J. Mol. Biol. (1982) **157**, 105–132

A Simple Method for Displaying the Hydropathic Character of a Protein

JACK KYTE AND RUSSELL F. DOOLITTLE

Department of Chemistry
University of California, San Diego, La Jolla, Calif. 92093, U.S.A.

(Received 8 August 1981, and in revised form 25 January 1982)

A computer program that progressively evaluates the hydrophilicity and hydrophobicity of a protein along its amino acid sequence has been devised. For this purpose, a *hydropathy scale* has been composed wherein the hydrophilic and hydrophobic properties of each of the 20 amino acid side-chains is taken into consideration. The scale is based on an amalgam of experimental observations derived from the literature. The program uses a moving-segment approach that continuously determines the average hydropathy within a segment of predetermined length as it advances through the sequence. The consecutive scores are plotted from the amino to the carboxy terminus. At the same time, a midpoint line is printed that corresponds to the grand average of the hydropathy of the amino acid compositions found in most of the sequenced proteins. In the case of soluble, globular proteins there is a remarkable correspondence between the interior portions of their sequence and the regions appearing on the hydrophobic side of the midpoint line, as well as the exterior portions and the regions on the hydrophilic side. The correlation was demonstrated by comparisons between the plotted values and known structures determined by crystallography. In the case of membrane-bound proteins, the portions of their sequences that are located within the lipid bilayer are also clearly delineated by large uninterrupted areas on the hydrophobic side of the midpoint line. As such, the membrane-spanning segments of these proteins can be identified by this procedure. Although the method is not unique and embodies principles that have long been appreciated, its simplicity and its graphic nature make it a very useful tool for the evaluation of protein structures.

1. Introduction

One of the most persistent and absorbing problems in protein chemistry has been the unraveling of the various forces involved in folding polypeptide chains into their unique conformations. Insight into this question has been gained both from a consideration of non-covalent forces as they apply in model systems and from a detailed examination of the actual structures of protein molecules. It is generally accepted that, to a rough approximation, two opposing, but not independent, tendencies are reflected in the final structure of a protein when it folds. The resulting compromise allows hydrophilic side-chains access to the aqueous solvent while at the same time minimizing contact between hydrophobic side-chains and

0022–2836/82/130105–28 $03.00/0 © 1982 Academic Press Inc. (London) Ltd.

Halozyme EX2082
Merck v. Halozyme
PGR2025-00052

water. Recently, however, it has been noticed that there are important subtle deviations from these expectations (Lifson & Sander, 1979; Janin & Chothia, 1980), suggesting that the extent to which residues are buried depends not only upon strict hydrophobicity but also upon steric effects that determine packing between the secondary structures in the crowded interior of the macromolecule. Nevertheless, if one could evaluate the contrary forces of hydrophobicity and hydrophilicity inherent within the residues themselves, then it would be possible, perhaps, at least to distinguish the exterior portions of a protein from the interior ones, on the basis of the amino acid sequence alone. Moreover, in the case of a protein that interacts directly with the alkane portion of a phospholipid bilayer in a membrane, there is general agreement that the amino acid side-chains involved are chiefly hydrophobic. Once again, an appropriate evaluation of a given amino acid sequence should be able to predict whether or not a given peptide segment is sufficiently hydrophobic to interact with or reside within the interior of the membrane.

Considerable effort has already been expended in devising schemes for predicting three-dimensional aspects from amino acid sequences alone. The most notable of these have dealt with the prediction of local secondary structure (Chou & Fasman, 1973; Wu & Kabat, 1973; Garnier *et al.*, 1978). These are empirical methods in that they utilize a library of known protein structures from which the distribution of the 20 amino acids among various conformational settings is rigorously tallied. If the frequency with which the individual amino acids or short peptides occur in α-helices, β-sheets or reverse turns is known, any sequence can be systematically scanned and the probability of those secondary structures can be evaluated. Interestingly, even earlier attempts had been made to predict the general shape of a protein on the basis of the types of amino acids it contained. Thus, in the light of the general observation that the interiors of water-soluble proteins are predominantly composed of hydrophobic amino acids, while the hydrophilic side-chains are on the exterior where they can interact with water, Fisher (1964) and Bigelow (1967) tried to correlate the sizes and shapes of proteins with their overall amino acid compositions.

Recently, a method for displaying the distribution of hydrophobicity over the sequence of a protein was presented by Rose (1978) and Rose & Roy (1980). This procedure combines the progressive-evaluation approach of the secondary structure predictions with the earlier empirical observation that the hydrophobic side-chains tend to be buried within the native structure (Chothia, 1976). Rose & Roy (1980) also have demonstrated convincingly that this approach can distinguish regions of interior sequence from regions of exterior sequence.

In this paper we describe a simple computer program, similar to that employed by Rose (1978) and Rose & Roy (1980), that systematically evaluates the hydrophilic and hydrophobic tendencies of a polypeptide chain. The present program uses a *hydropathy scale* in which each amino acid has been assigned a value reflecting its relative hydrophilicity and hydrophobicity. The program continuously determines the average hydropathy of a moving segment as it advances through the sequence from the amino to the carboxy terminus. As such, the procedure gives a graphic visualization of the hydropathic character of the

chain from one end to the other, tracking the hydrophilic and hydrophobic regions relative to a universal midline. We have examined in detail the profiles of several proteins whose three-dimensional structures are known, and have found excellent agreement between the observed interiors and the calculated hydrophobic regions, on the one hand, and the observed exterior portions and the calculated hydrophilic regions, on the other. We have also examined a number of membrane proteins and have been able to identify membrane-spanning segments, as well as those hydrophobic regions that anchor certain proteins in membranes.

2. Experimental Procedures

(a) *The computer program*

The computer program, SOAP, assigns the appropriate hydropathy value to each residue in a given amino acid sequence and then successively sums those values, starting at the amino terminal, within overlapping segments displaced from each other by one residue. Although a segment of any size can be chosen, ordinarily spans of 7, 9, 11 or 13 were employed, odd numbers being used so that a given sum could be plotted above the middle residue of the segment. Thus, in the case of SOAP-7 the first value corresponds to the sum of the hydropathies of residues 1 to 7 and is plotted at location 4, the second value corresponds to the sum for residues 2 to 8 and is plotted at location 5, and so on.

The program was written originally in the language C (Kernighan & Ritchie, 1978) for use in the software system Unix, which is leased from the Western Electric Co. Because Unix is now widely used, and because C compilers are now available for many computers, the original program is supplied as a short Appendix to this paper so that interested readers may employ it directly. Plots may be obtained from any terminal that prints a standard 100-character output. The program has also been modified for use with a more sophisticated system linked to a Zeta Plotter (we are grateful to S. Dempsey, Department of Chemistry, University of California, San Diego, for these modifications). In this latter format the values are presented as averages rather than sums, and all the figures accompanying this paper were obtained with this system.

(b) *Sequence library*

The characterization of a large number of different proteins was facilitated by the fact that we had two extensive libraries of amino acid sequences already stored in the computer. One of these was a set of 1081 sequences that can be purchased from the National Biomedical Research Foundation and that includes all the sequences from the *Atlas of Protein Sequence and Structure* (Dayhoff, 1978). The other, NEWAT, was a collection of approx. 200 sequences gleaned from the original literature and covering the period 1978 to 1980 (Doolittle, 1981).

(c) *Choice of hydropathy values*

Ideally, the most satisfying way to determine the hydrophobic or hydrophilic inclinations of a given amino acid side-chain (i.e. its hydropathy†) would be to measure its partition coefficient between water and a non-interacting, isotropic phase and to calculate from that partition coefficient a transfer free energy. For example, ethanol is a solvent that has been proposed as a phase resembling the interior of a protein (Nozaki & Tanford, 1971). In this case, however, the choice of solvent may have been a matter of convenience rather than design, since the solubilities of many amino acids in ethanol and water were already known (Cohn & Edsall, 1943), and the theoretical basis for deriving the transfer free energies of the

† Since hydrophilicity and hydrophobicity are no more than the two extremes of a spectrum, a term that defines that spectrum would be as useful as either, just as the term light is as useful as violet light or red light. Hydropathy (strong feeling about water) has been chosen for this purpose.

individual amino acid side-chains between ethanol and water from these values had already been formulated (Cohn & Edsall, 1943). The transfer free energies from water to ethanol for various amino acid side-chains are presented in Table 1.

The assumption that ethanol is a neutral, non-interacting solvent may not be warranted; however, nor is it clear that any solvent could ever meet that condition. This difficulty has been noted in the past, and it has been suggested that water–vapor transfer free energies would be a less complicated index of hydropathy (Hine & Mookerjee, 1975). The water–vapor partition coefficients for model compounds identical to each of 14 amino acid side-chains were assembled by Wolfenden *et al.* (1979,1981) from the Tables published by Hine & Mookerjee (1975). They reported as well experimental determinations for four additional, previously unavailable, values (Wolfenden *et al.*, 1979). All of these values, expressed as transfer free energies between an aqueous solution and the condensed vapor†, are also presented in Table 1.

It is possible with these free energies to examine the claim that ethanol is a useful solvent with which to model the interior of a protein (Nozaki & Tanford, 1971). The last column in Table 1 presents the transfer free energies for the model compounds between ethanol and the condensed vapor. If ethanol were an isotropic, non-interacting phase, these values would be fairly constant, representing only the dispersion forces lost during vaporization. It is clear, however, that this is not the case and that ethanol retains many of the unpredictable peculiarities of water itself. Therefore, contrary to the choice made by Rose (1978), no use will be made here of the hydrophobicity scale based on solubilities in ethanol.

Another source of information bearing upon the tendency of a given side-chain to prefer the interior of a protein to the exterior is the tabulation of residue accessibilities calculated by Chothia (1976) from the atomic co-ordinates of 12 globular proteins. Indeed, the ensemble average of the actual locations, inside or outside, of a side-chain should be a direct evaluation of its hydropathy when it is in a protein. Chothia presented two sets of values, the fraction of the total number of a given residue that is more than 95% buried in the native structures, on the one hand, and the fraction that is 100% buried, on the other. It has already been noted (Wolfenden *et al.*, 1979) that there is a strong correlation between the fraction 95% buried and the water–vapor transfer free energies. Interestingly, there is an even stronger correlation between the fraction 100% buried and the same water–vapor transfer free energies (Table 2). In Table 2 the transfer free energies, the fractions 100% buried, and the fraction 95% buried have all been normalized arbitrarily in order that the 3 values may be compared directly.

We have used both the water–vapor transfer free energies and the interior–exterior distribution of amino acid side-chains determined by Chothia (1976) in assigning the final hydropathy values (Table 2). Results presented later in this paper indicate clearly that the number in the second place of the hydropathy values is of little consequence to the

† In the present tabulation (Table 1) a small correction ($<1{\cdot}1$ kcal mol^{-1}) was applied to the former values (Wolfenden *et al.*, 1979). To eliminate any entropy of mixing from the values, the transfer must occur between standard states chosen in such a way that no changes in volume are involved. If the aqueous standard state is chosen as the infinitely dilute solution at 1·00 mole fraction, the volume of the solute in the aqueous phase by definition will be its apparent molal volume. The gas at 1·00 mole fraction can be compressed mathematically to this same molal volume. This is accomplished readily by employing the relationship: $\Delta G = -RT\ln(V_b/V_a)$ for the adiabatic change in the volume of an ideal gas from V_a to V_b. Specifically, the solute vapor at a standard state of 1·00 mole fraction at standard temperature (25°C) and pressure ($V_g = 24{\cdot}47$ l mol^{-1}) is contracted to a volume equal to its particular apparent molal volume (ϕ in cm^3 mol^{-1}). Therefore, the formula used for this correction was

$$\Delta G^{\circ}_{\text{transfer}} = -RT\ln\left(\frac{N_w V_g}{N_g \phi}\right) = -RT\ln\left(\frac{18{\cdot}07\gamma}{\phi}\right),$$

where N_w = equilibrium mole fraction in aqueous phase, N_g = equilibrium mole fraction in the vapor at standard temperature and pressure, and γ is the partition coefficient in the units M M^{-1} as tabulated by Hine & Mookerjee (1975). The advantage of this choice of standard states is that free energies of solvation are directly presented and only the molecular interactions between water and the solutes are reflected in the values.

TABLE 1

Free energies of transfer for the side-chains of the amino acids between various phases

Side-chain	ϕ ($cm^3\ mol^{-1}$)	$\Delta G^\circ_{transfer}$ (kcal mol^{-1}) Water into condensed vapor	Water into ethanol	Ethanol into condensed vapor
Leucine	84[b]	−3·20[c]	−2·41[e]	−0·79
Isoleucine	84[b]	−3·06[c]		
Valine	68[b]	−2·78[c]	−1·68[e]	−1·10
Alanine	35[b]	−2·34[c]	−0·73[e]	−1·61
Phenylalanine	97[b]	−0·23[c]	−2·65[e]	+2·42
Methionine	83[b]	+0·58[c]	−1·29[e]	+1·87
Cysteine	51[b]	+0·63[c]		
Threonine	56[b]	+4·23[c]	−0·44[e]	+4·67
Serine	38[b]	+4·63[c]	−0·04[e]	+4·67
Tryptophan	121[b]	+4·77[d]	−3·22[f]	+7·99
Tyrosine	101[a]	+5·10[c]	−2·39[f]	+7·49
Lysine	88[a]	+8·08[c,f]		
Glutamine	71[a]	+8·59[d]	+0·10[e]	+8·49
Asparagine	55[a]	+9·04[d]	+0·01[e]	+9·03
Glutamic acid	68[a]	+9·09[c,g]	+2·87[e,g]	+6·22
Histidine	69[b]	+9·57[d,g]	−0·45[f,g]	+10·02
Aspartic acid	51[a]	+10·04[c,g]	+3·42[e,g]	+6·62

The apparent molal volumes (ϕ) at 25°C of model compounds for the side-chains (Wolfenden *et al.*, 1979) of the structure RH, where R is the side-chain of a given amino acid ($^+H_3NCH(R)COO^-$) are either the observed values (a) tabulated by Cohn *et al.* (1934) or values calculated (b) by the methods of Cohn *et al.* (1934), which themselves were adapted from Traube (1899). The water–vapor partition coefficients for the various model compounds are available in the Tables published by (c) Hine & Mookerjee (1975) or (d) Wolfenden *et al.* (1979). The standard states chosen for the free energies are 1·00 mole fraction for the solution and the condensed vapor at a volume equal to its apparent molal volume (ϕ). The water–ethanol transfer free energies were copied directly from the tabulations of (e) Cohn & Edsall (1943) or (f) Nozaki & Tanford (1971). The standard states in each solvent are 1·00 mole fraction. The transfer free energies for the ionized side-chains (g) were corrected to pH 7·0 (Wolfenden *et al.*, 1979) using the following pK_a values (Tanford, 1968): lysine, 10·4; histidine, 6·4; glutamic acid, 4·5; and aspartic acid, 4·1.

hydropathy profiles, and as a result we did not hesitate to adjust the values subjectively when only this level of accuracy was in question. Nevertheless, we tried to derive the best numbers we could from the data listed in the last 3 columns of Table 2. The hydropathy values for valine, phenylalanine, threonine, serine and histidine were simple averages of the 3 other numbers in the Table. When 1 of the 3 numbers for a given amino acid was significantly different from the other 2, the mean of the other 2 was used. This was done for cysteine/cystine, methionine and isoleucine. After a good deal of futile discussion concerning the differences among glutamic acid, aspartic acid, asparagine and glutamine, we came to the conclusion that they all had indistinguishable hydropathies and set their hydropathy value by averaging all of the normalized water–vapor transfer free energies and the normalized fractions of side-chains 100% buried. Because the structural information was so uncertain, tryptophan was simply assigned its normalized transfer free energy. Glycine was arbitrarily assigned the hydropathy value which was the weighted mean of the hydropathy values for all of the sequences in our data base because it was clear from a careful analysis of the actual distribution of glycine that it is not hydropathic; that is to say, it does not have strong feelings about water. On the basis of both the transfer free energy scale and the fraction buried, alanine ought to be more hydrophobic on our scale, its value exceeding that

TABLE 2

Hydropathy scale and information used in the assignments

Side-chain	Hydropathy index	$\Delta G^{\circ}_{transfer}$ (water–vapor)[a]	Fraction of side-chains 100% buried[b]	Fraction of side-chains 95% buried[c]
Isoleucine	4·5	4·4	4·5	5·2
Valine	4·2	4·2	4·3	4·2
Leucine	3·8	4·5	3·2	2·8
Phenylalanine	2·8	2·5	2·5	3·5
Cysteine/cystine	2·5	1·9	6·0	3·2
Methionine	1·9	1·9	1·0	1·9
Alanine	1·8	3·9	5·3	1·6
Glycine	−0·4	—	4·2	1·3
Threonine	−0·7	−0·6	−0·5	−1·0
Tryptophan	−0·9	−0·9	−2·4	−0·3
Serine	−0·8	−0·8	−0·7	−1·0
Tyrosine	−1·3	−1·1	−3·3	−2·2
Proline	−1·6	—	−2·4	−1·8
Histidine	−3·2	−4·2	−3·6	−1·9
Glutamic acid	−3·5	−3·9	−2·8	−1·7
Glutamine	−3·5	−3·5	−4·0	−3·6
Aspartic acid	−3·5	−4·5	−2·5	−2·3
Asparagine	−3·5	−3·8	−3·1	−2·7
Lysine	−3·9	−3·2	—	−4·2
Arginine	−4·5	—	—	—

All values in the last 3 columns result from arbitrary normalization to spread them between −4·5 and +4·5. The normalization functions were:

[a] $-0{\cdot}679(\Delta G^{\circ}_{transfer}$; Table 1)$+2{\cdot}32$.

[b] 48·1(fraction 100% buried; Chothia, 1976)−4·50.

[c] 16·45(fraction 95% buried; Chothia, 1976)−4·71.

of leucine. We find it difficult to accept that a single methyl group can elicit more hydrophobic force than a cluster of 4 methyl groups, and for that reason we have arbitrarily lowered the hydropathy value of the alanine side-chain to a point half-way between the hydropathy value of glycine and the value determined for alanine when the transfer energy and its distribution were used. No suitable model exists for proline, and in terms of its tendency to become buried it is fairly hydrophilic. Its hydropathy value was made somewhat more hydrophobic than this consideration because of its 3 methylene groups. The hydropathy value for arginine was arbitrarily assigned to the lowest point of the scale. Because it was difficult to accept the fact that tyrosine is a hydrophilic amino acid, even though the available data in Table 2 indicate that it is, its hydropathy value was subjectively raised to one closer to the water–vapor transfer free energy than the structural data would have yielded. Similarly, the hydropathy value for leucine was also raised above the average of the structural data and the transfer free energy, and the hydropathy value for lysine was lowered. None of these last 3 adjustments, the result of personal bias and heated discussion between the authors, affects the hydropathy profiles in any significant way.

3. Results

(a) *Choice of parameters*

The effectiveness of the program and the progressive-evaluation approach in general depend upon two decisions. First, we had to determine how large a span of consecutive residues yields a hydropathy profile that most consistently reflects the

exterior and interior portions of proteins. Second, we had to determine how critical the hydropathy assignments for the individual amino acids are to the outcome of the calculations. For example, is the profile of a given sequence radically changed if the hydropathy values for one or more residues are changed by an arbitrary factor? We met these problems directly by examining the same protein sequences under a variety of conditions.

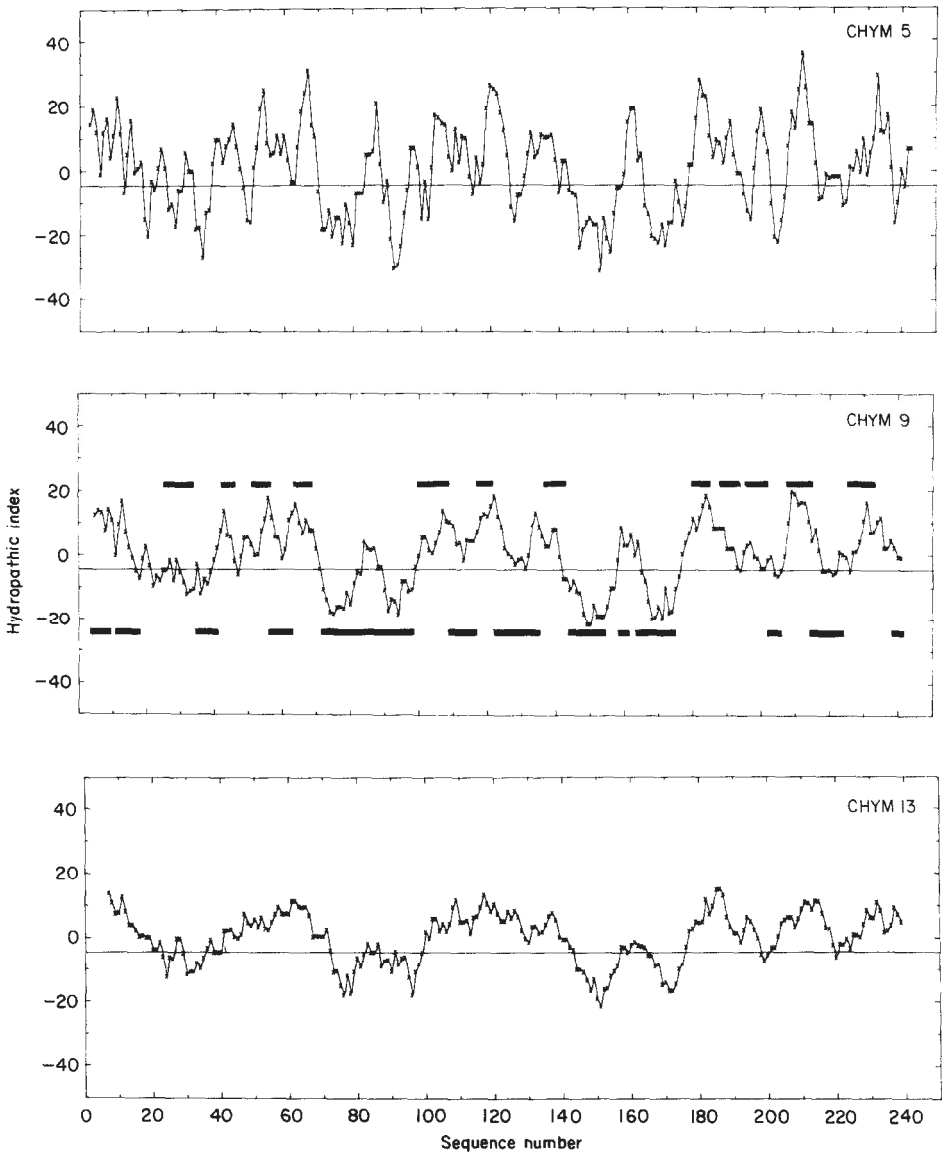


FIG. 1. SOAP profiles of bovine chymotrypsinogen (CHYM) at 3 different span settings (5, 9 and 13). The solid bars above the midpoint line on the SOAP-9 profile denote interior regions as determined by crystallography (Freer *et al.*, 1970). Similarly, the solid bars below the midpoint line indicate regions that are on the outside of the molecule.

With respect to the most effective choice of span, we compared the hydropathy profiles of a number of different proteins over a range of spans from 3 to 21 residues. Selected profiles from two of these surveys, for chymotrypsin and lactate dehydrogenase, respectively, are shown in Figures 1 and 2. Naturally, the hydropathy profiles using the shortest spans are noisier than intermediate spans, and runs employing spans less than seven residues were generally unsatisfactory. Long spans on the other hand tended to miss small, consistent features. Frequent and subjective analysis of the degree of correlation of the profiles with the exteriors and interiors of globular proteins (see below), as well as the resolution of the profile itself, revealed that information content was maximized when the spans were set at 7 to 11 residues.

The impact of the choice of hydropathy values was examined in two different

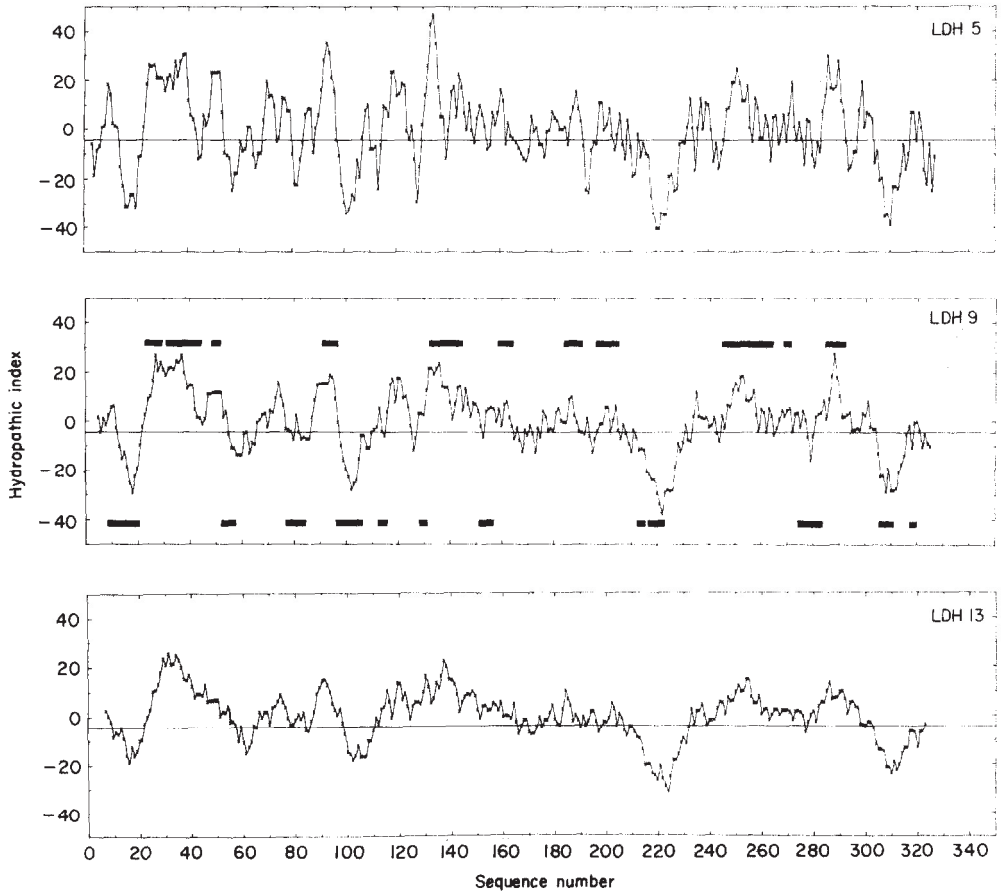


FIG. 2. SOAP profiles of dogfish lactate dehydrogenase (LDH) at 3 different span settings (5, 9 and 13). The solid bars above the midpoint line on the SOAP-9 profile denote interior regions as determined by the crystallographic study of the protein (Eventhoff *et al.*, 1977). Similarly, the solid bars below the midpoint line indicate regions that are known to be on the outside of the molecule.

ways. As an initial test, the 20 side-chains were assigned to three groups according to their rank on the hydropathy scale (Table 2). Thus, arginine, lysine, asparagine, aspartic acid, glutamine, glutamic acid and histidine were assigned to cluster I: proline, tyrosine, serine, tryptophan, threonine and glycine to cluster II; and alanine, methionine, cysteine/cystine, phenylalanine, leucine, valine and isoleucine to cluster III. The individual values contributing to each cluster were averaged (cluster I, −3·7, cluster II, −1·0 and cluster III, +3·0) and the mean values incorporated into a modified SOAP program called LARD. Comparisons of LARD against SOAP in the cases of chymotrypsinogen and lactic dehydrogenase are shown in Figures 3 and 4. Although the patterns exhibit some general similarities, as might be expected since the moving average itself tends to have a leveling aspect, an experimental approach loses nothing by using the best values available rather than settling for less precise estimates.

As a second test, the values of four of the most controversial assignments were shifted radically in order to assess the impact on the hydropathy profile. Thus, the values for tyrosine, histidine, proline and tryptophan, all of which have arguably (Nozaki & Tanford, 1971) low hydropathy scores (Table 2), were arbitrarily increased by 3·0 units. When the same two proteins were examined with this modified scale there was a noticeable if modest change in the patterns (Figs 3 and 4). That the change was modest is partly due to the fact that histidine and tryptophan are among the least common amino acids.

(b) *Exterior and interior segments of globular proteins*

Detailed comparisons between the hydropathy profiles of two globular proteins and their published three-dimensional structures are presented in Figures 1 and 2. In the case of bovine chymotrysinogen, a judgement was made about each side-chain on the basis of its position in the standard model that had been constructed in the laboratory of Professor J. Kraut, Department of Chemistry, University of California, San Diego, as a part of the crystallographic study of that protein (Freer *et al.*, 1970). A variety of hydropathy profiles of the chymotrypsinogen sequence were obtained and compared with the actual locations of the residues in the model structure (Fig. 1). The best agreement between strongly hydrophobic segments and interior regions and strongly hydrophilic segments and the exterior was obtained with a setting of nine residues.

Examination of the results reveals that, for the most part, agreement between the actual structure and the location expected from the hydropathy of a certain region is quite satisfactory. In particular, two of the regions that lie on the exterior of this protein, whose electron density is poorly defined and that show the greatest rearrangements during zymogen activation (Freer *et al.*, 1970), residues 72 to 77 and 144 to 152, exhibit very high hydrophilicity consistent with their loose, external attachment to the structure. The five major regions of the profile that lie below the midpoint line are all external sequences in the native protein and nine of the 11 major regions that lie above the midpoint line are internal.

A consideration of those few places where the correlation fails is also illuminating. In the case of residues 82 to 90, for example, this segment in the model

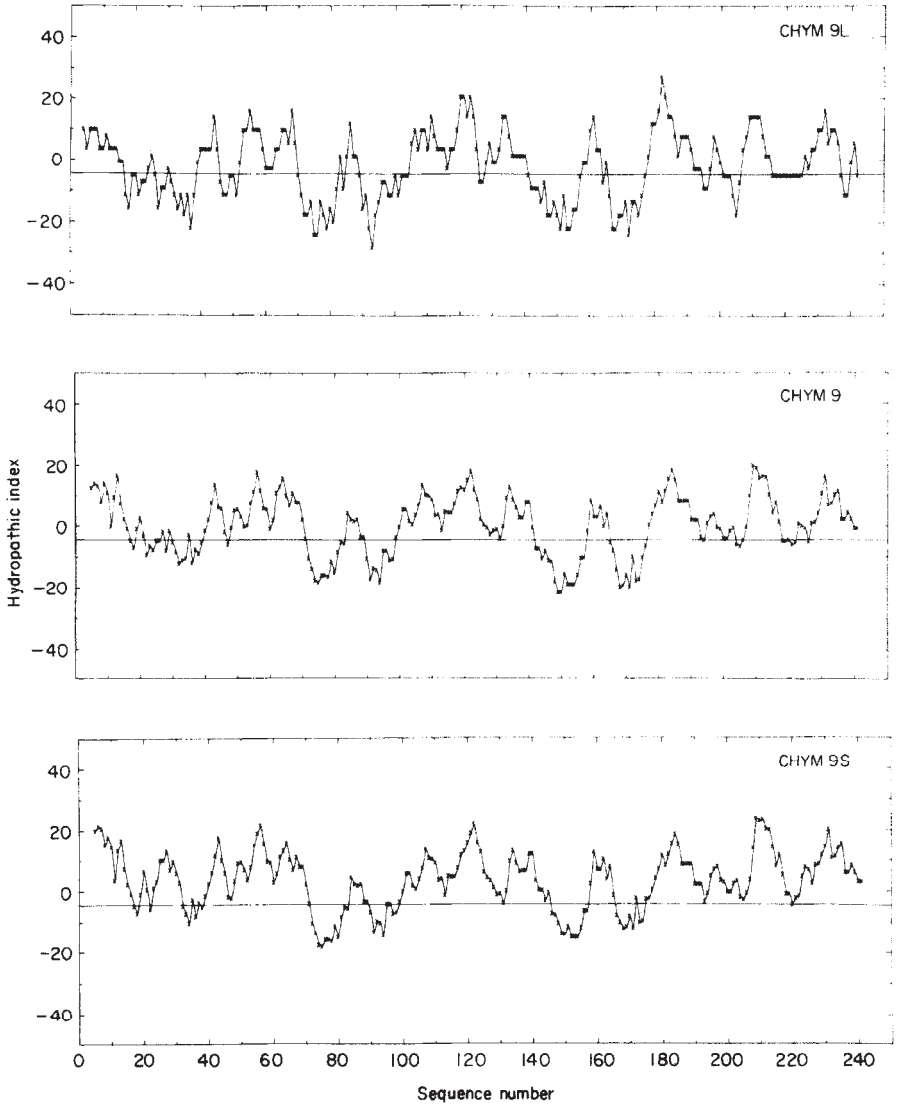


FIG. 3. SOAP profiles of bovine chymotrypsinogen (CHYM) using different hydropathy values for the 20 amino acids. In the top panel (9L) the program (LARD) used a set of clustered values in which case the 20 amino acids were divided into 3 sets (hydrophobic, neutral and hydrophilic). In the lower panel (9 S), the program used radically different weighting factors for some of the more controversial amino acid assignments, including those of histidine, tryptophan, tyrosine and proline. In the middle panel the program used the standard set of assignments presented in Table 2. All plots utilize a span setting of 9.

runs along the exterior of the protein, even though the hydropathy profile shows it to have a very hydrophobic character. This nine-residue sequence, Lys-Leu-Lys-Ile-Ala-Lys-Val-Phe-Lys, contains four positive charges intermingled with five very hydrophobic residues. The contradiction arises from the fact that the high concentration of positive charge does not weigh heavily enough in the moving

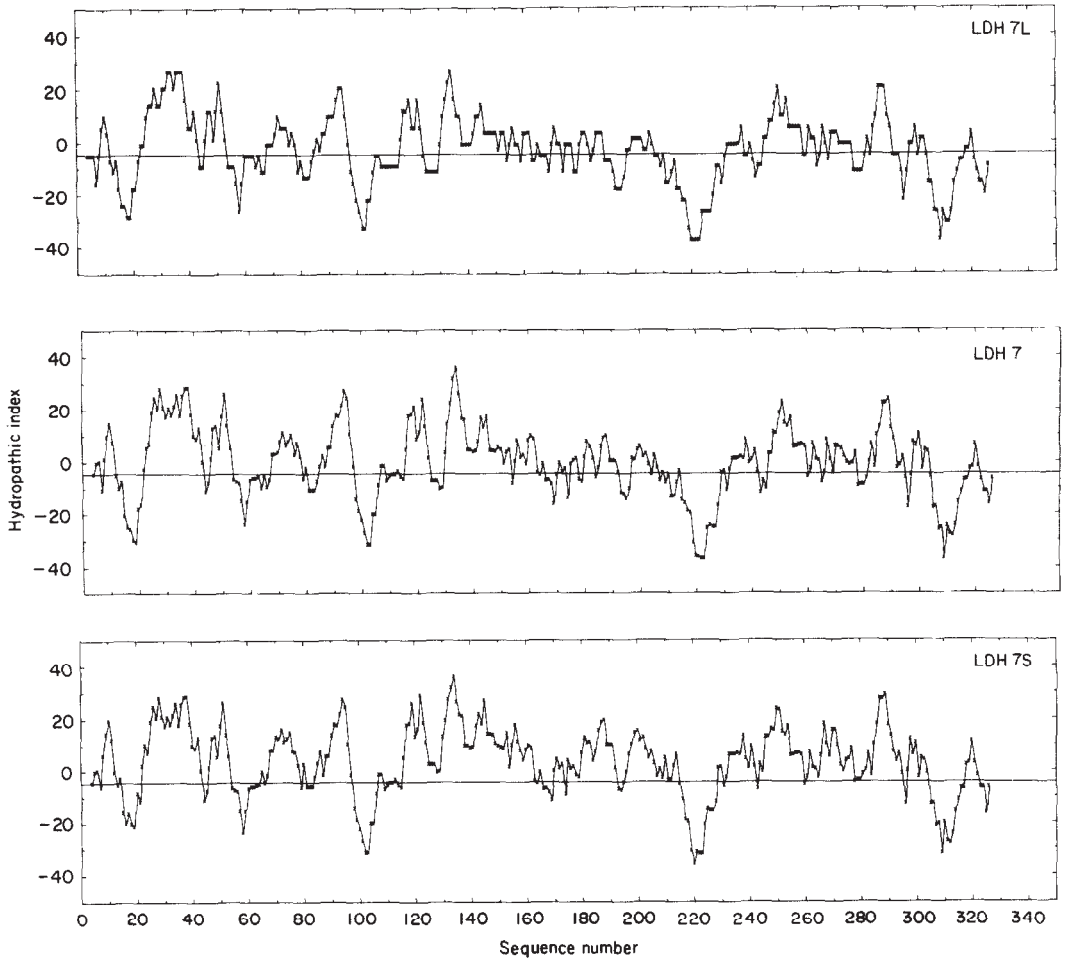


FIG. 4. SOAP profiles of dogfish lactate dehydrogenase in which different hydropathies were used for the 20 amino acids. All plots utilize a span setting of 7. See legend to Fig. 3 for meanings of 7L, 7S and 7.

average to offset the alkane side-chains present in this rather unusual sequence. A close examination of the model reveals that the five hydrophobic side-chains are all directed toward the interior while the lysine side-chains point out into the aqueous environment.

In the case of dogfish lactate dehydrogenase (Fig. 2), the designations of external and internal residues had already been made and published (Eventhoff *et al.*, 1977), and the hydropathy profile correlates well with these crystallographic findings. The five major regions of the profile that lie below the midpoint line are all external sequences in the native protein and six of the eight major regions above the midline are internal sequences. Again, as with chymotrypsinogen, the profile was least successful in evaluating those regions in which the main chain is only partly buried, such as the regions between residues 66 and 78, and 112 and 126, where the backbone repeatedly passes in and out of the aqueous phase.

(c) *Membrane-bound proteins*

The issue of the hydropathy of a particular sequence of amino acids assumes added significance when membrane-bound proteins are considered. The 3 nm of alkane that forms the bilayer is an invariant structural aspect with which such proteins must contend. It is generally accepted that the adaptation to this extremely hydrophobic environment is accomplished by the evolution of long hydrophobic sequences in those proteins whose destiny it is to become a component of a biological membrane. A hydropathy profile of the protein glycophorin (Tomita *et al.*, 1978) unmistakably identifies the archetypal membrane-spanning sequence, long ago noted by others (Tomita & Marchesi, 1975), that stretches from residues 72 to 95 (Fig. 5). In this example the polypeptide chain only crosses the bilayer once, polar segments extending into the aqueous environment on either side. A similar case is known to exist for vesicular stomatitis virus glycoprotein (Rose *et al.*, 1980), a partial profile of which is also shown in Figure 5.

Cytochrome b_5, on the other hand, is a protein for which the exact disposition of the hydrophobic sequence that anchors the protein in the membrane is less clear (Fig. 5). Although, as noted previously (Strittmatter *et al.*, 1972), the carboxy terminus is the general location of the embedded hydrophobic segment, the precise extent of the buried portion has not been determined unambiguously. The hydropathy profile indicates that the membrane-associated portion begins at residue 112. Fleming *et al.* (1979), on the other hand, have reported experiments that they believe suggest that the cluster of tryptophans at residues 108, 109 and 112 is deep within the membrane, a conclusion clearly at variance with the hydropathy profile (Fig. 5). The fact that the tryptophans are surrounded by aspartic acid, asparagine and serine residues, however, certainly strengthens the conclusion that this region is not within the alkane of the bilayer in the native structure.

In the case of bacteriorhodopsin (Khorana *et al.*, 1979), a protein that is located in the membranes of certain halophilic bacteria, five of the seven transmembrane shafts observed in the low-resolution electron density function (Henderson & Unwin, 1975) are clearly identified by the hydropathy profile (Fig. 6). The two segments nearest to the carboxy terminus, between residues 175 and 225, are not resolved from each other, although the profile clearly indicates that both are buried in the membrane. In this latter case, a point halfway along was arbitrarily chosen as the point at which the chain doubles back. The seven transmembrane sequences are aligned next to each other in Table 3.

(d) *Membrane-spanning sequences*

It was of considerable interest to explore whether or not the hydropathy profile could identify, within the linear sequence of a membrane-bound protein of unknown structure, those portions that span the bilayer and distinguish them from sequences that merely pass through the center of the protein itself. Certainly, casual observation of sequences known to be inserted into the alkane phase of the membrane (Figs 5 and 6) suggests that this should be possible. To this end, the hydropathy profiles of approximately 30 soluble proteins, chosen at random, were

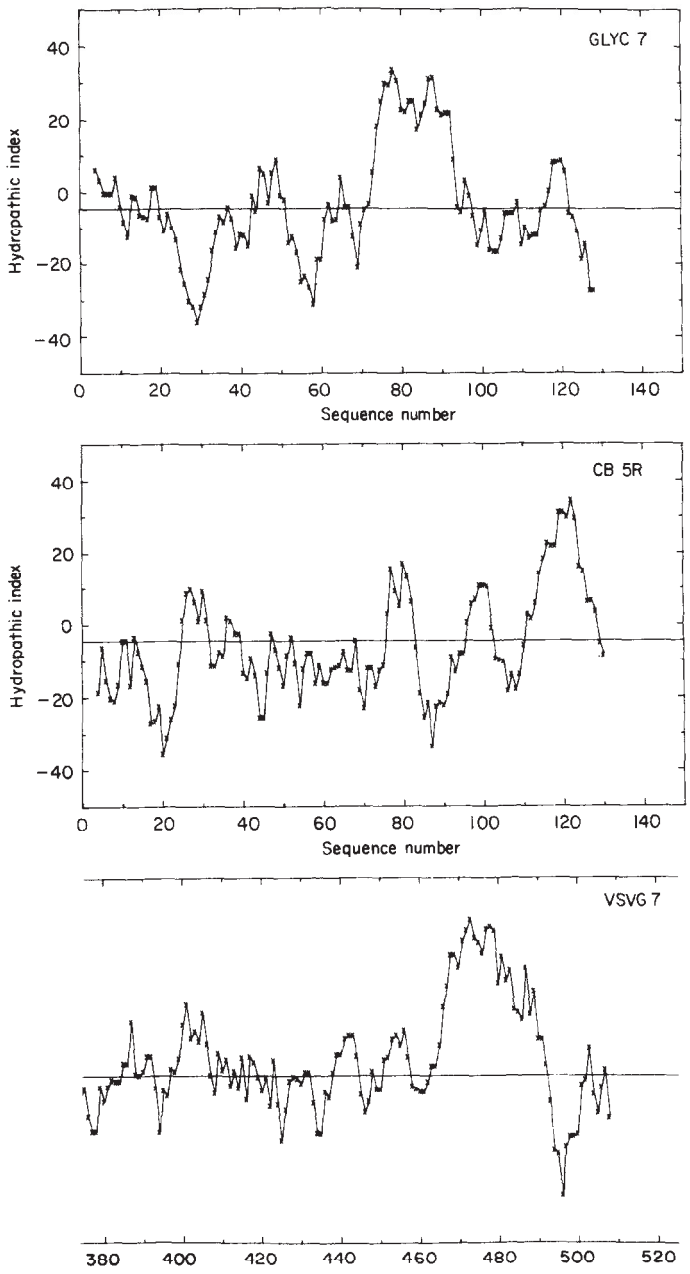


FIG. 5. SOAP profiles for 3 different proteins that have membrane affiliation. In the upper panel, the plot is that of erythrocyte glycophorin (GLYC), which has an easily recognized membrane-spanning segment in the region of residues 75 to 94. In the middle panel, rabbit cytochrome b_5 (CB5R) is depicted. In this case there is a membrane-anchoring unit involving the 20-residue carboxy-terminal segment. In the lower panel, the carboxy-terminal region of vesicular stomatitis virus glycoprotein (VSVG) is shown; a membrane-spanning segment is clearly evident from residues 470 to 490. All profiles are at span settings of 7.

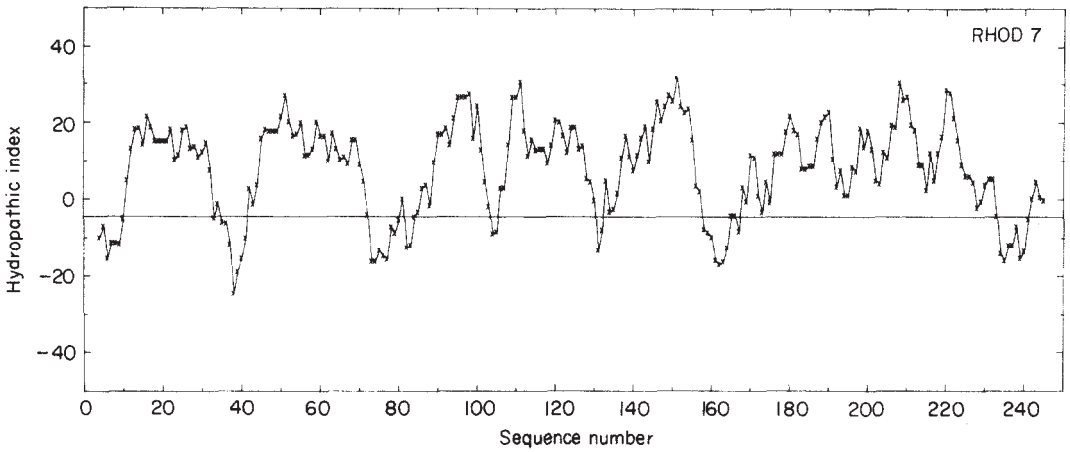


FIG. 6. SOAP profile of bacteriorhodopsin (RHOD) at a span setting of 7. Five of the well-known 7 transmembrane shafts (Henderson & Unwin, 1975) are clearly delineated; the separation point for the remaining 2 is not so clear and has been set arbitrarily at about residue 200.

TABLE 3

Transmembrane sequences of bacteriorhodopsin, aligned next to each other, in correct polarity

A	B	C	D	E	F	G
Trp_{10}	Met	Ile	Val_{130}	Arg	Val	Pro_{200}
Ile	Thr	Tyr	Lys	Phe_{135}	Ile	Leu
Trp	Leu	Trp_{80}	Thr	Val	Gly	Asn
Leu	Gly_{65}	Ala	Leu	Trp	Ala	Ile
Ala	Tyr	Arg	Ala	Trp	Gly_{195}	Glu
Leu_{15}	Gly	Tyr	Gly_{125}	Ala	Glu	Thr_{205}
Gly	Leu	Ala	Val	Ile_{140}	Ser	Leu
Thr	Leu	Asp_{85}	Leu	Ser	Gly	Leu
Ala	Met_{60}	Trp	Gly	Thr	Ile	Phe
Leu	Ser	Leu	Thr	Ala	Leu_{190}	Met
Met_{20}	Leu	Phe	Gly_{120}	Ala	Trp	Val_{210}
Gly	Tyr	Thr	Ile	Met_{145}	Val	Leu
Leu	Met	Thr_{90}	Met	Leu	Val	Asp
Gly	Thr_{55}	Pro	Ile	Tyr	Pro	Val
Thr	Phe	Leu	Gly	Ile	Tyr_{185}	Ser
Leu_{25}	Ala	Leu	Asp_{115}	Leu	Ala	Ala_{215}
Tyr	Ile	Leu	Ala	Tyr_{150}	Ser	Lys
Phe	Ala	Leu_{95}	Gly	Val	Trp	Val
Leu	Pro_{50}	Asp	Val	Leu	Leu	Gly
Val	Val	Leu	Leu	Phe	Val_{180}	Phe
Lys_{30}	Leu	Ala	Ala_{110}	Phe	Val	Gly_{220}
Gly	Thr	Leu	Leu	Gly_{155}	Thr	Leu
Met	Thr	Leu_{100}	Ile	Phe	Val	Ile
Gly	Ile_{45}	Val	Thr	Thr	Asn	Leu
Val	Ala	Asp	Gly	Ser	Arg_{175}	Leu

examined and the most hydrophobic region from each was picked. From this preliminary collection, a group of twelve 20-residue sequences, which were judged to be the most hydrophobic of the lot, was chosen for closer inspection. It was assumed that, since these were in each case the most hydrophobic region in the entire sequence of a given protein, they would serve as the most extreme models for a peptide that traverses the interior of a protein. From these 12 proteins, the most hydrophobic segment of each span-length from 9 to 21 residues was identified and its average hydropathy tabulated. The collected values for each span length were compared directly with those of the most hydrophobic segments of the same span length taken from bacteriophage M13 coat protein (Nakashima & Konigsberg, 1974), glycophorin, and the seven transmembrane sequences of bacterial rhodopsin (Table 3). These nine hydrophobic sequences, each of which is known to span the membrane, were chosen as models for a sequence which in the native protein is within the bilayer.

The discrimination between the segments from the soluble proteins as a group and those from the membrane-spanning sequences was most unequivocal when the span was lengthened to 19 residues (Fig. 7). This may be due to the fact that protein-spanning sequences passing through the interior are usually shorter than membrane-spanning sequences. Nevertheless, from an examination of Table 4 it can be concluded that when the hydropathy of a given 19-residue segment averages greater than +1·6 there is a high probability that it will be one of the sequences in a

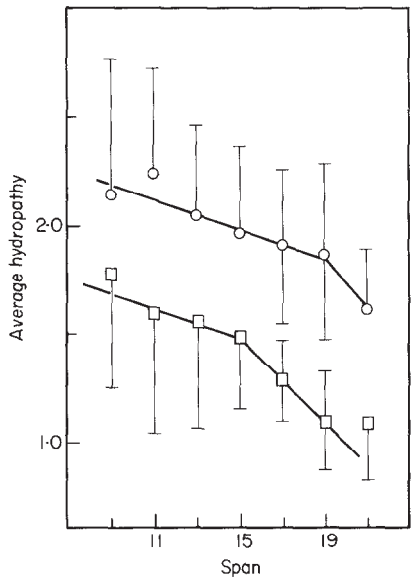


FIG. 7. Comparison between the hydropathy of sequences that span membranes and the hydropathy of those that span proteins. The most hydrophobic sequences from 9 globular proteins (Table 4), except for lactic dehydrogenase, were compared with 9 membrane-spanning sequences (Table 4). For each span length, the average hydropathies of the most hydrophobic segments were collected and the means and standard deviations of the 9 values were calculated. These are presented as a function of the span length for the membrane-spanning group (–○–○–) and the protein-spanning group (–□–□–).

TABLE 4

Nineteen-residue hydropathy averages for the most hydrophobic sequences from various proteins

Protein	Length	Sequence position	Mean hydropathy	
Soluble				
Dogfish lactate dehydrogenase	329	23–41	2·26	
Klebsiella aerogenes ribitol dehydrogenase	247	142–160	1·52	
Human transferrin	676	C53–C71	1·32	
Rabbit phosphorylase	841	139–157	1·18	
Bovine chymotrypsinogen A	245	51–69	1·14	1·09
Lobster glyceraldehyde-3-P dehydrogenase	333	14–32	1·04	±0·22
Bovine prothrombin	582	350–368	0·99	($n = 9$)
Bacillus stearothermophilus phosphofructokinase	316	213–231	0·96	
Human carbonic anhydrase B	260	135–153	0·88	
Escherichia coli dihydrofolate reductase	156	81–99	0·81	
Bovine carboxypeptidase A	307	95–113	0·81	
Bovine proalbumin	588	25–43	0·53	
Membrane-spanning				
M13 coat protein	50	21–39	1·92	
Human glycophorin	131	73–91	2·65	
Halobacterium halobium bacteriorhodopsin	248	11–29	1·79	
		44–62	1·79	1·86
		83–101	1·69	±0·38
		108–126	1·79	($n = 9$)
		136–154	1·85	
		177–195	1·22	
		206–224	2·07	

membrane-bound protein that spans the membrane. Furthermore, membrane-spanning sequences are more hydrophobic than sequences that pass through the center of a protein, and they can be distinguished from the latter by their hydropathy.

The sequences of subunits I to V and VII_{ser} of cytochrome oxidase were then examined as examples of membrane-spanning polypeptides (Fuller *et al.*, 1979) about which much less is known. All of the sequences that averaged greater than +1·6 over at least a 19-residue span were identified, as well as some even longer, more hydrophobic segments. All of these candidates for membrane-spanning sequences are presented in Table 5. The sequences of six of the seven or more subunits of cytochrome oxidase have been published, including all of the largest. They account for 1309 residues, more than 90% of the total protein. From this consideration and the information gathered in Table 5 it can be concluded that about 30% of the mass of cytochrome oxidase is located within the bilayer.

(e) *Grand averages of hydropathy*

In the past many claims have been made to the effect that information about the size and shape of a protein (Bigelow, 1967; Fisher, 1964) or its affiliation with a membrane (Capaldi & Vanderkooi, 1972) could be obtained from its amino acid

TABLE 5

Candidates for membrane-spanning sequences in cytochrome oxidase

Sequence	Average hydropathy ($n = 19$)	Sequence	Average hydropathy ($n = 19$)
Subunit I (yeast)[a]		Subunit III (yeast) [c]	
Ile_{14}-Ile_{36}†	2·04	Pro_{23}-Met_{41}	1·88
Leu_{56}-Ile_{74}	2·53	Leu_{89}-Trp_{107}	2·09
Ile_{99}-Val_{117}	2·09	Ser_{169}-Ile_{187}	1·95
Ile_{146}-Ile_{164}	1·65		
Leu_{182}-Leu_{210}†	2·25	Subunit IV (bovine)[d]	
Val_{242}-Tyr_{260}	1·62	Thr_{80}-Trp_{98}	2·15
Ile_{269}-Ser_{287}	1·73		
Leu_{332}-Gly_{350}	1·70	Subunit V (bovine)[e]	
Val_{451}-Ile_{469}	2·10	none	—
		Subunit VII_{ser} (bovine)[f]	
		Leu_{22}-Val_{40}	1·83
Subunit II (yeast)[b]		Subunit II (bovine)[g]	
Phe_{45}-Ile_{63}	2·30	Leu_{28}-Leu_{46}	2·69
Ile_{87}-Tyr_{105}	2·52	Ile_{67}-Tyr_{85}	2·26

[a] Bonitz *et al.* (1980).
[b] Coruzzi & Tzagoloff (1979).
[c] Thalenfeld & Tzagoloff (1980).
[d] Sacher *et al.* (1979).
† ($n > 19$).
[e] Tanaka *et al.* (1979).
[f] Buse & Steffens (1978).
[g] Steffens & Buse (1979).

composition alone. To explore this possibility in the present context, we compared the overall hydropathy of a large number of sequences by simply programming the computer to sum the hydropathy values of all the amino acids and dividing by the number of residues in the sequence to obtain a GRAVY score.

The GRAVY scores were plotted as a function of total sequence length, inasmuch as it has long been thought that larger globular proteins need more hydrophobic amino acids in order to fill up their interiors (Bigelow, 1967; Fisher, 1964). The distribution obtained when the overall hydropathies of 84 fully sequenced, soluble enzymes are plotted as a function of their total length indicates that the average hydropathy of soluble protein is independent of the sequence length, and earlier conclusions about possible correlations between the size and shape of a protein and its amino acid composition (Bigelow, 1967; Fisher, 1964) may have been overstated.

Included in Figure 8 are the GRAVY scores for several membrane-embedded proteins whose sequences have been established. These values lie well above those for the soluble proteins. GRAVY scores were also calculated for other membrane-spanning proteins on the basis of their amino acid compositions as determined from amino acid analysis of timed hydrolyses (Table 6). Although the GRAVY scores for these membrane-bound proteins are also quite high, in every case exceeding the mean of the compositions of sequenced soluble proteins (−0·4), when their spread is

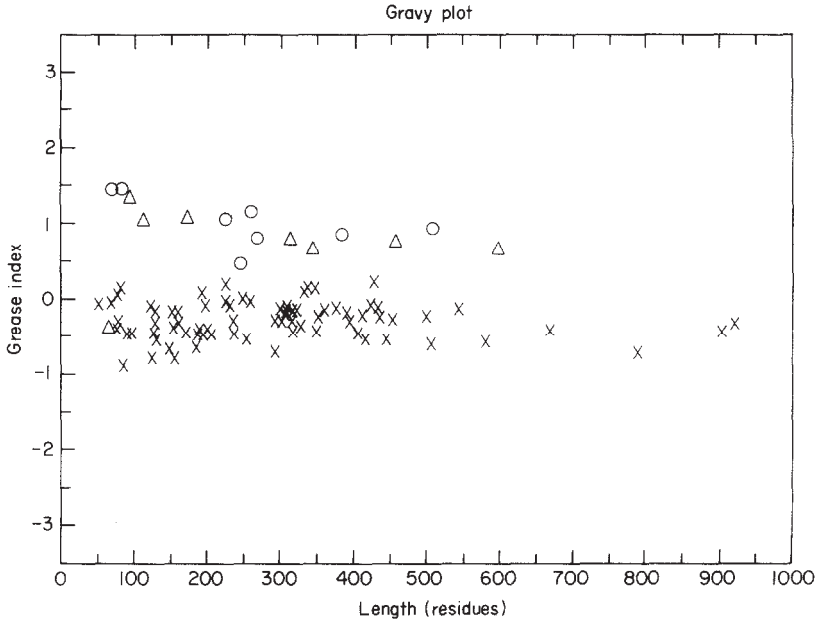


FIG. 8. Plot of mean hydropathies (GRAVY scores) of various proteins against their lengths: (×) 84 fully-sequenced soluble enzymes whose amino acid sequences have been taken from the recent literature; (⊙) 8 membrane-embedded proteins whose sequences have been determined (bacteriorhodopsin; yeast mitochondrial cytochrome oxidase subunits I to III, cytochrome *b*, and the (*oli*-2, *oli*-4) ATPase subunit; and 2 carbodiimide-sensitive mitochondrial proteins); (△) 8 putative proteins inferred from the unidentified reading frames found in the DNA of human mitochondria (Anderson *et al.*, 1981).

TABLE 6

Average hydropathy (GRAVY) for the entire amino acid composition of a collection of membrane-spanning proteins

Protein	GRAVY	References
Yeast cytochrome *b*[a]	0·79	Nobraga & Tzagoloff (1980)
Halobacterium halobium bacteriorhodopsin[a]	0·70	Khorana *et al.* (1979)
Cytochrome oxidase (yeast/bovine)[a]	0·37	[c]
Human glucose carrier[b]	0·37	Sogin & Hinkle (1978)
Bovine rhodopsin[b]	0·28	Heller (1968) Pober & Stryer (1975)
Human anion carrier[b]	0·04	Ho & Guidotti (1975) Drickamer (1977) Steck *et al.* (1978)
Canine Na^+, K^+-ATPase, α subunit[b]	−0·06	Kyte (1972)
Rabbit Ca^{2+}-ATPase[b]	−0·05	Allen *et al.* (1980)
Torpedo californica acetylcholine receptor[b]	−0·22	Vandlen *et al.* (1979)

[a] From sequence.
[b] From composition.
[c] From all subunits listed in Table 5.

compared to the spread of the values for an array of soluble proteins (Fig. 8) the claim that membrane-bound proteins can be distinguished from soluble proteins by their amino acid compositions alone (Capaldi & Vanderkooi, 1972) appears tenuous. There remains the possibility, however, that the unexpected hydrophilicity of the membrane-spanning proteins whose compositions are known only from amino acid analysis may actually be due to a failure to hydrolyze membrane-spanning sequences completely even after 72 hours at 108°C.

4. Discussion

The equilibrium that determines the unique molecular structure of a protein is the one that exists between it and a random coil (Anfinsen, 1973). It is generally assumed that this process can be described as a simple two-state equilibrium between the native structure and the random coil, and experimental results consistent with this assumption have been presented (Tanford, 1968). If this is indeed the case, the individual contributions to the overall free energy change for this isomerization would be the most critical factors in determining the outcome, rather than any kinetic features of the reaction. These thermodynamic forces, by the very nature of the process, must be non-covalent interactions. Several provocative discussions of these matters have been presented (Cohn & Edsall, 1943; Kauzmann, 1959; Jencks, 1969; Chothia, 1976). Moreover, it has been demonstrated definitively, by experimental observation, that neither hydrogen bonds (Klotz & Franzen, 1962), nor ionic interactions (Cohn & Edsall, 1943), nor dispersion forces (Deno & Berkheimer, 1963) can provide any *net* favorable free energy for the formation of the native structure in aqueous solution. Therefore, by exclusion, and perhaps for the lack of a better candidate, hydrophobic forces (Kauzmann, 1959) have attracted the most attention in discussions of this process.

Felicitously, this has drawn attention to the significant role of the aqueous solvent, *per se*. The hydrophobic force is simply that force, arising from the strong cohesion of the solvent, which drives molecules lacking any favorable interactions with the water molecules themselves from the aqueous phase (Jencks, 1969). In the case of the formation of the native structure from the random coil, this force participates in the reaction because hydrophobic side-chains, which are exposed to water in the extended coil, are removed to the interior of the protein during the folding of the native structure (Chothia, 1976). This transfer appears to provide the only favorable free energy available to drive the reaction to completion. Therefore, the more aversion water has for a given amino acid side-chain, the more free energy is gained when that residue or a portion of it ends up inside the native structure.

Conversely, and of equal importance, it is also the case that the more attraction water has for a functional group on an amino acid side-chain, the more free energy is lost when that functional group is removed from water during the folding process. This point becomes clear upon examination of the data in Table 1, when it is realized that most of the free energies of transfer from water to the condensed vapor are actually unfavorable, many by a considerable amount. This is due, of course, to the fact that water participates in strong interactions with hydrogen-bond donors and acceptors (Klotz & Farnham, 1968), as well as to the need to

neutralize charged side-chains. As a result, one of the major free energy deficits in the folding of a protein results from the requirement to unsolvate those hydrophilic functional groups destined for the interior. Some of this investment is returned when hydrogen bonds are formed in the interior. Nevertheless, because of geometric constraints, the hydrophilic side-chains in the center of a protein participate in far fewer hydrogen bonds than they would in the unfolded and exposed random coil, where both the donors and acceptors interact fully with water. As such, there is a high probability that significant free energy will be lost whenever a hydrophilic residue is removed from water during the folding process.

It is undeniably the case therefore that both the hydrophobicity and the hydrophilicity of a given sequence of amino acids affect the outcome of the equilibrium between the random coil and the native structure. Although one or the other of these two properties is often emphasized to make a particular point, neither is more important than the other. For example, it is often stated that the interior of a protein is formed from its hydrophobic sequences, but it is seldom pointed out that the interior of the protein is also formed because the hydrophilic sequences cannot be buried. Thus, it can be concluded that any description of the folding process that fails to consider either hydrophobicity or hydrophilicity is discarding half of the information contained within the sequence of the protein.

It has also been pointed out (Lifson & Sander, 1979; Janin & Chothia, 1980; Chothia & Janin, 1981) that the packing properties of residues such as leucine, isoleucine and valine might have an effect, independent of hydropathy, on the folding process as the interior of the protein is fitted together. Unfortunately, very little is known about the features of this steric interplay, and our understanding of the folding process does not extend beyond the conclusion that hydropathy is of central importance to it.

The conclusion that can be drawn from all of these considerations is that, to a first approximation, the native structure of a protein molecule will be that structure that permits the removal of the greatest amount of hydrophobic surface area and the smallest number of hydrophilic positions from exposure to water (Bigelow, 1967; Fisher, 1964; Chothia, 1976). The obvious prediction that follows from this conclusion is that the most hydrophobic sequences in a protein will be found in the interior of the native structure and the most hydrophilic sequences will be found on the exterior. In order to exploit this prediction with the greatest success, the most accurate evaluations of the hydrophobicity and hydrophilicity of each amino acid side-chain should be formulated.

To this end, a number of hydropathy scales have been proposed in other publications, but, in our view, they all suffer from serious drawbacks. Those based on water–ethanol transfer free energies (Nozaki & Tanford, 1971; Segrest & Feldman, 1974; Rose, 1978) are imperfect due to the peculiarities of ethanol as a solvent, which seem almost as unusual as those of water itself (Table 1). A scale based on the partition coefficient between the bulk aqueous phase and the air–water interface (Bull & Breese, 1974) also seems a poor choice, because the hydrogen bonds that must be broken and the charges that must be neutralized to remove a residue from the aqueous phase during the formation of the native structure probably remain intact at the air–water interface and are thus not a

factor in the overall reaction. In a more complicated attempt, Zimmerman *et al.* (1968) completely neglected the very large solvation energies associated with the hydrophilic side-chains (Jencks, 1969) in formulating their polarity ranking, which is based on electrostatic forces in a vacuum. Furthermore, the crystal lattice energies inherent in the amino acid solubilities that were used for hydrophobicity parameters are also disregarded by these authors. Finally, a scale proposed by von Heijne & Blomberg (1979), although sophisticated in its intent, relies entirely on theoretical calculations, with scant reference to any empirical observation.

The water–vapor partition free energies (Table 1), which were first applied to the problem of protein folding by Wolfenden *et al.* (1979), also have shortcomings. The use of the vapor as the reference state leads to the incorporation of the dispersion forces into the transfer free energies. Since there is, in all likelihood, only a negligible and unpredictable contribution of dispersion forces to the process of protein folding (Deno & Berkheimer, 1963), these in principle should be subtracted from each value. Unfortunately, it is not even clear at the moment what the order of magnitude of these free energies is, let alone their individual values (Jencks, 1969). If they are roughly the same for each side-chain then their only effect would be to shift the whole scale uniformly without affecting the relative position of each. If these forces are roughly proportional to the volume of the side-chains they should also be fairly constant (Table 1), but difficulties could arise with very large and very small side-chains, such as tryptophan or glycine and alanine, respectively. Furthermore, it is not clear whether the vacuum is an adequate model for the interior of a protein with its collection of heterogeneous polarizabilities and oriented dipoles. Nevertheless, as pointed out by Wolfenden *et al.* (1979), the values for these transfer free energies correlate remarkably well with the actual distribution of the side-chains between the interior and exterior of protein molecules (Chothia, 1976).

If it is assumed, based on the observed correlation of the transfer free energies and the actual distribution of the side-chains (Wolfenden *et al.*, 1979), that both of these parameters are, to the first approximation, measurements of the hydropathy of a given amino side-chain, then the best available hydropathy index should be based on a consideration of both of these quantities. This follows from the fact that each of them suffers from its own unique uncertainties. The transfer free energies incorporate dispersion forces of unknown magnitude. The distributions, based on examination of several protein structures, are calculated from a limited data base and are biased by steric features that are not yet understood. Since none of the drawbacks is shared by these two independent measures of hydropathy, the most satisfactory index should be formulated from a consideration of all of the available information, as has been done here. In addition, the hydropathy scale presented here, unlike many earlier ones, spans the entire hydropathy spectrum from the hydrophobic end to the hydrophilic. For the reasons discussed above, this is an essential aspect of any scale. It is a point that was also emphasized by Wolfenden *et al.* (1979).

The hydropathy values presented in Table 2, being singular numbers, do not have associated with them an indication of their uncertainty, such as for example, a standard deviation. In retrospect, some of these parameters are more reliable than

others. The most unequivocal values are those associated with leucine, isoleucine, valine, phenylalanine, methionine, threonine, serine, lysine, glutamine and asparagine. These ten residues together comprise slightly more than half (52% of the present census) of the amino acids found in proteins. Most of these side-chains have partial specific volumes between 50 and 100 $cm^3\ mol^{-1}$, which suggests that dispersion forces may not influence their rank. Their relative positions change little from the fraction 95% buried, to the fraction 100% buried, to the free energy of transfer (Table 2). As such, these residues anchor the scale and are probably those most responsible for its success.

There is a group of amino acids that are less reliable: cysteine is complicated by the problem of disulfide bonds; proline, by the lack of an adequate model compound in the transfer free energies as well as its tendency to participate in β-turns on the exterior; and aspartic acid, glutamic acid and tyrosine, by the large differences between their tendency to be buried and their free energies of transfer. Certain amino acids (tryptophan, tyrosine, glutamic acid and histidine) are very reluctant to bury the last 5% of their surface area while some (alanine, glycine and cysteine) are far more likely than the others to become fully buried. Finally, arginine was arbitrarily assigned a parameter of $-4{\cdot}5$, even though no arginine was found to be even as much as 95% buried (Chothia, 1976) and no model compound for arginine was employed in the water–vapor transfer studies of Wolfenden *et al.* (1979). It is possible that the parameter for this side-chain should be even more negative (Wolfenden *et al.*, 1981).

Glycine and alanine are especially difficult to categorize. Both lack satisfactory model compounds for phase-transfer studies. Because methane is such a small molecule, its relative hydrophobicity is probably seriously overestimated by water–vapor transfer energy, because of the ambiguity introduced by dispersion forces. Indeed, the use of hydrogen gas as a model compound for glycine (Wolfenden *et al.*, 1979) is such an extreme case of the problem that arises from the contributions of the dispersion forces when molecules of such radically different electron densities are compared, that its water–vapor transfer free energy is probably a meaningless number in this context, and it has not been included in Table 1. On the other hand, both alanine and glycine are quite insensitive to becoming fully buried (Table 2), which suggests that the side-chains contribute little energy one way or the other to protein folding because they are not hydropathic. The conclusion from these arguments is that the more alanine and glycine a segment contains the more equivocal its hydropathy becomes.

A rather interesting and unforeseen feature of the interaction between the various side-chains and water is that the aromatic amino acids, tryptophan, tyrosine and histidine, are far more polar than previously thought (Nozaki & Tanford, 1971). It has been noted that aromatic compounds are more soluble in water, by an order of magnitude, than their surface areas would indicate (Hermann, 1972). The phenylalanine side-chain, however, is much less hydrophilic than the other three and this suggests that it is the heteroatoms in the latter that are the major contributors to their hydrophilicity.

In this context, tryptophan is one of the most difficult residues to which to assign a hydropathy index. It has a fairly positive water–vapor transfer free energy

(Table 1), but much of this may result from large, favorable dispersion forces due to the residue's large volume, the opposite problem to the one experienced with glycine and alanine. Examination of the actual location of tryptophan in a number of proteins (Chothia, 1976), however, clearly indicates that this side-chain is infrequently totally buried (Table 2). In the specific case of the interaction of gramicidin with a phospholipid bilayer, it should be mentioned that its tryptophan residues are clustered at the two ends of the pore rather than being distributed evenly throughout, again suggesting an unexpected hydrophilicity for these residues and a reluctance to bury the last 5% of surface (Table 2). Although the hydropathy of tryptophan is relatively unimportant in considerations of soluble proteins, since its frequency is only about 1·2%, there are indications that it may be very significant in membrane-affiliated sequences. In particular, 18 of the 19 tryptophan residues in Ca^{2+}-ATPase seem to be within sequences directly associated with the bilayer (Allen *et al.*, 1980). Another instance is the tryptophan cluster in cytochrome b_5, noted above, which the hydropathy profile clearly positions at the aqueous interface (Fig. 5). Finally, earlier claims that tryptophan was the most hydrophobic of the amino acids were based entirely on transfer free energies between water and ethanol or dioxane (Nozaki & Tanford, 1971). It was not recognized at that time that when the tryptophan side-chain, which possesses a hydrogen-bond donor only, is transferred between water, a solvent with equal numbers of donors and acceptors, and ethanol or dioxane, solvents with excesses of hydrogen-bond acceptors, there is a net increase of one mole of hydrogen bond (mole indole)$^{-1}$ formed during the transfer, causing the side-chain to appear much more hydrophobic than it actually is. Using the same logic, it is clear that in a protein solution, which necessarily contains more acceptors than donors, the removal of the donor on tryptophan from access to the solvent is a significantly unfavorable reaction. It seems, when all points are considered objectively, that tryptophan is a fairly hydrophilic side-chain.

The particular values chosen for the amino acid hydropathies embody one of the major differences between the method presented here and a similar one proposed earlier by Rose (1978). He chose to employ water–ethanol transfer free energies in his scale, the disadvantages of which are noted above. He also chose to ignore, at least in principle, the hydrophilic force, the attraction that the aqueous solvent exhibits for many side-chains, by simply assigning a value of zero to all side-chains for which partition free energies were not listed in the Tables of Nozaki & Tanford (1971). In addition, Rose's curve-smoothing procedure, although mathematically sound, tends to remove a great deal of the simplicity and clarity of the unsmoothed moving average. In the program described in this paper the meaning of each value is clearly understood and a more distinct and graphic rendering of the sequence obtained.

In addition, we have extended the use of this approach to the area of membrane-spanning segments of protein sequences. In this regard, the most novel feature of the approach is that membrane-spanning segments can be identified and distinguished from sequences that merely pass through the interior of a protein (Table 4). Since it is these membrane-spanning portions of sequences that have proven to be most difficult to study (Allen *et al.*, 1980), a method for their

identification ought to be quite useful. For instance, the sequence of bacteriorhodopsin (Fig. 6) was correlated previously with an electron density map on the basis of several criteria (Engelman *et al.*, 1980). All of these earlier arguments were based, however, on an initial assignment of the transmembrane regions within the sequence. No explanation of how these decisions had been made was presented, and, *in lieu* of this, it can be assumed that the assignments made in Table 4 are based on more objective considerations. If the present assignments are correct, the criterion of ion-pairing used in the earlier study is no longer meaningful because the partners in the purported ion-pairs would be displaced from each other. The belief that the difficulty of buried charges can be overcome by forming an ion-pair is known to be naive inasmuch as virtually the same amount of free energy is required to bury an ion-pair as to bury a single fixed charge (Parsegian, 1969). The problem of burying charge in a protein is solved not by forming ion-pairs, but by titering the charge and forming a strong, internal hydrogen bond with the neutralized acid or base. In this light, the interactions proposed earlier between lysines and arginines, on the one hand, and carboxylates, on the other, are poor choices for two reasons. Carboxylates are weak bases and would be unable to withdraw the proton effectively from the very weak cationic acids. As a result, charge would be ineffectively neutralized. Furthermore, the difference in pK values between arginine or lysine and carboxylate is large, which would cause the hydrogen bond to be a weak one (Jencks, 1969). A much more favorable choice on both counts would be a hydrogen bond between lysine or arginine and tyrosinate anion. In fact, when the aligned sequences (Table 3) are examined, two potential hydrogen bonds of this type become immediately apparent; those between lysine 129 and tyrosine 79 on the one hand, and tyrosine 64 and arginine 82 on the other. In fact, neutral, strong hydrogen bonds of this type, in which the proton is retained preferentially by the phenolic oxygen, have been observed directly in lysine–tyrosine co-polymers (Kristof & Zundel, 1980).

The biological function of bacteriorhodopsin must also be considered. It has been suggested that this enzyme is a light-driven proton pump and it can be assumed that the protons are passed across the membrane along a relay system of lone pairs. The most reasonable possibility is that this relay system is a string of hydrogen-bonded carboxylic acids, since these groups can transfer protons efficiently through space by a simple rotation. Furthermore, only a string of carboxylic acids could shuttle protons fast enough to keep up with the turnover of the enzyme. A proton on lysine or arginine cannot be transferred to the lone pair of a water molecule in aqueous solution at a rate any greater than about 1 second^{-1} while the proton on a carboxylic acid can be transferred in the same reaction at 10^6 second^{-1} (Jencks, 1969). Although the former rate may be enhanced in a well-organized hydrogen-bonded network (Wang, 1968), it is unlikely that a proton traversing bacteriorhodopsin could pass through a hydrogen bond containing a basic amino acid side-chain. Again, examination of the aligned sequences (Table 3) suggests that the carboxylic acid side-chains at residues 204, 194, 85, 212, 115, 94 and 102, or some subset of these, are distributed appropriately across the membrane to form such a relay system.

Finally, the present assignment of the membrane-spanning sequences further

weakens the arguments used earlier (Engelman *et al.*, 1980) to correlate the sequence of bacteriorhodopsin with the electron density profile. In the first place, the total scattering power of the A sequence in Table 3 is not less than those of the others, and this would make its correlation with a low-intensity shaft unnecessary. More to the point of the present discussion, the model preferred in the earlier study (Engelman *et al.*, 1980) places sequence B, one of the most hydrophobic (Table 4), at a location where it is completely surrounded by protein; and sequence F, clearly the most hydrophilic (Table 4), at a location well-exposed to the alkane of the bilayer. These considerations demonstrate that an examination of the hydropathy of a given sequence may provide additional information to the crystallographer in situations where structural decisions are ambiguous. An assignment that is different from the previous one and that satisfies the demands of hydropathy more successfully would be to place sequence A into shaft 5, B into shaft 6, C into shaft 2, D into shaft 3, E into shaft 4, F into shaft 7 and G into shaft 1, in the enumeration of Engelman *et al.* (1980). This assignment positions the most hydrophilic sequence, F, in the location most shielded from the alkane, and the most hydrophobic, G, in the location most exposed to the alkane; juxtaposes sequences F, G, C and B, forming the proton relay system as well as permitting the strong hydrogen bonds mentioned earlier; and places the retinal that is attached to lysine 216 (Bayley *et al.*, 1981) in the very center of the carboxylate relay system, as well as at the location that it occupies within the projected neutron density profile (King *et al.*, 1980). Furthermore, no crossovers of the sequence are required, and the only long connection coincides fortuitously with the longest stretch of polar sequence, residues 158 to 175. This elaboration, as well as others presented earlier, is an example of the information that might be gained from an informed consideration of a hydropathy profile.

APPENDIX

A C program for evaluating the hydropathic character of sequence segments

```
main ()
        {
        int i,j,k,;
        float total;
        char residue;
          extern char code [23];
          extern float factor [23];
        char sequence[1099];
        float value[1099];

        j = 0;
        while (getchar() != '\n');
        while (j <1099) {
             for (i = 0; i <11; i++) getchar();
             while (j <1099){
             sequence[j++] = getchar();
        if (getchar() == '\n') break;
        }
        if (sequence[j - 1] == '*') break;
        }
```

```
j = (j - 1);
for (i = 0; i <j; i++) {
residue = sequence[i];
for (k = 0; k <23; k++)
if(residue == code[k]) value[i] = factor[k]; }
for (i = 0; i <(j - 6); i++) {
total = 0;
for (k = 0; k <7; k++) total = total + value[i + k];
printf("%4d  %c  %6.1f", i + 4, sequence[i+3], total);
for (k = 0; k <total; k++) {if(k == 29) printf(".");
else printf(" ");}
printf("X\n");
}
printf("\n");
}
 char code[]  "RKDBNSEHZQTGXAPVYCMILWF";
 float factor[] {0.0,0.6,1.0,1.0,1.0,3.6,1.0,1.3,1.0,1.0,3.8,4.1,4.1,6.3,
     2.9,8.7,3.2,7.0,6.4,9.0,8.2,3.6,7.2};
```

(See Experimental Procedures for more details.)

We thank many of our colleagues including J. Kraut and S. J. Singer for helpful discussions about many of the matters discussed in this paper. We are also grateful to S. Dempsey for assistance with various aspects of programming for the graph plotter. This work was supported by National Institutes of Health grants HL18576, HL26873, HL17879, RR00757 and by National Science Foundation grant PCM78-24284.

REFERENCES

Allen, G., Trinnaman, B. J. & Green, N. M. (1980). *Biochem. J.* **187**, 591–616.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperson, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981). *Nature (London)*, **290**, 457–465.

Anfinsen, C. B. (1973). *Science*, **181**, 223–230.

Bayley, H., Huang, K. S., Radhakrishnan, R., Ross, A. H., Takayaki, Y. & Khorana, H. G. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 2225–2229.

Bigelow, C. C. (1967). *J. Theoret. Biol.* **16**, 187–211.

Bonitz, S. G., Coruzzi, G., Thalenfeld, B. E. & Tzagoloff, A. (1980). *J. Biol. Chem.* **255**, 11927–11941.

Bull, H. B. & Breese, K. (1974). *Arch. Biochem. Biophys.* **161**, 665–670.

Buse, G. & Steffens, G. J. (1978). *Hoppe-Seyler's Z. Physiol. Chem.* **359**, 1005–1009.

Capaldi, R. A. & Vanderkooi, G. (1972). *Proc. Nat. Acad. Sci., U.S.A.* **69**, 930–932.

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

Chothia, C. & Janin, J. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 4146–4150.

Chou, P. Y. & Fasman, G. (1973). *J. Mol. Biol.* **74**, 263–281.

Cohn, E. J. & Edsall, J. T. (1943). *Proteins, Amino Acids, and Peptides as Ions and Dipolar Ions*, Reinhold, New York.

Cohn, E. J., McMeekin, T. L., Edsall, J. T. & Blanchard, M. H. (1934). *J. Amer. Chem. Soc.* **56**, 784–794.

Coruzzi, G. & Tzagoloff, A. (1979). *J. Biol. Chem.* **254**, 9324–9330.

Dayhoff, M. O. (1978). *Atlas of Protein Sequence and Structure*, vol. 5, supp. 1–3, National Biomedical Research Foundation, Washington, D.C.

Deno, N. C. & Berkheimer, H. E. (1963). *J. Org. Chem.* **28**, 2143–2144.

Doolittle, R. F. (1981). *Science*, **214**, 149–159.

Drickamer, L. K. (1977). *J. Biol. Chem.* **252**, 6909–6917.

Engelman, D. M., Henderson, R., McLachlan, A. D. & Wallace, B. A. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 2023–2027.
Eventhoff, W., Rossmann, M. G., Taylor, S. S., Torff, H.-J., Meyer, H., Keil, W. & Kiltz, H.-H. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 2677–2681.
Fisher, H. F. (1964). *Proc. Nat. Acad. Sci., U.S.A.* **51**, 1285–1291.
Fleming, P. J., Koppel, D. E., Lau, A. L. Y. & Strittmatter, P. (1979). *Biochemistry*, **18**, 5458–5464.
Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, Ng. H. (1970). *Biochemistry*, **9**, 1997–2008.
Fuller, S. D., Capaldi, R. A. & Henderson, R. (1979). *J. Mol. Biol.* **134**, 305–327.
Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). *J. Mol. Biol.* **120**, 97–120.
Heller, J. (1968). *Biochemistry*, **7**, 2906–2913.
Henderson, R. & Unwin, N. (1975). *Nature (London)*, **257**, 28–32.
Hermann, R. B. (1972). *J. Phys. Chem.* **76**, 2754–2759.
Hine, J. & Mookerjee, P. K. (1975). *J. Org. Chem.* **40**, 292–298.
Ho, M. K. & Guidotti, G. (1975). *J. Biol. Chem.* **250**, 675–683.
Janin, J. & Chothia, C. (1980). *J. Mol. Biol.* **143**, 95–128.
Jencks, W. P. (1969). *Catalysis in Chemistry and Enzymology*, McGraw-Hill, New York.
Kauzmann, W. (1959). *Advan. Protein Chem.* **14**, 1–63.
Kernighan, B. W. & Ritchie, D. M. (1978). *The C Programming Language*, Prentice-Hall. Englewood Cliffs, N.J.
Khorana, H. G., Gerber, G. E., Herlihy, W. C., Gray, C. P., Anderegg, R. J., Nihei, K. & Biemann, K. (1979). *Proc. Nat. Acad. Sci., U.S.A.* **76**, 5046–5050.
King, G. I., Mowery, P. C., Stoeckenius, W., Crespi, H. L. & Schoenborn, B. P. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 4726–4730.
Klotz, I. M. & Farnham, S. B. (1968). *Biochemistry*, **7**, 3879–3881.
Klotz, I. M. & Franzen, J. S. (1962). *J. Amer. Chem. Soc.* **84**, 3461–3466.
Kristof, W. & Zundel, G. (1980). *Biophys. Struct. Mech.* **6**, 209–225.
Kyte, J. (1972). *J. Biol. Chem.* **247**, 7642–7649.
Lifson, S. & Sander, C. (1979). *Nature (London)*, **282**, 109–111.
Nakashima, Y. & Konigsberg, W. (1974). *J. Mol. Biol.* **88**, 598–600.
Nobraga, F. G. & Tzagoloff, A. (1980). *J. Biol. Chem.* **255**, 9828–9837.
Nozaki, Y. & Tanford, C. (1971). *J. Biol. Chem.* **246**, 2211–2217.
Parsegian, A. (1969). *Nature (London)*, **221**, 844–846.
Pober, J. S. & Stryer, L. (1975). *J. Mol. Biol.* **95**, 477–481.
Rose, G. D. (1978). *Nature (London)*, **272**, 586–590.
Rose, G. D. & Roy, S. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 4643–4647.
Rose, J. K., Welch, W. J., Sefton, B. M., Esch, F. S. & Ling, N. C. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 3884–3888.
Sacher, R., Steffens, G. J. & Buse, G. (1979). *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 1385–1392.
Segrest, J. P. & Feldman, R. J. (1974). *J. Mol. Biol.* **87**, 853–858.
Sogin, D. C. & Hinkle, P. C. (1978). *J. Supramol. Struct.* **8**, 447–453.
Steck, T. L., Koziarz, J. J., Singh, M. K., Reddy, G. & Köhler, H. (1978). *Biochemistry*, **17**, 1216–1222.
Steffens, G. J. & Buse, G. (1979). *Hoppe-Seyler's Z. Physiol. Chem.* **360**, 613–619.
Strittmatter, P., Rogers, M. J. & Spatz, L. (1972). *J. Biol. Chem.* **247**, 7188–7194.
Tanaka, M., Haniu, M., Yasunobu, K. T., Yu, C. A., Yu, L., Wei, Y. H. & King, T. E. (1979). *J. Biol. Chem.* **254**, 3879–3885.
Tanford, C. (1968). *Advan. Protein Chem.* **23**, 121–282.
Thalenfeld, B. E. & Tzagoloff, A. (1980). *J. Biol. Chem.* **255**, 6173–6180.
Tomita, M. & Marchesi, V. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 2964–2968.
Tomita, M., Furthmayr, H. & Marchesi, V. (1978). *Biochemistry*, **17**, 4756–4770.
Traube, J. (1899). *Samml. Chem. Chem.-Tech. Vortr.* **4**, 19–332.

Vandlen, R. L., Wu, W. C. S., Eisenach, J. C. & Raftery, M. A. (1979). *Biochemistry*, **18**, 1845–1854.
von Heijne, G. & Blomberg, C. (1979). *Eur. J. Biochem.* **97**, 175–181.
Wang, J. H. (1968). *Science*, **161**, 328–334.
Wolfenden, R. V., Cullis, P. M. & Southgate, C. C. F. (1979). *Science*, **206**, 575–577.
Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. (1981). *Biochemistry*, **20**, 849–855.
Wu, T. T. & Kabat, E. A. (1973). *J. Mol. Biol.* **75**, 13–31.
Zimmerman, J. M., Eliezer, N. & Simha, R. (1968). *J. Theoret. Biol.* **21**, 170–201.

Edited by M. F. Moody

Note added in proof: A similar prediction method has recently been reported by Hopp & Woods (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 3824–3828.