Synthetic Peptides in Biology and Medicine (1985) 3-12
K. Alitalo, P. Partanen and A. Vaheri editors http://thomas-hopp.com/THscience2.html

PREDICTION OF PROTEIN SURFACES AND INTERACTION SITES FROM AMINO ACID SEQUENCES

THOMAS P. HOPP

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

The peptide chemist is often faced with a difficult problem: given only the amino acid sequence of a protein, how can one determine the critical sites for synthetic peptide studies? Where are the most immunogenic sites on a protein antigen? Which segments of a protein hormone are involved in receptor binding? If large quantities of the protein are available, these questions can be answered by X-ray crystallography or by chemically dissecting the molecule. However, many interesting antigens and hormones are available in very small quantities and in impure preparations. Often, through molecular cloning, an amino acid sequence is determined from nucleotide sequence data, making it possible to develop chemically synthesized peptides that possess the desired antigenicity or hormone activity. However, because antigenic sites and other interaction sites usually comprise only a minor portion of a given protein, it is essential to limit the amount of experimentation required, by selecting segments of the protein that are most important for antigenicity or other interactions. To this end, we have developed a method of computerized protein topological analysis that relates amino acid sequence to the distribution of surface oriented or buried portions of a protein, and selects the segments most likely for interactions with other proteins.

This analysis, which we call PROTO (for PROtein TOpology), is based on a simple averaging algorithm that requires very little computer time, but yields a surprising wealth of information about a protein's structure and interactions. In our procedure, each amino acid in a sequence is assigned two values, an acrophilicity value, and a hydrophilicity value (Table 1). When the acrophilicity values are averaged in groups of six, they yield an acrophilicity profile for the protein (see Figure 1). The hydrophilicity values may also be averaged to yield a similar profile (Figure 2). These two profiles are related, but emphasize two different aspects of the protein sequence. The acrophilicity profile is an accurate representation of the degree of surface exposure along a polypeptide chain, while the hydrophilicity profile indicates the locations of important interaction sites (antibody binding sites, receptor binding sites, proteolysis sites, etc.). By combining the information from the two profiles, it is possible to begin to understand the critical active sites of a protein, and their structural contexts as well.

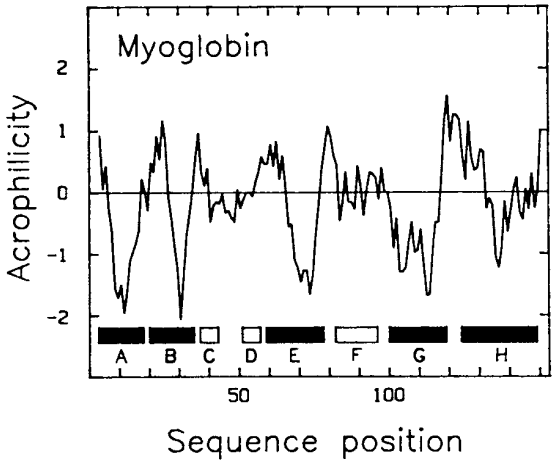

Fig. 1. Acrophilicity profile for myoglobin. The bars lettered A through H represent the 8 helices of myoglobin. The peaks on the acrophilicity profile occur between the helices and at their highly exposed ends. The five major acrophobic valleys are associated with the five largest helices of myoglobin (dark bars) that, together, constitute the core of the molecule. The three shorter helices C, D, and F (light bars) are not as tightly associated with the center of the molecule, and are correspondingly less acrophobic.

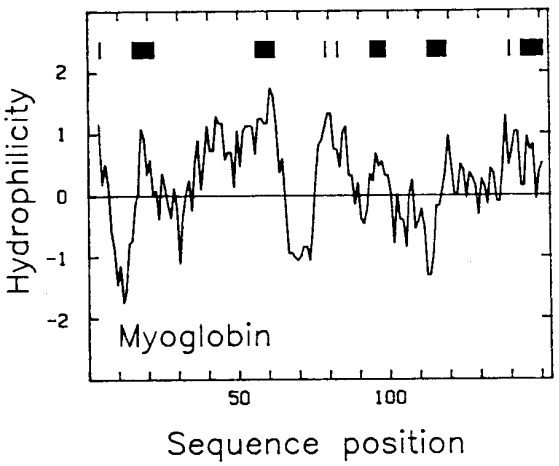

Fig. 2. Hydrophilicity profile for myoglobin. The locations of antigenic sites are indicated at the top of the profile. Vertical lines represent single antigenic residues; bars represent groups of contiguous antigenic residues. All of the largest hydrophilicity peaks have been associated with important antigenic sites.

ACROPHILICITY

The term acrophilicity (literally, "height-loving") refers to the frequency of occurrence of the amino acids in highly exposed, protruding portions of the folded structures of proteins. The acrophilicity scale (Column 1, Table 1) was determined by analysis of 49 protein X-ray structures, to find all protruding regions, then identifying the amino acids present at the apex of each protrusion

(T.P. Hopp and J.E. Merriam, submitted). The resulting scale is similar to the β-bend scale of Chou & Fasman(1) but is more successful in our averaging procedure, probably because it is not limited to β-bends or any other secondary structure.

TABLE 1. COMPARISON OF ACROPHILICITY AND HYDROPHILICITY VALUES

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

Acrophilicity profiles contain within them an unexpectedly large amount of useful information that becomes apparent when they are correctly interpreted. The peaks on the profiles represent the most highly exposed projections of proteins, as expected. In addition, we have observed that the lowest valleys are almost always found in the core of a protein. Furthermore, these "acrophobic" segments usually identify β-stranded or α-helical segments of a protein (Fig. 1, Fig. 3). Although the profile cannot indicate which of these two secondary structures is present, it seems sufficient to know that an acrophobic segment must be in the packed core of a molecule. Interestingly the ends of helices and the end-strands of β-sheets are usually *acrophilic.* This is appropriate, because these are usually high relief features of proteins.

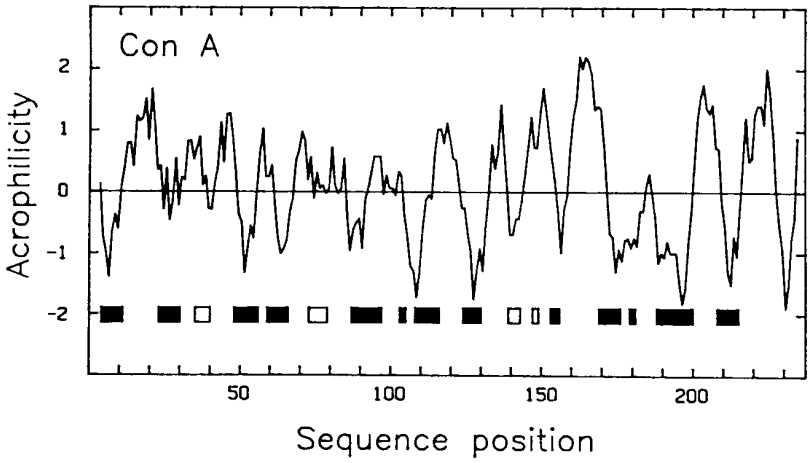

Fig. 3. Acrophilicity profile for concanavalin A. The 3-dimensional structure of this protein incorporates two large β-pleated sheets. The individual strands of these sheets are indicated by the bars below the acrophilicity plot. Most of the internal strands of the sheets (dark bars) have deep acrophobic valleys associated with them. The edge strands (light bars) are further from the core and are less acrophobic. The most acrophilic regions comprise highly exposed portions of the folded structure that connect the β-strands.

Acrophilicity profiles are also capable of identifying membrane-spanning segments of proteins. Thus, both signal peptide and transmembrane "anchor" segments of membrane proteins are clearly seen as long, low regions of the profiles. These membrane-spanning acrophobic valleys can be enhanced by giving special treatment to Gly and Ser residues occurring in them. The PROTO program lowers the values of Gly and Ser when they occur in likely membrane-spanning regions, but not in other parts of a protein. Using this procedure, we consistently get longer, lower membrane-spanning valleys than is possible by other methods, including the "hydropathy" method of Kyte and Doolittle (2), thus improving the reliability of our analysis.

HYDROPHILICITY

Ever since Tanford established the notion of amino acid hydrophobicity (3), it has been recognized as a major contributing factor to the folding patterns of proteins. There is a tendency for hydrophobic amino acids to be buried inside a protein, away from contact with water, while hydrophilic amino acids coat the surface of a protein. However, attempts to predict protein 3-dimensional structures based on hydrophobicity / hydrophilicity have been of limited usefulness (2,4,5). This is probably due to the fact that such methods generally ignore the ability of partially buried amino acids to extend their side chains inward or outward, depending on their hydrophilic or hydrophobic nature. For this reason, the hydrophilicity analysis in PROTO is not used for 3-dimensional information, but only to locate the subset of surface sites that have concentrations of charged and polar amino acids. It is these sites that were shown to be the most immunogenic parts of proteins (6) as can be seen in Figure 2. We have recently found that these are also the most likely sites for other types of protein interactions (T.P. Hopp and K.R. Woods, in

preparation).

Our ongoing survey of proteins indicates strongly that the most hydrophilic sites on a protein are the preferred locations for a number of types of protein binding sites and reactive sites (7). Interaction sites correlated with hydrophilicity include: phosphorylation sites (Ser, Thr, Tyr), acetylation sites, glycosylation sites, sites of limited proteolysis and sulfation sites and probably many other sites where one protein serves as a substrate for another (an example of proteolysis at hydrophilic sites is given in Figure 4). Proteins often bind other macromolecules via their most hydrophilic segments, even when no direct catalytic activity is involved. For example, apolipoprotein E is bound by its cellular receptor by its most hydrophilic segment. A mutation there causes type III hyperlipoproteinemia. Immunoglobulin binds complement at its most hydrophilic site. DNA polymerase binds DNA at its most hydrophilic segment. These, and many other examples not cited here, emphasize that hydrophilicity analysis as carried out by PROTO is potentially very useful in finding critical interaction sites in the sequences of proteins.

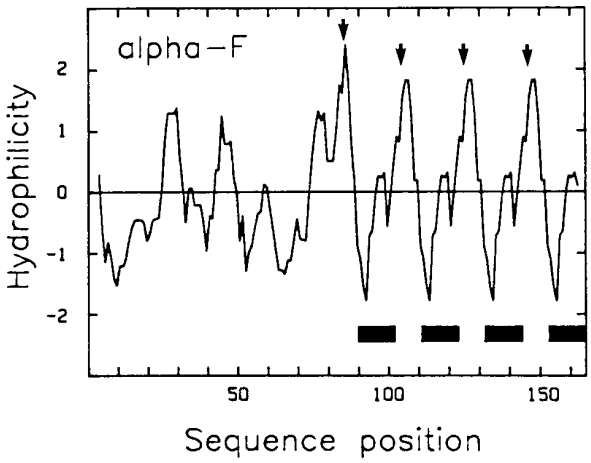

Fig. 4. Hydrophilicity profile for preproalpha-factor. Four tandem copies of the yeast mating pheromone, alpha-factor, occur at the C-terminus of the primary translation product, preproalpha-factor. These mature hormone segments (bars) are released from the precursor by proteolytic processing that is initiated by cleavage at the hydrophilic sites between the copies of the hormone (arrows).

A DISTINCTION BETWEEN ACROPHILICITY AND HYDROPHILICITY

We initially developed the acrophilicity scale in an attempt to improve our ability to identify antigenic sites and other sites of protein interactions. We were therefore surprised that the acrophilicity method, which is better at locating highly exposed sites, is less accurate in locating protein interaction sites than the hydrophilicity method. This probably reflects the fact that bonding between charged and polar amino acids are major sources of binding energy when proteins interact with each other, while other, less well characterized forces dictate the folding of an individual protein. One effect that may be important to protein folding is apparent in the

acrophilicity scale. As is seen in Table 1, acrophilicity is, with minor exceptions, a *size* scale. Glycine, the smallest amino acid, is most often highly exposed on proteins, and tryptophan, the largest, is most often buried. While it is likely that the greater hydrophobicity of the larger amino acids plays some part in this, the simple size correlation also may be important. This is apparent when members of groups of similar amino acids are considered, for example, among the charged amino acids (Asp, Lys, Glu, Arg). The smallest, Asp, is the most acrophilic while the largest, Arg, is the most acrophobic. It seems possible that protein stability may depend on packing of large side chains in low relief and internal regions while smaller side chains are more appropriate where the main chain loops outward into highly exposed segments. Regardless of the answer to this question, it is probable that the distinction between acrophilicity and hydrophilicity in some way reflects the different forces that dictate protein folding and protein interactions.

In light of the foregoing discussion we have begun to use the PROTO program to generate plots like Figure 5, where profiles for both acrophilicity and hydrophilicity are presented. In such a plot, it is possible to identify a large number of features, including probable surface sites and interaction sites, as well as the signal and transmembrane segments. With these capabilities, the PROTO program should be a most useful source of information for identifying important regions of proteins to be studied by chemical peptide synthesis.

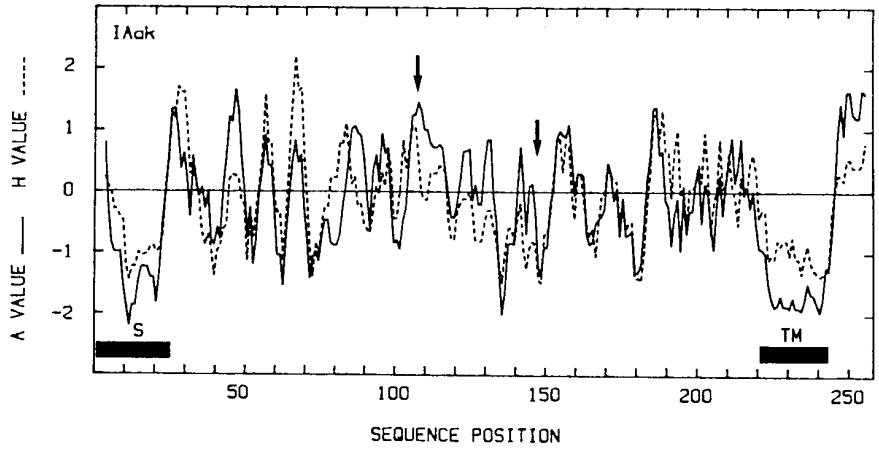

Fig. 5. PROTO analysis of the murine class II major histo-compatibility antigen IAα^K chain. The solid line is the acrophilicity profile; the dashed line is the hydrophilicity profile. The widest and lowest acrophobic valleys correspond to the signal peptide (S) and transmembrane segment (TM). The other acrophobic valleys probably represent the internal β-strands of this immunoglobulin-like molecule. The highest peaks on the hydrophilicity profile, at positions 25-30, 54-59 and 64-69, probably represent sites of major antigenicity, and would be considered as likely candidate regions for participating in interactions with other molecules to achieve the immunoregulatory functions of the IA system. The arrows indicate the two sites of N-linked glycosylation. Of these, the first is considerably more hydrophilic than the second. Consistent with this, it has recently been demonstrated that the first site is more extensively glycosylated than is the second (8).

REFERENCES

1. Chou PY, Fasman GD (1978) Adv Enzymol 47:45-148.
2. Kyte J, Doolittle RF (1982) J Mol Biol 157:105-132.
3. Nozaki Y, Tanford C (1971) J Biol Chem 246:2211-2217.
4. Rose G, Roy S (1980) Proc Natl Acad Sci USA 77:4643-4647.
5. Chothia C (1976) J Mol Biol 105:1-14.
6. Hopp T, Woods K (1981) Proc Natl Acad Sci USA 78:3824-3828.
7. Hopp TP (1985) In: Rein R (ed) Molecular Basis of Cancer Part B: Macromolecular Recognition, Chemotherapy and Immunology. Alan R Liss Inc. New York, pp 367-377.
8. Swiedler SJ, Freed JH, Tarentino AL, Plummer TH, Hart GW (1985) J Biol Chem 260:4046-4054.

APPENDIX

Derivation of the Acrophilicity Values

The 49 protein crystalographic structures used in this study were: adenylate kinase, alcohol dehydrogenase, alpha lytic protease, carbonic anhydrase c, carboxypeptidase A, tobacco mosaic virus coat protein, cobratoxin, concanavalin A, cytochrome c, elastase, erabutoxin B, ferredoxin, flavodoxin, glucagon, influenza hemagglutinin (HA1 and HA2), hemerythrin, hemoglobin β chain, Bence Jones IgG dimer, IgG (New) heavy and light chains, IgG fc fragment, lysozyme, lysozyme T4, muscle calcium binding protein, myoglobin, myohemerythrin, staphylococcal nuclease, papain, penicillopepsin, phospholipase, phosphorylase, plastocyanin, prealbumin, proinsulin, protease b, rhodanese, ribonuclease, ribosomal protein L7/L12, rubredoxin, subtilisin, subtilisin inhibitor, superoxide dismutase, thermolysin, thioredoxin, triose phosphate isomerase, trypsin inhibitors (bovine and soybean), and trypsinogen; available from the Protein Data Bank (Brookhaven National Laboratory, Upton, NY 11973).

TABLE 2. DERIVATION OF ACROPHILICITY VALUES

Amino Acid	Number in exposed positions	Total number in sample	Frequency(f) in exposed positions	Normalized acrophilicity value (A)
Gly	161	818	0.195	3.0
Pro	73	399	0.182	2.6
Asn	80	457	0.174	2.3
Asp	83	496	0.166	2.1
Ser	114	729	0.155	1.8
Lys	82	567	0.144	1.4
Glu	54	471	0.114	0.5
Arg	38	354	0.107	0.3
Thr	57	590	0.096	-0.1
Gln	31	340	0.091	-0.2
His	18	207	0.086	-0.4
Ala	58	697	0.083	-0.5
Val	30	706	0.042	-1.7
Met	5	122	0.041	-1.8
Tyr	11	350	0.031	-2.0
Leu	10	655	0.015	-2.5
Ile	7	457	0.015	-2.5
Cys	3	213	0.014	-2.6
Phe	3	348	0.009	-2.7
Trp	0	138	0.000	-3.0
Totals	918	9114		

Stereo pair images of α-carbon drawings were visually inspected to identify all protruding segments of the peptide chain. No side chain information was used. The α-carbon at the apex of each chain protrusion was selected and its number logged. A total of 918 highly exposed amino acids were selected from the total of 9114 amino acids in the 49 proteins. The identities of the amino acids at the selected positions were then determined and summed to yield the numbers in Column 1 of Table 2. The ratio of the number in column 1 to the total occurrences of a given amino acid (Column 2) yields the frequency of occurrence of each amino acid in highly exposed positions (f).

These frequency values (Column 3) were then normalized according to the following equation: $A = 30.77x(f) - 3.0$ to yield the final acrophilicity scale (Column 4).