In: Proteins, Structure and Function, 1987
J.J. L'Italien, ed, pp 437-443, Plenum, New York.
http://thomas-hopp.com/THscience2.html

IDENTIFICATION OF PROTEIN SURFACES AND INTERACTION SITES BY HYDROPHILICITY ANALYSIS

Thomas P. Hopp

Immunex Corporation, 51 University Street, Seattle, WA 98101

The packing of hydrophobic amino acids into the interior of proteins has been recognized as a major factor in the creation of a three dimensional structure from a linear amino acid sequence. A related realization is that regions of protein sequences that contain clusters of charged and polar amino acids are usually highly exposed at the protein surface, and are therefore likely to be involved in protein interactions with other molecules, for example, binding to antibodies. Long hydrophobic segments have been recognized as probable membrane spanning sequences. In 1981, we published a method for averaging the hydrophilicity values of the amino acids in protein sequences (Hopp and Woods, ref. 1). The hydrophilicity profiles made by this procedure are useful in locating antigenic sites on native protein antigens, because major antigenicity is always found in the highest peaks (1,2). In 1982, Kyte and Doolittle (3) published a procedure that is virtually identical to ours, but made the additional observation that regions of wide hydrophobic maxima (valleys on our plots) are usually associated with membrane spanning segments of peptide chain. In this paper we compare our method to other similar procedures and describe several improvements to our method.

Comparison of hydrophilicity/hydrophobicity methods. Figure 1 shows the great similarity of the profiles generated using the different available scales of amino acid values. Almost without exception, the profiles show the same set of peaks and valleys, only differing in the magnitude of displacement from zero. Table 1 shows the success rate of each method in locating antigenic sites, determined as previously described (1). This is, to some extent, a measure of the correctness of a method in locating highly exposed segments. The methods all show some ability to locate antigenic sites, although our method, which was developed specifically for this purpose, is best. Furthermore, all procedures show a correlation of hydrophobic valleys with segments of secondary structure, as was pointed out earlier by Rose and Roy (4). A number of other such scales have been published, but were omitted from this comparison because many were very poor predictors of antigenic sites, and others did not have a full set of 20 amino acid values. The acrophilicity scale was developed by us (5) from direct observations on protein 3-dimensional structures. Interestingly, it locates secondary structure elements as well or better than any of the other methods, but is less successful in locating antigenic sites. This underscores our previous conclusions (1), that antigenic sites are a subset of surface locations, and that some highly exposed sites will not be immunogenic.

The major differences between scales occur at their top (hydrophilic) ends. Two groups of amino acids figure prominently: the charged amino acids, and the small amino acids. Our hydrophilicity method gives the four highly charged amino acids the same maximum value of +3.0, while the other scales spread these amino acids over a wide range. We found that the success rate for locating antigenic sites was improved by making all four values equal (1). This discrepancy

probably contributes to the lower prediction success rates of the other scales, because all of the charged residues are likely to contribute equally to protein-protein interactions by forming charge pairs. Furthermore, as seen in Table 1, the other scales are biased in favor of positively charged sites.

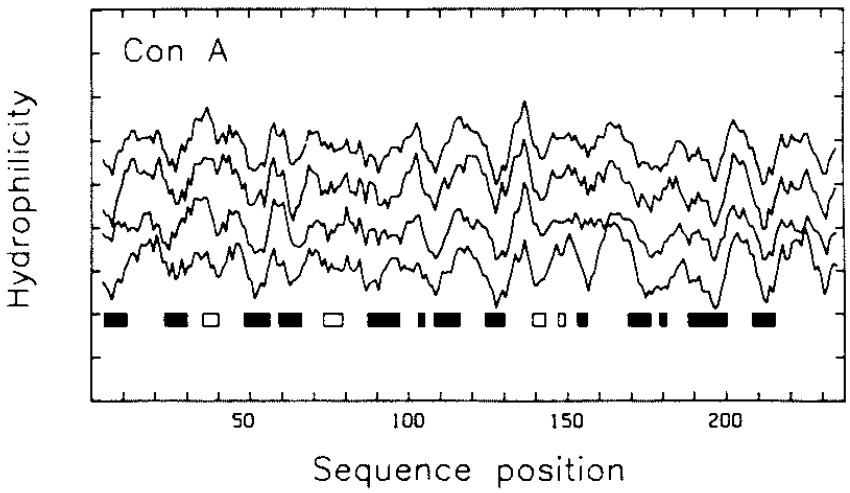


Fig. 1. Hydrophilicity analysis of concanavalin A by four sets of values. The scale has been compressed to facilitate comparisons, and each profile is offset by two hydrophilicity units from adjacent profiles. Top profile: values of Rose et al. (4); second profile: values of Kyte and Doolittle (3); third: Hopp & Woods (1); fourth: acrophilicity values (5). Strands of the two β-pleated sheets are indicated at bottom. Solid bars, internal strands; open bars, edge strands.

Table 1. Comparison of hydrophilicity methods.

A. Antigenic site selectivity. Percent correct among verifiable predictions.[a]

Peaks	Methods[b]							
	HYDRO4	H&W	Welling	Rose	K&D	Acro	C&F	Random
Highest	100	100	89	75	71	63	44	56
Top 3	88	75	76	68	65	59	43	57

B. Charge selectivity. Expressed as number of charges per six amino acids in peak hexapeptides.[c]

Peaks	Methods														
	H&W			K&D			Rose			Acro			HYDRO4		
	+	−	+/−	+	−	+/−	+	−	+/−	+	−	+/−	+	−	+/−
Highest	1.7	2.1	0.8	1.8	1.2	1.5	1.7	1.2	1.4	0.8	0.8	1.0	1.9	1.9	1.0
Top 3	1.4	1.8	0.8	1.5	1.3	1.2	1.5	1.2	1.3	0.7	0.8	0.9	1.6	1.7	0.9

[a]Determined as in reference 1. [b]Methods are: HYDRO4, hydrophilicity with N, C and amino acid adjustments as described in this paper; H&W, original hydrophilicity (1); Welling, Welling et al. (8); Rose, Rose et al. (6); K&D, Kyte and Doolittle (3); Acro, acrophilicity (5); C&F, Chou and Fasman (9); Random, numbers from 3 to –3.4 assigned to the amino acids randomly. [c]Averaged results for 70 proteins.

Relationship to secondary structure. As mentioned earlier, all of the hydrophilicity / hydrophobicity scales show large hydrophobic (downward) deflections in regions of secondary structure. Our experience with a large number of proteins has led to the following observations: (1) the internal strands of β-pleated sheets are always correlated with deep hydrophobic valleys, while edge strands can have much higher values (cf. Figure 1). (2) similarly, large helices usually correlate with deep valleys because their central regions are packed against the core of the folded protein. In contrast, the ends of large helices, and the whole extent of small helices are much less hydrophobic (cf. Figure 2). This is appropriate, because these segments, as well as the edges of β-pleated sheets, are necessarily highly exposed on protein surfaces. (3) even where valleys exist in the absence of secondary structure, the segment is usually buried. For example, the last valley of Con A is not involved in secondary structure, but in fact represents a random coil segment that is indeed buried inside the molecule. In cytochrome c (Figure 3), valleys corresponding to residues 27-35 and 80-83 have no secondary structure, but actually represent critical regions of hydrophobic amino acids lining the heme binding pocket.

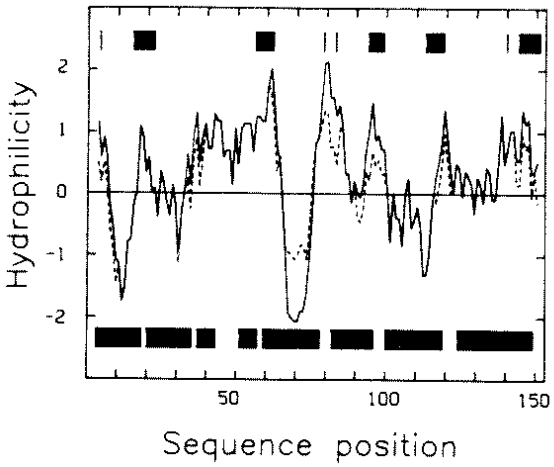
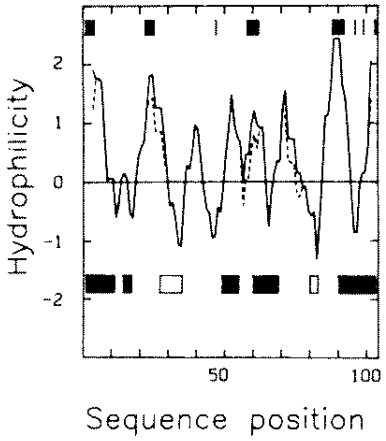


Fig. 2. Hydrophilicity profile for myoglobin. The solid profile was made by the HYDRO4 procedure; the dotted profile, by the original method (1). Bars below the profile represent the 8 helices of myoglobin. Slashes and bars above represent antigenic residues and segments.

The observations above lead to the following generalization concerning hydrophilicity plots: the parts of the profile above the zero line represent the edge strands of β-pleated sheets, the ends of helices, and highly exposed loop regions connecting them, while the parts below zero represent the internal strands of β-pleated sheets, the central residues of large helices and occasional buried coils. It should also be emphasized that an averaging group length of 6 amino acids is optimal for extracting such information from protein sequences. This was established in our original hydrophilicity paper (1) and has been reconfirmed in our recent work in developing the acrophilicity scale (T. Hopp and J. Merriam, in preparation).

Fig. 3. Hydrophilicity profile for cytochrome c. Solid bars below profile represent helices. The open bars are two segments in contact with the heme moiety. Antigenic sites are indicated above the profile.



Improvements to the method. We have been investigating ways to improve the usefulness of hydrophilicity analysis, both in predicting antigenic sites and in locating structural elements of proteins. Several useful adjustments have been found: 1) because N and C termini are usually highly exposed, raising the value of the first and last hexapeptide averages increases prediction success rates. 2) the amino acids Gly, Ser, Thr, His, Tyr, and Trp are somewhat ambiguous in their hydrophilic/hydrophobic behavior (6). We found that lowering the values of selected Gly, Ser, and Thr residues (when four neighboring residues are hydrophobic) increases the ability to see transmembrane segments as broad valleys (Figure 4) while leaving the rest of the profile unchanged. Increasing the values of His, Tyr, and Trp residues (when four neighboring residues are hydrophilic) on the other hand, increases success in predicting antigenic sites (Figures 2 and 3, Table 1).

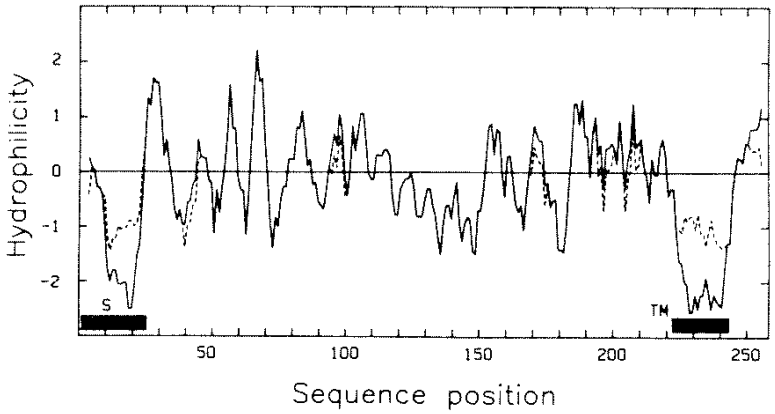


Fig. 4. Hydrophilicity profile for the class II histocompatibility antigen chain, IAa^k. The solid profile was made by the HYDRO4 procedure; the dotted profile, by the original method. The Gly, Ser, and Thr downward adjustments cause dramatic lowering of the profile in the regions of the signal peptide (S) and the transmembrane anchor segment (TM). The highest peaks probably represent sites of major antigenicity and other protein-protein interactions. The intervening valleys are probably the β-strands of this immunoglobulin-like molecule.

Hydrophilicity in the future. The redundancy of the profiles made with four different scales shown in Figure 1 implies that the development of new hydrophilicity/hydrophobicity scales is not likely to produce much new information. Rather, new uses for the existing scales may lead to improvements like those seen for our N and C terminal, and amino acid adjustments described above. At the same time, lessons already learned should not be forgotten. The use of hexapeptide averages has been advocated based upon experimental findings (1) but has not been generally applied. This is unfortunate because the use of other average sizes can obscure secondary structure information and, at the least, has served to obscure the redundancy of the methods developed to date.

In a recent survey of the immunological literature (5), we have uncovered an interesting fact. All of the major disease organisms for which detailed information is available, have proven to have major antigenicity associated with the most hydrophilic sites on their surface antigens. Thus, hydrophilicity analysis correctly predicted the location of neutralizing and/or strain specific antigenic sites on proteins from influenza, polio, foot-and-mouth disease, hepatitis B, herpes and common-cold viruses, as well as on streptococcal M protein and gonococcal pilin. Most of these studies were carried out without using our method as a guide, so the outcome strongly indicates that using hydrophilicity analysis in the future will lead to an accelerated success rate with other organisms. On the other hand, there recently have been a few reports of failed attempts at raising anti-protein responses using hydrophilic peptides. Experience in our own laboratories has

demonstrated that it is indeed easy to fail to get a useful immunization using peptides, but also that minor changes in the structure of the immunogen can drastically alter the outcome in generating anti-protein antibodies. Therefore, it is clear that hydrophilicity analysis will only live up to its full potential to immunologists after ways of assuring the proper immunogenicity of the synthetic peptides have been found. Experimentation directed toward this goal is underway in our laboratories.

TABLE 2. Comparison of hydrophilicity and acrophilicity values.

Hydrophilicity		Acrophilicity	
Asp	3.0	Gly	3.0
Glu	3.0	Pro	2.6
Lys	3.0	Asn	2.3
Arg	3.0	Asp	2.1
Ser	0.3	Ser	1.8
Asn	0.2	Lys	1.4
Gln	0.2	Glu	0.5
Gly	0.0	Arg	0.3
Pro	0.0	Thr	–0.1
Thr	–0.4	Gln	–0.2
His	–0.5	His	–0.4
Ala	–0.5	Ala	–0.5
Cys	–1.0	Val	–1.7
Met	–1.3	Met	–1.8
Val	–1.5	Tyr	–2.0
Leu	–1.8	Leu	–2.5
Ile	–1.8	Ile	–2.5
Tyr	–2.3	Cys	–2.6
Phe	–2.5	Phe	–2.7
Trp	–3.4	Trp	–3.0

Some of the most exciting developments may arise simply through an increased understanding of the information already present in hydrophilicity profiles. We have recently observed that a number of important protein interactions other than antibody-antigen interactions can occur at the most hydrophilic sites (5,7). These include the complement-binding site on IgG, and receptor-binding sites on apolipoprotein E, fibronectin, and interleukin 2, to name a few (reviewed in reference 5). Sites of phosphorylation, proteolysis, and other post-translational modifications also occur with great frequency at the most hydrophilic sites. In all probability, the acceptance of a standard hydrophilicity scale and averaging group length would go a long way toward fostering new discoveries, especially because it would enable investigators to communicate results and to understand each other more readily.

The hydrophilicity analysis methods described in this paper (HYDRO4 and ACRO3) are available from the author on request.

REFERENCES

1. T.P. Hopp and K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, Proc. Natl. Acad. Sci. USA 78:3824 (1981).
2. T.P. Hopp and K.R. Woods, A computer program for predicting protein antigenic determinants, Molec. Immuno1. 20:483 (1983).
3. J. Kyte and R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, J. Mol. Biol. 157:105 (1982).
4. G.D. Rose and S. Roy, Hydrophobic basis of packing in globular proteins, Proc. Natl. Acad. Sci. USA 77:4643 (1980).
5. T.P. Hopp, Protein Antigen Conformation, in: "International Conference on Synthetic Antigens," Annali Sclavo, 1(2):47 (1985).
6. G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee, and M.H. Zehfus, Hydrophobicity of amino acid residues in globular proteins, Science 229:834 (1985).
7. T.P. Hopp, Computer Prediction of Protein Surface Features and Antigenic Determinants, in: "Molecular Basis of Cancer, Part B," R. Rein, ed., Alan R. Liss, New York (1985).
8. G.W. Welling, W.J. Weijer, R.v.d. Zee, and S. Welling-Wester, Prediction of sequential antigenic regions in proteins, FEBS Letters 188:215 (1985).
9. P.Y. Chou and G.D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, Adv. Enzymol. 47:45 (1978).