Protides of the Biological Fluids
Volume 34, 1986
Pergamon Press

PROTEIN INTERACTION SITES PREDICTABLE BY HYDROPHILICITY ANALYSIS

T. P. HOPP

Protein Chemistry Department, Immunex Corporation, 51 University Street, Seattle, Washington 98101, USA

ABSTRACT

The original hydrophilicity plotting procedure of Hopp and Woods remains the most useful of such methods for locating the portions of protein sequences that are involved in interactions at the molecular surface. Methods are described for locating membrane spanning segments, antigenic sites, sites of protein-nucleic acid interactions, proteolytic sites and regions of post-translational modifications.

KEYWORDS

Hydrophilicity: hydrophobicity: hydropathy: acrophilicity.

METHODS

Hydrophilicity (or hydrophobicity; hydropathy) plotting methods all use the same principle. The amino acids of a protein are first assigned numerical values from a set of 20 parameters that correspond to the 20 amino acids. These values are then averaged down the length of the protein in order to derive a profile that expresses the average hydrophilicity or hydrophobicity of each segment of the protein being studied. Figure 1 shows such a profile for the histocompatibility antigen, IAβ^d. These plots are known to be useful in locating segments of secondary structure (Rose and Roy, 1980), antigenic sites (Hopp and Woods, 1981), and membrane associated regions (Kyte and Doolittle, 1982). Methods for predicting protein helices, strands and turns (Chou and Fasman, 1978), yield plots that are very similar to hydrophilicity plots. Recently, a method for predicting segmental mobility has been proposed (Karplus and Schulz, 1985), which also gives similar plots, because segmental mobility is directly correlated with hydrophilicity.

Although the different methods have been developed for a variety of purposes, a direct comparison of the profiles indicates that all methods yield essentially the same information regarding a protein (Hopp, 1986a). Some confusion has resulted because different authors have chosen to orient the hydrophobic end of their plots either at the top or at the bottom of their profiles. It has been suggested (Hopp, 1986b) that consistently placing the hydrophobic region at the bottoms of plots would make them easier to comprehend, because surface features would be correlated with regions where the profile trends upward (up = exposed), while interior segments and transmembrane regions would appear as low segments (down = buried). We have used that convention throughout the work presented below.

WINDOWS

The choice of an averaging group length, or "window" is an important factor in any of the methods. We demonstrated that a window of six residues performs optimally in locating protein antigenic sites (Hopp and Woods, 1981), with a five residue window giving the second best success rate. In a method published subsequently to ours, Kyte and Doolittle (1982) advocated

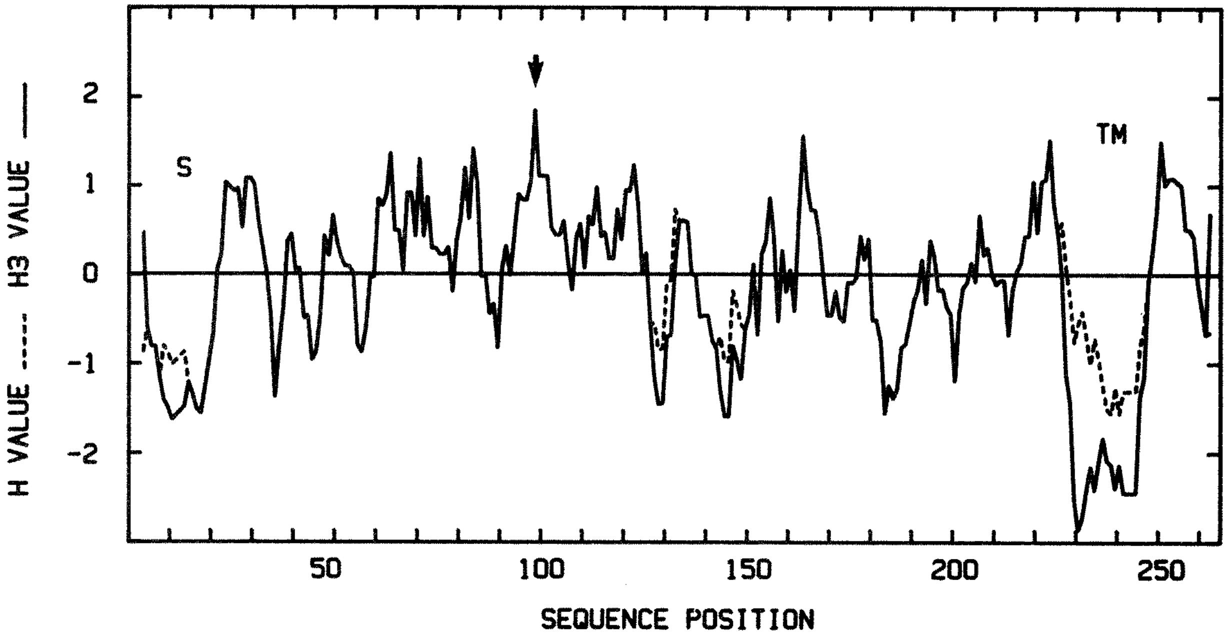

Fig. 1. Hydrophilicity profile for H-2 Class II histocompatibility antigen A^d β chain (Malissen et al., 1983). The solid profile was generated using the HYDRO3 procedure (Hopp, 1986b) that emphasizes membrane spanning regions. The dotted segments were generated using the original method (Hopp and Woods, 1981). S, signal peptide; TM, the transmembrane anchor segment; arrow, a major antigenic site (Brown et al., 1986). The valleys in the region between residues 28 and 220 probably represent the many β-strands of this immunoglobulin-like molecule.

windows of seven or more residues, but provided no experimental evidence to support this contention. We therefore reinvestigated the problem, this time using the independent criterion of secondary structure correlation with hydrophobic valleys. Because α-helices and β-strands comprise the overwhelming majority of segments that are packed in the cores of proteins (Rose and Roy, 1980) it is reasonable that any procedure that maximizes the agreement of hydrophobic valleys with these structures, must in turn be optimal in its correlation with the three dimensional structures of proteins.

Using a data set of 70 proteins for which both sequence and X-ray structure are known (Bernstein et al., 1977), we determined the correlation of valleys with helices and strands. Typical results are shown in figure 2. It is clear that each of these methods correlates best with secondary structure when a window of six is used. The next best window is five. Even the method of Kyte and Doolittle works best with a six residue window, and it is further clear that the longer windows that they advocated are less accurate. Similar results to those shown were obtained with all of the currently available hydrophilicity/hydrophobicity methods. The 15 methods tested included our hydrophilicity and acrophilicity procedures (Hopp, 1986b) as well as procedures developed by Chothia (1976), Janin (1979), von Heijne (1981), Sweet and Eisenberg (1983), Rose et al. (1985), Fauchere and Pliska (1983), and Welling et al. (1985) as well as the secondary structure parameters of Chou and Fasman (1976) and Garnier et al. (1978).

MEMBRANE ASSOCIATED SEQUENCES

Many investigators use long windows in order to emphasize transmembrane segments. However, this results in the loss of all other useful information from the plots. In order to emphasize transmembrane segments in plots that use a six residue window, we developed a new algorithm (Hopp, 1986b). This procedure identifies Gly, Ser and Thr residues that are likely to be buried, and lowers their hydrophilicity values. This results in a dramatic deepening of transmembrane segments and signal peptides, without any major changes to the remainder of the profile. As seen in figure 1, the membrane and signal sequences are readily identified, while the profile retains its usefulness in identifying an important antigenic site.

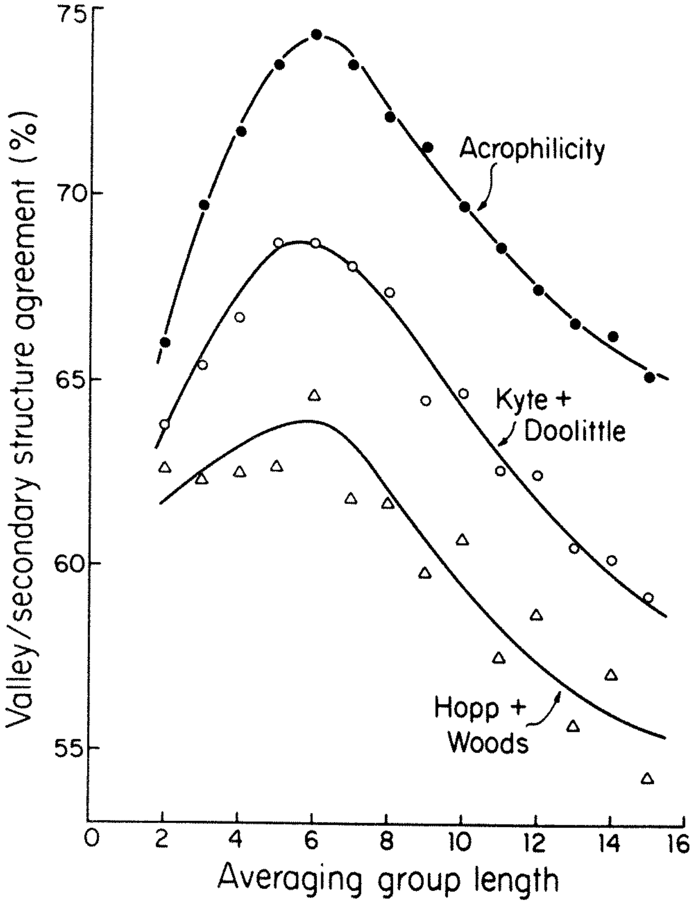

Fig. 2. Effect of window length on correlation with secondary structures. The length of the averaging group was varied from 2 to 15. The 70 protein data set was then used to compare the match of the lowest 20% of the plots (bottoms of valleys) with helices and strands. Each time a low average hydrophilicity value was located directly on a secondary structure residue (for odd averages) or between two such residues (even averages), a correct prediction was counted. The sum of correct predictions for the whole dataset was divided by the sum of all points in the lowest 20% of the plots, and multiplied by 100 to achieve the values shown in the figure. The amino acid scales used were: acrophilicity (Hopp, 1986a), hydropathy (Kyte and Doolittle, 1982), and hydrophilicity (Hopp and Woods, 1981).

ANTIGENIC SITES

We recently tested all available methods for their ability to predict protein antigenic sites (Hopp, 1986b). The original hydrophilicity values of Hopp and Woods (1981) were better for this purpose than any other procedure. We speculate that this results from the balanced values of the charged residues in our procedure, in contrast to the other methods, which usually favor positively charged residues over negatively charged residues. Furthermore, despite allegations to the contrary (Kyte and Doolittle, 1982), the water/ethanol free energy values of Nozaki and Tanford (1971) upon which our values are based, are probably a very accurate measure of the orientation of amino acid side chains in proteins.

OTHER INTERACTION SITES

Antigen-antibody interactions probably represent a special case of the more general interactions of proteins with other macromolecules. We have found that many other types of interactions occur at the most hydrophilic segments of proteins, as determined by the procedure of Hopp and Woods. I have previously reported that many cases of limited proteolysis occur at the most hydrophilic segments of proteins (Hopp, 1984). Post-translational modifications of proteins also occur with great frequency at the most hydrophilic sites. These include phosphorylation, acetylation and methylation reactions as well as many others; protein-nucleic acid interactions also occur at hydrophilic sites, and hormones often bind to their receptors via maximally hydrophilic regions (Hopp, 1984). In our experience, these important interaction sites are also best seen with a window of six. Therefore it appears that there is a need to retain this window for any protein sequence, while allowing the identification of signal and membrane anchor segments by using the Gly, Ser and Thr adjustments.

COMPUTER PROGRAMS

One facet of hydrophilicity plots that is not often mentioned is that they are easy to generate, compared to other types of protein conformational analysis. They require much less computer time than methods such as energy minimization or secondary structure predictions. In fact they can if necessary be generated by one or two hours of manual calculations. Furthermore, they typically yield a single line plot for a protein, allowing for easy analysis of the results.

However, even with such a simple system, variations between users can occur. Differences can result from the way that ends of proteins are treated, in the averaging group length (as described in this paper), in the interpretation of peak heights and sizes, and in special procedures such as the Gly, Ser and Thr adjustments. Methods for standardizing these aspects of hydrophilicity analysis have been described and expressed in published computer algorithms (Hopp and Woods, 1983 Hopp, 1986b). In addition, a variety of commercial sources provide hydrophilicity analysis as part of their computer software packages. However, because none of these organizations provides an authorized version, they should be used with some caution at this time.

REFERENCES

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977). J. Mol. Biol. 112, 535-542.

Brown, M.A., Glimcher, L.A., Nielsen, E.A., Paul, W.E. and Germain, R.N. (1986). Science 231, 255-258.

Chothia, C. (1976). J. Mol. Biol. 105, 1-14.

Chou, P.Y. and Fasman, G.D. (1978). Adv. Enzymol. 47, 45-148.

Fauchere, J.L. and Pliska, V. (1983). Eur. J. Med. Chem.-Chim. Ther. 18, 369-375.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978). J. Mol. Biol. 120, 97-120.

Hopp, T.P. (1984). Ann. Sclavo 2, 47-60.

Hopp, T.P. (1986a). In: Modern methods in protein chemistry, Vol. 1, J.J. L'Italien, ed., Plenum Press, New York, in press.

Hopp, T.P. (1986b). J. Immunol. Methods, in press.

Hopp, T.P. and Woods, K.R. (1981). Proc. Natl. Acad. Sci. USA 78, 3824-3828.

Hopp, T.P. and Woods, K.R. (1983). Mol. Immunol. 20, 483-489.

Janin, J. (1979). Nature (Lond.) 277, 491-492.

Karplus, P.A. and Schulz, G.E. (1985). Naturwissenschaften 72, 212-213.

Kyte J. and Doolittle, R.F. (1982). J. Mol. Biol. 157, 105-132.

Malissen, M., Hunkapiller, T. and Hood, L. (1983). Science 221, 750-754.

Nozaki, Y. and Tanford, C. (1971). J. Biol. Chem. 246, 2211-2217.

Rose, G. and Roy, S. (1980). Proc. Natl. Acad. Sci. USA 77, 4643-4647.

Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H. and Zehfus, M.H. (1985a). Science 229, 834-838.

Sweet, R.M. and Eisenberg, D. (1983). J. Mol. Biol. 171, 479-488.

Von Heijne, G. (1981). Eur. J. Biochem. 116, 419-422.

Welling, G.W., Weijer, W.J., Zee, R.v.d. and Welling-Wester, S. (1985). FEBS. Ltrs 188, 215-218.