Molecular Basis of Cancer, Part B: Macromolecular Recognition,
Chemotherapy, and Immunology 1985 R. Rein, ed. pp 367-377
http://thomas-hopp.com/THscience2.html

COMPUTER PREDICTION OF PROTEIN SURFACE FEATURES AND ANTIGENIC DETERMINANTS

Thomas Hopp, Ph.D.

Immunex Corporation

51 University Street, Seattle, WA 98101

For many years it has been appreciated that the arrangement of amino acids in the linear sequence of a protein is responsible for the three dimensional structure of the folded protein. However, until recently very little practical information could be obtained from amino acid sequences, because of our imperfect understanding of the way that the individual amino acids influence the conformation of a peptide chain. At present there are several useful ways of predicting conformation from sequence, including methods based on energy minimization, frequency of occurrence of amino acids in particular secondary structures, and consideration of the solubility of the amino acids in aqueous and organic solvents. These methods seldom generate information that is useful on a practical level, partly because they attempt to predict too much detailed information from a given sequence. In the development of my method, I asked a simpler question, namely, is it possible to predict the locations of antibody binding sites on a protein sequence, regardless of any consideration of the precise conformation of the peptide chain? Using this criterion and a data set of twelve well-known protein antigens, I developed a simple hydrophilicity analysis that reliably predicts the locations of antigenic residues in protein sequences. A listing of experiments where the outcome was predictable by my method is presented in Table I.

TABLE I. Protein Surface Features Predicted by Hydrophilicity Analysis

A. Antigenic Determinants

1. Influenza hemagglutinin: Site #1 synthetic peptide immunogen is protective in mice. Muller et al., PNAS 79, 569 (1982).
2. Influenza hemagglutinin: Sites #1, 2, and 3 contain antigenically important amino acids. Wiley et al., Nature 289, 373 (1981).
3. Influenza hemagglutinin: Sites #1 and 3 synthetic peptide immunogens are protective in mice. Shapira et al., PNAS 81, 2461 (1984).
4. Influenza hemagglutinin: Site #1 (X31 strain) is contained in a synthetic peptide immunogen recognized by T-cells. Lamb et al., Nature 300, 66 (1982).
5. Streptococcal M protein: Site #1 synthetic peptide immunogen is protective. Beachey et al., PNAS 81, 2203 (1984).
6. Poliovirus VP1: Sites #1, 3, and 5 synthetic peptide immunogens stimulate neutralizing sera. Emini et al., Nature 304, 699 (1983).

7. Poliovirus VP1: Site #3 is a neutralizing epitope. Evans et al., Nature 304, 459 (1983).
8. Poliovirus VP1: Site #3 synthetic peptide reacts with neutralizing antibodies. Wychowski et al., EMBO J. 2, 2019 (1983).
9. Foot and mouth disease virus VP1: Sites #1 and 3 of A24 strain synthetic peptide immunogens raise neutralizing antisera. Bittle et al., Nature 298, 30 (1982).
10. Foot and mouth disease virus VP1: Site #3 synthetic peptide immunogen raises neutralizing antisera. Pfaff et al., EMBO J. 1, 869 (1982).
11. Hepatitis B surface antigen: Site #1 reacts with antisera raised against HBsAg. Hopp and Woods, PNAS 78, 3824 (1981).
12. Hepatitis B surface antigen: Site #1 synthetic peptides are immunogenic in mice. Prince et al., PNAS 79, 579 (1982).
13. Hepatitis B surface antigen: Site #1 synthetic peptide is immunogenic. Bhatnagar et al., PNAS 79, 4400 (1982).
14. Hepatitis B surface antigen: Sites #2, 4, and 5 are contained in synthetic peptides that stimulate precipitating sera to HBsAg. Lerner et al, PNAS 78, 3403 (1981).
15. Hepatitis B surface antigen: Site #3 contained in synthetic peptide that stimulates anti-subtype antibodies. Dreesman et al., Nature 295, 158 (1982).
16. Hepatitis B surface antigen: Site #3 synthetic peptide immunogen is partially protective. Gerin et al., PNAS 80, 2365 (1983).
17. Influenza neuraminidase: Sites #1 through 4 all contain antigenically important residues or are immediately adjacent to them. Colman et al., Nature 303, 41 (1983).
18. Ragweed allergen protein RA5: Site #1 is an important allergenic determinant. Roebber et al., J. Allergy and Clin. Immunol. 71, 162 (1983).
19. Herpes virus gpD: Site #3 of the extracytoplasmic portion, synthetic peptide raises neutralizing antisera. Cohen et al., J. Virology 49, 102 (1984).
20. Histocompatibility antigen H2 K^b: Site #1 in second domain is the "gene conversion" site causing the antigenic specificity change in the H2 K^{bm1} mutant. Schulze, et al., PNAS 80, 2007 (1983).
21. Histocompatibility antigen HLA B7: Site #2 in second domain is an alloantigenic site. Lopez de Castro et al., Biochem 22, 3961 (1983).
22. Histocompatibility antigen DR alpha chain: Site #2 synthetic immunogen used to make hybridoma. Niman et al., PNAS 80, 4949 (1983).
23. Histocompatibility antigen DR beta chain: Site #2 synthetic immunogen used to make hybridoma. Niman et al., PNAS 80, 4949 (1983).
24. Beta 2 microglobulin: Site #2 recognized by a monoclonal antibody. Parham et al., J.B.C. 258, 6179 (1983).
25. Myelin basic protein: Site #1 is an encephalitogenic determinant. Hashim, Immunol. Rev. 39, 60 (1978).
26. Scorpion toxin II: Sites #1 and 2 are antigenic. Granier et al., Int. J. Peptide Protein Res. 23, 187 (1984).
27. Immunoglobulin gamma chain: Site #1 of third constant domain is G1m (a) allotypic marker. Kehoe and Kehoe, Immunochemistry of Proteins 3, 87 (1979).
28. Interferon alpha 1: Site #1 synthetic peptide raises antibody to interferon. Arnheiter et al., PNAS 80, 2539 (1983).
29. Interleukin 2: Site #1 synthetic peptide raises antibody to Interleukin 2: Altman et al., PNAS 81, 2176 (1984).
30. Myoglobin: Site #1 synthetic peptide causes production of macrophage inhibitory factor (MIF) by cultured lymph node cells. Stavitsky et al., Immunochem. 12, 959 (1975).
31. Myoglobin: Site #1 synthetic peptide used to raise a hybridoma antibody. Schmitz et al., Molec. Imm. 20, 719 (1983).
32. Cytochrome c: Site #1 causes delayed type hypersensitivity and T-cell transformation. Wang and Reichlin, Molec. Imm. 16, 805 (1979).
33. Metallothionein: Sites #1 and 3 are autoimmune antigenic sites. Winge and Garvey, PNAS 80, 2472 (1983).
34. Rous sarcoma virus transforming protein (src): Site #2 synthetic peptide immunogen raises

sera that neutralize tyrosine kinase activity, cross react with yes-transforming protein (where it is Site #1) and precipitate possible cellular analogs. Gentry et al., J.B.C. 258, 11219 (1983).

35. Polyoma virus middle T antigen transforming protein: Site #1 synthetic peptide immunogen raises sera that react with middle T as well as a normal cellular protein. Ito et al., J. Virology 48, 709 (1983).

B. Interaction Sites

36. Immunoglobulin gamma chain: Site #1 of second constant domain is the C1q binding site. Prystowsky et al., Biochem. 20, 6349 (1981).
37. Calmodulin: Sites #3 and 4 are calcium binding sites. Waterson et al., J.B.C. 255, 962 (1980); Sasagawa et al., Biochem. 21, 2565 (1982).
38. Influenza hemagglutinin: Site #1 of X31 strain is the proteolytic processing site for cell fusion activity. Hopp and Woods, Molec. Immunol. 20, 483 (1983).
39. Fibronectin: Site #3 synthetic peptide has the cell binding activity of the whole molecule. Pierschbacher and Ruoslahti, Nature 309, 30 (1984).
40. Hepatitis B surface antigen: Site #1 contains the asparagine residue that is preferentially glycosylated over other asparagines. Peterson, J.B.C. 256, 6975 (1981).
41. Histocompatibility antigens HLA B7 and H2 K^b: Site #1 in the cytoplasmic domain contains the phosphorylatable threonine or serine residue. Pober et al., PNAS 75, 6002 (1978); Bregegere, et al., Nature 292, 78 (1981).
42. Polyoma virus middle T antigen transforming protein: Site #1 is immediately adjacent to tyrosine 315, which is phosphorylated. Hunter et al., EMBO J. 3, 73 (1984).

C. Miscellaneous

43. Fava bean lectin: Site #1 of the beta chain is the location of the annealing site for circular permutation of the genes for favin and Con A. Cunningham et al., PNAS 76, 3218 (1979).

It is clear from the number of entries in this list and the variety of the objectives achieved, that my prediction method is highly successful and has broad applicability to problems in immunology as well as the field of general protein chemistry. While no attempt was made to present an exhaustive list, the examples cited here demonstrate the potential for using hydrophilicity analysis in many areas, including the elucidation of the antigenic structures of pathological organisms. The examples in Table I include all of the well characterized major disease organisms presently under investigation. The predictable antigenic sites have been found to possess the full range of known immunological phenomena, including antibody production (precipitins, neutralizing and protective sera), delayed-type hypersensitivity, allergic responses, autoimmunity, encephalitic responses, T-cell proliferative responses, graft rejection, and lymphokine production. In addition, many unexpected examples of other protein surface sites have been listed. These include protein-protein interaction sites such as the complement binding region of immunoglobulin, the protein-cell surface interaction site of fibronectin and the protein-metal interaction sites of calmodulin. Other interaction sites comprise locations of post-translational modification of peptide chains, including proteolytic processing sites, and sites of phosphorylation and carbohydrate attachment.

When the wealth of information listed above is considered, it is clear that hydrophilicity analysis should be extremely useful in the analysis of molecular phenomena related to carcinogenesis. In particular, there has become available a huge body of sequence information on the proteins involved in transformation, both in spontaneously occurring tumors and in virally induced tumors. These sequences, as well as the genomic sequences of a variety of oncogenic viruses have been obtained for the most part by nucleic acid sequencing, and therefore little or nothing is known about the structures of the proteins produced from the sequences. In this regard it is instructive to consider several examples of experiments suggested by hydrophilicity analysis of tumor virus genome encoded proteins.

The first example is an investigation of the hydrophilicity properties of the two protein products of the env region of the recently described adult thymic leukemia virus (ATLV) genome. In Figure 1, the hydrophilicity profile for the heavy chain of the env translation product is compared to two other viral envelope proteins, those of the influenza virus (HA1) and the hepatitis B virus (HBsAg). This comparison seems especially interesting because it has been noted that the heavy and light chains of the retroviral env translation products bear a general resemblance to the two chains of the influenza hemagglutinin (HA1 and HA2). A comparison of the ATLV env light chain to HA2 and to a membrane glycoprotein product of the early region of adenovirus (ADVE16) is shown in Figure 2. The most striking similarity among these plots is between the HA1 and env1 heavy chains. A number of common features lead me to conclude that these two proteins share closely similar three-dimensional structures, even though they have not been reported to be homologous.

The HA1 and env1 chains are obviously similar in length, although env1 is slightly shorter. More significantly, the hydrophilicity profiles show a great number of common characteristics. Each has a broad hydrophobic valley near the N-terminus. This region of HA1 makes up the central strand of the membrane associated globular domain. At their C-termini, the two proteins again show a similar feature in the large terminal peak. This feature is associated with the known proteolytic processing site of the influenza hemagglutinin and the proposed processing site of the env product. The profile for HBsAg is included in Figure 1 to show that not all envelope proteins share such hydrophilicity profiles. While HBsAg does have an N-terminal hydrophobic valley, it clearly lacks a C-terminal peak. It is also obviously different in being much shorter than HA1 or env1, and in having a broad central hydrophobic valley that may actually be a membrane-spanning segment. Although HA1 and env1 both have a number of hydrophobic valleys in their central regions, neither has one of sufficient length to span a membrane. This central region of the HA1 chain is known to comprise the globular domain at the distal end of the hemagglutinin spike and to contain the binding site for cell-surface sialic acid. The core of this cell-binding domain is formed into an eight stranded β pleated sheet with several associated short helical stretches. These repeating secondary structures are reflected in the hydrophilicity profile by the series of large peaks and valleys in the central part of the plot. Most of the valleys are related to the different β strands, while the peaks represent the highly exposed chain turns at the ends of the strands. In this light, it is most interesting to note that the env1 hydrophilicity plot also shows a series of large peaks and valleys in its central region, suggesting that it, too contains a globular domain composed of repeating β-strands. It is not possible to align all of the peaks and valleys of HA1 and env1, so there must be significant differences between the two proteins as well. It is not possible, from these findings, to make a case in favor of any homology or evolutionary relationship between these two proteins, however, given the number of general similarities, it would not be surprising if they had indeed evolved from a common ancestor. Most significantly, if these two proteins do share similar overall three-dimensional structures, then important antigenic determinants for the neutralization of ATL virus must be located on the hydrophilic peaks found throughout the central portion of the env1 protein. In particular, amino acids 90-95, 142-147, and 230-235, comprising the second, third, and fourth highest peaks for env1 should be explored for usefulness as synthetic peptide vaccines against this disease. Past experience indicates that the homologous portions of the env products of the related AIDS associated viruses should also be considered as vaccines against that disease.

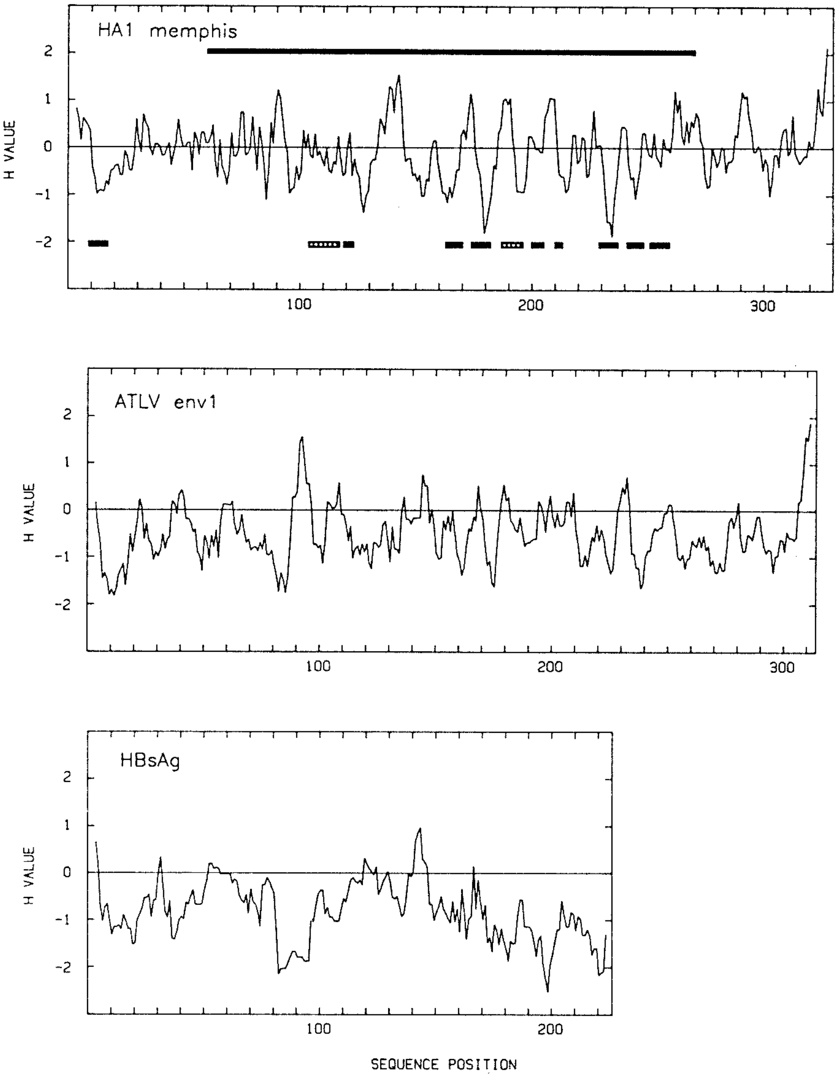

Figure 1. Hydrophilicity analysis of viral surface antigens. The bar above the profile for HA1 indicates the extent of the globular cell-binding domain. Below the profile, the solid bars represent β-strands and the hatched bars represent helices.

In Figure 2, several similarities are apparent between env2 and HA2. Each protein has an N-terminal hydrophobic valley, which is associated with the membrane fusion activity in HA2. Both proteins have long hydrophobic stretches near their C-termini, likely to be membrane-anchoring regions. It is also clear that there may be substantial differences between the two, because HA2 is significantly longer, and its profile shows more of the short-period spikiness associated with the large helix content of the molecule. The shorter env2 protein may have less of this helix, although it does contain some of the short-period spikes in addition to the pronounced peaks and valleys that may imply a greater content of β-strands. Interestingly, a more convincing structural similarity is suggested by comparing the plot for env2 to that for adenovirus E16. Although the E16 membrane protein has not been implicated as a viral structural protein, it is intriguing to note the apparent similarity of the hydrophilicity plots for membrane-associated proteins of two very different types of oncogenic viruses.

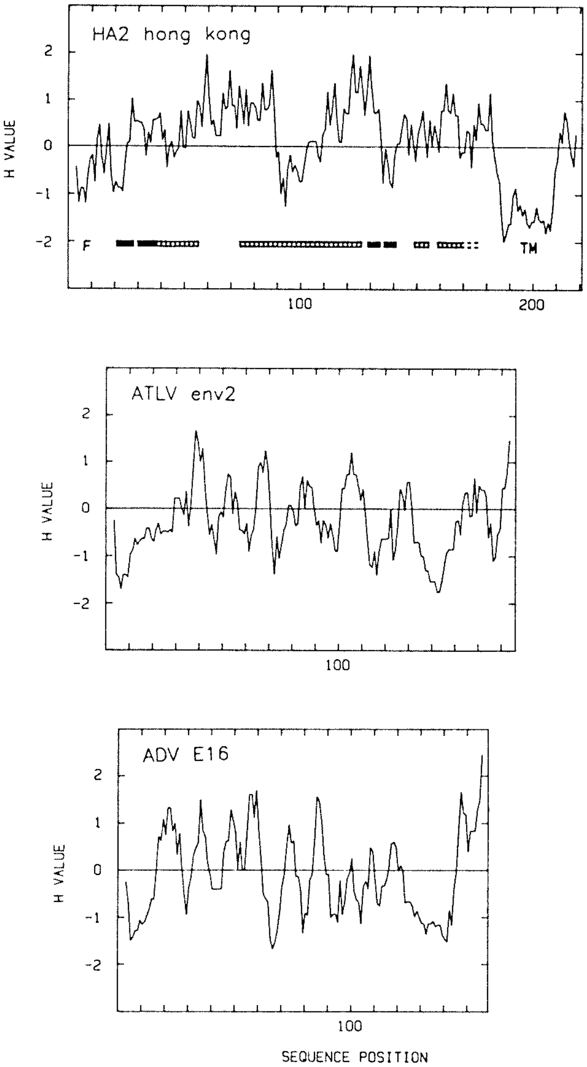

Figure 2. Hydrophilicity analysis of surface antigen light chains. In the HA2 plot, solid bars represent β-strands and hatched bars represent helices; F, membrane fusion region; TM, transmembrane anchoring segment.

Examples of hydrophilicity analysis of another important family of cancer related proteins are shown in Figure 3. The two retroviral oncogene products, src and erbB each contain a region of homology to each other, to protein kinase, and to the cytoplasmic portion of epidermal growth factor receptor. Antibodies raised against a synthetic peptide comprising the second highest prediction peak for src (residues 498 to 512) have been shown to neutralize its tyrosine kinase activity, and to precipitate the homologous yes transforming protein as well as the normal cellular analog of src. It would be interesting to raise antibodies specific to the highest peak for src (residues 155 to 160) because this region of the molecule is not homologous to the erbB product or EGF receptor. Such an antiserum should be useful in characterizing the function of this region of the src molecule because it should cross react strongly with the recently characterized cellular c-src gene product, which is identical in this region, but may be non-cross reactive with yes, which has many amino acid substitutions at this site. It is likely that studies of the antigenic peptides predicted for the erbB protein would also be quite useful, because predicted site #1 has been proposed as a major site of tyrosine phosphorylation in erbB, and site #3 is homologous to the tyrosine phosphorylation site in src. Antibodies to these two sites would clarify the role of tyrosine phosphorylation in erbB, and, because the amino acids around site #1 are highly conserved between erbB and EGF receptor, the same antibodies should cross-react with the receptor as well. Another region of interest on erbB is at the N-terminus, where the first 60-70 residues have been proposed to reside on the outside of the cell plasma membrane. Antibodies specific to this region could be raised using synthetic peptides comprising one or more of the hydrophilic regions found there. Using these antibodies and the ones described above it should be possible to begin a molecular dissection of this important group of transforming proteins.

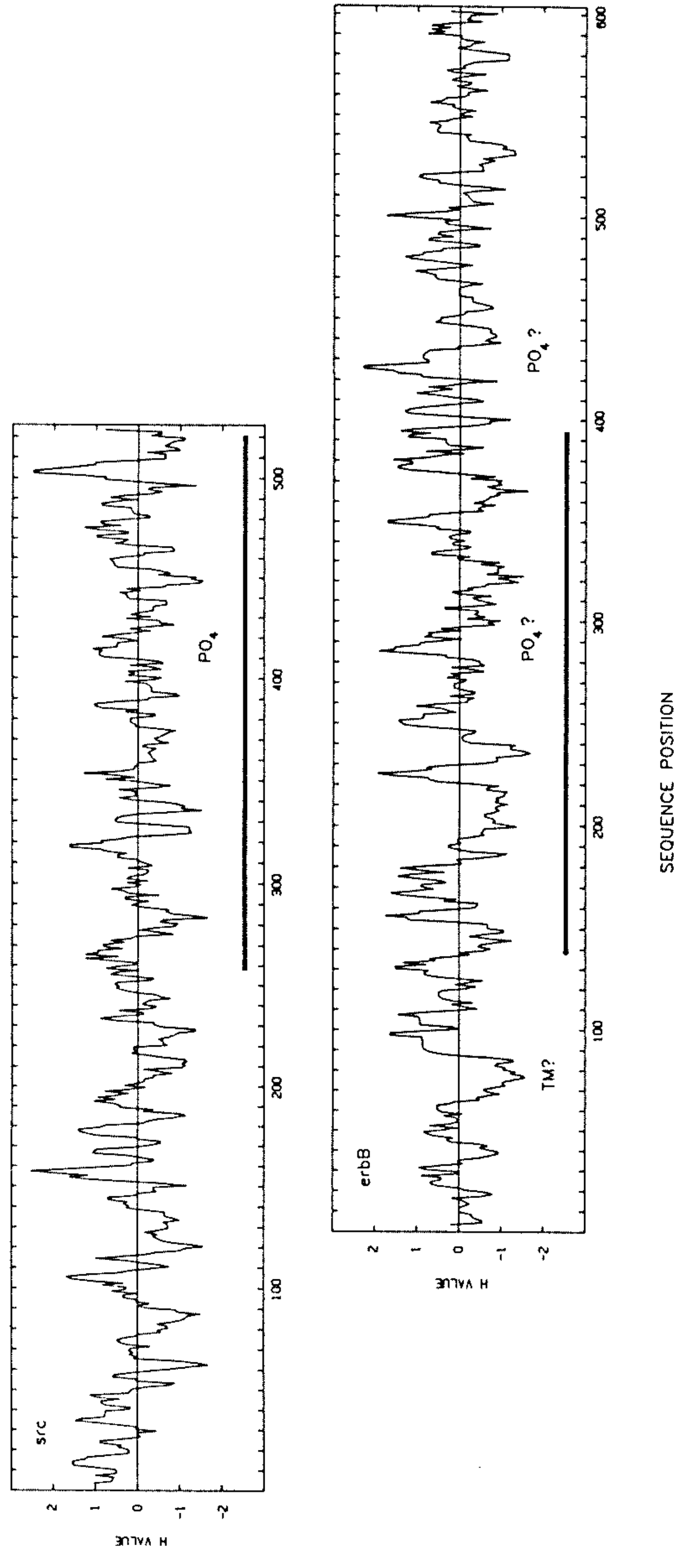

Figure 3. Hydrophilicity analysis of src and erbB oncogene proteins. The horizontal bars represent regions of shared homology between src, erbB, protein kinase, and EGF receptor. PO_4 indicates a site of known or proposed tyrosine phosphorylation; TM, proposed transmembrane segment.