a molecule such as the influenza hemagglutinin.

Careful application of the hydrophilicity program and synthesis strategy described in this paper should enable much more rapid elucidation of protein antigenic structures than has been possible in the past. For example, the antigenic structure of sperm whale myoglobin was worked out by Atassi and co-workers (Atassi, 1975) over a period of 11 years, and required gram quantities of the protein. More recently, several groups have been able to do away with the need for purified antigen by using synthetic peptides corresponding to poly peptides coded for by nucleotide sequences, as immunogens (Walter et al., 1980; Sutcliffe et al., 1980). However, in a study on HBsAg where synthesis targets were chosen based on a prediction method that was not specifically developed for antigenic determinant prediction (Kyte & Doolittle, 1982) 9 out of 13 synthetic peptides failed to produce significant anti-HBsAg responses, and antigenic activity was not detected in the major (a) determinant region (Lerner et al., 1981). The addition of this prediction method to the modern methods of immunoochemistry should enable a further acceleration of the pace of molecular immunology and synthetic vaccine development.

Acknowledgement: We thank Dorothea Gamarra for excellent technical assistance.

REFERENCES

Atassi M. Z. (1975) Antigenic structure of myoglobin, the complete immunoochemical anatomy of a protein and conclusions relating to antigenic structures in proteins, Immunoochemistry 12, 473-488.


*Note added in proof* Results of experiments published recently indicate that this method may also be useful for locating T-cell recognition sites on proteins [Lamb J. R., Eckels D. D., Lake P., Woody J. N. & Green N. (1982) Human T-cell clones recognize chemically synthesized peptides of influenza hemagglutinin. *Nature, Lond.* **300**, 66-69]. In that paper, the major T-cell stimulating peptide (peptide 20) overlaps the proposed synthetic peptide for the X31 strain (Table 1, step 5) by ten amino acid residues.