Molecular Immunology Vol. 20, No. 4, pp. 483–489, 1983
Printed in Great Britain.

0161-5890/83/040483-07$03.00/0

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. natn. Acad. Sci. U.S.A.* **78**, 3824–3828.] The method is suitable for use in smaller personal computers, and 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 Glx). 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-

*Correspondence to be addressed to Thomas P. Hopp, Ph.D., Immunex Corporation, 51 University Street, Seattle, Washington 98101, U.S.A.

```
10 REM        ** HYDRO **
20 0=0
30 SHORT A(500),B(500),C(500),D
   (501),K(25),P(10)
40 DIM P$[70],L$[70],M$[30]
50 L$="ASPASNTHRSERGLUGLNPROGLY
   ALACYSVALMETILELEUTYRPHETRPL
   YSHISARG"
60 M$="DNTSEQPGACVMILYFWKHR"
70 FOR I=1 TO 20
80 READ K(I)
90 DATA 3,.2,-.4,.3,3,.2,0,0,-.
   5,-1,-1.5,-1.3,-1.8,-1.8,-2.
   3,-2.5,-3.4,3,-.5,3
100 NEXT I
110 DISP "NAME OF PROTEIN";
120 INPUT P$
130 DISP "# OF AMINO ACIDS";
140 INPUT N
150 IF N>500 THEN 130
160 FOR I=1 TO N
170 DISP "AMINO ACID ";I;
180 INPUT A$
190 J=POS(M$,A$)
200 IF J=0 THEN 170
210 A(I)=J
220 NEXT I
230 FOR I=1 TO N
240 Y=A(I)
250 B(I)=K(Y)
260 NEXT I
270 PRINT P$
280 PRINT
290 PRINT "HYDROPHILICITY ANALYS
   IS"
300 PRINT
310 PRINT
320 PRINT "       FIRST     AVERAG
   E"
330 PRINT "        AA       H VALU
   E"
340 PRINT
350 C(1)=-3 @ C(2)=-3 @ C(3)=-3
   @ C(N-1)=-3 @ C(N)=-3 @ C(N+
   1)=-3
360 FOR I=1 TO N-5
370 X=B(0+1)+B(0+2)+B(0+3)+B(0+4
   )+B(0+5)+B(0+6)
380 0=0+1
390 Z=X/6
400 C(I+3)=Z
410 W=A(I)
420 A$=L$[3*W-2,3*W]
430 PRINT USING 440 ; I,A$,C(I+3
   )
440 IMAGE 3D,4X,3A,5X,DD.DDD
450 NEXT I
460 FOR I=N-4 TO N
470 W=A(I)
480 A$=L$[3*W-2,3*W]
490 PRINT USING 500 ; I,A$
500 IMAGE 3D,4X,3A
510 NEXT I
520 PRINT
530 PRINT
540 REM                PEAK FINDER
550 FOR I=1 TO N+1
560 D(I)=0
570 NEXT I
580 FOR I=4 TO N-2
590 W=C(I) @ X=C(I-1) @ Y=C(I-2)
    @ Z=C(I-3) @ X1=C(I+1) @ Y1
   =C(I+2) @ Z1=C(I+3)
600 IF W<=X THEN 670
610 IF W<=Y THEN 670
620 IF W<=Z THEN 670
630 IF W<X1 THEN 670
640 IF W<Y1 THEN 670
650 IF W<Z1 THEN 670
660 D(I)=1
670 NEXT I
680 FOR J=1 TO 10
690 P(J)=0
700 Y=-3 @ Z=0
710 FOR I=4 TO N-2
720 IF D(I)=0 THEN 760
730 W=C(I)
740 IF W>Y THEN Z=I
750 IF W>Y THEN Y=W
760 NEXT I
770 P(J)=Z @ D(Z)=0
780 NEXT J
790 PRINT "  PEAKS"
800 PRINT
810 FOR I=1 TO 10
820 IF P(I)=0 THEN 860
830 X=P(I)-3
840 PRINT USING 845 ; I,X
845 IMAGE DD,2X,DDD
850 NEXT I
860 PRINT
870 PRINT
880 PRINT
890 PRINT
900 END
```

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.

ommended 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 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. Where two averages have the same value, the

MYOGLOBIN

HYDROPHILICITY ANALYSIS

	FIRST AA	AVERAGE H VALUE
1	VAL	.500
2	LEU	.183
3	SER	.517
4	GLU	.167
5	GLY	-.583
6	GLU	-.883
7	TRP	-1.467
8	GLN	-1.150
9	LEU	-1.750
10	VAL	-1.533
11	LEU	-.783
12	HIS	-.733
13	VAL	-.150
14	TRP	.017
15	ALA	1.083
16	LYS	.917
17	VAL	.333
18	GLU	.583
19	ALA	0.000
20	ASP	.083
21	VAL	-.383
22	ALA	.367
23	GLY	.150
24	HIS	-.150
25	GLY	-.367
26	GLN	.133
27	ASP	-.200
28	ILE	-1.117
29	LEU	-.317
30	ILE	.933
31	ARG	.250
32	LEU	-.250
33	PHE	.550
34	LYS	.900
35	SER	.100
36	HIS	.550
37	PRO	1.133
38	GLU	.717
39	THR	.717
40	LEU	1.283
41	GLU	1.167
42	LYS	1.167
43	PHE	.583
44	ASP	.700
45	ARG	.700
46	PHE	.133
47	LYS	1.050
48	HIS	.467
49	LEU	1.050
50	LYS	1.133
51	THR	1.133
52	GLU	1.117
53	ALA	.667
54	GLU	1.250
55	MET	1.250
56	LYS	1.167
57	ALA	1.167
58	SER	1.750
59	GLU	1.617
60	ASP	1.117
61	LEU	.367
62	LYS	.600
63	LYS	-.150
64	HIS	-.950
65	GLY	-.933
66	VAL	-1.017
67	THR	-1.067
68	VAL	-1.000
69	LEU	-.833
70	THR	-.833
71	ALA	-1.067
72	LEU	-.483
73	GLY	.317
74	ALA	.817
75	ILE	.900
76	LEU	1.117
77	LYS	1.333
78	LYS	1.333
79	LYS	.750
80	GLY	.750
81	HIS	.450
82	HIS	1.033
83	GLU	1.117
84	ALA	.317
85	GLU	.317
86	LEU	-.150
87	LYS	.200
88	PRO	-.383
89	LEU	-.467
90	ALA	-.233
91	GLN	.350
92	SER	.233
93	HIS	.683
94	ALA	.467
95	THR	.550
96	LYS	.317
97	HIS	.317
98	LYS	.017
99	ILE	-.783
100	PRO	.017
101	ILE	-.400
102	LYS	-.400
103	TYR	-.850
104	LEU	.033
105	GLU	.250
106	PHE	-.550
107	ILE	-.433
108	SER	-.217
109	GLU	-.517
110	ALA	-1.317
111	ILE	-1.317
112	ILE	-.967
113	HIS	-.167
114	VAL	-.167
115	LEU	.083
116	HIS	.383
117	SER	.500
118	ARG	.033
119	HIS	-.467
120	PRO	-.467
121	GLY	.033
122	ASN	-.050
123	PHE	-.050
124	GLY	.367
125	ALA	.283
126	ASP	.150
127	ALA	-.317
128	GLN	.267
129	GLY	.150
130	ALA	-.150
131	MET	.433
132	ASN	.350
133	LYS	-.100
134	ALA	-.100
135	LEU	.483
136	GLU	1.283
137	LEU	.483
138	PHE	.700
139	ARG	1.033
140	LYS	1.033
141	ASP	.150
142	ILE	.150
143	ALA	.950
144	ALA	.733
145	LYS	.817
146	TYR	-.067
147	LYS	.350
148	GLU	-.150
149	LEU	
150	GLY	
151	TYR	
152	GLN	
153	GLY	

PEAKS

1	58
2	77
3	40
4	136
5	54
6	50
7	83
8	15
9	93
10	3

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

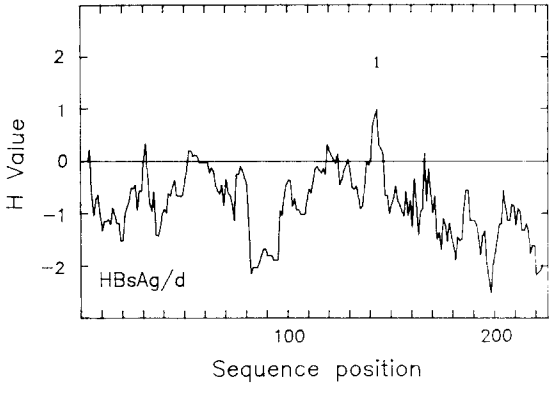

Fig. 3. Hydrophilicity profile for hepatitis B surface antigen. Average hydrophilicity values are plotted against sequence position in the peptide chain. Each average is plotted in the middle of the six positions from which it was derived (i.e. the value for positions 1–6 is plotted at position 3.5). The number, 1, indicates the maximum hydrophilicity value, located at position 143.5, corresponding to amino acids 141–146.

HBsAg (Hopp, 1981) and furthermore to elicit anti-HBsAg responses in animals (Prince *et al.*, 1982). More recently Bhatnagar *et al.*, (1982) 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 hemagglutinins

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

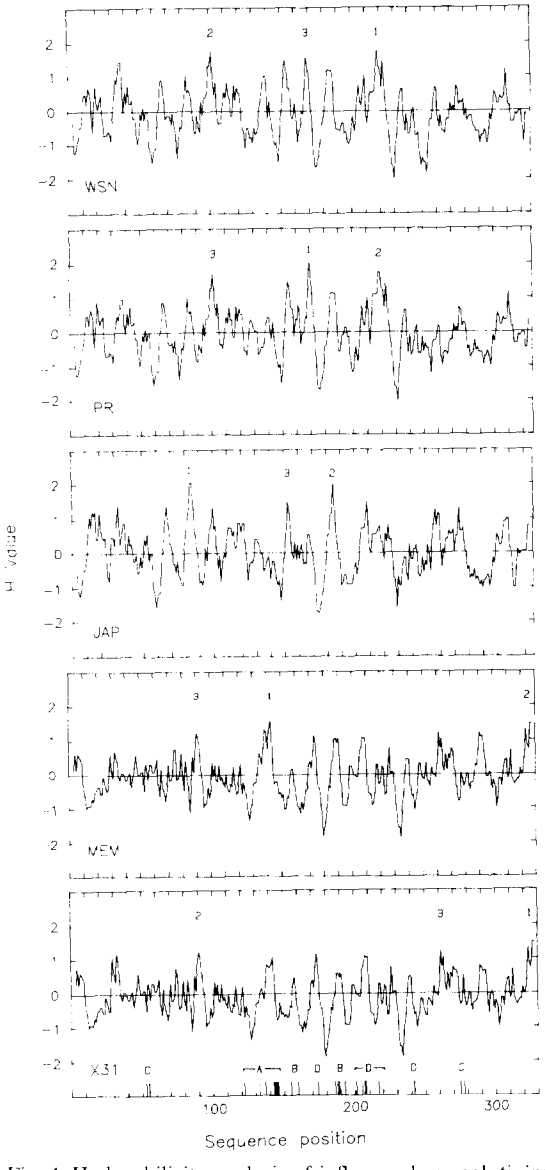

Fig. 4. Hydrophilicity analysis of influenza hemagglutinins (HA1 subunit). The letters A to D refer to the antigenic sites described by Wiley *et al.*, (1981); vertical lines in the bottom panel represent individual antigenic residues. The hemagglutinins are: WSN, A/WSN/33 (Hiti *et al.*, 1981); PR, A/PR/8/34 (Winter *et al.*, 1981); JAP, A/Japan/305/57 (Gething *et al.*, 1980); MEM, A/Memphis/102/72 (Ward & Dopheide 1980); X31, A/Aichi/2/68i (Ward & Dopheide, 1981). Peaks are numbered according to their height in the hydrophilicity profile.

Table 1. Strategy for identifying antigenic sites on influenza hemagglutinin

Step	Hydrophilicity peak	Residues to be synthesized	Known antigenic residues incorporated
1	WSN-1	217–228	217, 220
2	PR-1	167–178	174
3	JAP-1	86–97	none[b]
4	MEM-1	137–148	137, 143–146
5	X31-1	321–332	none[c]
6	WSN-2	103–114	none
7	JAP-2	184–195	186, 188, 189, 193
8	JAP-3	152–163	155, 160

[a]All residues are numbered according to the numbering scheme for A/Aichi/2/68 (X31) of Ward & Dopheide (1981).
[b]Synthetic peptides from this region have been shown to be immunogenic despite the absence of known naturally occurring antigenic residues.
[c]This peptide would overlap the site of proteolytic activation of the hemagglutinin.

previous peptide had proved to be antigenically active or inactive).

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 a 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 immunochemistry 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 immunochemistry 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 polypeptides 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 immunochemistry 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 immunochemical anatomy of a protein and conclusions relating to antigenic structures in proteins. *Immunochemistry* **12**, 423–438.

Atassi M. Z. & Saplin, B. J. (1968) Immunochemistry of sperm whale myoglobin. I. The specific interaction of some tryptic peptides and of peptides containing all the reactive regions of the antigen. *Biochemistry* **7**, 688–698.

Bhatnagar P. K., Papas E., Blum H. E., Milich D. R., Nitecki D., Karels M. J. & Vyas G. N. (1982) Immune response to synthetic peptide analogues of hepatitis B surface antigen specific for the *a* determinant *Proc. natn. Acad. Sci. U.S.A.* **79**, 4400–4404.

Crumpton M. J. (1974) Protein antigens: the molecular bases of antigenicity and immunogenicity. In *The Antigens* (Edited by Sela M.) Vol. 2, pp. 1–78. Academic Press, New York.

Dayhoff M. O. (1976) *Atlas of Protein Sequence and Structure*, Vol. 5 Suppl. 2, p. 22. The National Biomedical Research Foundation, Silver Spring, Maryland.

Gething M. J., Bye J., Skehel J. & Waterfield M. (1980) Cloning and DNA sequence of double-stranded copies of haemagglutinin genes from H2 and H3 strains elucidates antigenic shift and drift in human influenza virus. *Nature, Lond.* **287**, 301–306.

Hiti A. L., Davis A. R. & Nayak D. P. (1981) Complete sequence analysis shows that the hemagglutinins of the HO and H2 subtypes of human influenza virus are closely related. *Virology* **111**, 113–124.

Hopp, T. P. (1981) A synthetic peptide with hepatitis B surface antigen reactivity. *Molec. Immun.* **18**, 869–872.

Hopp T. P. & Woods K. R. (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc. natn. Acad. Sci. U.S.A.* **78,** 3824–3828.

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

Lerner R. A., Green N., Alexander H., Liu F. T., Sutcliffe J. G. & Shinnick T. M. (1981) Chemically synthesized peptides predicted from the nucleotide sequence of the hepatitis B virus genome elicit antibodies reactive with the native envelope protein of Dane particles. *Proc. natn. Acad. Sci. U.S.A.* **78,** 3403–3407.

Muller G. M., Shapira M. & Arnon R. (1982) Anti-influenza response achieved by immunization with a synthetic conjugate. *Proc. natn. Acad. Sci. U.S.A.* **79,** 569–573.

Prince A. M., Ikram H. & Hopp T. P. (1982) Hepatitis B virus vaccine: identification of HBsAg/a and HBsAg/d but not HBsAg/y subtype antigenic determinants on a synthetic immunogenic peptide. *Proc. natn. Acad. Sci. U.S.A.* **79,** 579–582.

Sutcliffe J. G., Shinnick T. M., Green N., Liu F. T., Niman H. L. & Lerner R. A. (1980) Chemical synthesis of a polypeptide predicted from nucleotide sequence allows detection of a new retroviral gene product. *Nature, Lond.* **287,** 801–805.

Walter G., Scheidtmann K. H., Carbone A., Laudano A. P. & Doolittle R. F. (1980) Antibodies specific for carboxy- and amino-terminal regions of simian virus 40 large tumor antigen. *Proc. natn. Acad. Sci. U.S.A.* **77,** 5197–5200.

Ward C. W. & Dopheide T. A. (1980) Completion of the amino acid sequence of a Hong Kong influenza hemagglutinin heavy chain: sequence of cyanogen bromide fragment CN1. *Virology* **103,** 37–53.

Ward C. W. & Dopheide T. A. (1981) Amino acid sequence and oligosaccharide distribution of the haemagglutinin from an early Hong Kong influenza virus variant A/Aichi/2/68(X-31). *Biochem. J.* **193,** 953–962.

Wiley D. C., Wilson I. A. & Skehel J. J. (1981) Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature, Lond.* **289,** 373–378.

Wilson I. A., Skehel J. J. & Wiley D. C. (1981) Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3A resolution. *Nature, Lond.* **289,** 366–373.

Winter G., Fields S. & Brownlee G. G. (1981) Nucleotide sequence of the haemagglutinin gene of a human influenza virus H1 subtype. *Nature, Lond.* **292,** 72–75.

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.