Journal of Immunological Methods, 88 (1986) 1-18. http://thomas-hopp.com/THscience.html

Review Article

Protein Surface Analysis

Methods for identifying antigenic determinants and other interaction sites

Thomas P. Hopp

Protein Chemistry Department, Immunex Corporation,
51 University Street, Seattle, WA 98101, U.S.A.

ABSTRACT: A variety of methods are currently employed in attempts to identify antigenic determinants and other features of proteins. Most of these involve calculations based on scales of values for the 20 amino acids that reflect their likelihood of occurrence at the surface of proteins or as part of secondary structures such as helices or β-bends. The most successful of these procedures all use scales related to the water solubility of the individual amino acids. In particular, the highest success rates are obtained using hydrophilicity scales that emphasize the charged and polar amino acids, but are not overly selective for either positive or negative charges. Such a method can correctly pinpoint major antigenic sites on the proteins of most well characterized infectious organisms. The hydrophilicity profiles also contain information concerning other types of protein interaction sites and membrane spanning segments.

Introduction

The advent of gene cloning procedures has led to a peculiar phenomenon of the times. Primary sequence data are accumulating at a rate that, for once, exceeds our ability to carry out thorough chemical characterizations on the many proteins of interest that are now becoming known. X-ray crystallographic structures and structure-function dissection work lag ever farther behind because these time-honored procedures have not yet had, and are unlikely to soon see, major accelerations in their rates of progress. For these and other reasons many investigators have turned to computer-aided analytical methods that allow some insight into the 3-dimensional structures coded in the primary sequences they are concerned with. These methods offer the advantage of quickly deriving useful information about protein antigenic sites and other sites of interaction, often before any significant quantity of purified protein is available.

This review will concentrate on a comparison of the many methods of structure analysis now in use, and will consider the type of information generated by the analyses. Two major themes will be developed. First, most of the methods yield similar information, with newer methods for the most part being redundant on older ones. Second, much more information is available in the analyses than is generally recognized, and in particular, the hydrophilicity/hydrophobicity plotting methods offer a simple means to arrive at a wealth of insight concerning protein surfaces, interaction sites and folding patterns. Emphasis will be placed on simple methods, which can be accomplished quickly using microcomputers, because more complicated procedures (for example energy minimization) require longer running times and much more expensive computer facilities. Additionally, it will be pointed out that standardization of methods would be helpful in promoting better communication of results between investigators.

Concepts

What is an antigenic site? This subject has been the source of considerable controversy over the years, so some operational definitions are in order. Also referred to as antigenic determinants or epitopes, antigenic sites are quite simply the portions of the surface of a protein that are in contact with the antibodies specific for that protein. Because work with hybridomas (reviewed by Benjamin et al., 1984) and synthetic peptides (Green et al., 1982) has recently proven that virtually any portion of a protein, inside or out, is capable of stimulating the production of specific antibodies under one condition or another, it is important to emphasize that the discussion in this review will be concerned only with *major* antigenic sites. These are locations on the surface of a protein, which bind a larger proportion of the antibodies produced in a normal immune response against a native protein antigen than do other surface areas. These major antigenic determinants are thought of as *immunodominant,* in the sense that antibody-producing B-cell clones specific for them are stimulated to proliferate more frequently than clones reacting to other sites, resulting in a greater proportion of antibody specific for these sites. Typical globular proteins usually possess 3–6 major antigenic sites. These may be *continuous* (comprising a single segment of peptide chain) or *assembled* (comprising two or more chain segments brought together in the tertiary structure of the protein) (Berzofsky, 1985).

One feature of major antigenic sites that was recognized some time ago is that they usually contain a preponderance of charged and polar amino acids (Sela and Mozes, 1966; Atassi, 1975). This does not deny that hydrophobic interactions may add to the stability of the binding between antibodies and antigens, in fact, hydrophobic effects are well established in many antibody binding site interactions. However, it is likely that the highly directional nature of charge-charge and other polar interactions (e.g., hydrogen bonds) contributes the critical element of *geometric specificity* to the binding sites. Hence the polar residues are an absolute requirement in protein interactions, whereas nonpolar groups are helpful but not always necessary.

Current usage

A survey of recent literature shows that many laboratories are using some form of prediction method to help locate antigenic sites on proteins. Most investigators indicate that they are guided by predictions of hydrophilic sites as likely regions of high antigenicity, and many include considerations of secondary structures (helices, strands and β-bends) in their analyses. The hydrophilicity (or hydrophobicity) methods most often cited include those of Hopp and Woods (1981), Rose and Roy (1980) and Kyte and Doolittle (1982). Secondary structures are most often predicted by the methods of Chou and Fasman (1978) or Garnier et al. (1978). Another source of useful information is found in the amino acid substitutions that occur in protein antigens between species or strains of an organism because most substitutions occur at the surfaces of proteins. Sites of insertion/deletion mutations are also useful, although they have been neglected for the most part. In the sections that follow, each of the methods mentioned above will be considered, and an attempt will be made to discuss the value of each approach, based on experience in our own laboratories, as well as the results reported in the recent literature.

Secondary structure and hydrophilicity methods

It is appropriate to consider secondary structure prediction methods simultaneously with hydrophilicity methods, because protein secondary structures (α-helices, β-strands and β-bends) all depend on the hydrophilic or hydrophobic nature of local protein chain segments, and are therefore inextricably related to the hydrophilicity methods (Rose and Roy, 1980; Rose et al., 1985b). At this point, it is worthwhile to note that the choice of scale orientation is optional, and has no effect on the predictive capabilities of any method. There has been a tendency to place the hydrophobic end at the top. This probably reflects the fact that the early

work in this field (Nozaki and Tanford, 1971; Chothia, 1976) was concerned with the important influence of amino acid hydrophobicity on protein packing. Later authors have oriented their scales so that the quality most important to them is at the top. This results in a mixed group of scales, and some potential for confusion, especially for those who are unacquainted with the nuances of terms such as 'hydropathy', 'hydrophilicity' or 'segmental mobility'. Therefore it seems reasonable to argue for standardization of scale orientation, with the hydrophilic end at the top, as has been done in Fig. 1 and Table I. This should decrease the level of confusion in reporting and discussing the results of the analyses, and help to underscore the similarity of the results. Furthermore, for researchers less experienced in using these methods, this orientation makes the relationship to structure more apparent: when viewing a hydrophilicity profile, the peaks represent exposed segments of peptide chain (up = surface), while the valleys are packed in the core of the protein (down = buried). This allows an easier intuitive grasp of the information conveyed by the plots.

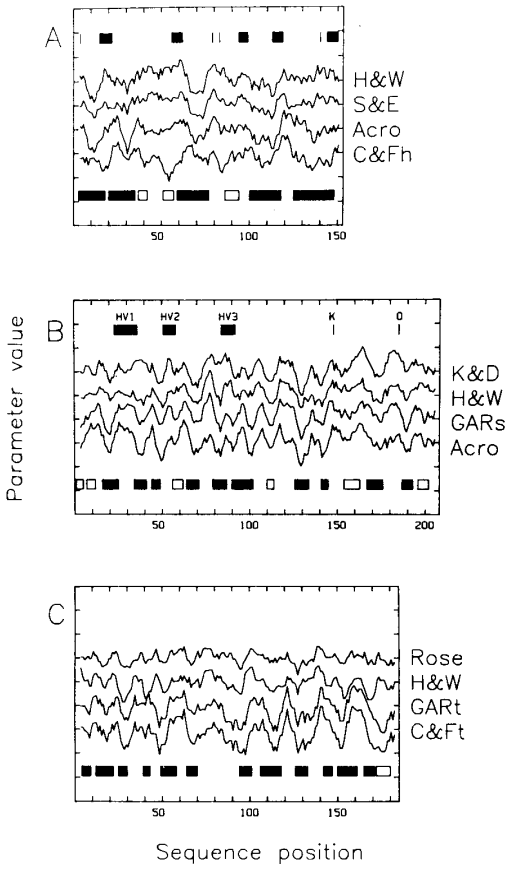

Fig. 1. Comparison of prediction profiles made by a variety of methods. All profiles represent hexapeptide averages. The scale has been compressed to allow multiple profiles on each panel and each profile is plotted 2 hydrophilicity units below the previous one. In order to facilitate comparison several of the scales have been inverted so that all scales are hydrophilic at the top and hydrophobic at the bottom. Horizontal bars represent secondary structure elements and antigenic sites. A: Sperm whale myoglobin. Bars above profiles represent known major antigenic sites, vertical slashes represent single antigenic residues. Bars below the profiles are helices. The larger helices that form the hydrophobic core of the molecule are represented by solid bars; open bars represent smaller, more exposed helices. The profiles are, from top to bottom: H&W, hydrophilicity according to Hopp and Woods (1981); S&E, hydrophobicity (Sweet and Eisenberg, 1983) inverted; Acro, acrophilicity (Hopp, 1984a); C&Fh, helix parameters (Chou and Fasman, 1978) inverted. B: immunoglobulin light chain NEW. Above the profiles, bars represent the hypervariable, antigen-binding segments, HV1, 2 and 3; vertical slashes are the allotypic antigenic markers: K, Kern; O, Oz. Below the profile, solid bars represent internal β-strands; open bars represent the edge strands that are exposed at the ends of the multiple β-sheets of this molecule. The profiles are: K&D, hydropathy (Kyte and Doolittle, 1982) inverted; H&W, hydrophilicity; GARs, β-strand parameters (Garnier et al., 1978) inverted; Acro, acrophilicity. C: proteinase B of *Streptomyces griseus*. Solid bars represent β-strands; open bar, the C-terminal helix. The profiles are: Rose, fractional accessibility (Rose et al., 1985a) inverted; H&W, hydrophilicity; GARt. turn parameters (Garnier et al., 1978); C&Ft, β-turn parameters (Chou and Fasman, 1978).

Table I compares the amino acid scales discussed in this review. The amino acids have been arranged across the table in order of their values, from the most hydrophilic to the most hydrophobic, with the hydrophilic end at the left. This involved inversion of some scales, while others retained their original orientation. The scales of other investigators were normalized to a range from 3.0 to –3.0 in order to facilitate comparisons with the hydrophilicity and acrophilicity scales. The charged amino acids are printed in bold type in order to emphasize the distribution of this important group on each of the scales. The scales themselves have been ranked, roughly in order of their usefulness in predicting antigenic sites, with the most successful at the top. The success rates were determined by a computerized procedure that automatically checked the validity of each method's predictions for a data set of 12 known protein antigens (Hopp and Woods, 1981).

Table I. Comparison Of Hydrophilicity / Hydrophobicity And Other Scales

The scales are arranged with their most hydrophilic amino acids at left. The second and third columns compare the scales' ability to predict antigenic sites among the top three peaks of hydrophilicity profiles. Scales are abbreviated as follows: H&W, Hopp and Woods (1981); HYDRO3, HYDRO4, see notes below; Welling, Welling et al. (1985); Janin, Janin (1979) inverted; Rose, Rose et al. (1985a) inverted; C&Fh, Chou and Fasman (1978) helix parameters, inverted; v. Heijne, Von Heijne (1981) inverted; S&E, Sweet and Eisenberg (1983) inverted; K&D, Kyte and Doolittle (1982) inverted; Chothia, Chothia (1976) inverted; K&S, Karplus and Schulz (1985); Acro, acrophilicity (Hopp, 1984a); Random, amino acid values selected at random; GARs, Garnier et al. (1978) strand parameters, inverted; C&Ft, Chou and Fasman (1978) turn parameters; −(H&W), Hopp and Woods (1981) inverted. Amino acids are identified by the single letter code. Charged residues are printed in bold type for emphasis.

Scale	% of predictions correct [a]		Ranking of amino acids on scales (value) [b]								
	Top peak	Top 3 peaks	1	2	3	4	5	6	7	8	9
H&W	100	75	**R 3.0**	**D 3.0**	**E 3.0**	**K 3.0**	S 0.3	N 0.2	Q 0.2	G 0.0	P 0.0
(HYDRO3) [c]	(100)	(84)	**R 3.0**	**D 3.0**	**E 3.0**	**K 3.0**	S 0.3	N 0.2	Q 0.2	G 0.0	P 0.0
(HYDRO4) [d]	(100)	(88)	**R 3.0**	**D 3.0**	**E 3.0**	**K 3.0**	S 0.3	N 0.2	Q 0.2	G 0.0	P 0.0
Welling	89	76	H 3.0	**K 2.1**	A 1.3	L 1.0	**D 0.9**	**R 0.8**	Y 0.4	Q 0.2	V 0.2
Janin	80	68	**K 3.0**	**R 2.1**	**E 0.6**	Q 0.6	**D 0.3**	N 0.1	Y -0.1	P -0.3	T -0.6
Rose	75	68	**K 3.0**	D **1.5**	**E 1.5**	Q 1.5	N 1.3	**R 1.1**	P 1.1	S 0.8	T 0.2
C&Fh	75	68	G 3.0	P 3.0	N 2.4	C 2.2	Y 2.2	S 1.7	T 1.3	**R 0.4**	H 0.3
v. Heijne	75	64	**R 3.0**	**D 1.4**	**E 0.8**	**K 0.1**	H -0.2	P -0.3	N -0.4	Q -0.6	S -1.0
S&E	80	56	**R 3.0**	**K 1.4**	**D 0.5**	Q 0.4	N 0.3	**E 0.3**	H -0.3	S -0.6	T -0.8
K&D	71	65	**R 3.0**	**K 2.6**	N 2.3	**D 2.3**	**E 2.3**	Q 2.3	H 2.1	P 1.1	Y 0.9
Chothia	71	60	**R 3.0**	**K 2.8**	Q 2.4	N 1.9	D **1.6**	Y 1.6	H 1.4	**E 1.3**	P 1.3
K&S	60	70	S 3.0 [e]	Q 2.9	G 2.3	N 1.7	**E 1.2**	**K 1.1**	T 0.6	P 0.2	A -0.1
Acro.	63	59	G 3.0	P 2.6	N 2.3	**D 2.1**	S 1.8	**K 1.4**	**E 0.5**	**R 0.3**	T -0.1
Random [f]	56	57	S 1.5	V 1.3	G 0.9	**E 0.8**	I 0.4	**D 0.4**	W 0.4	H 0.2	A -0.1
GARs	57	55	**E 3.0**	**D 2.7**	G 2.6	N 2.5	**K 2.1**	H 1.7	A 1.6	P 1.4	S 1.3
C&Ft	50	43	N 3.0	G 3.0	P 2.8	**D 2.4**	S 2.3	C 1.0	Y 0.7	**K 0.0**	Q -0.2
−(H&W) [g]	17	25	W 3.4	F 2.5	Y 2.3	L 1.8	I 1.8	V 1.5	M 1.3	C 1.0	H 0.5

Table I continued

Ranking of amino acids on scales (value)

10	11	12	13	14	15	16	17	18	19	20
T -0.4	A -0.5	H -0.5	C -1.0	M -1.3	V -1.5	I -1.8	L -1.8	Y -2.3	F -2.5	W -3.4
T -0.4	A -0.5	H -0.5	C -1.0	M -1.3	V -1.5	I -1.8	L -1.8	Y -2.3	F -2.5	W -3.4
T -0.4	A -0.5	H -0.5	C -1.0	M -1.3	V -1.5	I -1.8	L -1.8	Y -2.3	F -2.5	W -3.4
S 0.1	P -0.1	T -0.1	N -0.3	**E -0.3**	C -0.7	W -0.7	F -0.9	G -1.3	I -2.2	M -3.0
H -0.8	S -0.8	A -1.7	G -1.7	W -1.7	M -1.9	L -2.1	F -2.1	V -2.3	I -2.6	C -3.0
G -0.1	A -0.4	Y -0.7	H -1.0	L -2.1	M -2.1	W -2.1	V -2.3	I -2.5	F -2.5	C -3.0
D 0.2	V -0.1	I -0.3	W -0.3	Q -0.4	F -0.6	**K -0.8**	L -1.1	A -2.4	M -2.6	**E -3.0**
T -1.2	Y -1.2	G -1.6	A -2.0	C -2.2	W -2.4	V -2.4	I -2.6	L -2.6	M -2.7	F -3.0
P -1.1	C -1.3	Y -1.3	G -1.6	A -1.8	M -1.9	W -2.1	L -2.5	V -2.5	F -2.7	I -3.0
W 0.6	S 0.5	T 0.5	G 0.3	A -1.2	M -1.3	C -1.7	F -1.9	L -2.5	V -2.8	I -3.0
S 0.9	T 0.8	W 0.4	G -0.6	A -0.8	M -1.0	L -1.5	C -2.0	F -2.0	V -2.4	I -3.0
R -0.2	**D -0.3**	I -1.1	H -1.6	V -1.6	L -2.0	C -2.1	Y -2.1	M -2.5	F -2.9	W -3.0
Q -0.2	H -0.4	A -0.5	V -1.7	M -1.8	Y -2.0	I -2.5	L -2.5	C -2.6	F -2.7	W -3.0
Q -0.1	M -0.2	**K -0.2**	F -0.4	Y -0.4	C -0.6	P -0.8	L -1.0	**R -1.3**	N -1.4	T -1.5
W 1.0	**R 0.3**	Q -0.2	T -0.2	L -0.7	M -0.7	F -0.9	Y -1.6	C -1.8	I -2.9	V -3.0
T 0.3	W -0.3	**R -0.4**	H -0.4	**E -1.5**	A -2.0	L -2.3	M -2.3	F -2.3	V -2.8	I -3.0
A 0.5	T 0.4	P 0.0	G 0.0	Q -0.2	N -0.2	S -0.3	**K -3.0**	**E -3.0**	**D -3.0**	**R -3.0**

[a] Determined as in Hopp and Woods (1981).

[b] All scales (except the last) have been matched so that the hydrophilic end is to the left. This required inverting some scales by multiplying by −1.0. To facilitate comparison to the hydrophilicity scale, all others were normalized to range between +3.0 and −3.0 according to the equation: new value = (old value × 6) / (max − min) − 3.0 where max and min refer to the maximum and minimum amino acid values on the old scale.

[c] Numbers in parentheses indicate improvements obtained when the HYDRO3 subroutine (Fig. 3A) is added. The amino acid values are the same as for H&W.

[d] As in [c] above, but with the HYDRO4 subroutines (Fig. 3A and B) added.

[e] Three separate scales are required in this procedure. The scale shown here is the BNORMO scale.

[f] Values between 3.0 and −3.4 were randomly assigned to the amino acids and prediction success rate determined. The scale values and success rate percentages represent the averaged results of 8 repetitions of this procedure.

[g] Multiplied by −1.0 so that the hydrophobic end is at the top of the scale (left).

The percentage of correct or wrong predictions at the highest peak for each of the 12 proteins, and also for the top three peaks for each protein, are shown in Table I. Although the small size of the data set makes this ranking of the scales somewhat tentative, it represents the only available means for such a comparison until more antigenic structures become known in sufficient detail to be added to the 12 proteins that are now known. The table also includes an entry for the success rate obtained when numbers were assigned randomly to the amino acids. Scales ranked above this level are considered to be selective for antigenic sites, while scales ranked lower are selective for non-antigenic sites. Several scales have been omitted from consideration because they do not offer values for the complete set of 20 amino acids (Rose and Roy, 1980; Wolfenden et al., 1981).

In Fig. 1, the secondary structures and antigenic sites of three proteins are depicted along with multiple data plots generated using a variety of hydrophilicity, hydrophobicity and secondary structure prediction methods. All of the profiles show a significant relatedness, with most peaks and valleys occurring in the same places. The deepest valleys invariably occur at or near the centers of the largest helices, or in β-strands, which constitute the tightly packed hydrophobic cores of these molecules. For sperm whale myoglobin (Fig. 1A) three hydrophilicity profiles are compared to each other and to the helix prediction profile of Chou and Fasman (1978). A comparison of the plots suggests that the helix prediction profile has some similarity to the hydrophilicity profiles although not as much as strand- and turn-predicting profiles (see below). That is, to some extent, helix predictions are made based on the hydrophobicity of a given chain segment. This is appropriate, because the majority of large helices must have a substantial hydrophobic face to allow stable packing of the helix against the core of the molecule. However, the superior helix prediction ability of the helix parameter scales probably rests on their capacity to identify the shorter helices. These smaller helices (indicated by white bars) are often much more surface exposed than the longer helices, so that they are missed by the hydrophilicity methods; that is, they do not appear as hydrophobic valleys on the three hydrophilicity profiles in Fig. 1A. However, two out of three of these helices are correctly identified by valleys in the Chou and Fasman plot. At the same time, it seems appropriate that the hydrophilicity profiles indicate that these short helices are highly exposed segments, and therefore are likely to contain antigenic sites. Indeed, two of the antigenic sites of myoglobin are located at the ends of short helices (segments 56–62, and 94–99).

Fig. 1B shows that a strong relationship also exists between hydrophilicity plots and β-strand identification profiles. In fact, hydrophilicity and β-strand profiles are always more strongly correlated than hydrophilicity and helix profiles, regardless of the helical or stranded nature of the protein in question. Furthermore, the deepest valleys in hydrophilicity plots always correlate with the central (most buried) strands of β-pleated sheets while plots made by the β-strand predictive methods do not always show this useful effect. In Fig. 1B, an immunoglobulin light chain is used to compare three hydrophilicity profiles to the β-strand prediction profile of Garnier et al. (1978). It has been pointed out recently (Garratt et al., 1985) that β-strand predictive methods work better for buried strands than surface strands. This can be seen in Fig. 1B, where all methods show deep valleys for internal strands (dark bars below profiles), but a rather shallow valley for the strand at positions 154–163, which is a highly exposed, edge strand of a β-pleated sheet. Interestingly this result can be viewed two ways: as a failure to locate a β-strand, or as a success in indicating the more surface-oriented nature of this segment of the polypeptide chain. Most of the profiles in Fig. 1B correctly associate the known interaction sites of this molecule with peaks. The hypervariable regions HV1–3 bind to antigens, while the allotypic markers Kern and Oz are themselves bound by antibodies. All are found on outward projecting loops of the molecule, and the profiles project upward in agreement with this.

Finally, Fig. 1C compares several hydrophilicity profiles to the turn and β-bend predicting profiles of Garnier et al. (1978) and Chou and Fasman (1978), using proteinase B as an example. These profiles are also similar, but this is not surprising because the turn predicting scales of these authors are very nearly the inverses of their β-strand predicting scales. It can therefore be concluded that the secondary structure methods all incorporate a substantial influence of hydrophilicity and hydrophobicity on the resulting predictions.

From observations such as those described above, it is possible to establish the following generalizations. Although hydrophilicity profiles are not typically used to assign secondary structure types (helices or strands) they give very useful information concerning the degree of surface exposure or burial of these structures in a protein. Furthermore, if one assumes that all regions where a hydrophilicity profile falls below the zero line are involved in secondary structures, one would be correct most of the time. The hydrophilic peaks then correspond to the surface loops and chain turns that connect these buried or partially buried secondary structure elements, as well as a few highly exposed short helices and edge strands of β-sheets. Perhaps not surprisingly, these regions of highly exposed peptide chain have been recognized recently not only as the most significant antigenic regions, but also as the most probable locations for a wide variety of other protein interactions (Hopp, 1984a; Rose et al., 1985b).

To further emphasize the agreement of hydrophilicity analysis with surface location, Fig. 2 shows the correlation of hydrophilicity values as determined by the method of Hopp and Woods (1981) with the 3-dimensional packing of myoglobin and concanavalin A. These drawings were made by shading the bars that connect the amino acids, so that darker bars represent hydrophobic segments and lighter bars are more hydrophilic regions. The cores of these proteins are apparent as dark (hydrophobic) regions that correspond to secondary structures associated with the deepest valleys observed in their hydrophilicity profiles. Myoglobin has a core of packed large helices, and concanavalin A has a core made up of two β-pleated sheets. The highly exposed connecting loops appear as light grey and white (hydrophilic) segments of polypeptide chain.

Another important observation concerns the information conveyed by the peak regions of the hydrophilicity profiles. While peaks usually occur in the same places on different profiles, the different methods show a considerable variability in their relative heights (see Fig. 1). This results from the different emphasis that each scale places on the most hydrophilic amino acids. It is in this area that the hydrophilicity scale of Hopp and Woods has its major advantage over the other procedures. As can be seen in Fig. 2A, most of the antigenic residues of myoglobin (white circles) are found in highly exposed loops of peptide chain, and these segments have been correctly identified as hydrophilic peaks, as is evidenced by the whiter shading of the connecting bars near the antigenic residues. However, some regions are highly exposed but not antigenic, as is the case with the loop projecting to the lower right side of myoglobin. Such regions are often lacking in charged and polar amino acids. That this is the case in this loop is shown by the darker shading of the connecting bars in this region. Methods that yield large peaks for such regions would be wrong more often when used to predict major antigenic sites. As Table I shows, the various scales have widely varying success rates, and no other method is as successful as the original hydrophilicity method in identifying antigenic sites. Our recent efforts to develop an improved antigenic determinant prediction scale have shed some additional light on the importance of the ranking of the hydrophilic amino acids.

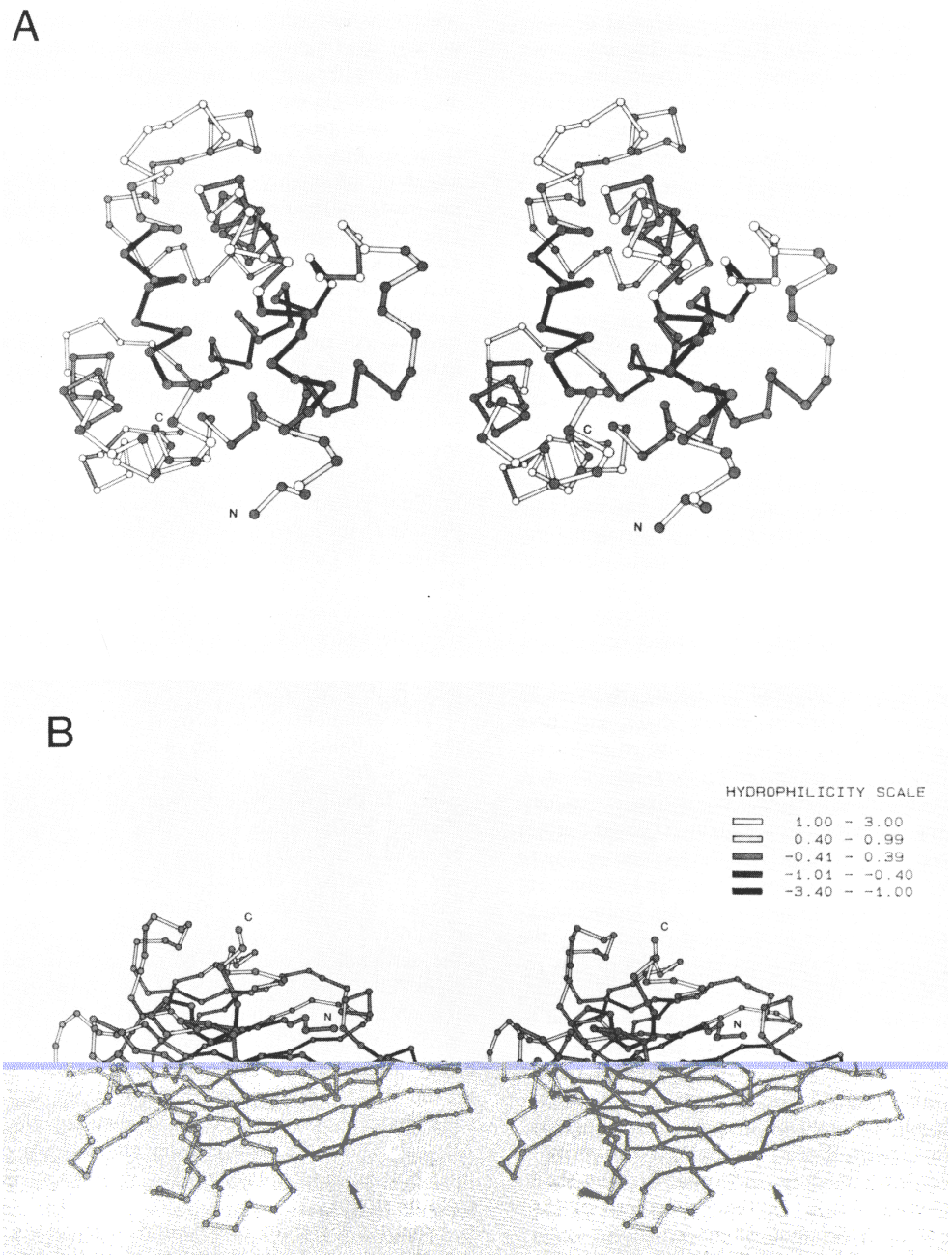

Fig. 2. Relationship of hydrophilicity to protein 3-dimensional structure. Crystallographically determined stereo views of the alpha carbon tracings of myoglobin (A) and concanavalin A (B) are shaded according to the local hydrophilicity value of the polypeptide chain. As indicated in the key, the most hydrophilic regions are white, the most hydrophobic regions are black, with intermediate regions in shades of grey. Each hydrophilicity average was assigned to the connecting bar that joins the two central alpha carbons of the hexapeptide segment from which it was derived. For myoglobin, the known antigenic residues are indicated as white alpha carbons, non-antigenic residues, grey. No antigenic sites have been determined for concanavalin A. The arrow indicates the region that is buried in the concanavalin A dimer. Images were generated with the aid of the NAMOD procedure, available from the Protein Data Bank, Brookhaven National Laboratory, Upton, New York. The hydrophilicity values were determined by the HYDRO3 computer program.

In the thirteenth row of Table I are shown the amino acid values of a newly developed scale called acrophilicity (literally, 'height-loving'). This scale was derived by visual observations to determine the frequency of occurrence of the amino acids in highly exposed locations on the surface of 49 proteins (Hopp, 1984a). In brief, the method used stereo paired α-carbon tracings similar to those in Fig. 2 (without shading) to allow the selection of highly exposed residues without knowledge of their identities. A subsequent compilation of the selected residues yielded a scale of acrophilicity values that was unbiased by any notion of side chain

orientation or amino acid water solubility. The results were surprising for several reasons. As expected, the acrophilicity scale is better than the hydrophilicity scale in its correlation of valleys with secondary structure, and peaks with exposed loops and turns (cf. Fig. 1), but at the same time, it is one of the least capable methods in locating antigenic sites. Thus, there is a seeming paradox that *high exposure* of polypeptide chain loops is not sufficient to guarantee antigenicity. One explanation is suggested by comparing the top ends of the hydrophilicity and acrophilicity scales. In contrast to the hydrophilicity scale, which is based on the water solubility of the amino acids, the acrophilicity scale is, for the most part, a *size* scale.

This dichotomy is at its most extreme between the hydrophilicity and acrophilicity scales, with most of the other scales falling somewhere between. In the hydrophilicity scale, all of the highly charged amino acids (Asp, Glu, Lys, Arg) are grouped at the maximum value of 3.0 (His, which can be positively charged, is almost entirely uncharged at physiological pH and is therefore not considered to be a highly charged amino acid). In the acrophilicity scale, all of the smallest amino acids (Gly, Pro, Asn, Ser, Asp) are ranked in the top positions. The small amino acids have intermediate values on the hydrophilicity scale and the highly charged amino acids have intermediate values on the acrophilicity scale. Most other scales have mixtures of these two groups at their top ends. All scales rank the hydrophobic amino acids in their bottom halves, also roughly in order of increasing size.

From the foregoing observations it is apparent that the exposed loops and turns of proteins may be generated by concentrations of small amino acids while antigenic sites are likely to occur in regions where there are concentrations of charged and polar residues, which may or may not be chain turn segments. This could explain the lower prediction success rates of some of the other scales, because they are effectively mixtures of the hydrophilicity and acrophilicity scales, and therefore do not fully express the unique qualities of either. This also represents one of the pitfalls of using concensus scales, as several of these authors have done.

Segmental or atomic mobility

It has been pointed out recently (Tainer et al., 1984; Westhof et al., 1984) that there probably is a correlation of antigenic sites with segments of proteins that have a greater than average flexibility. To some extent, this is not surprising, because the most flexible segments of a protein usually occur in the most highly exposed, hydrophilic loops of peptide chain. Although these authors are critical of the use of hydrophilicity prediction methods, they fail to note that application of their approach requires prior solution of the crystallographic structure of the particular protein in question. This process may require years of effort to obtain suitable crystals, and always requires vastly more complicated and expensive computer calculations. These constraints make this approach currently unfeasible for most researchers. An interesting attempt to rectify this situation was reported by Karplus and Schulz (1985). They developed amino acid scales for the purpose of predicting the segmental mobility of amino acid sequences. However, perhaps not surprisingly, these scales turn out to be essentially hydrophilicity scales, and, as seen in Table I, they are somewhat less successful in locating antigenic sites on the data set of native protein antigens.

Amino acid sequence variability

A number of investigators have reported success in locating antigenic sites by identifying regions of high variability in sequences derived from homologous proteins of two or more species or subtypes of microorganisms. This criterion was used either alone, or in concert with hydrophilicity analysis. This approach is actually quite old, and a large body of literature exists to support its applicability, most notably, using cytochrome c and the hemoglobins (Reichlin, 1975) as antigens. Several limitations exist, however. In many cases, the protein in question is unique so that comparison to a homologue is impossible. On the other hand, many

viral antigens are so highly variable that any two strains may show so many differences that it is impossible to single out a particular segment that is more likely to be a major antigenic site. Hydrophilicity analysis can help in both cases. The highest peaks of hydrophilicity are often correlated with the most variable segments on proteins (Hopp and Woods, 1983). Therefore, if only one species of a protein antigen is available, one is likely to be correct in assuming that its most hydrophilic segments are also sites of amino acid variability. On the other hand, when confronted with a large number of amino acid substitutions, it is unlikely that all are involved in major antigenic sites. One then simply relies on the fact that the most hydrophilic sites are probably involved in major antigenic determinants. Another, unrelated observation concerns sites of deletion or insertion mutations. It has been pointed out (Greer, 1981) that these sites almost always occur in external, looping segments of a protein, because shortening or lengthening the polypeptide chain can be accommodated better on the outside of a protein. Therefore, inserted segments are very likely to be highly exposed and hence, immunogenic. Furthermore, the sequences surrounding a deletion site are also likely to be surface oriented, and are therefore also likely to have immunogenic potential.

Additional information found in hydrophilicity profiles

The discussion above should make it clear that hydrophilicity analysis is an extremely effective way to predict the locations of antigenic sites, and that most other procedures are more or less related to it. As has been pointed out in a recent review on protein chain turns (Rose et al., 1985b), the hydrophilic, looping regions that cover most of a protein's surface are obvious places to look for protein interactions such as antibody binding. It should be emphasized, however, that it is the subset of exposed chain turns that are covered with concentrations of charged and highly polar amino acids that are most likely to be involved in such interactions. Appropriately, major antigenicity has been located at the most hydrophilic segments of all major disease organism surface antigens that have been studied in detail. These include influenza, polio, foot and mouth disease, hepatitis B, herpes and common cold viruses as well as streptococcus and gonococcus (reviewed by Hopp, 1984a).

Rose et al. (1985b) also point out that other types of protein interactions occur on exposed loops. I have reported similar observations (Hopp, 1984a) including a long list of such sites too numerous to report here. In general, hydrophilic loops appear to be the most probable locations for binding by other macromolecules. For example, immunoglobulins bind complement at the most hydrophilic segments of the Fc region (Prystowsky et al., 1981), apolipoprotein E is bound by cellular receptors at its most hydrophilic site (Innerarity et al., 1983), and, in the area of protein-nucleic acid interactions, both tobacco mosaic virus coat protein (Bloomer et al., 1978) and DNA polymerase II (Ollis et al., 1985) contact their respective nucleic acids via their most hydrophilic segment. The immunosuppressive effects of a human leukemia virus have recently been traced to a short segment of its p15E protein that contains the second most hydrophilic site of the molecule (Cianciolo et al., 1985). Hydrophilic loops often act as substrates in enzyme catalyzed reactions. I reported a large number of cases of phosphorylations, acetylations, and other addition reactions at the most hydrophilic sites on a variety of proteins, as well as a long list of instances of limited proteolysis taking place at these sites (Hopp, 1984a). Although it is beyond the scope of this review to provide all the relevant information on these observations, it should be apparent that hydrophilic sites deserve special consideration as likely sites for these other types of protein interactions. Again, it should be emphasized that it is probably the charged and highly polar amino acids that give specificity to these sites.

Another important class of interaction sites visible in hydrophilicity profiles are the membrane-spanning hydrophobic segments that comprise the signal peptide and membrane anchor segments of integral membrane proteins. In describing their 'hydropathy' method, Kyte and Doolittle (1982) emphasized that transmembrane segments appear as broad peaks of

hydrophobicity (valleys on hydrophilicity profiles). However, they apparently failed to note that such regions are readily discerned by other previously published methods as well (Rose and Roy, 1980; Hopp and Woods, 1981). Hydrophilicity profiles by many of the methods included in this review are capable of indicating membrane-associated sequences, which are recognized as particularly wide, low valleys. This is intuitively satisfying because these downward projecting profile segments can be thought of as being 'buried' in the lipid bilayer.

Notes on applications

Given that a substantial amount of information can be gained by hydrophilicity analysis, it seems appropriate to clarify certain of the procedural aspects that can impact the quality of the information obtained. Two important considerations are the choice of an averaging group length, and choice of the amino acid scale.

Averaging group length

Despite the demonstration that an averaging group length of six amino acids gives an optimal match of sequence to structure (Hopp and Woods, 1981), many investigators have used a window of five or seven, and even up to 18 amino acids to generate their profiles. This is unfortunate, because windows greater than six tend to cause adjacent valleys to coalesce into single, longer valleys, giving the wrong impression that a single secondary structure element exists where, in reality, two or more may exist, with intervening exposed loops. In a similar way, hydrophilic peaks can be averaged into adjacent valleys. This usually causes a substantial rearrangement of the locations of the highest peaks on profiles, and hence, of the predicted interaction sites.

Using a wider window also leads to fewer correct antigenic site predictions, because the method is then less capable of distinguishing internal from external segments. The use of a wide window in order to emphasize transmembrane segments, advocated by Kyte and Doolittle, is not necessary, especially in light of observations concerning Gly, Ser, and Thr residues, discussed below. The tendency of many investigators to use odd numbered window sizes probably stems from a desire to locate the center of the averages directly on the central amino acid rather than between amino acids, as is the case with even numbered windows. However, this choice has little meaning, because the averages represent information pertinent to all of the amino acids in the averaging group, not just the central residues. An additional incentive for retaining the 6 amino acid window, is the convenience of making images like figures 2A and 2B, where the bar between amino acids is shaded, rather than the amino acids themselves. This is also useful when viewing 3-dimensional alpha carbon line drawings of proteins on computer video display systems. Finally, and most importantly, a standard window size is essential for communication between investigators. If windows are chosen arbitrarily, then two researchers studying the same protein could refer to entirely different polypeptide segments as the most hydrophilic site on the molecule.

Choice of scale

Like so many aspects of science, the number of hydrophilicity/hydrophobicity scales has been growing exponentially. However, Fig. 1 should help to demonstrate that little new information is likely to be forthcoming in additional scales developed in the future. In fact, it appears that the oldest information available in this field, namely the original hydrophobicity scale determined experimentally by Nozaki and Tanford (1971), remains the best for the purposes described in this paper, albeit in its modified form as the hydrophilicity scale (Hopp and Woods, 1981). The following is a brief comparison with the other commonly referenced scales listed in Table I.

The distinguishing feature of the hydrophilicity scale is the clustering of all four of the highly charged amino acids at the maximum value of 3.0. Charge-charge interactions have long been

known to be of major importance to antigen-antibody interactions (Sela and Mozes, 1966; Atassi, 1975). We experimentally established, in our original paper (Hopp and Woods, 1981) that this equivalence of the charged amino acids improves antigenic site selection. The lesser ability of the other scales to locate antigenic sites is probably due in large part to their neglecting this charge balance. Most of the other methods have tended to overemphasize positively charged residues over negatively charged residues, apparently to their detriment. The occurrence of the small amino acids as well as the polar but uncharged amino acids in the middle of the hydrophilicity scale assures that the highest peaks will be composed of mixtures of charged, polar and small amino acids. These are precisely the combinations that have been found to comprise the highly exposed loops of peptide chain that usually represent major antigenic sites. The other scales all vary somewhat from the organization of the hydrophilicity scale, and are correspondingly less accurate in locating antigenic sites.

Of the other scales, that of Kyte and Doolittle is probably the most often used in correlating sequence to structure. Hydropathy plots are frequently encountered in the literature, but many of these may be somewhat compromised in their usefulness, both because of the inappropriate averaging window sizes used, and because this scale has one of the greater discrepancies among the charged amino acids. This is due to arginine and lysine having values that are unrealistically elevated above all other values on the scale. In fact, the value of arginine was arbitrarily raised above the already high value of lysine without experimental substantiation by Kyte and Doolittle (1982), who state that this choice was '... the result of personal bias and heated discussion between the authors'. It is likely that this arbitrary choice is responsible for making their method the most strongly selective for positively charged sites among all methods that have been used for antigenic site selection (Hopp, 1985). This leads to errors by overlooking negatively charged or mixed-charge sites, which are frequently antigenic. The scale of Sweet and Eisenberg (1983) is a consensus scale compiled by averaging the values of other investigators. Because of the selection of other scales included in the consensus, this scale also suffers from overemphasis on positive charges.

The scales of Janin (1979) and Rose et al. (1985a) were derived by consideration of the degree of burial or surface exposure of amino acids in proteins. They are reasonably successful in locating antigenic sites, and this probably reflects the clustering of charged amino acids to the left (top) end of the scale. However, the process of deriving a scale by accessibility calculations can be greatly affected by the more or less arbitrary definition of surface exposure of amino acids that is chosen by various experimenters. This can be seen in the much lower success rates obtained using the values of Chothia (1976) which were obtained by a method similar to those of Rose and Janin.

The scale of Welling et al. (1985), which is ranked as the second most successful prediction scale in Table I, should be viewed with caution. It is the result of optimizing the values for success in predictions with a database of antigenic residues in 20 completely or partially characterized proteins. Thus, it has been 'matched' to the rather limited number of currently known antigenic residues and may be strongly biased as a result. Because all 12 proteins of our database (Hopp and Woods, 1981) were incorporated into this process, it is not surprising that these values do well here. A further problem could arise because much of the additional data were from studies using synthetic peptide immunogens, which cannot be considered as data on native protein antigens. It seems likely that, as a larger database of native protein antigens accumulates, this method will be most likely to move to a lower ranking, as its match to the available data changes.

Finally, it is common to see references to predictions of antigenic sites by including consideration of secondary structures, particularly, β-bends. As can be seen in Table I (row 16), β-bend predictions are actually anti-predictive for antigenic sites, doing less well than random selection. This is not to say that the methods do not locate β-bends, but rather

that many β-bends actually occur in low-relief or semi-buried locations and are therefore not antigenic. On the other hand, the helix prediction scale of Chou and Fasman (1978) shows some promise. Because the helix parameters have been inverted for the purposes of this review, it should be emphasized that in this case, it is the low end of the scale that is predictive of helices, and that the success in locating antigenic sites is due to the regions predicted as being non-helical by this method.

Improvements to the method

In the future, it is likely that improvements in interaction site predictions will come not by inventing new scales, but by modifying the applications of scales already available. Using the basic hydrophilicity profile as a starting point, additional information can be considered separately, or added to the profile. Examples of potential improvements are shown in Figs. 3 and 4. It has been observed (Thornton and Sibanda, 1983) that the N- and C-termini of proteins are typically more highly exposed than might be expected. Reasoning that this probably makes them more antigenic, we tested the predictive improvements made by increasing the hydrophilicity values of the first and last hexapeptides in protein sequences.

A

```
265 GOSUB 910

900 STOP
910 REM  ADJUSTMENT SUBROUTINE
920 REM  N,C ADJUSTMENTS
930 B(l)=B(l)+4 @ B(N)=B(N)+4
940 REM GLY,SER,THR ADJUSTMENTS
950 FOR I=3 TO N-2
960 IF A(I)=3 THEN GOTO 980
970 IF A(I)#4 AND A(I)#8 THEN 1080
980 FOR J=-2 TO 2
990 IF J=0 THEN 1060
1000 IF A(I+J)=1 OR A(I+J)=2 THEN 1080
1010 IF A(I+J)=4 THEN GOTO 1100
1020 IF A(I+J)=5 OR A(I+J)=6 THEN 1080
1030 IF A(I+J)=7 OR A(I+J)=15 THEN 1080
1040 IF A(I+J)=17 OR A(I+J)=18 THEN 1080
1050 IF A(I+J)=19 OR A(I+J)=20 THEN 1080
1060 NEXT J
1070 B(I)=-3.4
1080 NEXT I
1090 RETURN
1100 FOR K=-2 TO 2
1110 IF K=0 THEN 1150
1120 IF K=J THEN 1150
1130 IF A(I+K)=3 OR A(I+K)=4 THEN 1080
1140 IF A(I+K)=8 AND A(I)#8 THEN 1080
1150 NEXT K
1160 GOTO 1020
1170 END
```

B

```
266 GOSUB 1170

1170 REM HIS,TYR,TRP ADJUSTMENTS
1180 FOR I=3 TO N-2
1190 IF A(I)=15 THEN GOTO 1210
1200 IF A(I)#17 AND A(I)#19 THEN 1310
1210 S=0
1220 FOR J=-2 TO 2
1230 IF J=0 THEN 1290
1240 IF A(I+J)=1 OR A(I+J)=2 THEN S=S+1
1250 IF A(I+J)=4 OR A(I+J)=5 THEN S=S+1
1260 IF A(I+J)=6 OR A(I+J)=7 THEN S=S+1
1270 IF A(I+J)=18 OR A(I+J)=19 THEN S=S+1
1275 IF A(I+J)=20 THEN S=S+1
1280 IF A(I+J)=3 OR A(I+J)=8 THEN S=S+.5
1290 NEXT J
1300 IF S>2.4 THEN B(I)=B(I)+2.4
1310 NEXT I
1320 RETURN
1330 END
```

Fig. 3. Adjustment subroutines for hydrophilicity analysis. A: When these lines are added to the hydrophilicity program (Hopp and Woods, 1983), the N- and C-terminal hexapeptide averages are adjusted upward, and Gly, Ser, and Thr values are adjusted downward when they occur in hydrophobic segments. This results in a more realistic representation of the degree of exposure of N- and C-termini, and to longer, lower valleys in transmembrane regions. B: These lines raise the values of His, Tyr, and Trp residues when they are in hydrophilic surroundings. This may provide a more accurate representation of their probability of occurrence in protein interaction sites, but caution is advised because the appropriateness of this adjustment has not been definitively established. For clarity, we refer to the original program as HYDRO, to the program with section A added as HYDRO3, and to the program with sections A and B added as HYDRO4.

Using the database of antigenic proteins (Hopp and Woods, 1981) we found that prediction success increases when this adjustment is made, verifying that N- and C-termini are indeed more antigenic. We also found that too great an increase led to lower prediction success rates

owing to over-prediction of antigenicity at the termini. An optimum occurred when the terminal averages were raised by 2/3 of a hydrophilicity unit. This change is made on lines 920 and 930 of the subroutine shown in Fig. 3A. The remaining lines of that subroutine serve to locate Gly, Ser, and Thr residues that are likely to be buried in the interiors of proteins or in transmembrane segments. These are picked out by their occurrence between four adjacent hydrophobic neighbors, two on each side. Lowering the values of these Gly, Ser, and Thr residues to the bottom of the hydrophilicity scale (−3.4) results in additional emphasis on transmembrane valley regions as shown in Fig. 4. The Gly, Ser, Thr adjustment subroutine never changes the highest peaks, and so has no effect on antigenic site predictions. This procedure has the advantage that it readily emphasizes transmembrane segments without resorting to wider windows, which, as already mentioned, can cause the loss of other useful information.

Additional improvements of this nature seem likely in the future. For example we have observed that His, Tyr, and Trp residues, which are often buried, may sometimes be surface oriented and therefore antigenic. This probably stems from the fact that, while each is aromatic and has substantial hydrophobic character, each also has one or more polar atoms capable of interacting strongly with complementary groups on an antibody. Attempts to identify candidate His, Tyr, and Trp residues by their occurrence between hydrophilic neighbors look promising, as judged by the increase in success rate indicated in Table I. However, until it can be demonstrated that this does not represent 'matching' of results to the database, this particular procedure (Fig. 3B) should be used with caution. Additional work on this, and other analyses should result in significant improvements to antigenic determinant predictions in the future.

Computer programs

Several recent papers have included copies of computer program code that enable readers to carry out calculations exactly as specified by the authors. These are potentially very useful in allowing a wider use of the available methods, and to assure that they are applied correctly. They are written in Fortran, C, and Basic computer languages, and are considered separately below. Several commercial organizations provide hydrophilicity/hydrophobicity calculations on a fee basis, but because none have consulted this author, the validity of their procedures cannot be assessed.

Rose et al., (1985b)

Fortran. Included at the end of an excellent review on chain turns in proteins, this procedure allows the user to select between three amino acid scales: that of Nozaki and Tanford (1971), Hopp and Woods (1981), or Kyte and Doolittle (1982). This method is capable of generating a data plot in addition to numerical output, although the line-printer plots so obtained are crude and would require re-plotting by other means for publication quality copy. A subroutine is included to allow selection of an averaging group size. This can be considered one of the weak points of the program, both because variable windows lead to unstandardized results, and more importantly, because the program issues error messages when even numbered windows are requested. This obviously rules out windows of six. It is therefore advised that users revise this section before use, at least to allow a six-residue window, and preferably, to eliminate all others, so that results from one laboratory can be compared more easily to the results from others. The hydrophilicity scale (Hopp and Woods, 1981) has been inverted; this can be corrected if desired by changing the signs on all values.

Kyte and Doolittle (1982)

'C' language. Contains only the 'hydropathy' scale. A choice of windows is available, including six, which should be used because this method is also best at locating antigenic sites using this length. Hydrophobic is at the top of this scale, but inversion by changing signs

is possible. As discussed, this scale suffers somewhat from an overemphasis on positively charged residues. Other scales could be substituted (in the appropriate order) on the last line of the program.

Karplus and Schulz (1985)

Fortran. This procedure appears to be of less general use than the others, for several reasons. First, when tested on the protein antigen database, it yielded a rather low success rate in locating antigenic sites compared to the other methods (Table I). Secondly, it uses a method that is significantly more complicated than simple averaging, and requires a window size of seven. It also requires three separate scales of values in order to allow neighbor effects to be computed. The authors also did not provide sufficient data to establish a true correlation of their profiles with the desired objective of predicting the mobility of chain segments in proteins. For these reasons, this method should be used with caution until further data are obtained to support its validity.

Hopp and Woods (1983)

Basic. This program (HYDRO) is written for use with a specific computer, the Hewlett-Packard HP-85 microcomputer. Although it is necessary to change some lines in order to run it on other microcomputers, this is the only program available in Basic, and therefore it is probably the best guide for users writing programs for other microcomputers. Alternatively, the author will provide copies of this procedure in Fortran or Apple Basic on request.

There are several unique features of the HYDRO program that are worth mentioning. First, the only averaging group length is six amino acids. This assures that different users of the method will obtain the same results for a given protein. While the window could be changed relatively easily, this is not advised, for the reasons cited above. Second, a peak finder routine is included, which ranks the peaks in order of decreasing height. While this might at first seem trivial and merely for convenience, it is not. The notion of what a ‘peak’ is, can be the subject of considerable variation. The ‘peakfinder’ subroutine in this program avoids confusion by only considering as peaks, those points that have no higher neighbors within three sequence positions. This procedure prevents redundant identification of the same antigenic site caused by shoulders on a main peak. Furthermore, by ranking the peaks in height order, it gives information concerning which sites should be considered as priority targets for investigation. This automatic peak ranking system also makes possible the impartial assessment of prediction success rates shown in Table I. Finally, such a procedure eliminates the possibility that different investigators might miscommunicate because of different judgments concerning what constitutes the second or third highest peak of a particular molecule.

As mentioned earlier, the subroutines shown in Fig. 3A can be added to this program in order to increase antigenic site prediction success and to emphasize the hydrophobicity of membrane-associated segments and signal peptides. A version so updated is referred to as HYDRO3. The subroutine in Fig. 3B can subsequently be added in order to adjust His, Tyr, and Trp values, if desired (the program is then referred to as HYDRO4).

Practical aspects

Structure and antigenicity predictions

Fig. 4 demonstrates the information available in hydrophilicity profiles, using the circumsporozoite (CS) antigen of the malarial parasite, *Plasmodium falciparum* (Dame et al., 1984). Two striking features are the hydrophobic valleys at either end of the molecule, representing the signal peptide (S), responsible for the export of this protein to the exterior of the organism, and the membrane anchor segment (M) that binds the CS protein to the surface of the sporozoite. Both of these valleys have been dramatically deepened by adjusting Ser

and Gly residues downward by the HYDRO3 procedure. Other, minor changes in the profile result from the HYDRO4 adjustments of several His, Tyr, and Trp residues (dotted segments elsewhere on the profile) but are without significant effects in this case. Peaks on the profile indicate major potential antigenic sites at around residues 92, 108, 327, and 370. Interestingly, antigenicity at the second and fourth of these sites has recently been reported, including partial neutralizing ability (Nussenzweig, 1984). However, the majority of antibodies produced in response to the organism are directed against the peculiar repetitive sequence (R) in the center of the CS protein. It is intriguing to note that while the hydrophilicity profile falls to near zero in this region, suggesting poor antigenicity, the acrophilicity (Ac) profile in this region runs at a continuously elevated level.

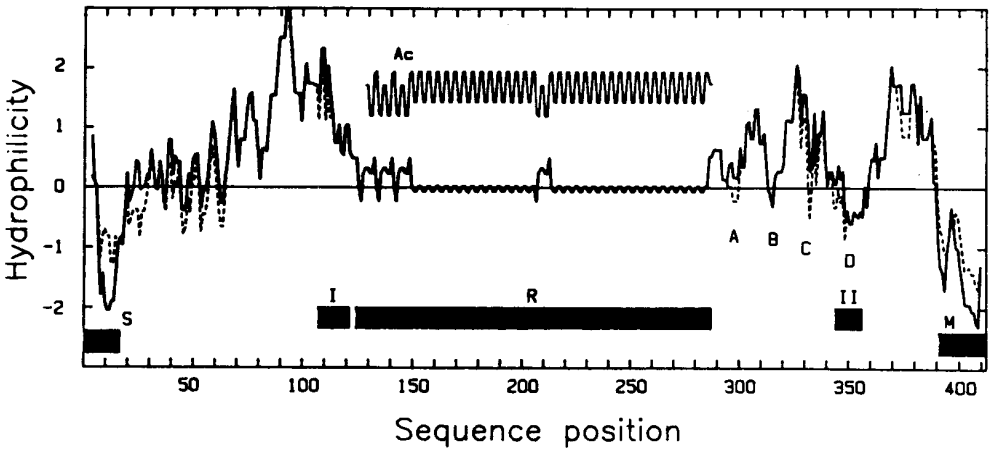

Fig. 4. Hydrophilicity profile for the circumsporozoite antigen of *Plasmodium falciparum*. The solid profile represents the hexapeptide averages obtained using the original hydrophilicity values (Hopp and Woods, 1981) and including the adjustment routines from Fig. 3. The dotted segments represent the results obtained without the adjustment routines. Bars below the profile represent: S, the signal peptide; I and II, the regions of sequence conservation between the antigens of different *Plasmodium* species; R, the area of tandem sequence repeats, containing 37 copies of the tetrapeptide, Asn-Ala-Asn-Pro; M, the membrane anchor segment; A through D, valley regions (probably secondary structures) where hydrophobicity is conserved even though the sequence may vary. The segment (Ac) above the hydrophilicity profile is a portion of the acrophilicity profile, emphasizing the great difference between acrophilicity and hydrophilicity in this region.

It is tempting to speculate that this region serves to give the organism a highly solvated surface that is less immunogenic than it would be if it were covered with charged amino acids and other highly polar groups. The fact that it represents a major antigenic site may then simply reflect its large extent and repetitive nature, which would allow it to bind many antibody molecules having the same specificity. However, because acrophilicity is a poor predictor of antigenic sites in more typical proteins, it seems possible that the immune system might react slowly or atypically with this exposed region of the CS protein, while at the same time overlooking the potentially more useful sites elsewhere on the molecule.

Furthermore, antibodies to this region would lack the strong stabilizing charge-charge and hydrophobic binding interactions that usually characterize antibody-antigen interactions, and therefore might be expected to have lower affinities for the CS protein. Such a situation might allow the organism to escape some of the antibody-mediated immune processes that would otherwise be expected to neutralize it and eliminate the infection. Supporting the foregoing argument is the fact that, of all CS proteins sequenced to date, none break the rule of high acrophilicity and medium hydrophilicity in the repeat region, even though all have unique repeating sequences.

Finally, there are 4 valley regions (marked A-D, Fig. 4) that are conserved in all known CS proteins. Their presence implies secondary structure in the C-terminal region although a distinction between helix or strand cannot be made based on this profile. However, it can be

assumed that these elements pack into a globular domain in this region, and will probably have a conserved tertiary structure between different species of *Plasmodium*. Several valleys at the N-terminus could conceivably participate in this globular domain, although the lack of conservation in this region makes this assessment more dubious. From the preceding analysis, a picture of the CS protein emerges that suggests a membrane anchored globular protein that is protected by a highly exposed looping segment of repeating sequences that offers little to the immune system other than relatively weak hydrogen bonding interactions.

Synthetic peptide immunogens

As a conclusion to this review, it is appropriate to address the applicability of hydrophilic peptides in obtaining anti-protein antisera. While reports of successful applications are numerous, it is probable that an even greater number of failed experiments exist. However, it should be possible to get a useful response from any hydrophilic segment given the right approach. Experience in our own laboratories provides several illustrative examples. We have been very successful in raising antisera to hepatitis B virus using a synthetic peptide comprising the most hydrophilic site in the surface antigen (Hopp, 1984b). On the other hand, another group reported negligible titers using this segment (Lerner et al., 1981). In this case, we used a slightly longer version of the peptide, and a new carrier, dipalmitoyl lysine, whereas they used a keyhole limpet hemocyanine conjugate of their shorter peptide. In contrast, the results were just the opposite using the most hydrophilic segment of interleukin 2. In that case, Lerner's collaborators (Altman et al., 1984) reported obtaining very useful antisera reactive with native interleukin 2 while our efforts, using essentially the same peptide on the dipalmitoyl lysyl carrier were unsuccessful. The preceding results imply that one should be open minded concerning the immunization protocols used, and that exploration of a variety of peptide sizes and carrier types can improve the prospects for success.

Another concern for many immunochemists is the notion of continuous versus assembled antigenic sites (Benjamin et al., 1984; Berzofsky, 1985). Two points should be emphasized here. First, it may seem that hydrophilicity analysis cannot give information regarding assembled antigenic determinants because it considers only the linear sequence of amino acids, while assembled determinants are, by definition, composed of two or more peptide segments brought together in the 3-dimensional folding of the protein. However, most assembled determinants are made up of multiple hydrophilic segments, so that hydrophilicity analysis must be useful to some extent in their identification. Furthermore, the very first case of a crystallographically determined structure of an antibody-protein interaction (Amit et al., 1985) bears this out quite emphatically. In this case, an anti-lysozyme hybridoma is seen to bind lysozyme in such a way that it is in contact with an assembled determinant that includes *both* the most hydrophilic, and second most hydrophilic segments of the protein.

The second problem is more serious, however. It currently seems unlikely that an assembled determinant can be constructed properly from two or more short peptides. Thus, it appears impossible to use synthetic peptides to generate this subset of antibodies that are readily produced by intact protein antigens. On the other hand, as was mentioned before, it is possible to achieve very different results by varying immunization protocols and carriers. It may be possible to reconstruct some assembled determinants by judicious use of disulfide connections between peptide segments, or by disposing two or more peptides on a carrier that allows them to align with each other in the orientation that they have in the native antigen. We are optimistic that time and experience will bring changes in our ability to use synthetic peptide immunogens to produce desired outcomes for the majority of antigenic sites. Furthermore, because other types of protein interaction sites are also associated with hydrophilic loops, it may sometimes be possible to use hydrophilicity analysis as a guide in producing peptide analogs of hormone-receptor binding sequences, and other portions of proteins that are of interest.

Summary

This paper has utilized several rather complicated figures in order to compare hydrophilicity methods and to clarify the fine points of the information conveyed by hydrophilicity plots. However, it should be emphasized that the lesson to be learned from these comparisons is relatively simple. The most useful information concerning a protein's antigenicity and other structural features can be obtained from a single, simple plot made by the HYDRO3 procedure. The data points in such a plot represent hexapeptide averages of the original hydrophilicity values, with upward adjustments at the N- and C-termini and downward adjustments of buried Gly, Ser, and Thr residues. The highest peaks represent antigenic segments and other types of protein interaction sites, while the valleys represent the secondary structure elements packed in the core of the molecule. The longest, lowest valleys represent membrane anchor segments or signal peptides.

References

Altman, A., J.M. Cardenas, R.A. Houghten, F.J. Dixon and A.N. Theofilopoulos, 1984, Proc. Natl. Acad. Sci. U.S.A. 81, 2176.
Amit, A.G., R.A. Mariuzza, S.E.V. Phillips and R.J. Poljak, 1985, Nature 313, 156.
Atassi, M.Z., 1975, Immunochemistry 12, 423.
Benjamin, D.C., J.A. Berzofsky, I.J. East, F.R.N. Gurd, C. Hannum, S.J. Leach, E. Margoliash, J.G. Michael, A. Miller, E.M. Prager, M. Reichlin, E.E. Sercarz, S.J. Smith-Gill, P.E. Todd and A.C. Wilson, 1984, Ann. Rev. Immunol. 2, 67.
Berzofsky, J.A., 1985, Science 229, 932.
Bloomer, A.C, J.N. Champness, G. Bricogne, R. Staden and A. Klug, 1978, Nature (London) 276, 362.
Chothia, C, 1976, J. Mol. Biol. 105, 1.
Chou, P.Y. and G.D. Fasman, 1978, Adv. Enzymol. 47, 45.
Cianciolo, G.J., T.D. Copeland, S. Orozlan and R. Snyderman, 1985, Science 230, 453.
Dame, J.B., J.L. Williams, T.F. McCutchan. J.L. Weber, R.A. Wirtz, W.T. Hockmeyer, W.L. Maloy, J.D. Haynes, I. Schneider, D. Roberts, G.S. Sanders, E.P. Reddy, C.L. Diggs and L.H. Miller, 1984, Science 225, 593.
Garnier, J., D.J. Osguthorpe and B. Robson, 1978, J. Mol. Biol. 120, 97.
Garratt, R.C., W.R. Taylor and J.M. Thornton. 1985, FEBS Lett. 188, 59.
Green, N., H. Alexander, A. Olson, S. Alexander, T.M. Shinnick, J.G. Sutcliffe and R.A. Lerner, 1982, Cell 28, 477.
Greer, J., 1981, J. Mol. Biol. 153, 1027.
Hopp, T.P., 1984a, Ann. Sclavo 2, 47.
Hopp, T.P., 1984b, Mol. Immunol. 21, 13.
Hopp, T.P., 1986, in: Proteins, structure and function, ed. J.J. L'Italien (Plenum Press, New York) in press.
Hopp, T.P. and K.R. Woods, 1981. Proc. Natl. Acad. Sci. U.S.A. 78, 3824.
Hopp, T.P. and K.R. Woods, 1983, Mol. Immunol. 20, 483.
Innerarity, T.L., E.J. Friedlander, S.C. Rall, K.H. Weisgraber and R.W. Mahley, 1983, J. Biol. Chem. 258, 12341.
Janin, J., 1979, Nature (London) 277, 491.
Karplus, P.A. and G.E. Schulz, 1985, Naturwissenschaften 72, 212.
Kyte, J. and R.F. Doolittle, 1982, J. Mol. Biol. 157, 105.
Lerner, R.A., N. Green, H. Alexander, F.T. Liu, J.G. Sutcliffe and T.M. Shinnick, 1981, Proc. Natl. Acad. Sci. U.S.A. 78, 3403.
Nozaki, Y. and C. Tanford, 1971, J. Biol. Chem. 246, 2211.
Nussenzweig, V., 1984, Ann. Sclavo 2, 187.
Ollis, D.L., P. Brick, R. Hamlin, N.G. Xuong and T.A. Steitz, 1985, Nature (London) 313, 762.
Prystowsky, M.B., J.M. Kehoe and B.W. Erickson, 1981, Biochemistry 20, 6349.
Reichlin, M., 1975, Adv. Immunol. 20, 71.
Rose, G.D. and S. Roy, 1980, Proc. Natl. Acad. Sci. U.S.A. 77, 4643.
Rose, G.D., A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, 1985a, Science 229, 834.
Rose, G.D., L.M. Gierasch and J.A. Smith, 1985b, Adv. Protein Chem. 37, 1.
Sela, M. and E. Mozes, 1966, Proc. Natl. Acad. Sci. U.S.A. 55, 445.
Sweet, R.M. and D. Eisenberg, 1983, J. Mol. Biol. 171, 479.
Tainer, J.A., E.D. Getzoff, H. Alexander, R.A. Houghten, A.J. Olson, R.A. Lerner and W.A. Hendrickson, 1984, Nature (London) 312, 127.
Thornton, J.M. and B.L. Sibanda, 1983, J. Mol. Biol. 167, 443.
Von Heijne, G., 1981, Eur. J. Biochem. 116, 419.
Welling, G.W., W.J. Weijer, R. v.d. Zee and S. Welling-Wester, 1985, FEBS Lett. 188, 215.
Westhof, E., D. Altschuh, D. Moras, A.C. Bloomer, A. Mondragon, A. Klug and M.H.V. Van Regenmortel, 1984, Nature (London) 311, 123.
Wolfenden, R., L. Andersson, P.M. Cullis and C.C.B. Southgate, 1981, Biochemistry 20, 849.