J. Mol. Biol. (1987) 195, 659–685

Hydrophobicity Scales and Computational Techniques for Detecting Amphipathic Structures in Proteins

James L. Cornette[1,3], Kemp B. Cease[2], Hanah Margalit[1], John L. Spouge[1]
Jay A. Berzofsky[2] and Charles DeLisi[1]

[1]*Laboratory of Mathematical Biology and* [2]*Metabolism Branch*
National Cancer Institute
National Institutes of Health
Bethesda, MD 20892, U.S.A.
[3]*Department of Mathematics*
Iowa State University
Ames IA 50011, U.S.A.

(Received 24 July 1986, and in revised form 6 January 1987)

Protein segments that form amphipathic α-helices in their native state have periodic variation in the hydrophobicity values of the residues along the segment, with a 3·6 residue per cycle period characteristic of the α-helix. The assignment of hydrophobicity values to amino acids (hydrophobicity scale) affects the display of periodicity. Thirty-eight published hydrophobicity scales are compared for their ability to identify the characteristic period of α-helices, and an optimum scale for this purpose is computed using a new eigenvector method. Two of the published scales are also characterized by eigenvectors.

We compare the usual method for detecting periodicity based on the discrete Fourier transform with a method based on a least-squares fit of a harmonic sequence to a sequence of hydrophobicity values. The two become equivalent for very long sequences, but, for shorter sequences with lengths commonly found in α-helices, the least-squares procedure gives a more reliable estimate of the period. The analog to the usual Fourier transform power spectrum is the "least-squares power spectrum", the sum of squares accounted for in fitting a sinusoid of given frequency to a sequence of hydrophobicity values.

The sum of the spectra of the α-helices in our data base peaks at 97·5°, and approximately 50% of the helices can account for this peak. Thus, approximately 50% of the α-helices appear to be amphipathic, and, of those that are, the dominant frequency at 97·5° rather than 100° indicates that the helix is slightly more open than previously thought, with the number of residues per turn closer to 3·7 than 3·6. The extra openness is examined in crystallographic data, and is shown to be associated with the C terminus of the helix.

The alpha amphipathic index, the key quantity in our analysis, measures the fraction of the total spectral area that is under the 97·5° peak, and is a characteristic of hydrophobicity scales that is consistent for different sets of helices. Our optimized scale maximizes the amphipathic index and has a correlation of 0·85 or higher with nine previously published scales. The most surprising feature of the optimized scale is that arginine tends to behave as if it were hydrophobic; i.e. in the crystallographic data base it has a tendency to be on the hydrophobic face of the amphipathic helix. Although the scale is optimal only for predicting α-amphipathicity, it also ranks high in identifying β-amphipathicity and in distinguishing interior from exterior residues in a protein.

We factor the expressions for the power spectra into a matrix product so that the helical sequence information is isolated from the hydrophobicity scale. The largest eigenvalue of the matrix containing only helical sequence information also identifies the 97·5° frequency, thus confirming a 3·7 residue per turn spacing independently of hydrophobicity scales.

0022-2836/87/110659-27 $03.00/0

1. Introduction

We examine procedures for associating amphipathic secondary structure in a protein molecule with the sequence of hydrophobicity values of its residues. It has long been recognized that the regular, organized structure of a protein embedded in a non-isotropic environment would be reflected in recognizable patterns in the sequence of chemical properties of the residues in the protein. Perutz *et al.* (1965) observed that on α-helices in myoglobin and hemoglobin, hydrophobic residues tend to cluster on one side, with hydrophilic residues on the opposite side, resulting in what is called an amphipathic α-helix. Because one turn of the α-helix requires approximately 3·6 residues, along any peptide that constitutes an amphipathic α-helix there will be periodic variation in the hydrophobicity values with a period of approximately 3·6 residues per cycle. Lim (1974*a*) identified hydrophobic pairs and triplets and hydrophobic–hydrophilic pairs of appropriately spaced amino acids that would be characteristic of α-helices, and of β-structural fragments. In α-tropomyosin, McLachlan & Stewart (1976) identified repetition at a period of $19\frac{2}{3}$ residues per cycle, of both non-polar and charged amino acids, that they associated with a supercoiled structure of two α-helices. Inherent to such analyses is the hypothesis that the anisotropic environment partly induces the secondary structure of the protein, a hypothesis that has been tested experimentally (Kaiser & Kézdy, 1983).

Several qualitative, algorithmic and quantitative techniques have been introduced to model and detect periodic variation in chemical properties along the protein sequence that is characteristic of secondary structural features. Qualitatively, Schiffer & Edmundson (1967) introduced helical wheels, circles on which are marked the hydrophobicity character of the residues in a protein segment at 100° intervals around the unit circle (100° = 360°/3·6 residues per cycle = 100°/residue for an α-helix). If the protein segment makes up an α-helix with hydrophobic residues on one side and hydrophilic residues on the other, an amphipathic α-helix, then, on the circle, characters signaling hydrophobic residues will be clustered to one side with characters signaling hydrophilic residues on the opposite side (see Fig. 1 (a)). A similar qualitative geometric model is the helical net introduced by Dunnill (1968), in which the helix is visualized as being cut along one side, parallel to its axis and flattened onto a plane (Fig. 1 (b)). Lim (1974*b*) used his α-helical and anti-α-helical pairs and triplets to develop an algorithm for localization of α-helical regions. Quantitatively, the auto correlation function, $\Sigma h_k h_{k+p}$, where h_k is the hydrophobicity of the kth residue, is frequently used to detect periodicity in sequences of hydrophobicity values; a value of p for which $\Sigma h_k h_{k+p}$ is "large" is a characteristic period. McLachlan & Stewart (1976) were among the first to use discrete Fourier analysis (described below) to analyze protein sequences, and they developed a statistical analysis to distinguish significant Fourier intensities from intensities that may occur by chance. Eisenberg *et al.* (1982*a*) introduced the mean helical hydrophobic moment, which gave a quantitative interpretation to the helical wheels, as follows. For each residue plotted on the wheel, multiply the unit vector from the center of the wheel to the position of the residue on the wheel by the hydrophobicity of the residue. Sum the vectors; the magnitude of the resultant is the mean helical hydrophobic moment†. Eisenberg *et al.* (1984) interpret the

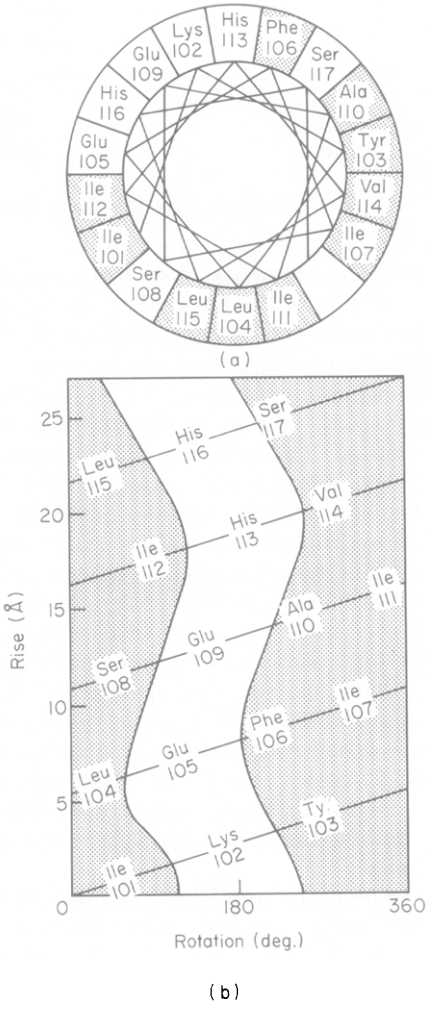

Figure 1. A helix of sperm whale myoglobin (usually labeled helix G) is identified by the Kabsch–Sander algorithm as residues Ile101 to Ser117. The residues are shown distributed (a) around the helical wheel at 100° intervals and (b) along the trace of a helix marked on the helical net. The hydrophilic path shown in the unshaded portion of the net corresponds to the arc from Glu105 to His113 of the wheel. Arg118 is also assigned to helix G by the crystallographers and would be placed on the helical wheel in the one remaining blank, surrounded by hydrophobic residues (see Discussion).

† To introduce the mean helical moment, Eisenberg *et al.* actually use for each residue the unit vector pointing from the nucleus of the α-carbon to the geometric center of the side-chain; their computation, however, is based on the unit vector we have described.

hydrophobic moment as the modulus of the discrete Fourier transform evaluated at 100°, and examine the modulus of the Fourier transform for the frequencies 80° to 180°, corresponding to alternative spacing of residues around the helical wheel. They sum the hydrophobic moments of 157 α-helices of length greater than or equal to 7 and observe a very distinct peak around the 100° per residue frequency that is interpreted as typical of the amphipathic α-helix. A reproduction of that curve (with only 145 helices) appears as the continuous-line curve in Figure 2(a). It is important to recognize that it is the *amphipathic* α-helices that contribute to the formation of the characteristic 100° peak; the other helices presumably contribute randomly to the sum of the hydrophobic moments (but see Results, section (b)). Eisenberg *et al.* (1984) also plot the sum of the hydrophobic moments of 220 segments of β-structure of length greater than or equal to 4 and observe a peak around 180° that, however, is not so clearly defined. Graphs similar to their beta-segments curve appear at the end of this paper in Figure 12.

Also shown in Figure 2(a) (broken-line graph) is the comparable curve for the same set of α-helices computed using the hydrophobicity scale of Kyte & Doolittle (1982) instead of the Eisenberg *et al.* (1982*b*) "consensus" scale that was used to calculate the continuous-line curve. Both curves are normalized to be 1000 at the frequency angle of 100°. It is clear that the choice of hydrophobicity scale has an influence on the sharpness of the signal around 100°, and a comparison of 38 published hydrophobicity scales for their effectiveness in identifying characteristic α-helical and β-strand periodicities is given in Results, section (a). We introduce an "amphipathic index" as an objective measure of the sharpness of the 100° signal and compute hydrophobicity scales that maximize the amphipathic index for various sets of helices and computational techniques. These different scales are highly correlated and scales that maximize the amphipathic index for one set of helices yield high values of the amphipathic index for other sets of helices. An interesting characteristic of the "optimum" scales is that all of them assign a hydrophobic value to arginine, suggesting that arginine tends to appear on the hydrophobic side of helices. The amphipathic index is used to identify amphipathic helices within our data sets of helices, and we conclude that approximately one-half of the helices we study are amphipathic.

We extract from the Fourier transform power spectrum a parameter (the largest eigenvalue of a certain matrix) computed only from sequence information in a set of helices and independent of hydrophobicity information, which also detects the natural frequency inherent in helices. The frequency detected by this parameter is approximately 97·5°, suggesting that the typical α-helix is slightly more "open" than is usually thought, with 3·7 residues per turn being typical. The extra "openness" is shown to be associated with the C terminus of the helix.

The converse of the Eisenberg *et al.* (1984) result is that a protein segment that exhibits a peak in hydrophobic moment at or near 100° has a likelihood of being an amphipathic α-helix and a segment that exhibits a peak in hydrophobic moment near 180° has a likelihood of being a β-segment. These implications have been applied in algorithms to identify secondary structure in the acetylcholine receptor (Finer-Moore & Stroud, 1984), to associate amphipathic structures with

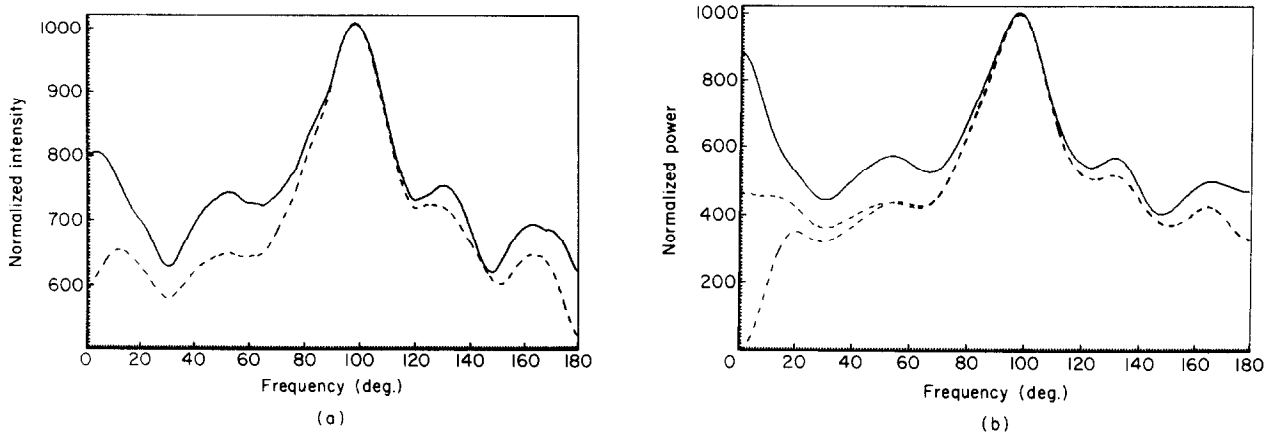

Figure 2. (a) The sum of the Fourier intensities of 147 α-helices computed by Eisenberg *et al.* (1984) using the Eisenberg *et al.* scale (continuous line) and the Kyte–Doolittle scale (broken line). The Fourier intensity of a single helix at the frequency ω is

$$\left\{\left[\sum_{k=0}^{l-1} h_k \cos k\omega\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin k\omega\right]^2\right\}^{1/2}$$

where $h_0, \ldots, h_{l-1}$ are the hydrophobicity values of the residues in the helix. (b) The sum of the Fourier powers (= square of the intensity) for the same helices as in (a). Also, branching to the origin for the Kyte–Doolittle scale is the sum of the Fourier powers of $\{h_k - \bar{h}\}_{k=0}^{l-1}$ for those helices, where $\bar{h}$ is the average hydrophobicity of the helix. All curves are normalized to be 1000 at 100° frequency. The Kyte–Doolittle scale identifies a sharper 100° signal.

T-cell antigenic sites (DeLisi & Berzofsky, 1985), and to classify proteins as α-rich, β-rich, etc. (Klein & DeLisi, 1986). The selection of a hydrophobicity scale that sharply defines the characteristic signals of amphipathic structures can improve the success of these algorithms.

We describe the computational procedures of the discrete Fourier transform and of an alternative least-squares procedure that, for short peptides, provides a more reliable estimate of periodicity. Both the Fourier transform and least-squares power functions are formulated in the language and notation of linear algebra. A novel approach using matrices and eigenvectors is introduced to enable computation of optimum scales and to detect periodicity in helices without reference to hydrophobicity values. Matrices and eigenvectors also are used to characterize two scales from the literature, Sweet & Eisenberg (1983) and Ponnuswamy *et al.* (1980). Our analytical discussion of Fourier transform and least-squares is brief; for more detail, the reader is referred to the excellent exposition of these topics presented by Bloomfield (1976).

2. Data Bases and Mathematical Methods

(a) *Proteins and secondary structure*

Our primary *set* of α-helices is based on the 23 proteins used by Eisenberg *et al.* (1984) in their study of α-helices. According to the crystallographers' designation of helices in the specific versions that we used of the 23 proteins, there were 145 helices having 7 or more residues. Kabsch & Sander (1983) have an alternative, more systematic algorithm for assigning residues to secondary structure from the crystallographic data of a protein, and for the same set of proteins there were 115 Kabsch–Sander helices of length 7 or more residues in the α-helical conformation in 21 of the proteins and no such helices in 2 of the proteins (2B5C cytochrome *b*5, and 1MHR myohemerythrin). Fig. 3 is a comparison of the sum of

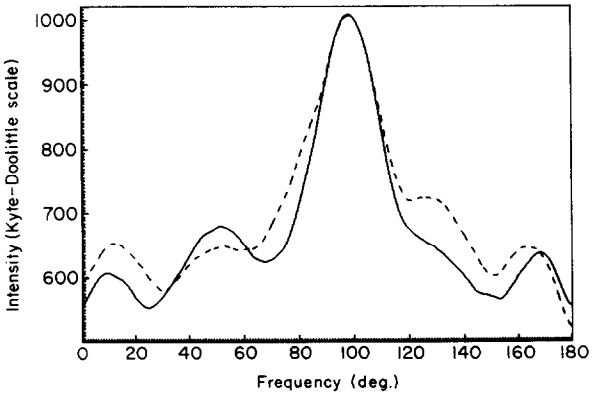

Figure 3. The sum of the Fourier intensities of 147 helices identified by the crystallographers (broken line) and of 115 helices identified by the Kabsch–Sander algorithm (continuous line). The helices are of length 7 or more residues in the 21 proteins noted under Primary list in Table 1 together with 2B5C cytochrome *b*5 and 1MHR myohemerythrin. The Kabsch–Sander helices exhibit a sharper amphipathic signal than do the crystallographers' helices.

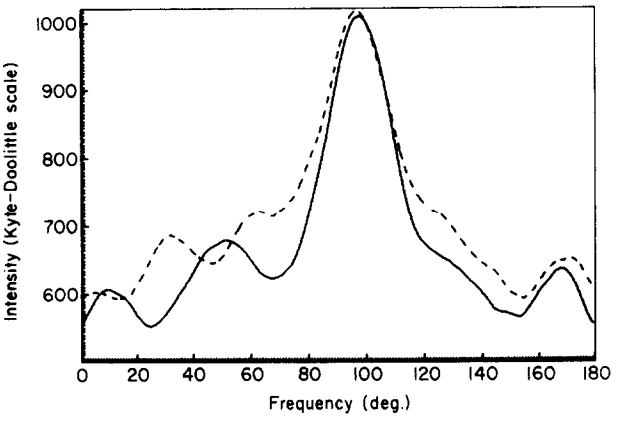

Figure 4. The sum of the Fourier intensities for the helices in the *primary* set (continuous line) and in the alternative set (broken line) of helices using the Kyte–Doolittle hydrophobicity scale. The amphipathic signal for the primary helices is sharper than that of the alternative helices.

the hydrophobic moments of the set of 145 helices with the sum of the hydrophobic moments of the 115 Kabsch–Sander helices, using the Kyte–Doolittle hydrophobicity scale. A small but distinct sharpening of the peak around $\omega = 100°$ for the Kabsch–Sander helices can be seen, and the same improvement was observed for almost all of the hydrophobicity scales that we examined. As a consequence, our primary set of α-helices consists of the 115 Kabsch–Sander helices in the 21 proteins listed in Table 1 under Primary list. Residues in the 3_{10} conformation are not included.

We selected an alternative set of α-helices to test the consistency of our results. There are 21 proteins, noted under Alternative list in Table 1, from which the alternative set of 135 Kabsch–Sander α-helices were selected; each had at most a small amount of homology with any of the proteins for the primary set and each contained at least 2 Kabsch–Sander helices of length 7 or more residues. It will be seen that the amphipathic character of the alternative set of helices is not so sharply defined as for the primary set of helices. (See the hydrophobic moment profiles shown in Fig. 4, and in Table 6, observe the systematically lower amphipathic indices shown under Alternative list as contrasted with those shown under Primary list.) For a few computations, we combine the primary and alternative sets of helices into the total set of helices.

To test the hydrophobicity scales for identification of β-strands, we used the Kabsch–Sander-designated β-strands in the 23 proteins from which Eisenberg *et al.* (1984) selected β-strands. Of the 23 proteins that they used, 22 are listed under Beta list in Table 1 (the 23rd, 1ABP L-arabinose-binding protein has no Kabsch–Sander β-strand of length 4 or more residues). In the 23 proteins, the crystallographers identify 220 β-strands of length 4 or more residues, whereas the Kabsch–Sander algorithm identifies 161. Our β-strands are the Kabsch–Sander β-strands of length 4 or more residues in the proteins noted under Beta list in Table 1.

As a brief comparison of the hydrophobicity scales for identifying interior *versus* exterior residues, we have quantitatively compared their "hydropathy profiles" (as described by Kyte & Doolittle (1982)) for the protein, dogfish M_4 (muscle) lactate dehydrogenase. This protein was used for illustration by Kyte & Doolittle (1982), and

Table 1
Proteins used

Protein data bank		Primary list	Alternative list	Beta list	Nearest neighbor matrix
2APE	Acid proteinase, endothiapepsin			B	
2APP	Acid proteinase, penicillopepsin				N
2ACT	Actinidin (sulfhydryl proteinase)		A	B	N
2ADK	Adenylate kinase	P		B	N
4ADH	Apo-liver alcohol dehydrogenase	P		B	N
1ALP	Alpha lytic protease				N
2TAA	Taka-*amylase A		A		N
1ABP	L-*Arabinose-binding protein	P		B	N
2ATC	Aspartate carbamoyltransferase (*E. coli*)		A		
3CPV	Calcium-binding parvalbumin B	P			N
1CAC	Carbonic anhydrase form C (human)	P		B	N
4CPA	Carboxypeptidase A (bovine)/potato inhibitor	P		B	N
3CAT	Catalase		A		
3CHA	Alpha chymotrypsin A			B	
2GCH	Gamma chymotrypsin A				N
1CTS	Citrate synthase—citrate complex		A		
3CNA	Concanavalin A			B	
1CRN	Crambin		A		
156B	Cytochrome *b*562 (*E. coli* oxidized)	P			
3CYT	Cytochrome $*c* (oxidized)	P			
4CYT	Cytochrome $*c* (reduced)				N
1CCY	Cytochrome $*c* (prime)	P			N
1CYP	Cytochrome $*c* peroxidase (yeast)		A		
2C2C	Cytochrome $*c* = 2 = (oxidized)				N
351C	Cytochrome $*c* = 551 = (oxidized)	P			
3DFR	Dihydrofolate reductase (*Lactobacillus casei*)				N
4DFR	Dihydrofolate reductase (*E. coli*)	P		B	N
1EST	Tosyl-elastase			B	N
1ECD	Hemoglobin (erythrocruorin, deoxy)		A		N
2FD1	Ferredoxin				N
3FXN	Flavodoxin (oxidized form)	P		B	
4FXN	Flavodoxin (semiquinone form)				N
2GRS	Glutathione reductase		A		
1GPD	D-Gyceraldehyde-3-phosphate dehydrogenase		A		N
1HMQ	Hemerythrin (met)		A		
2MHB	Hemoglobin (horse, aquo met)	P			N
1LHB	Hemoglobin (lamprey, met)		A		N
1REI	Bence-*Jones immunoglobulin/REI var. portion				N
1FC1	Immunoglobulin FC fragment (IgG1 class)				N
4LDH	Lactate dehydrogenase apo enzyme M_4	P		B	N
1LH1	Leghemoglobin (acetate, met) (yellow lupine)		A		N
1LZM	Lysozyme (bacteriophage T4)		A		N
2LYZ	Lysozyme (hen egg-white)		A		N
2MDH	Cytoplasmic malate dehydrogenase		A		
1MLT	Melittin	P			
2MBN	Myoglobin (met)	P			
1MBD	Myoglobin (deoxy, $p*H 8·4)				N
8PAP	Papain		A	B	N
3PGK	Phosphoglycerate kinase (yeast)		A		N
3PGM	Phosphoglycerate mutase (yeast)		A		N
1BP2	Phospholipase A2 (bovine)	P		B	N
1PCY	Plastocyanin (Cu^{2+}, $p*H 6·0)			B	
2PAB	Prealbumin (human plasma)			B	
2SGA	Proteinase A (*Streptomyces griseus*)				N
1RHD	Rhodanese	P		B	N
1RN3	Ribonuclease A				N
2SNS	Staphylococcal nuclease		A		N
2SBT	Subtilisin novo	P		B	
2SOD	Cu, Zn superoxide dismutase			B	N
3TLN	Thermolysin	P		B	N
1TIM	Triose phosphate isomerase	P		B	N
2PTN	Trypsin (orthorhombic, 2·4 M-ammonium sulfate)		A		
3PTP	Beta trypsin, diisopropylphosphoryl inhibited			B	
2PTC	Trypsin/trypsin inhibitor complex				N

The primary and alternative sets of helices are the helices identified from crystallographic data by the Kabsch–Sander algorithm in the proteins noted under Primary list or Alternative list. The set of β-strands used was the Kabsch–Sander designated β-strands in the proteins noted under Beta list. The nearest neighbor matrix described in Results, section (e), was computed using the crystallographic co-ordinates of the residues in the proteins noted under Nearest neighbor matrix.

a designation of interior and exterior residues is available in the literature (Eventhoff *et al.* (1977)). For the computation, the residues labeled "internal cavity" by Eventhoff *et al.* (1977) were considered to be internal residues.

For computation of the "nearest-neighbor matrix", described in Results, section (e), we used the proteins that are the basis of the Miyazawa & Jernigan (1985) study, noted under Nearest neighbor matrix in Table 1.

(b) *Detection of periodicity in hydrophobicity values*

Given a sequence $h_0, h_1, \ldots, h_{l-1} = \{h_k\}_{k=0}^{l-1}$ of hydrophobicity values of residues along a protein segment of length l, the Fourier transform power spectrum, $P(\omega)$ of eqn (1), may be used to detect periodic variation in the sequence:

$$P(\omega) = \left[\sum_{k=0}^{l-1} h_k \cos k\omega\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin k\omega\right]^2. \quad (1)$$

In Appendix I we interpret $P(\omega)$ as

$$\max_{\phi}\left[\sum_{k=0}^{l-1} h_k \cos(k\omega+\phi)\right],$$

the maximum (non-normalized) correlation of $\{h_k\}_{k=0}^{l-1}$ with any sinusoid of frequency ω. A sequence $\{h_k\}_{k=0}^{l-1}$ that is highly correlated with a periodic function tends to be periodic also. Thus, the value $\omega = \hat{\omega}$ for which $P(\omega)$ is maximum is a characteristic frequency of $\{h_k\}_{k=0}^{l-1}$, and $2\pi/\hat{\omega}$ is a characteristic period.

The continuous-line graph in Fig. 2(a) is the sum of the square-roots of 145 different power spectra, $P(\omega)$, each of which is computed for the sequence of Eisenberg *et al.* (1982*b*) hydrophobicity values in 1 of 145 helices. For algebraic simplicity we prefer to avoid the square-root; the continuous-line graph in Fig. 2(b) is the sum of the same powers without taking square-roots. In both curves the maximum power near the characteristic frequency $\hat{\omega} = 100°$ is evident.

Of the 38 published hydrophobicity scales that we have examined, 15 assign only positive numbers as hydrophobicity values to all 20 amino acids. For those scales particularly, the power spectra have an unusually large value at 0 that (1) may mask important periodicities, and (2) will distort the amphipathic index defined in the next section. To avoid this, we look for periodicity in $\{h_k - \bar{h}\}_{k=0}^{l-1}$, where:

$$\bar{h} = \left(\sum_{k=0}^{l-1} h_k\right)/l$$

is the mean hydrophobicity of the sequence. Thus, our computations are based on:

$$P_m(\omega) = \left[\sum_{k=0}^{l-1} (h_k - \bar{h}) \cos k\omega\right]^2 + \left[\sum_{k=0}^{l-1} (h_k - \bar{h}) \sin k\omega\right]^2. \quad (2)$$

It is shown in Appendix I that $P_m(0) = 0$, so that the previous problem is avoided. The contrast may be seen in Fig. 2(b), where one of the graphs for the Kyte–Doolittle scale branches down to the origin. A subscript m always signals that the mean $\bar{h}$ has been subtracted, so that, for example, $\hat{\omega}_m$ is the value of ω that maximizes $P_m(\omega)$.

An alternative to eqn (1) for comparing $\{h_k\}_{k=0}^{l-1}$ to a sinusoid is to find the best least-squares fit of a sinusoid to $\{h_k\}_{k=0}^{l-1}$. An analysis of this procedure is presented in Appendix I, where a least-squares spectrum $Q(\omega)$ similar to $P(\omega)$ is developed. For short, individual helices, the least-squares procedure more accurately detects periodicity, as shown in Results, section (f). However, for long helices or for an aggregate of helices, the procedures yield similar results. Therefore, except for Results, section (f), we have used the more familiar Fourier transform, and the remainder of the paper may be read without reference to Appendix I.

(c) *Amphipathic index and interior/exterior correspondence*

We have defined an amphipathic index (AI) for any power spectrum, written generically as $P(\omega)$, either of a single helix or the composite spectrum of a collection of helices as:

$$\mathrm{AI}[P(\omega)] = \frac{\frac{1}{25°}\int_{85°}^{110°} P(\omega)\,d\omega}{\frac{1}{180°}\int_{0°}^{180°} P(\omega)\,d\omega}. \quad (3)$$

The rationale for centering the interval of integration in the numerator at 97·5° is presented in Results, sections (b) to (d). The width of that interval corresponds roughly to the distance between half maxima on each side of the peak of the spectra. Briefly, the amphipathic index is a measure of how much of the power spectrum is concentrated around 97·5°, as compared to the total area under the spectrum. A helix whose power spectrum exhibits a large amphipathic index is very likely to be amphipathic according to the scale used to compute the power spectrum. The nature of the amphipathic index is demonstrated by its values for 3 consecutive α-helices in citrate synthase (residues 300 to 310, 328 to 340 and 345 to 363) evaluated with the Kyte–Doolittle hydrophobicity scale. The sequences, the sequences of Kyte–Doolittle hydrophobicity values, and the amphipathic indices are shown in Table 2. The helix 300–310 has predominantly hydrophilic residues (negative hydrophobicity values), but 3 well-placed hydrophobic residues produce an amphipathic helix, and the amphipathic index is unusually high (3·364). Similarly, the helix 345–363, although predominantly hydrophobic, has hydrophilic residues well-placed for creating an amphipathic helix, and the amphipathic index is also quite high (3·172). The intervening helix 328–340, however, has a long hydrophilic segment followed by a hydrophobic segment (according to the Kyte–Doolittle scale). It clearly is not amphipathic as a helix, and it has an unusually low amphipathic index (0·244).

We also define an amphipathic index for the power spectrum of a collection of β-strands as:

$$\frac{\frac{1}{20°}\int_{160°}^{180°} P(\omega)\,d\omega}{\frac{1}{180°}\int_{0°}^{180°} P(\omega)\,d\omega}.$$

This index is referred to as β-amphipathic index. Unmodified, "amphipathic index" refers to eqn (3).

We also compare the scales for consistency of performance in the context of tertiary structure, using as an objective test a computation based on the "hydropathy profile" of dogfish lactate dehydrogenase as described by Kyte & Doolittle (1982). Kyte & Doolittle (1982) assign a hydropathy (hydrophobicity) value to

Table 2
Amphipathic indices of $P_m(\omega)$ *and* $Q_m(\omega)$ *for three helices in citrate synthase*

Helical peptide	Sequence of residues Kyte–Doolittle hydrophobicities																			Amphipathic index of $P_m(\omega)$
300–310	Lys	Leu	Arg	Asp	Tyr	Ile	Trp	Asn	Thr	Leu	Asn									3·364
	−3·9	3·8	−4·5	−3·5	−1·3	4·5	−0·9	−3·5	−0·7	3·8	−3·5									
328–340	Pro	Arg	Tyr	Thr	Cys	Gln	Arg	Glu	Phe	Ala	Leu	Lys	His							0·244
	−1·6	−4·5	−1·3	−0·7	−0·4	−3·5	−4·5	−3·5	2·8	1·8	3·8	−3·9	−3·2							
345–363	Pro	Met	Phe	Lys	Leu	Val	Ala	Gln	Leu	Tyr	Lys	Ile	Val	Pro	Asn	Val	Leu	Leu	Glu	3·172
	−1·6	1·9	2·8	−3·9	3·8	4·2	1·8	−3·5	3·8	−1·3	−3·9	4·5	4·2	−1·6	−3·5	4·2	3·8	3·8	−3·5	

The hydrophobic and hydrophilic residues in helices 300–310 and 345–363 of citrate synthase alternate with approximately 3·6 residue per cycle periodicity characteristic of amphipathic α-helices, and the amphipathic indices of both helices is quite high. The intermediate helix, 328–340, however, is not amphipathic and the amphipathic index is quite low.

each residue of the enzyme, compute the mean hydropathy of each block of 9 contiguous residues, and assign the mean value to the mid-residue of the block. The hydropathy profile is a plot of the mean hydropathy value *versus* residue position number. They marked on the profile the locations of the interior and exterior residues as specified by Eventhoff *et al.* (1977), and observed a "remarkable correspondence between the interior portions of the 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". We quantify the correspondence by computing the "interior/exterior correspondence" defined to be the sum of the 9-residue block mean hydrophobicity values assigned to the interior residues minus the sum of the mean values assigned to the exterior residues. In order to compare different scales equally, it is necessary to use the normalized versions of the scales (see Hydrophobicity Scales, and Table 4).

(d) *Quadratic forms for* $P(\omega)$, $P_m(\omega)$, $Q(\omega)$, *and* $Q_m(\omega)$

For any 20 by 1 vector, $\boldsymbol{\eta}$, representing a hydrophobicity scale and any collection, S, of helices, the power spectrum of eqn (1) may be written:

$$P(\omega) = \boldsymbol{\eta}^{\mathrm{T}} W_S(\omega) \boldsymbol{\eta},$$

where $W_S(\omega)$ is a 20 by 20 matrix that is characteristic of S and the type of power spectrum and independent of $\boldsymbol{\eta}$. The factorization is also possible for $P_m(\omega)$ of eqn (2) or either least-squares spectrum $Q(\omega)$ or $Q_m(\omega)$ described in Appendix I. We present the analysis for $P(\omega)$ and state the results for $P_m(\omega)$, $Q(\omega)$ and $Q_m(\omega)$.

It is very useful to recognize from eqn (1) that for any fixed frequency ω, $P(\omega)$ is simply a quadratic expression in hydrophobicity values that is never negative. That is, with respect to the 20 hydrophobicity values as variables, $P(\omega)$ is a non-negative second-degree polynomial, commonly referred to as a non-negative definite quadratic form. The algebra of non-negative definite quadratic forms is well-understood and provides excellent insight into the periodic information contained in a set of α-helices.

Assume that the amino acids are ordered into a specific list (we use the order of Table 3, but any order is acceptable), and let $\boldsymbol{\eta}$ denote a 20 by 1 column vector representing a hydrophobicity scale listing the hydrophobicity values for each amino acid in the same order. If a protein segment, σ, has residues $\{R_0, \ldots, R_{l-1}\}$, the corresponding sequence $\{h_k\}_{k=0}^{l-1}$ of hydrophobicity values of the residues will be determined by the rule: h_k is η_j if R_k is the jth amino acid in the established order. Accordingly, we construct an l by 20 "selection" matrix $E_\sigma = [e_{k,j}]$ so that $e_{k,j} = 1$ if the kth residue of σ is the jth amino acid in the order, and $e_{k,j} = 0$ otherwise. Then the sequence $\{h_k\}_{k=0}^{l-1}$, thought of as an l by 1 column vector, $\mathbf{h}$, is represented by $\mathbf{h} = E_\sigma \boldsymbol{\eta}$. We let $\mathbf{c}$ and $\mathbf{s}$ denote, respectively, 20 by 1 column vectors $[\cos k\omega]$ and $[\sin k\omega]$, where the index k ranges from 0 to $l-1$. Then we compute $P(\omega)$ for a protein segment σ as:

$$\begin{aligned} P_\sigma(\omega) &= \left(\sum_{k=0}^{l-1} h_k \cos k\omega\right)^2 + \left(\sum_{k=0}^{l-1} h_k \sin k\omega\right)^2 \\ &= (\mathbf{c}^{\mathrm{T}}\mathbf{h})^2 + (\mathbf{s}^{\mathrm{T}}\mathbf{h})^2 \qquad (\mathbf{v}^{\mathrm{T}} = \mathbf{v} \text{ transpose}) \\ &= (\mathbf{c}^{\mathrm{T}} E_\sigma \boldsymbol{\eta})^2 + (\mathbf{s}^{\mathrm{T}} E_\sigma \boldsymbol{\eta})^2 \\ &= (\boldsymbol{\eta}^{\mathrm{T}} E_\sigma^{\mathrm{T}} \mathbf{c})(\mathbf{c}^{\mathrm{T}} E_\sigma \boldsymbol{\eta}) + (\boldsymbol{\eta}^{\mathrm{T}} E_\sigma^{\mathrm{T}} \mathbf{s})(\mathbf{s}^{\mathrm{T}} E_\sigma \boldsymbol{\eta}) \\ &= \boldsymbol{\eta}^{\mathrm{T}} \{E_\sigma^{\mathrm{T}} (\mathbf{c}\mathbf{c}^{\mathrm{T}} + \mathbf{s}\mathbf{s}^{\mathrm{T}}) E_\sigma\} \boldsymbol{\eta} \\ &= \boldsymbol{\eta}^{\mathrm{T}} E_\sigma^{\mathrm{T}} F(\omega) E_\sigma \boldsymbol{\eta}, \end{aligned} \tag{4}$$

where the k,j entry of $F(\omega)$ is $\cos k\omega \cos j\omega + \sin k\omega \sin j\omega = \cos(k-j)\omega$. Because $\cos(k-j)\omega = \cos(j-k)\omega$, $F(\omega)$, and therefore $E_\sigma^T F(\omega) E_\sigma$, is symmetric, and because $P_\sigma(\omega)$ is non-negative, $P_\sigma(\omega)$ is a non-negative definite quadratic form, and $E_\sigma^T F(\omega)\, E_\sigma$ is a non-negative definite matrix (and therefore has real non-negative eigenvalues). Furthermore, for any collection, S, of protein segments, the sum, $P_S(\omega)$ of the power spectra for all the segments in S (of which Fig. 2(b) is an example) is:

$$\begin{aligned} P_S(\omega) &= \sum_{\sigma \varepsilon S} P_\sigma(\omega) = \sum_{\sigma \varepsilon S} \boldsymbol{\eta}^T E_\sigma^T F(\omega) E_\sigma \boldsymbol{\eta} \\ &= \boldsymbol{\eta}^T \left\{ \sum_{\sigma \varepsilon S} E_\sigma^T F(\omega) E_\sigma \right\} \boldsymbol{\eta} \\ &= \boldsymbol{\eta}^T W_S(\omega) \boldsymbol{\eta} \qquad (W_S(\omega) \text{ implicitly defined}). \quad (5) \end{aligned}$$

It can be seen that $P_S(\omega)$ is also a non-negative definite quadratic form, and $W_S(\omega)$ is a non-negative definite matrix.

The matrix $W_S(\omega)$ is an important characteristic of the set S of peptides that is independent of hydrophobicity scales. It will be seen in Results, section (c), that $W_S(\omega)$ *alone* identifies α-helical periodicity when S is a collection of α-helices.

A representation similar to that of eqn (6) for $P_m(\omega)$, $Q(\omega)$, and $Q_m(\omega)$ is also possible, for which we give the k,j entry of the matrix corresponding to F:

Let:

$$C1 = \sum_{k=-l_1}^{l_2} \cos(k'\omega), \quad CC = \sum_{k=-l_1}^{l_2} \cos^2(k'\omega),$$

and:

$$SS = \sum_{k=-l_1}^{l_2} \sin^2(k'\omega),$$

where l_1, l_2 and k' are defined in eqns (A4) of Appendix I.

$$P_m(\omega): f_{k,j} = \left(\cos(k'\omega) - \frac{C1}{l}\right)\left(\cos(j'\omega) - \frac{C1}{l}\right) + \sin(k'\omega)\sin(j'\omega)$$

$$Q(\omega): f_{k,j} = \frac{1}{l} + \frac{\left(\cos(k'\omega) - \frac{C1}{l}\right)\left(\cos(j'\omega) - \frac{C1}{l}\right)}{CC - (C1)^2/l} + \frac{\sin(k'\omega)\sin(j'\omega)}{SS}$$

$$Q_m(\omega): f_{k,j} = \frac{\left(\cos(k'\omega) - \frac{C1}{l}\right)\left(\cos(j'\omega) - \frac{C1}{l}\right)}{CC - (C1)^2/l} + \frac{\sin(k'\omega)\sin(j'\omega)}{SS}.$$

(e) *Computation of optimum scales*

By definition, an optimum scale maximizes the amphipathic index of a power spectrum of some set of helices. Presented here are the procedures used to compute optimum scales.

From eqns (3) and (5) it can be seen that for any hydrophobicity scale $\boldsymbol{\eta}$ and any collection S of helices and power spectrum $P(\omega)$ computed from $\boldsymbol{\eta}$ and S, the amphipathic index computed for the triple $\boldsymbol{\eta}$, S and P is:

$$\begin{aligned} \mathrm{AI}[\boldsymbol{\eta},S,P] &= \frac{\frac{1}{25^\circ}\int_{85^\circ}^{110^\circ} \boldsymbol{\eta}^T W_S(\omega) \boldsymbol{\eta}\, d\omega}{\frac{1}{180^\circ}\int_{0^\circ}^{180^\circ} \boldsymbol{\eta}^T W_S(\omega) \boldsymbol{\eta}\, d\omega} \\ &= 7{\cdot}2 \frac{\boldsymbol{\eta}^T \left[\int_{85^\circ}^{110^\circ} W_S(\omega)\, d\omega\right] \boldsymbol{\eta}}{\boldsymbol{\eta}^T \left[\int_{0^\circ}^{180^\circ} W_S(\omega)\, d\omega\right] \boldsymbol{\eta}} \\ &= 7{\cdot}2 \frac{\boldsymbol{\eta}^T U \boldsymbol{\eta}}{\boldsymbol{\eta}^T V \boldsymbol{\eta}} \\ &\qquad (U \text{ and } V \text{ implicitly defined}). \quad (6) \end{aligned}$$

Eqn (6) was used to compute the amphipathic indices shown in Table 6; the integrals are computed numerically using Simpson's rule with an interval size of 2·5°.

Eqn (6) expresses the amphipathic index as the quotient of 2 non-negative definite quadratic forms, and for any fixed set, S, of helices it is possible to compute a vector $\boldsymbol{\eta}$ that yields the global maximum of the amphipathic index for that set of helices. We consider only the case in which each of the 20 amino acids occurs in at least 1 of the helices in S (which is true for both the primary and alternative sets of helices). Then the matrices U and V in eqn (6) computed for the power spectra $P(\omega)$ and $Q(\omega)$ are both positive definite, with all positive eigenvalues. For positive definite V, there is an orthogonal matrix X (X^TX = the identity matrix) and a positive diagonal matrix L (the eigenvalues of V are on the diagonal of L) such that $V = X^TLX$. Let D be the diagonal matrix whose diagonal entries are the positive square-roots of the diagonal entries in L. Then $L = D^2$ and $V = X^TLX = X^TD^2X$. With a change of variables $\boldsymbol{\xi} = DX\boldsymbol{\eta}$ we obtain:

$$\begin{aligned} \mathrm{AI}[\boldsymbol{\eta},S,P] &= 7{\cdot}2 \frac{\boldsymbol{\eta}^T U \boldsymbol{\eta}}{\boldsymbol{\eta}^T X^T D^T D X \boldsymbol{\eta}} \\ &= 7{\cdot}2 \frac{\boldsymbol{\xi}^T D^{-1} X U X^T D^{-1} \boldsymbol{\xi}}{\boldsymbol{\xi}^T \boldsymbol{\xi}}. \quad (7) \end{aligned}$$

We can assume that $\boldsymbol{\xi}$ is of Euclidean length 1 so that $\boldsymbol{\xi}^T\boldsymbol{\xi} = 1$, and then seek the unit vector $\boldsymbol{\xi}$ that maximizes the numerator of eqn (7). That vector is simply the (unit) eigenvector associated with the largest eigenvalue of $D^{-1}XUX^TD^{-1}$, and is readily computed. Having computed $\boldsymbol{\xi}$, one then computes $\boldsymbol{\eta} = X^TD^{-1}\boldsymbol{\xi}$ to obtain the hydrophobicity scale that maximizes the amphipathic index for the set S of helices. Six such scales are listed in Table 3; PRIFT, ALTFT and TOTFT maximize the amphipathic index of the Fourier transform power spectrum $P_S(\omega)$ computed for S = primary helices, for S = alternative helices and for S = total set of helices, respectively. PRILS, ALTLS and TOTLS are analogous scales for the least-square power $Q(\omega)$. All eigenvectors and eigenvalues were computed using IMSL, Inc. (1984) FORTRAN subroutines.

We also computed the scales that maximize the amphipathic index of $P_m(\omega)$ and $Q_m(\omega)$. These scales are not shown (see Discussion), but 2 of these scales were used to compute the maximum possible amphipathic indices shown at the bottom of Table 6. We note that the scale with all entries equal to 1 gives an indeterminate form $(\frac{0}{0})$ for this amphipathic index. An optimum scale can be found in the 19-dimensional subspace perpendicular to the vector $\mathbf{1}$ with each entry 1. It may be

computed by substituting

$$\eta_{20} = -\sum_{i=1}^{19} \eta_i$$

into the 20-variable quadratic forms $\boldsymbol{\eta}^T U \boldsymbol{\eta}$ and $\boldsymbol{\eta}^T V \boldsymbol{\eta}$ of eqn (6) to obtain 19-variable quadratic forms $\boldsymbol{\zeta}^T U_{19} \boldsymbol{\zeta}$ and $\boldsymbol{\zeta}^T V_{19} \boldsymbol{\zeta}$, and proceeding as in eqn (7).

3. Hydrophobicity Scales

The assignment of hydrophobicity values to the individual amino acids is at least as important as the computational technique used to detect periodicity in the hydrophobicity values along a protein segment. As shown in Figure 2(b), the power spectrum for the collection of α-helices computed with the Kyte–Doolittle hydrophobicity scale is distinctly sharper around 100° than the spectrum computed with the Eisenberg *et al.* "consensus" scale. There are a large number of hydrophobicity scales in the literature; we have examined 38 of them for their effectiveness in organizing the (approximately) 100° peak for both the primary and alternative sets of α-helices. A number of these scales were examined by Hopp (1986) for their ability to detect antigenic determinants bound by antibodies. The 38 scales are listed in Table 3 as they were originally published (or, in 2 cases, as computed from data in the original papers). A brief description of the basis for each scale is given in Appendix II. Because of the variety of methods used to specify the scales, they are difficult to compare in their original forms, and in Table 4 are linear images of the scales, normalized so that the hydrophobicity of glycine is zero (with 5 exceptions), the hydrophobicity of leucine is positive, and the average of the absolute values of the hydrophobicities assigned is 2·5. In five of the scales, the hydrophobicity of glycine was so anomolous that translating it to zero distorted the other values, and another center for the scale was selected. Six of the scales do not assign values to all 20 amino acids, and, it will be seen below, were at a disadvantage in identifying the peak at 100°.

Another way of determining similarity in two scales $\{x_i\}_{i=1}^{20}$ and $\{y_i\}_{i=1}^{20}$ is to compute the correlation:

$$\sum_{i=1}^{20} (x_i - \bar{x})(y_i - \bar{y}) \Big/ \left[\sum_{i=1}^{20} (x_i - \bar{x})^2 \sum_{i=1}^{20} (y_i - \bar{y})^2 \right]^{1/2}.$$

Shown in Table 5 are the correlations between the scales of Table 3; when some values are omitted, the correlations are based on the amino acids to which both scales assign values. The correlation of two scales of Table 3 is also the correlation of the linear images of the two scales in Table 4.

The scales divide into two groups: (1) those based on experimental measurements of the chemical behavior of the particular amino acid (solubility in water, partition between water and an organic solvent, chromatographic migration, effect on surface tension); and (2) statistical scales, based on the positions of the amino acid in the tertiary structure of proteins as observed from crystallographic data or, in one case, based on mutation data. Some of the scales combine the results of experimental measurements with tertiary structure information. The review article by Rose *et al.* (1985*b*) contains an excellent description of the types of scales. Four scales, published by Rekker (1977), von Heijne & Blomberg (1979), Frömmel (1984) and Eisenberg & McLachlan (1986), compute the hydrophobicity of an amino acid in terms of the substituent parts of the acid or of the side-chain.

4. Results

(a) *Comparison of hydrophobicity scales and optimum scales*

The amphipathic index of an α-helix or a collection of α-helices varies according to the hydrophobicity scale used, and we rank the various hydrophobicity scales according to the amphipathic index they assign to the primary helices, and check the ranking with the alternative set of helices. The implicit assumption is that α-helices tend to be amphipathic (although only about half of them are), and the hydrophobicity scales that best detect the amphipathicity will yield a high amphipathic index for a set of helices. Hydrophobicity scales that actually maximize the amphipathic index for different sets of helices are computed, and two additional tests of all of the scales are made to examine their performance outside the context of α-helices.

Shown in Table 6 are the amphipathic indices assigned by the various scales to the primary set of helices (column 2), to the alternative set of helices (column 3) and the β-amphipathic index assigned to the β-strands (column 4), all computed for the Fourier transform power spectrum $P_m(\omega)$ (with $\bar{h}$ subtracted). When the amphipathic index for each scale computed for the primary set of helices is plotted against the corresponding amphipathic index for the alternative set of helices, the result is very nearly linear, demonstrating strong consistency in the amphipathic index as a gauge of the scales. The data are in columns 2 and 3 of Table 6 and are plotted as diamonds in Figure 5(a). The correlation coefficient for the data is 0·96.

The hydrophobicity scales that actually maximize the amphipathic indices associated with the primary helices, the alternative helices, and the total set of helices (primary plus alternative) were computed for the two power spectra $P(\omega)$ and $Q(\omega)$ by the methods described above (Data Bases and Mathematical Methods, section (e)). These "optimum" scales are shown at the bottom of Table 3 and in normalized form in Table 4. The set of helices used to compute the optimum scales has more influence than the power spectrum selected. In Figure 5 (a) there are two squares near each of the labels PRI, ALT and TOT corresponding to the two scales at the bottom of Table 3 computed with,

Table 3

Hydrophobicity scales as originally published

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
EXP	ZIMMR	0·83	0·83	0·09	0·64	1·48	0·00	0·65	0·10	1·10	2·52	3·07	1·60	1·40	2·75	2·70	0·14	0·54	0·31	2·97	1·79
EXP	N TAN	500	****	****	****	****	****	****	0	500	****	1800	****	1300	2500	****	−300	400	3400	2300	1500
EXP	NTANR	500	****	****	****	1650	****	****	0	500	1800	1800	****	1300	2500	****	****	400	3400	2300	1500
EXP	JONES	0·87	0·85	0·09	0·66	1·52	0·00	0·67	0·10	0·87	3·15	2·17	1·64	1·67	2·87	2·77	0·07	0·07	3·77	2·67	1·87
X/S	LEVIT	−0·5	3·0	0·2	2·5	−1·0	0·2	2·5	0·0	−0·5	−1·8	−1·8	3·0	−1·3	−2·5	−1·4	0·3	−0·4	−3·4	−2·3	−1·5
X/S	HOPPW	−0·5	3·0	0·2	3·0	−1·0	0·2	3·0	0·0	−0·5	−1·8	−1·8	3·0	−1·3	−2·5	0·0	0·3	−0·4	−3·4	−2·3	−1·5
EXP	YUNGD	−2·74	−4·08	****	****	****	****	−3·69	−3·11	−1·95	−1·69	−1·79	−3·05	−1·87	−1·43	−2·54	−3·07	−2·94	−1·11	−2·05	−2·26
EXP	FAUPL	0·31	−1·01	−0·60	−0·77	1·54	−0·22	−0·64	0·00	0·13	1·80	1·70	−0·99	1·23	1·79	0·72	−0·04	0·26	2·25	0·96	1·22
EXP	ZASLZ	1·41	3·74	−0·48	−6·11	****	0·33	−8·19	0·00	****	3·85	3·30	4·74	3·07	7·06	0·33	−0·48	0·74	17·22	****	2·41
EXP	WOLF	1·94	−19·92	−9·68	−10·95	−1·24	−9·38	−10·20	2·39	−10·27	2·15	2·28	−9·52	−1·48	−0·76	****	−5·06	−4·88	−5·88	−6·11	1·99
EXP	KUNTZ	1·5	3·0	2·0	6·0	1·0	2·0	7·5	1·0	4·0	1·0	1·0	4·5	1·0	0·0	3·0	2·0	2·0	2·0	3·0	1·0
EXP	ABODR	5·1	2·0	0·6	0·7	0·0	1·4	1·8	4·1	1·6	9·3	10·0	1·3	8·7	9·6	4·9	3·1	3·5	9·2	8·0	8·5
EXP	MEEK	0·5	0·8	0·8	−8·2	−6·8	−4·8	−16·9	0·0	−3·5	13·9	8·8	0·1	4·8	13·2	6·1	1·2	2·7	14·9	6·1	2·7
EXP	BULDG	−200	−120	80	−200	−450	160	−300	0	−120	2260	2460	−350	1470	2330	−980	−390	−520	2010	2240	1560
AVE	EISEN	0·25	−1·76	−0·64	−0·72	0·04	−0·69	−0·62	0·16	−0·40	0·73	0·53	−1·10	0·26	0·61	−0·07	−0·26	−0·18	0·37	0·02	0·54
AVE	KYTDO	1·8	−4·5	−3·5	−3·5	2·5	−3·5	−3·5	−0·4	−3·2	4·5	3·8	−3·9	1·9	2·8	−1·6	−0·8	−0·7	−0·9	−1·3	4·2
STA	CHOTH	0·38	0·01	0·12	0·15	0·50	0·07	0·18	0·36	0·17	0·60	0·45	0·03	0·40	0·50	0·18	0·22	0·23	0·27	0·15	0·54
STA	WERSC	0·52	0·49	0·42	0·37	0·83	0·35	0·38	0·41	0·70	0·79	0·77	0·31	0·76	0·87	0·35	0·49	0·38	0·86	0·64	0·72
STA	JANIN	1·70	0·10	0·40	0·40	4·60	0·30	0·30	1·80	0·80	3·10	2·40	0·05	1·90	2·20	0·60	0·80	0·70	1·60	1·50	2·90
STA	OLSEN	1·38	0·00	0·37	0·52	1·43	0·22	0·71	1·34	0·66	2·32	1·47	0·15	1·78	1·72	0·85	0·86	0·89	0·82	0·47	1·99
STA	MEIRO	0·93	0·98	0·98	1·01	0·88	1·02	1·02	1·00	0·89	0·79	0·85	1·05	0·84	0·78	1·00	1·02	0·99	0·83	0·93	0·81
X/S	PONNU	12·28	11·49	11·00	10·97	14·93	11·28	11·19	12·01	12·84	14·77	14·10	10·80	14·33	13·43	11·19	11·26	11·65	12·95	13·29	15·07
STA	NNEIG**	50·76	48·66	45·80	43·17	58·74	46·09	43·48	50·27	49·33	57·30	53·89	42·92	52·75	53·45	45·39	47·24	49·26	53·59	51·79	56·12
STA	ROBOS	−1·0	0·3	−0·7	−1·2	2·1	−0·1	−0·7	****	1·1	4·0	2·0	−0·9	1·8	2·8	0·4	−1·2	−0·5	3·0	2·1	1·4
STA	CHDLG	−0·27	2·00	0·61	0·50	−0·23	1·00	0·33	−0·22	0·37	−0·80	−0·44	1·17	−0·31	−0·55	0·36	0·17	0·18	0·05	0·48	−0·65
STA	WSDLG	0·05	0·12	0·29	0·41	−0·84	0·46	0·38	0·31	−0·41	−0·69	−0·62	0·57	−0·38	−0·45	0·46	0·12	0·38	−0·98	−0·25	−0·46
STA	JADLG	0·3	−1·4	−0·5	−0·6	0·9	−0·7	−0·7	0·3	−0·1	0·7	0·5	−1·8	0·4	0·5	−0·3	−0·1	−0·2	0·3	−0·4	0·6
STA	GUY	0·10	1·91	0·48	0·78	−1·42	0·95	0·83	0·33	−0·50	−1·13	−1·18	1·40	−1·59	−2·12	0·73	0·52	0·07	−0·51	−0·21	−1·27
AVE	GUY M	0·06	0·84	0·48	0·80	−1·36	0·73	0·77	0·41	−0·49	−1·31	−1·21	1·18	−1·27	−1·68	0·70	0·50	0·27	−0·88	−0·33	−1·09
X/S	KRIDG	300	−540	−170	−60	1320	−170	−150	−100	100	2400	1280	1500	2300	2700	1060	100	250	3200	1900	1300
X/S	KRIGK	4·32	6·55	6·24	6·04	1·73	6·13	6·17	6·09	5·66	2·31	3·93	7·92	2·44	2·59	7·19	5·37	5·16	2·78	3·58	3·31
STA	NIOII	0·23	−0·26	−0·94	−1·13	1·78	−0·57	−0·75	−0·07	0·11	1·19	1·03	−1·05	0·66	0·48	−0·76	−0·67	−0·36	0·90	0·59	1·24
STA	MIJER	5·33	4·18	3·71	3·59	7·93	3·87	3·65	4·48	5·10	8·83	8·47	2·95	8·95	9·03	3·87	4·09	4·49	7·66	5·89	7·63
STA	ROSEF	0·74	0·64	0·63	0·62	0·91	0·62	0·62	0·72	0·78	0·88	0·85	0·52	0·85	0·88	0·64	0·66	0·70	0·85	0·76	0·86
STA	SWEET	−0·40	−0·59	−0·92	−1·31	0·17	−0·91	−1·22	−0·67	−0·64	1·25	1·22	−0·67	1·02	1·92	−0·49	−0·55	−0·28	0·50	1·67	0·91
STA	SWEIG**	−0·414	−0·584	−0·916	−1·310	0·162	−0·905	−1·218	−0·684	−0·630	1·237	1·215	−0·670	1·020	1·938	−0·503	−0·563	−0·289	0·514	1·699	0·899
X/S	REKKR	0·53	****	−1·05	−0·02	0·93	−1·09	−0·07	0·00	−0·23	1·99	1·99	0·52	1·08	2·24	1·01	−0·56	−0·26	2·31	1·70	1·46
X/S	VHEBL	−12·04	39·23	4·25	23·22	3·95	2·16	16·81	−7·85	6·28	−18·32	−17·79	9·71	−8·86	−21·98	5·82	−1·54	−4·15	−16·19	−1·51	−16·22
X/S	FROMM	79·1	93·7	48·4	31·1	119·9	59·4	61·5	43·1	129·6	163·5	156·4	108·8	164·6	215·0	144·8	56·8	97·2	225·4	176·3	140·2
X/S	EIMCL	0·67	−2·10	−0·60	−1·20	0·38	−0·22	−0·76	0·00	0·64	1·90	1·90	−0·57	2·40	2·30	1·20	0·01	0·52	2·60	1·60	1·50
STA	PRIFT**	−0·96	0·75	−1·94	−5·68	4·54	−5·30	−3·86	−1·28	−0·62	5·54	6·81	−5·62	4·76	5·06	−4·47	−1·92	−3·99	0·21	3·34	5·39
STA	PRILS**	−0·26	0·08	−0·46	−1·30	0·83	−0·83	−0·73	−0·40	−0·18	1·10	1·52	−1·01	1·09	1·09	−0·62	−0·55	−0·71	−0·13	0·69	1·15
STA	ALTFT**	−0·73	−1·03	−5·29	−6·13	0·64	−0·96	−2·90	−2·67	3·03	5·04	4·91	−5·99	3·34	5·20	−4·32	−3·00	−1·91	0·51	2·87	3·98
STA	ALTLS**	−1·35	−3·89	−10·96	−11·88	4·37	−1·34	−4·56	−5·82	6·54	10·93	9·88	−11·92	7·47	11·35	−10·86	−6·21	−4·83	1·80	7·61	8·20
STA	TOTFT**	−0·56	−0·26	−2·87	−4·31	1·78	−2·31	−2·35	−1·35	0·81	3·83	4·09	−4·08	3·11	3·67	−3·22	−1·85	−1·97	−0·11	2·17	3·31
STA	TOTLS**	−1·37	−1·33	−6·29	−8·93	4·47	−3·88	−4·04	−3·39	1·65	7·92	8·68	−7·70	7·13	7·96	−6·25	−4·08	−4·02	−0·79	4·73	6·94

Appendix II interprets the acronyms in column 2 and describes the origins of the scales.

EXP, Scales based on experimental measurements. STA, Scales based on statistical studies. X/S, Scales that combine experimental and statistical data. AVE, Scales that are averages of other scales. ** Scales calculated in this paper. **** Values omitted in the original scales.

Table 4

Normalized linear images of the hydrophobicity scales

		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
EXP	ZIMMR	−0·49	−0·49	−2·64	−1·04	1·39	−2·90	−1·01	−2·61	0·29	4·41	6·00	1·74	1·16	5·07	4·93	−2·49	−1·33	−2·00	5·71	2·29
EXP	N TAN	0·95	****	****	****	****	****	****	0·00	0·95	****	3·41	****	2·47	4·74	****	−0·57	0·76	6·45	4·36	2·84
EXP	NTANR	0·85	****	****	****	2·80	****	****	0·00	0·85	3·06	3·06	****	2·21	4·25	****	****	0·68	5·78	3·91	2·55
EXP	JONES	−0·33	−0·38	−2·29	−0·86	1·31	−2·52	−0·83	−2·27	−0·33	5·42	2·95	1·61	1·69	4·71	4·46	−2·34	−2·34	6·98	4·21	2·19
X/S	LEVIT	0·83	−4·98	−0·33	−4·15	1·66	−0·33	−4·15	0·00	0·83	2·99	2·99	−4·98	2·16	4·15	2·33	−0·50	0·66	5·65	3·82	2·49
X/S	HOPPW	0·84	−5·05	−0·34	−5·05	1·68	−0·34	−5·05	0·00	0·84	3·03	3·03	−5·05	2·19	4·21	0·00	−0·51	0·67	5·72	3·87	2·53
EXP	YUNGD	0·20	−4·37	****	****	****	****	−3·04	−1·06	2·90	3·79	3·45	−0·85	3·18	4·68	0·89	−0·92	−0·48	5·77	2·56	1·84
EXP	FAUPL	0·85	−2·78	−1·65	−2·12	4·24	−0·61	−1·76	0·00	0·36	4·95	4·68	−2·72	3·38	4·92	1·98	−0·11	0·72	6·19	2·64	3·36
EXP	ZASLZ	0·94	2·50	−0·32	−4·09	****	0·22	−5·48	0·00	****	2·58	2·21	3·17	2·06	4·73	0·22	−0·32	0·50	11·53	****	1·61
EXP	WOLF	4·49	−5·38	−0·76	−1·33	3·05	−0·62	−0·99	4·69	−1·02	4·58	4·64	−0·69	2·94	3·27	****	1·33	1·41	0·96	0·85	4·51
EXP	KUNTZ	0·18	−2·58	−0·74	−8·12	1·11	−0·74	−10·89	1·11	−4·43	1·11	1·11	−5·35	1·11	2·95	−2·58	−0·74	−0·74	−0·74	−2·58	1·11
EXP	ABODR	0·83	−1·74	−2·90	−2·82	0·00	−2·24	−1·91	0·00	−2·07	4·31	4·89	−2·32	3·81	4·56	0·66	−0·83	−0·50	4·23	3·23	3·65
EXP	MEEK	0·21	0·34	0·34	−3·51	−2·91	−2·05	−7·23	0·00	−1·50	5·95	3·77	0·04	2·05	5·65	2·61	0·51	1·16	6·38	2·61	1·16
EXP	BULDG	−0·55	−0·33	0·22	−0·55	−1·24	0·44	−0·82	0·00	−0·33	6·21	6·76	−0·96	4·04	6·40	−2·69	−1·07	−1·43	5·52	6·15	4·29
AVE	EISEN	0·43	−9·17	−3·82	−4·20	−0·57	−4·06	−3·72	0·00	−2·67	2·72	1·77	−6·02	0·48	2·15	−1·10	−2·01	−1·62	1·00	−0·67	1·81
AVE	KYTDO	2·18	−4·07	−3·08	−3·08	2·88	−3·08	−3·08	0·00	−2·78	4·86	4·17	−3·47	2·28	3·17	−1·19	−0·40	−0·30	−0·50	−0·89	4·56
STA	CHOTH	0·29	−5·16	−3·54	−3·10	2·06	−4·28	−2·65	0·00	−2·80	3·54	1·33	−4·87	0·59	2·06	−2·65	−2·06	−1·92	−1·33	−3·10	2·65
STA	WERSC	1·43	1·04	0·13	−0·52	5·45	−0·78	−0·39	0·00	3·77	4·94	4·68	−1·30	4·55	5·97	−0·78	1·04	−0·39	5·84	2·99	4·03
STA	JANIN	−0·24	−4·16	−3·42	−3·42	6·85	−3·67	−3·67	0·00	−2·44	3·18	1·47	−4·28	0·24	0·98	−2·93	−2·44	−2·69	−0·49	−0·73	2·69
STA	OLSEN	0·16	−5·46	−3·95	−3·34	0·37	−4·56	−2·57	0·00	−2·77	3·99	0·53	−4·85	1·79	1·55	−2·00	−1·96	−1·83	−2·12	−3·55	2·65
STA	MEIRO	2·13	0·61	0·61	−0·30	3·66	−0·61	−0·61	0·00	3·35	6·40	4·57	−1·52	4·88	6·71	0·00	−0·61	0·30	5·18	2·13	5·79
X/S	PONNU	0·54	−1·03	−2·01	−2·07	5·81	−1·45	−1·63	0·00	1·65	5·49	4·16	−2·41	4·61	2·82	−1·63	−1·49	−0·72	1·87	2·54	6·08
STA	NNEIG**	0·32	−1·04	−2·89	−4·59	5·48	−2·70	−4·39	0·00	−0·61	4·55	2·34	−4·75	1·60	2·06	−3·16	−1·96	−0·65	2·15	0·98	3·78
STA	ROBOS	−1·74	0·52	−1·22	−2·09	3·65	−0·17	−1·22	****	1·91	6·96	3·48	−1·57	3·13	4·87	0·70	−2·09	−0·87	5·22	3·65	2·44
STA	CHDLG	0·22	−9·59	−3·59	−3·11	0·04	−5·27	−2·38	0·00	−2·55	2·51	0·95	−6·01	0·39	1·43	−2·51	−1·69	−1·73	−1·17	−3·03	1·86
STA	WSDLG	1·39	1·02	0·11	−0·54	6·16	−0·80	−0·38	0·00	3·86	5·36	4·98	−1·39	3·70	4·07	−0·80	1·02	−0·38	6·91	3·00	4·13
STA	JADLG	0·00	−7·14	−3·36	−3·78	2·52	−4·20	−4·20	0·00	−1·68	1·68	0·84	−8·82	0·42	0·84	−2·52	−1·68	−2·10	0·00	−2·94	1·26
STA	GUY	0·63	−4·31	−0·41	−1·23	4·77	−1·69	−1·36	0·00	2·26	3·98	4·11	−2·92	5·23	6·68	−1·09	−0·52	0·71	2·29	1·47	4·36
AVE	GUY M	1·06	−1·30	−0·21	−1·18	5·36	−0·97	−1·09	0·00	2·72	5·21	4·90	−2·33	5·08	6·33	−0·88	−0·27	0·42	3·90	2·24	4·54
X/S	KRIDG	0·92	−1·01	−0·16	0·09	3·26	−0·16	−0·11	0·00	0·46	5·74	3·17	3·67	5·51	6·43	2·66	0·46	0·80	7·58	4·59	3·21
X/S	KRIGK	2·63	−0·68	−0·22	0·07	6·49	−0·06	−0·12	0·00	0·64	5·62	3·21	−2·72	5·43	5·21	−1·64	1·07	1·38	4·92	3·73	4·14
STA	NIOH	1·02	−0·64	−2·95	−3·59	6·26	−1·69	−2·30	0·00	0·61	4·27	3·72	−3·32	2·47	1·86	−2·34	−2·03	−0·98	3·28	2·23	4·43
STA	MIJER	1·18	−0·42	−1·07	−1·24	4·80	−0·85	−1·15	0·00	0·86	6·05	5·55	−2·13	6·22	6·33	−0·85	−0·54	0·01	4·42	1·96	4·38
STA	ROSEF	0·50	−2·01	−2·26	−2·51	4·77	−2·51	−2·51	0·00	1·51	4·02	3·27	−5·03	3·27	4·02	−2·01	−1·51	−0·50	3·27	1·01	3·52
STA	SWEET	0·81	0·24	−0·75	−1·91	2·50	−0·72	−1·64	0·00	0·09	5·72	5·64	0·00	5·04	7·72	0·54	0·36	1·16	3·49	6·98	4·71
STA	SWEIG**	0·80	0·30	−0·69	−1·85	2·50	−0·65	−1·58	0·00	0·16	5·68	5·62	0·04	5·04	7·76	0·54	0·36	1·17	3·54	7·05	4·68
X/S	REKKR	1·32	****	−2·62	−0·05	2·32	−2·72	−0·17	0·00	−0·57	4·96	4·96	1·30	2·69	5·59	2·52	−1·40	−0·65	5·76	4·24	3·64
X/S	VHEBL	0·82	−9·24	−2·37	−6·10	−2·31	−1·96	−4·84	0·00	−2·77	2·05	1·95	−3·44	0·20	2·77	−2·68	−1·24	−0·73	1·64	−1·24	1·64
X/S	FROMM	1·22	1·71	0·18	−0·41	2·60	0·55	0·62	0·00	2·93	4·08	3·84	2·22	4·11	5·82	3·44	0·46	1·83	6·17	4·51	3·29
X/S	EIMCL	1·45	−4·55	−1·30	−2·60	0·82	−0·48	−1·65	0·00	1·39	4·12	4·12	−1·24	5·20	4·98	2·60	0·02	1·13	5·64	3·47	3·25
STA	PRIFT**	0·22	1·42	−0·46	−3·08	4·07	−2·81	−1·81	0·00	0·46	4·77	5·66	−3·04	4·23	4·44	−2·23	−0·45	−1·90	1·04	3·23	4·67
STA	PRILS**	0·49	1·67	−0·21	−3·13	4·27	−1·49	−1·15	0·00	0·76	5·21	6·67	−2·12	5·18	5·18	−0·76	−0·52	−1·08	0·94	3·79	5·39
STA	ALFT**	1·36	1·15	−1·84	−2·43	2·32	1·20	−0·16	0·00	4·00	5·41	5·32	−2·33	4·22	5·53	−1·16	−0·23	0·53	2·23	3·89	4·67
STA	ALTLS**	1·43	0·62	−1·64	−1·94	3·26	1·43	0·40	0·00	3·95	5·36	5·02	−1·95	4·25	5·49	−1·61	−0·12	0·32	2·44	4·29	4·48
STA	TOTFT**	0·81	1·12	−1·56	−3·03	3·20	−0·98	−1·02	0·00	2·21	5·30	5·57	−2·79	4·56	5·14	−1·91	−0·51	−0·63	1·27	3·60	4·77
STA	TOTLS**	1·00	1·02	−1·43	−2·73	3·88	−0·24	−0·32	0·00	2·49	5·58	5·95	−2·13	5·19	5·60	−1·41	−0·34	−0·31	1·28	4·01	5·10

The scales in Table 3 are translated to have the hydrophobicity of glycine equal to 0 (5 exceptions) and the average of the absolute values of all entries shown equal to 2·5. Appendix II interprets the acronyms in column 2 and describes the origins of the scales.

EXP. Scales based on experimental measurements. STA. Scales based on statistical studies. X/S. Scales that combine experimental and statistical data. AVE. Scales that are averages of other scales. **Scales calculated in

Table 5

Correlations between the hydrophobicity scales (100 times the absolute value)

ZIMMR	ZI
N TAN	53	N
NTANR	45	100	NT
JONES	72	99	95	JO
LEVIT	45	100	98	64	LE
HOPPW	40	100	98	59	98	HO
YUNGD	48	90	87	75	89	90	YU
FAUPL	55	94	90	76	92	91	91	FA
ZASLZ	20	90	89	68	61	66	67	64	ZA
WOLF	40	5	13	33	66	66	52	69	18	WO
KUNTZ	22	17	16	25	68	74	52	64	55	66	KU
ABODR	65	83	80	78	83	83	81	95	61	71	63	AB
MEEK	49	88	70	71	72	73	74	69	82	43	67	79	ME
BULDG	55	88	76	68	66	71	76	73	58	46	45	87	69	BU
EISEN	46	62	50	54	86	85	78	87	34	91	62	84	58	63	EI
KYTDO	53	24	21	47	68	70	60	81	31	89	71	86	52	60	88	KY
CHOTH	42	20	14	42	63	64	58	78	24	88	64	79	41	54	87	96	CH
WERSC	46	82	79	66	73	78	84	86	66	49	55	84	56	75	67	71	72	WE
JANIN	43	46	27	44	62	64	67	78	38	76	63	88	31	47	74	87	89	77	JA
OLSEN	42	6	2	40	60	60	54	73	18	87	60	76	42	51	87	94	97	65	80	OL
MEIRO	52	76	68	68	75	78	82	86	58	57	59	85	63	79	75	79	80	94	74	76	ME
PONNU	54	56	43	55	67	71	72	82	42	66	59	86	42	68	73	85	84	88	90	80	89	PO
NNEIG	42	73	51	51	73	78	69	84	57	67	73	88	52	63	73	85	84	87	92	76	86	94	NN
ROBOS	60	91	77	81	75	76	81	87	65	48	54	84	67	80	66	64	65	88	70	61	88	82	82	RO
CHDLG	35	16	10	37	69	68	64	75	14	92	52	72	38	50	95	88	93	61	76	93	69	71	67	54	CH
WSDLG	40	83	77	65	72	76	83	86	69	47	51	82	52	70	65	68	69	98	78	60	90	87	88	87	58	WS
JADLG	27	28	20	35	78	78	68	83	27	83	67	75	39	47	90	87	91	74	85	87	76	79	83	64	91	73	JA
GUY	50	60	51	51	78	80	84	86	39	73	62	83	48	67	85	84	86	87	81	82	90	88	82	75	84	80	87	GU
GUY M	52	73	66	60	78	82	86	90	52	65	64	88	55	75	79	84	84	96	84	79	96	93	90	86	76	91	84	97	GU
KRIDG	59	93	91	89	71	70	84	82	75	47	42	83	73	74	65	58	53	73	53	52	74	63	60	81	52	69	46	69	72	KR
KRIGK	37	82	72	54	74	78	76	86	50	64	60	87	47	69	76	80	81	90	85	74	87	89	90	79	72	88	82	88	93	70	KR
NIOII	46	75	57	57	71	75	73	85	57	64	62	90	45	65	70	81	80	89	91	71	86	96	98	83	64	91	79	81	90	62	90	NI
MIJER	54	74	67	65	75	78	81	90	55	65	65	93	61	79	78	86	84	94	82	80	95	92	89	87	73	89	80	92	98	77	93	89	MI
ROSEF	45	77	65	56	79	83	82	90	52	67	65	89	52	68	81	84	86	94	88	80	93	94	95	85	77	92	89	93	98	65	94	94	95	RO
SWEET	74	78	73	73	76	79	76	83	59	57	62	93	74	88	72	75	66	80	65	64	83	80	77	84	58	73	58	80	84	81	79	76	87	79	SW
SWEIG	74	78	74	73	76	79	76	83	60	56	61	92	74	88	71	75	65	80	64	63	83	79	77	84	58	73	58	80	84	81	79	76	87	79	100	SW
REKKR	75	94	91	93	69	66	81	85	66	56	37	91	70	77	77	68	63	75	61	59	78	70	67	82	62	73	52	68	74	89	68	71	79	71	85	85	RE
VHEBL	34	57	43	45	83	85	78	79	48	88	71	78	65	65	92	80	76	59	60	75	68	61	64	57	84	55	76	77	71	64	65	60	70	69	69	68	64	VH
FROMM	67	95	92	90	75	74	86	83	79	33	41	83	76	73	57	51	44	78	45	42	78	64	62	87	39	74	42	66	74	90	64	65	76	69	84	84	88	55	FR
EIMCL	56	90	80	73	92	90	95	92	60	70	58	90	73	73	88	73	67	74	57	68	79	70	67	76	74	70	72	82	81	84	74	68	81	77	82	82	83	86	84	EI
PRIFT	60	58	53	54	63	68	60	76	42	57	64	85	53	77	64	81	77	88	80	71	87	92	89	80	60	84	70	81	90	59	83	89	91	88	86	86	72	55	65	64	PR
PRILS	66	56	49	56	64	68	61	77	40	57	64	86	54	77	65	81	75	85	77	71	86	91	86	81	59	80	67	81	89	60	80	87	91	86	88	88	73	57	68	67	99	PR
ALTFT	57	62	51	53	69	74	71	77	45	52	56	80	52	79	66	73	69	86	65	66	88	87	82	83	58	80	66	81	88	58	78	83	87	86	85	85	67	61	72	73	88	90	AL
ALTLS	56	66	56	53	69	74	72	78	44	53	53	80	47	79	67	72	70	88	69	66	88	89	83	84	59	83	67	83	90	60	83	86	88	88	84	84	66	61	71	72	88	89	99	AL
TOTFT	61	58	50	54	67	72	66	78	42	58	62	85	53	80	68	80	76	89	75	73	90	92	88	83	62	83	70	84	92	60	83	88	92	90	88	88	71	60	69	70	97	97	97	96	TO
TOTLS	63	57	48	55	66	71	66	78	40	58	59	85	50	79	68	80	76	88	76	73	89	93	87	83	62	82	69	85	92	61	83	88	92	89	88	88	71	60	70	71	96	97	97	97	100
	ZI	N	NT	JO	LE	HO	YU	FA	ZA	WO	KU	AB	ME	BU	EI	KY	CH	WE	JA	OL	ME	PO	NN	RO	CH	WS	JA	GU	GU	KR	KR	NI	MI	RO	SW	SW	RE	VH	FR	EI	PR	PR	AL	AL	TO

Appendix II interprets the acronyms in column 1 and describes the origins of the scales.

Table 6

Scale	Amphipathic index Fourier transform			Int/ext corresp	
	Prima	Alt	Beta	LDH	$\hat{\omega}_m$†
ZIMMR	1·691	1·589	1·623	63·27	98·75
N TAN	1·613	1·545	1·526	72·17	96·00
NTANR	1·864	1·761	1·575	93·92	97·25
JONES	1·556	1·509	1·565	70·89	98·25
LEVIT	1·748	1·731	1·484	119·88	98·25
HOPPW	1·784	1·763	1·437	123·29	97·75
YUNGD	1·081	1·118	1·263	81·16	95·00
FAUPL	1·943	1·837	1·517	146·87	98·75
ZASLZ	1·293	1·132	1·256	68·80	100·25
WOLF	1·497	1·452	1·402	172·96	99·00
KUNTZ	1·654	1·520	1·318	80·48	98·50
ABODR	1·980	1·854	1·747	125·49	98·25
MEEK	1·663	1·416	1·414	95·97	98·50
BULDG	1·877	1·735	1·556	117·28	98·00
EISEN	1·691	1·625	1·458	85·80	98·25
KYTDO	1·906	1·764	1·427	155·81	98·25
CHOTH	1·828	1·703	1·320	95·48	98·25
WERSC	2·098	1·888	1·532	167·71	98·00
JANIN	1·980	1·730	1·514	110·72	98·25
OLSEN	1·737	1·647	1·197	87·74	98·25
MEIRO	2·034	1·905	1·374	179·64	97·50
PONNU	2·100	1·907	1·437	172·32	97·50
NNEIG‡	2·061	1·869	1·456	144·11	98·00
ROBOS	1·876	1·819	1·381	100·12	98·75
CHDLG	1·562	1·503	1·232	82·46	98·25
WSDLG	2·090	1·847	1·503	175·75	98·00
JADLG	1·710	1·622	1·331	99·11	98·25
GUY	1·910	1·800	1·306	154·38	97·50
GUY M	2·078	1·926	1·418	168·81	97·75
KRIDG	1·833	1·778	1·444	113·07	97·00
KRIGK	1·905	1·782	1·370	177·50	97·50
NIOII	2·077	1·901	1·617	151·84	98·00
MIJER	2·107	1·933	1·509	169·97	98·00
ROSEF	2·054	1·916	1·481	148·27	98·00
SWEET	2·024	1·867	1·651	152·37	97·75
SWEIG‡	2·019	1·866	1·650	151·27	97·50
REKKR	1·814	1·654	1·837	123·48	98·00
VHEBL	1·593	1·523	1·308	58·96	98·00
FROMM	1·665	1·660	1·416	106·92	98·75
EIMCL	1·761	1·713	1·453	124·27	98·00
PRIFT‡	2·253	1·934	1·680	164·49	97·75
PRILS‡	2·224	1·958	1·646	170·32	98·00
ALTFT‡	2·052	2·072	1·496	141·94	98·25
ALTLS‡	2·028	2·053	1·500	139·49	98·25
TOTFT‡	2·196	2·039	1·602	156·10	98·00
TOTLS‡	2·172	2·039	1·583	160·75	98·00
§	2·259	2·079			

Appendix II interprets the acronyms in column 1 and the origins of the scales. Columns 2 and 3 show the (α) amphipathic indices of the Fourier transform power spectra for, respectively, the primary helices and the alternative helices, and column 4 shows the (β) amphipathic index of the β-strands. Column 5 shows the interior/exterior correspondence of the hydrophobicity profiles of dogfish M_4 (muscle) lactate dehydrogenase computed with the normalized scales of Table 4. The profile for each normalized scale is computed as 9-residue block averages, as in Kyte & Doolittle (1982). Each number in column 5 is the sum of the profile values of the interior residues minus the sum of the profile values for the exterior residues; interior and exterior residues are identified by Eventhoff *et al.* (1977). Column 6 shows the frequencies at which the Fourier transform power spectrum, $P_m(\omega)$, is maximized, when computed with the primary helices.

respectively, the primary, alternative and total sets of helices. Note that the amphipathic indices of the PRI scales computed with the alternative helices rank relatively higher than do the amphipathic indices of the ALT scales computed with the primary helices.

Shown in Figure 5 (b) is a graph of the β-amphipathic index of the collection of β-strands *versus* the amphipathic index for the primary helices (column 4 *versus* column 2 of Table 6). There is an apparent trend, but the data are quite scattered, probably due to the shortness of β-strands and the relatively small fraction of β-strands that are amphipathic. The important result illustrated by Figure 5(b) is that the PRI scales that maximally distinguish the 100° frequency for the primary helices also distinguish the 180° frequency for the β-strands as well as all but two of the 38 scales from the literature.

Shown in Figure 5(c) is a plot of interior/exterior correspondence *versus* amphipathic index of primary helices (column 5 *versus* column 2 of Table 6). Except for an interesting outlier due to the Wolfenden *et al.* (1981) scale, there is a very clear linear relationship between the interior/exterior correspondence and the amphipathic index. Again, in this test the PRI scales perform better than the ALT scales, and perform well compared to the majority of the scales in the literature. The interior/exterior correspondence was computed for lactate dehydrogenase (Data Bases and Mathematical Methods, section (a)).

(b) *Characteristics of helices*

It will be observed that the amphipathic index actually is centered around 97·5°, and not around 100° as might be expected. Listed in Table 6, column 6, are the frequencies of the actual peaks of the composite Fourier transform power spectrum $P_m(\omega)$ for the primary helices and for each of the scales. The average of those peaks for the 38 previously published scales is 98·1°. The corresponding average computed with the alternative helices is 97·3°. The peaks at 97·25° and 97·5° of the graphs in Figure 7 (explained in section (c), below) and the comparison with crystallographic data in section (d), below, provide additional evidence that the dominant frequency of α-helices is consistently around 97·5°. These results suggest that the side-chains of the typical helix are arranged in a slightly more open conformation than is usually thought, with the number of residues per turn being closer to 3·7 than to 3·6. However, we also find that the standard deviation of the location of the peaks for individual amphipathic helices is

LDH, Lactate dehydrogenase.

† Computed with the primary helices. The average $\hat{\omega}_m$ for previously published scale is 98·1°. The corresponding average computed with the alternative helices 97·3°.

‡ Scales calculated in this paper.

§ Maximum possible value for this column.

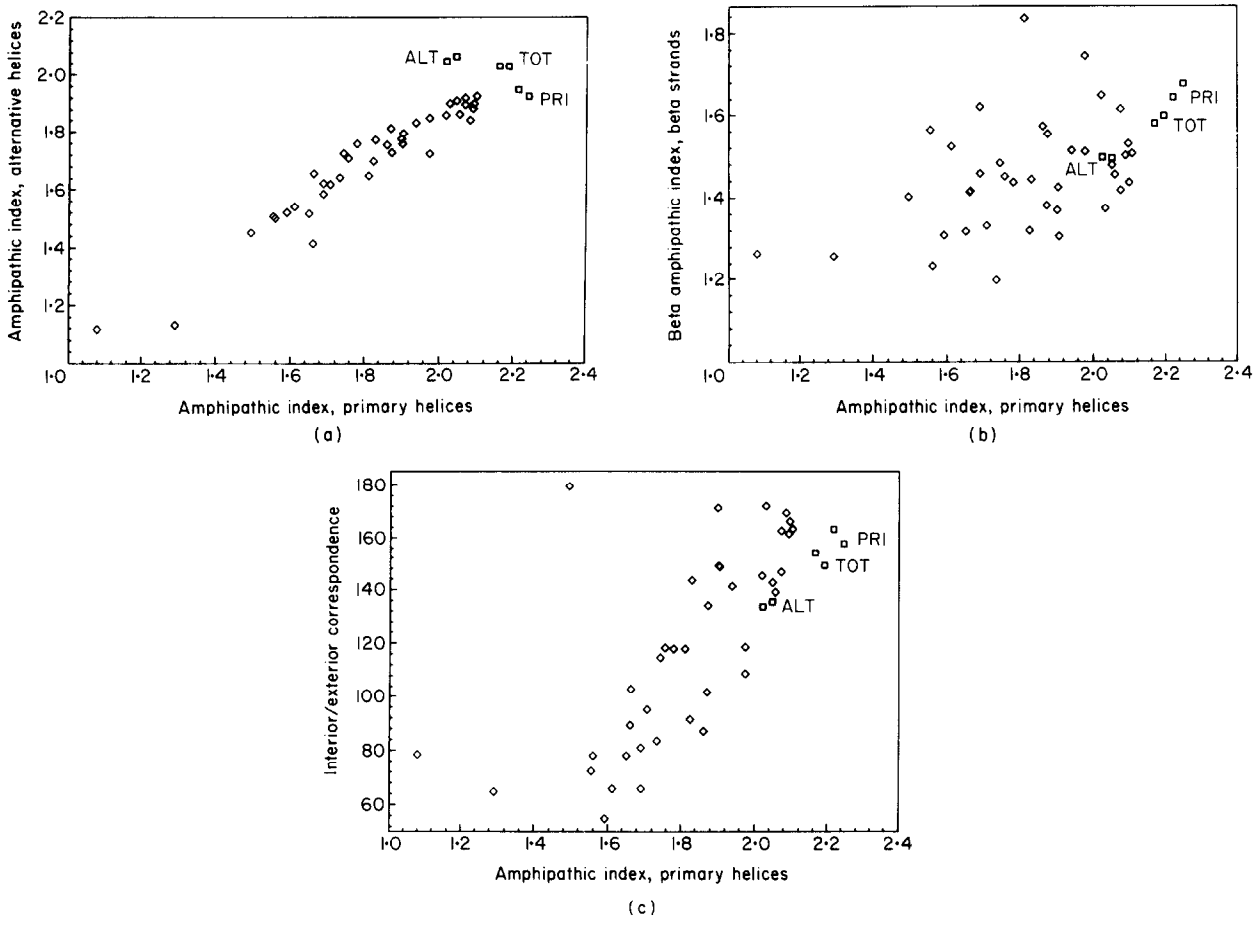

Figure 5. (a) The graph of the amphipathic index of $P_m(\omega)$ computed with the alternative helices *versus* the amphipathic index of $P_m(\omega)$ computed with the primary helices, for the different scales listed in Table 3. (b) The graph of the β-amphipathic index of $P_m(\omega)$ computed with the β-strands *versus* the amphipathic index of $P_m(\omega)$ computed with the primary helices, for the different scales listed in Table 3. (c) The graph of the interior/exterior correspondence of hydrophobicity for dogfish lactate dehydrogenase *versus* the amphipathic index of $P_m(\omega)$ computed with the primary helices, for the different scales listed in Table 4. The diamonds represent previously published scales, the squares represent scales that optimize an amphipathic index computed with the primary helices (PRI), with the alternative helices (ALT), and with the total set of helices (TOT).

approximately 8°, so there is substantial variation among helices.

What fraction of all helices are amphipathic? Even though the boundary between amphipathic and non-amphipathic helices is not finely drawn, we can approximate the fraction of helices in our data sets that are amphipathic, as follows. We used the amphipathic index to examine the primary and the secondary peaks of the composite power spectrum of the primary set of α-helices computed with the PRIFT hydrophobicity scale of this paper, the Miyazawa–Jernigan scale, the Kyte–Doolittle scale, and the Eisenberg *et al.* scale. The Eisenberg spectrum is the continuous-line graph in Figure 6 (see also Fig. 11). The questions posed are, how many helices contribute to the primary peak, and are the rather distinct secondary peaks also characteristic of an amphipathic α-helix, or are they due to a contribution from non-amphipathic helices.

The primary list of helices was partitioned into two sets using the PRIFT scale, those with amphipathic indices greater than 2·0 (72 helices out of the 115 helices in the primary set) and the complement (43 helices). In similar analyses, using the Miyazawa–Jernigan scale, 64 of the 115 primary set helices had amphipathic indices greater than 2·0; using the Kyte–Doolittle scale, 59 helices had amphipathic indices greater than 2·0; and using the Eisenberg *et al.* consensus scale, 50 helices had amphipathic indices greater than 2·0. The composite spectrum computed with the Eisenberg *et al.* scale for the primary set helices was decomposed into the sum of the contributions from the 50 helices with amphipathic index greater than 2·0 and the sum of the contributions from the remaining 65 helices. The results are illustrated by Figure 6, where it is apparent that the 50 clearly amphipathic helices account for almost all of the primary peak. Similar analysis for the other three scales showed, of course, even a larger portion of the primary peak attributed to the larger numbers of clearly amphipathic helices. Although the distinction between an amphipathic helix and a non-amphi-

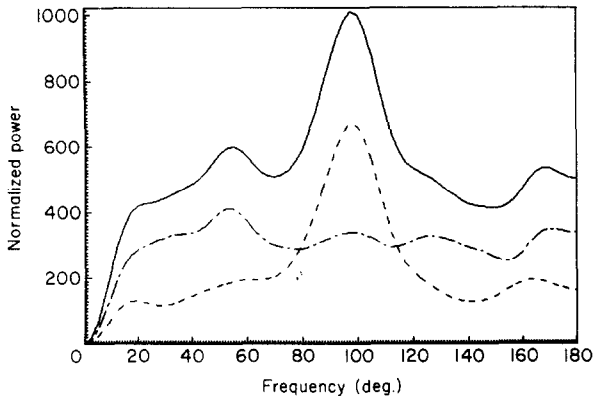

Figure 6. The sum of the power spectra for the 115 helices in the primary set using the Eisenberg *et al.* scale (continuous line) is partitioned into the sum for 50 helices with amphipathic index > 2·0 (broken line) and the sum for the remaining 65 helices (long dashes). Almost all of the amphipathic signal (97·5° peak of ———) is contributed by the 50 helices (97·5° peak of – – –). The approximately 50° peak of (———) is mostly associated with the 65 non-amphipathic helices (– –).

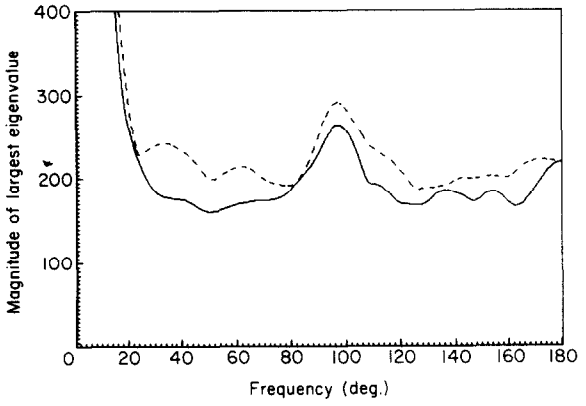

Figure 7. The graphs of the largest eigenvalue of $W_S(\omega)$ (eqn (5)) *versus* ω, where S is the primary set of helices for the continuous line and S is the alternative set of helices for the broken line. No hydrophobicity scale is involved. Both curves have a distinct local maximum near 97·5°.

pathic helix has not been sharply defined, we think it a reasonable conclusion that approximately one-half of the helices in the primary set of helices are amphipathic. Similar results were observed with the alternative set of helices, with the exception that a smaller percentage of the helices are amphipathic (e.g. with the Miyazawa–Jernigan scale, 49% of the alternative helices had an amphipathic index greater than 2·0, and 56% of the primary helices had an amphipathic index greater than 2·0).

It can also be seen in Figure 6 that the clearly amphipathic helices make little or no contribution to the secondary peaks. The 50° local peak is a consistent feature of the composite power spectra for essentially all hydrophobicity scales, and is generally due to the non-amphipathic helices. Presumably the non-amphipathic helices contribute largely random noise to the amphipathic index (e.g. the scale (not shown) that maximizes the amphipathic index computed with only the 72 clearly amphipathic helices according to PRIFT has 0·98 correlation with PRIFT). The noise is not completely random, but we have no physical basis for the 50° peak.

(c) *Eigenvalues and eigenvector formulations: 97·5° periodicity independent of the hydrophobicity scale*

The matrix $W_S(\omega)$ is a characteristic of the collection S of protein segments that is totally independent of hydrophobicity scales, and exhibits an interesting phenomenon when S is a collection of α-helices. Figure 7 shows graphs of the largest eigenvalue of $W_S(\omega)$ as a function of ω for S the primary collection of helices (continuous-line graph) and for S the alternative collection of helices (broken-line graph). A very distinct peak around $\omega = 100°$ is apparent in both curves, reflecting the tendency of certain pairs of amino acids to be located at intervals of approximately 3·6 residue spacing.

The largest eigenvalue $\lambda(\omega)$ of $W_S(\omega)$ is distinguished by two important characteristics. First, $\lambda(\omega)$ is the largest possible value of $\boldsymbol{\eta}^T W_S(\omega)\boldsymbol{\eta}$ for any hydrophobicity scale $\boldsymbol{\eta}$ (normalized to have Euclidean length = 1). Each curve shown in Figure 7, then, is the upper envelope of all possible power spectra for the respective set of helices. Second, because $W_S(\omega)$ is positive definite, $\lambda(\omega)$ is precisely the Euclidean norm of $W_S(\omega)$ (see Faddeev & Faddeev (1963)), and in that sense is a good measure of the "size" of $W_S(\omega)$.

We note that an eigenvector, $\boldsymbol{\xi}$, of $W_S(100°)$ associated with the largest eigenvalue of $W_S(100°)$ is the value of $\boldsymbol{\eta}$ that maximizes $P_S(100°)$. That is, $\boldsymbol{\xi}$ is the hydrophobicity scale that maximizes the Fourier transform at 100°, again assuming that all candidate scales are normalized to have Euclidean length = 1. The dominant eigenvector, $\boldsymbol{\xi}$, of $W_S(100°)$ also gives comparatively large values to $P_S(\omega)$ for ω away from 100°. For that reason, the optimum scales shown at the bottom of Table 3 are vectors that maximize the amphipathic index, and thus maximize the contrast between $P_S(100°)$ and $P_S(\omega)$ for ω away from 100°.

A 0·25° analysis of the graphs in Figure 7 shows that the local peak occurs at $\omega = 97·25°$ for the primary set of helices and at $\omega = 97·5°$ for the alternative set of helices. This is additional evidence that the dominant frequency of α-helices is around 97·5°, the center chosen for the amphipathic index. Moreover, this frequency is an inherent property of α-helices, independent of hydrophobicity scales (see Discussion).

(d) *Comparison with crystallographic co-ordinates*

We examined the crystallographic co-ordinates of a set of helices to see whether the 97·5° frequency correlates with observed structure, and to test a conjecture (Robert Jernigan, personal communication) that the more open conformation would be

associated with the C terminus of a helix. Our data base consisted of the crystallographic co-ordinates of 158 helices of length ten or more residues in the total set of helices (the proteins 3CAT, 156B and 1LHB were not used). Because the side-chain determines the hydrophobic character of an amino acid, it is the location and per residue rotation of the side-chain around the helix that determines the frequency of interest. We measured the per residue rotation for each helix as follows. For each seven-residue block of the helix we computed an axis for the block and the centroids of the side-chains, and for each two consecutive residues we computed the rotation angle for the two centroids about the axis (an angle between 2 planes, each containing the axis and 1 of the centroids). The rotation for two consecutive residues of the helix was chosen to be the rotation determined in a block centered as nearly as possible on the two residues (the end residues of the helix necessarily are end residues of the block). We then computed the average rotation angle (1) for the helix, (2) for the seven residues at the N terminus of the helix, and (3) for the seven residues at the C terminus. The axis chosen for each seven-residue block passes through the centroid of the α-carbon atoms of the seven residues and has the direction for which the projections of those α-carbon atoms onto the axis are most evenly distributed (the variance of the lengths between projections of consecutive residue α-carbon atoms onto the axis is minimized).

For the 158 helices, the average of the per residue rotations is 97·4° (S.D. = 2·3°) and the average for the N terminus and C terminus seven-residue blocks are 98·1° (S.D. = 3·4°) and 96·6° (S.D. = 3·9°), respectively. Although the standard deviations are rather large, the mean of 97·4° is clearly consistent with the observations above, the standard error of the mean is small (0·2°), and the C terminus side-chain rotation appears to be less than the N terminus rotation. The 158 paired data (N terminus rotation, C terminus rotation) were examined using a one-tailed Student's *t*-test. The null hypothesis that there is no difference between the ends is rejected with a significance of $p = 0{\cdot}0002$ ($t = -3{\cdot}52$).

Our helices include only residues in the α-helical conformation and exclude residues in the 3_{10} conformation. Adjacent to the 158 helices are 15 segments of 3_{10} structure. Inclusion of these segments makes less than 0·2° difference in the averages noted above; so they are too few to change the overall conclusions.

(e) *Eigenvectors related to two previously published scales*

Matrices and eigenvectors clearly are important in our analysis and have been important in at least two other scales, although not explicitly stated. Starting with the mutation matrix of amino acid replacements among closely related proteins of Dayhoff *et al.* (1978) (Fig. 80, page 346), Sweet & Eisenberg (1983) compute hydrophobicity values "which best represent the observed substitutions" as follows. The hydrophobicity of each amino acid is taken to be the average hydrophobicity of all amino acids found to replace it, weighted according to the frequency of replacement. The circular meaning of "hydrophobicity" in the previous statement is avoided as follows. Beginning with any hydrophobicity scale to make the first assignment of hydrophobicity, Sweet & Eisenberg compute a new, average, hydrophobicity scale based on the point mutation matrix. From that scale they average again to obtain a third scale, and they continue the iteration. It is necessary to normalize the new scale at each iteration to have mean zero and standard deviation of 1. Sweet & Eisenberg find that by doing so the normalized scales converge to a final scale that is *independent* of the starting scale of hydrophobicities. This interesting scale can be interpreted in terms of eigenvectors, as follows.

Let L be the (lower triangular) Dayhoff *et al.* (1978) matrix of numbers of accepted point mutations and let M be $L+L^{T}$, normalized to have row sums equal to 1. The largest eigenvalue of M is 1 with corresponding eigenvector **1** (the vector with each entry equal to 1). The second largest eigenvalue is 0·6004, and its corresponding eigenvector, when normalized to have mean zero and appropriate length is the Sweet & Eisenberg scale. See Table 3 where SWEET is the Sweet & Eisenberg scale and the normalized eigenvector is listed just below it as SWEIG. Clearly they are the same scale. Iterations such as they perform but without normalization to have mean zero will converge to the dominant eigenvector (**1** in this case; clearly an uninteresting scale because all hydrophobicity values would be the same!). The normalization at each iteration to have mean zero insures that the iterates are perpendicular to **1**, so the Sweet & Eisenberg iterates converge to the component perpendicular to **1** of the second dominant eigenvector. We extend this idea to another context in the next paragraph.

In the Ponnyswamy *et al.* (1980) hydrophobicity scale, the "average surrounding hydrophobicity of a residue" is computed by examining the neighbors of each residue in the folded state of a protein in a certain set of 21 proteins, as follows. Let $n'_{j,k}$ be the number of residues of type k (1 of the 20 amino acids) in any one of the 21 proteins distinct from but within 8 Å (1 Å = 0·1 nm) of any residue of type j in the same protein, and let $n_{j,k}$ be $n'_{j,k}$ divided by the total number of residues of type j among all 21 proteins. We refer to the matrix $N = [n_{j,k}]$ as the nearest neighbor matrix. For computation of the nearest neighbor matrix, we used the 42 proteins that are the basis of the Miyazawa & Jernigan (1985) study noted under Nearest neighbor matrix in Table 1 and the 8 Å radius noted above.

Let $\boldsymbol{\eta}_0$ be the Jones (1975) hydrophobicity scale. Then the Ponnuswamy *et al.* scale is the vector $\boldsymbol{\eta}_1$ computed by $\boldsymbol{\eta}_1 = N\boldsymbol{\eta}_0$. By iterations similar to that of Sweet & Eisenberg, one can obtain a self-

consistent scale, so that the "average surrounding hydrophobicity of a residue (type)" is proportional to the hydrophobicity of that type. To do so, one computes $\eta_2 = N\eta_1$, $\eta_3 = N\eta_2$, . . ., normalizing for length at each step, but no other normalization is necessary. The scales thus computed converge to a dominant eigenvector. A dominant eigenvector of the nearest neighbor matrix is listed under the code NNEIG in Table 3. Note from Table 5 that NNEIG correlates well with the scales of Nishikawa & Ooi (0·98), Ponnuswamy *et al.* (0·94), Krigbaum & Komoriya (0·90) and Miyazawa & Jernigan (0·89), all of which involve the nearest neighbor matrix or something similar. There are also correlations of 0·92 or higher with three other scales, Janin, Guy mean and Rose *et al.*

(f) *Comparison of Fourier transform and least-squares*

The frequency $\hat{\omega}_m$ that maximizes the Fourier power spectrum $P_m(\omega)$ often is an accurate estimate of a characteristic frequency in a series of hydrophobicity values $\{h_k\}_{k=0}^{l-1}$ for large values of l (say $l \geq 25$), but protein sequences as short as seven residues (or even 4) frequently are analyzed for periodicity in hydrophobicity values, in which case some error may be introduced. To explore the importance of this effect, we have tested the procedure on harmonic sequences of length seven residues defined by $h_k = \cos(k\omega_0 + \pi/4)$, $k = 0, 1, \ldots, 6$ that could represent hypothetical seven-residue hydrophobicity segments with the natural frequency ω_0. At 1° intervals for ω_0 in $0° \leq \omega_0 \leq 180°$, we computed $P_m(\omega)$ for the sequence $\{\cos(k\omega_0 + \pi/4)\}_{k=0}^{6}$ and the value $\hat{\omega}_m$ of ω for which $P_m(\omega)$ is maximum. Figure 8 is a graph of $\hat{\omega}_m$ *versus* ω_0 from which it is clear that, for these short sequences, the predicted frequency $\hat{\omega}_m$ may deviate substantially from the input frequency ω_0. Note particularly that $\hat{\omega}_m = 180°$ for $\omega_0 \geq 155°$. In contrast to the test of $\hat{\omega}_m$ illustrated by Figure 8, a least-squares fit of $\{D + A\cos(k\omega) + B\sin(k\omega)\}_{k=0}^{l-1}$ to any harmonic sequence $\{\cos(k\omega_0 + \phi)\}_{k=0}^{l-1}$ will be an exact fit when $\omega = \omega_0$. That is, the value $\omega = \tilde{\omega}$ that maximizes the least-squares power spectrum $Q(\omega)$ (or $Q_m(\omega)$) will be exactly ω_0. The diagonal line in Figure 8 may be considered to be the graph of the least-squares optimum frequency $\tilde{\omega}$ *versus* ω_0. Fortunately for the analysis of α-helices, the deviation of the Fourier optimum frequency $\hat{\omega}$ from ω_0 is small near 100° compared to the extremes at $\omega_0 < 40°$ or $\omega_0 > 150°$.

Figure 9 shows comparisons of the Fourier and least-squares power spectra $(2/l)P_m(\omega)$ and $Q_m(\omega)$, respectively, for four different α-helices, computed with the Kyte–Doolittle hydrophobicity scale ((a), (b) and (c)) and (d) with the Miyazawa–Jernigan scale. The most common comparison is shown in Figure 9 (a) in which $(2/l)P_m(\omega)$ and $Q_m(\omega)$ are quite similar, and for which $\hat{\omega}_m \doteq \tilde{\omega}_m$. However, the contrast of Figure 9 (b) is important, is not uncommon, and demonstrates that the least-

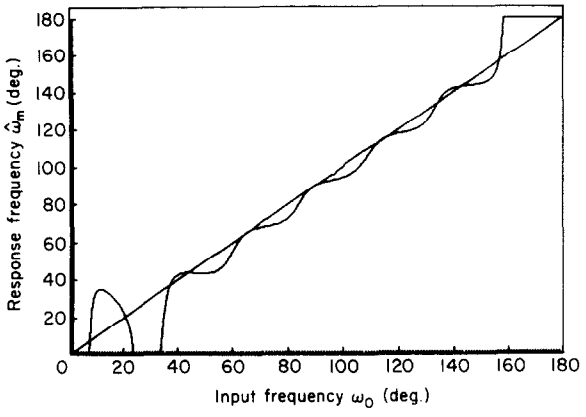

Figure 8. The graph of $\hat{\omega}_m$ *versus* ω_0, where $\hat{\omega}_m$ is the (response) frequency that maximizes the Fourier transform power spectrum $P_m(\omega)$ of the hypothetical hydrophobicity sequence $\{\cos(k\omega_0 + \pi/4)\}_{k=0}^{6}$ with intrinsic (input) frequency ω_0. The diagonal line represents response frequency = input frequency, which is true for the least-squares measure of frequency, $\tilde{\omega}_m$.

squares optimum frequency $\tilde{\omega}_m$ may detect important periodicity that the Fourier optimum frequency $\hat{\omega}_m$ misses. As stated above, for long sequences, $\tilde{\omega}_m$ usually is $\hat{\omega}_m$, but Figure 9 (c) illustrates the power spectra for an α-helix of 21 residues for which $\tilde{\omega}_m$ and $\hat{\omega}_m$ are quite different, and $\tilde{\omega}_m$ is probably more relevant. Figure 9(d) illustrates an instance in which $\hat{\omega}_m$ and $\tilde{\omega}_m$ differ substantially due to slight differences in the relative magnitudes of the two peaks. The actual curves differ very little, however, and both frequencies are probably important.

Figure 10 is a plot of $\tilde{\omega}_m$ *versus* $\hat{\omega}_m$ for the 115 helices of the primary set, in which it can be seen that $\tilde{\omega}_m$ is usually close to $\hat{\omega}_m$, and that both tend to cluster around 100°, but that there is significant scatter away from 100°. However, the triangles on the right-hand edge show that there is a tendency for $\hat{\omega}_m$ to be 180°. The manner in which this happens is clearly illustrated by Figures 9 (b) and (c), and in such instances we have a greater confidence in $\tilde{\omega}_m$. The least-squares spectra $Q(\omega)$ and $Q_m(\omega)$ have a peculiarity at $\omega = 0$, in that they are usually discontinuous there. It is fairly easy to see that $Q_m(0) = 0$. However:

$$\lim_{\omega \to 0} Q_m(\omega) = 0 \text{ only if } \sum_{k=-l}^{l_2} k'(h_k - \bar{h}) = \sum_{k=-l}^{l_2} (k')^2(h_k - \bar{h}) = 0.$$

Our program computes $Q_m(0)$ as $\lim_{\omega \to 0} Q_m(\omega)$.

The influence of this discontinuity is a tendency for $\tilde{\omega}_m$ to be zero, and is shown by the three triangles on the horizontal axes of Figure 10 at 16°, 34° and 43°. This low-frequency difference is not so serious, however; the 16° and 34° frequencies associate periods of length 22 and 11, respectively, that are longer than the peptides from which the frequencies were computed.

In the total collection of 250 helices in the

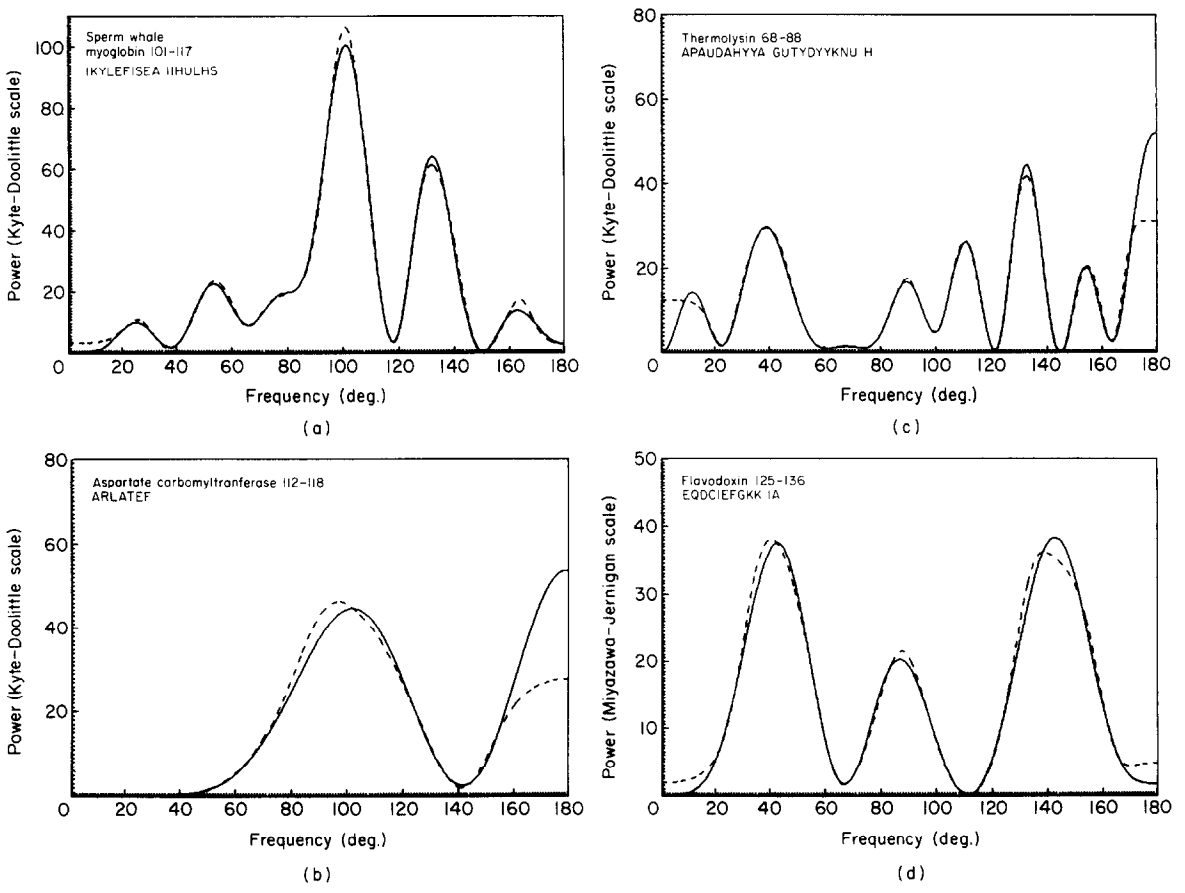

Figure 9. Graphs of the Fourier transform (continuous line) and least-squares (broken line) power spectra $(2/l)P_m(\omega)$ and $Q_m(\omega)$ of 4 helices computed with (a) to (c) the Kyte–Doolittle scale and (d) with the Miyazawa–Jernigan scale. The most common comparison (a) is that the 2 spectra are quite similar and the frequencies $\hat{\omega}$ and $\tilde{\omega}$ at which the maxima occur are quite close. However, (b) and (c) exhibit important differences that occur in approximately 1 out of 16 helices and in which the least-squares $\tilde{\omega}$ is more useful than the Fourier transform $\hat{\omega}$. In (d) there is a large difference in $\hat{\omega}$ and $\tilde{\omega}$, although both curves show almost equal peaks; possibly both frequencies are important characteristics of the sequence. (a) $\hat{\omega}_m = 101°$, $\tilde{\omega} = 101°$; (b) $\hat{\omega}_m = 180°$, $\tilde{\omega}_m = 97°$; (c) $\hat{\omega}_m = 180°$, $\tilde{\omega}_m = 133°$; (d) $\hat{\omega}_m = 143°$, $\tilde{\omega}_m = 41°$.

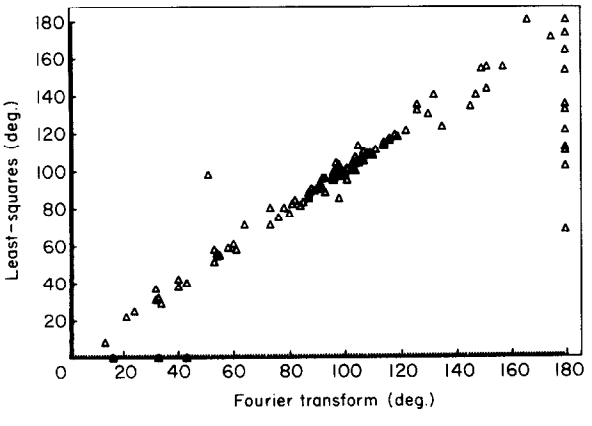

Figure 10. The graph of the frequency $\tilde{\omega}_m$ at which the least-squares spectrum $Q_m(\omega)$ is maximized *versus* the frequency $\hat{\omega}_m$ at which the Fourier transform spectrum $P_m(\omega)$ is maximized. Each triangle represents a helix in the primary set and the spectra were computed with the Kyte–Doolittle hydrophobicity scale. Most triangles are near the diagonal, indicating that $\tilde{\omega} \doteq \hat{\omega}$. The triangles on the right edge indicate a tendency for $\hat{\omega}$ to be 180°, as illustrated by Fig. 9(b) and (c). The triangles on the lower edge reflect a tendency for $\tilde{\omega}$ to be 0°.

primary and alternative sets, the least-squares peak frequency $\tilde{\omega}_m$ and the Fourier transform peak frequency $\hat{\omega}_m$ differ by no more than 5° in 195 of them and by more than 5° in 55 of them.

5. Discussion

The amphipathic index is a useful, objective measure of the ability of hydrophobicity scales to identify amphipathic helices. Intuitively, a successful scale in this context assigns similar numbers to residues that, on average, appear on the same side of α-helices and dissimilar values to amino acids that, on average, appear on opposite sides of α-helices. Thus hydrophobic forces are partly and indirectly involved in the test. The name "association" scale more accurately describes all of the scales computed in this paper (including the nearest neighbor eigenvector NNEIG and the Sweet & Eisenberg eigenvector SWEIG) and some of the scales from the literature. Two amino acids are "associated" if one frequently substitutes for the other in mutation; or if they are frequently near one another in the native state of proteins; or if they

frequently appear on the same side of α-helices. The scales measure the degree of association.

The consistency of the amphipathic index test is demonstrated by the linear relation between the amphipathic index computed with the primary helices and the amphipathic index computed with the alternative helices. The relevance of the amphipathic index is also demonstrated by the performance of the optimum scales, scales that maximize an amphipathic index, in settings outside the context in which they are optimum. This is particularly true of the PRI scales that maximize the amphipathic index of a power spectrum computed with the primary helices. These scales have high β-amphipathic indices and high interior/exterior correspondences.

Of the scales computed in this paper, we suggest the normalized form of PRIFT in Table 4 to be the most useful in algorithms searching for potential amphipathic structure in proteins. We have used this scale in an algorithm similar to that of Delisi & Berzofsky (1985) for identification of T-cell antigenic sites in proteins with reasonable success. The data base of T-cell antigenic sites is quite small, however, and possibly only three-fourths of the T-cell antigenic sites form helical amphipathic structures (Margalit *et al.*, 1987), so that the test is preliminary and inconclusive.

The primary helices have two advantages over the alternative helices. First, the crystallographic data for the primary list proteins are more certain than for the alternative list proteins (for the primary list proteins there are fewer "conflict in residue" notices at the beginning of the data bank files and fewer proteins with resolution greater than 2·5 Å). Second, a larger fraction of the primary helices are amphipathic than for the alternative helices, as demonstrated by the partitions of the two sets shown in Results, section (b). Clearly, in Figure 5 (b) and (c), the PRI scales perform better than the ALT scales (optimized with the alternative helices), and it appears that the inclusion of the alternative helices in the total set reduces the performance of the TOT scales. Clearly there is little difference in PRIFT (optimized with Fourier transform) and PRILS (optimized with least-squares) with correlation between them equal to 0·99, and we consider them equivalent. We also computed the scale that maximizes the amphipathic index of $P_m(\omega)$ for the primary set of helices. However, subtraction of the mean $\bar{h}$ from the hydrophobicity values in a sequence before computing a power spectrum is an accommodation to scales not balanced around zero; it eliminates a possible peak around zero that would distort the amphipathic index for such scales. When computing optimum scales, that accommodation is unnecessary, and for technical reasons we prefer the optimum scales based on $P(\omega)$ or $Q(\omega)$ over those based on $P_m(\omega)$ or $Q_m(\omega)$ (the optimizations based on $P_m(\omega)$ and $Q_m(\omega)$ are constrained to a 19-dimensional subspace of the potential 20-dimensional space of hydrophobicity scales). The scales based on $P_m(\omega)$ and $Q_m(\omega)$ correlate slightly less well with themselves and with previously published scales than do the scales based on $P(\omega)$ and $Q(\omega)$. Thus we suggest PRIFT as the most useful.

The normalized versions of the optimum scales shown at the bottom of Table 4 have values reasonably consistent with other scales and show high correlation with some of them. Particularly, PRIFT, PRILS, TOTFT and TOTLS all have correlations of 0·85 or higher with the experimental scale of Aboderin and eight of the statistical scales. Because the optimum scales also are statistical scales, it is, perhaps, not surprising that the highest correlations are with other statistical scales.

The only uniform abberation in the optimum scales of Table 4 is the positive (hydrophobic) assignment to arginine, suggesting that on α-helices arginine tends to be on the same side as hydrophobic residues, at least more so than does glycine. This observation, not previously noted, may reflect a physical or chemical property of arginine, or its interaction with other residues, that would be useful to explore experimentally. An interesting illustration of the location of arginine on helices is provided by sperm whale myoglobin residue 118, which is included in the crystallographers helix G of sperm whale myoglobin, although not the Kabsch–Sander designation. If placed on the helical wheel of Figure 1(a), it would be located in the one remaining blank space, entirely surrounded by hydrophobic residues. Arginine does not tend to appear at the end of our helices, but rather is distributed evenly along the helices. It may be noted that the normalized versions of the previously published scales assign to arginine a value ranging from −9·11 to 2·36, a spread almost as wide as that for tryptophan.

The normalized scales in Table 4 identify some interesting characteristics among the scales. For example, the experimentally determined scales assign proline a positive (hydrophobic) value (1 exception) reflecting its chemical behavior in solution, whereas the statistical scales generally assign to proline a negative (hydrophilic) value, reflecting its tendency to be exposed in a protein molecule, despite its hydrophobic character. Valine is generally more hydrophobic in the statistical scales than in the experimental scales. The ambiguity of threonine and of histidine is apparent, with both positive and negative assignment for experimental scales and also for statistical scales for both amino acids.

When analyzing information from a large collection of helices, the methods of the Fourier transform and least-squares analysis yield almost the same results. The correlation between optimum scales computed for corresponding power spectra of the two methods is 0·99 or 1·00. Graphs (not shown) based on $Q_m(\omega)$ are very similar to those shown in Figure 5(a), (b) and (c) that are based on $P_m(\omega)$. When examining single peptides, however, the two methods occasionally yield different information, as

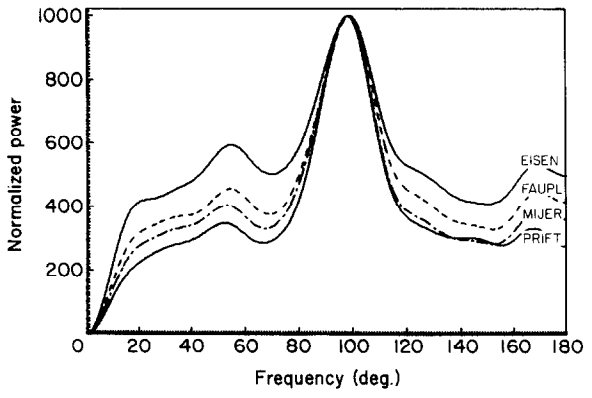

Figure 11. Graphs of the sum of the power spectra $P_m(\omega)$ for the primary helices using the Eisenberg *et al.* scale (EISEN), the Fauchère–Pliška scale (FAUPL), the Miyazawa–Jernigan scale (MIJER), and an optimum scale (PRIFT).

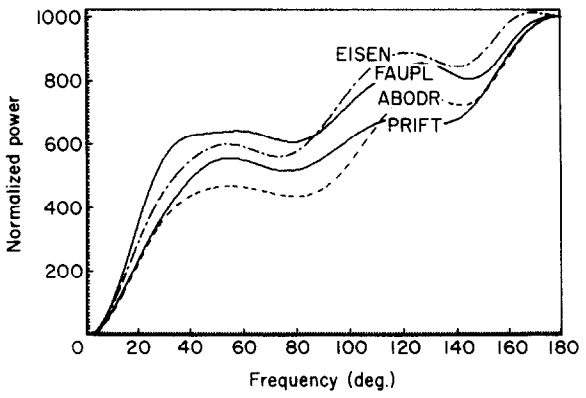

Figure 12. Graphs of the sum of the power spectra $P_m(\omega)$ for 161 Kabsch–Sander β-strands in the 22 proteins noted under Beta list in Table 1. The hydrophobicity scales used are Eisenberg *et al.* (EISEN), Miyazawa & Jernigan (MIJER), Aboderin (ABODR), and an optimum scale (PRIFT).

shown in Figures 8, 9 and 10, and under such circumstances the least-squares analysis appears to be more reliable, particularly for peptides of length, say, less than 25 residues.

The quadratic form matrix $W_S(\omega)$ extracted from $P_S(\omega)$ in Data Bases and Mathematical Models, section (d), carries sequence information contained in the set, S, of α-helices that is independent of hydrophobicity scales. The local maximum close to 97·5° of the largest eigenvalue $\lambda(\omega)$ of $W_S(\omega)$ expresses an important aspect of that information (Fig. 7). The natural frequency at 97·5° suggests that the typical α-helix is slightly more "open" than is usually considered, with 3·7 residues per turn being typical. Computations with crystallographic data confirm the more open conformation and indicate greater openness at the C terminus of the helix. We have found that $W_S(\omega)$ is also useful in identifying periodic variation in DNA sequences, where there are only four residue types.

Ben-Naim (1980) and Charton & Charton (1982) have seriously challenged the idea that a single hydrophobicity scale can account for the energy of folding of proteins and the several other chemical characteristics of amino acids attributed to their interaction with water molecules. Given the variety of scales already published, one might conclude that they are correct. We have examined the scales for one very specific characteristic, and find that some of them are distinctly better than others for identifying amphipathic helical structure. Several of the statistical scales perform better than all of the experimental scales, perhaps because the test itself is statistical. Among the strongest scales from the literature are Miyazawa & Jernigan (1985), Ponnuswamy *et al.* (1980), Guy, mean (1985), Wertz & Scheraga (1978), Nishikawa & Ooi (1980), and Rose *et al.* (1985*a*). It is interesting that the experimental scale of Aboderin (1971) based on relatively simple measurements of paper chromatography ranks at the top of the experimental scales. A close second among the experimental scales is that of Fauchère & Pliška (1983), and we have found this scale to be useful in other studies (Margalit *et al.*, 1987). Shown in Figures 11 and 12 are Fourier transform profiles, $P_m(\omega)$, for the primary helices using the Eisenberg scale, a strong experimental scale (FAUPL or ABODR), a strong statistical scale (MIJER), and PRIFT. Because the mean is subtracted, all of the profiles pass through the origin (0°, 0).

The scale PRIFT is based on the primary helices, correlates (0·996) with the optimum scale for $P_m(\omega)$ computed with the primary helices, and naturally the PRIFT spectrum displays the most distinct peak of any of the spectra shown in Figure 11. The conclusion is similar, however, for the alternative set of helices (not shown). In a different test, the β-strand spectrum of PRIFT shown in Figure 12 also has a 180° peak almost as distinct as that for any other scale. In a quite different test, the scale PRIFT also ranks reasonably high in the interior/exterior correspondence measurement shown in Table 6. We conclude that the amphipathic index is an important gauge of the performance of hydrophobicity scales, and that the scale PRIFT that maximizes an amphipathic index will be useful in other settings.

The authors gladly express their appreciation to Dr Robert L. Jernigan and Dr Lila M. Gierasch for their interesting and helpful comments on an early version of this paper.

APPENDIX I

Discrete Fourier Transform and Least-squares Fit of Harmonic Functions

A standard method for detecting periodicity in a finite sequence $h_0, h_1, h_2, \ldots, h_{l-1} = \{h_k\}_{k=0}^{l-1}$ of numbers is to compare the sequence with a test sequence of known periodicity, typically a harmonic sequence:

$$\{A\cos(k\omega) + B\sin(k\omega)\}_{k=0}^{l-1}$$

(or, equivalently, $\{R\cos(k\omega+\phi)\}_{k=0}^{l-1}$)†

of frequency ω and period $2\pi/\omega$. Two ways of comparing the tested sequence with the test sequence are (1) to compute a correlation between them (associated with the discrete Fourier transform) and (2) to compute the sum of the squares of the differences of corresponding entries (used in the least-squares analysis).

(a) *Discrete Fourier transform*

For any finite sequence of numbers $\{h_k\}_{k=0}^{l-1}$ (hydrophobicity values along a protein segment) and angular frequency ω measured in radians per residue, define $A(\omega)$, $B(\omega)$ and $P(\omega)$ by:

$$A(\omega) = \sum_{k=0}^{l-1} h_k \cos(k\omega)$$
$$B(\omega) = \sum_{k=0}^{l-1} h_k \sin(k\omega) \qquad \text{(A1)}$$

$$P(\omega) = [A(\omega)]^2 + [B(\omega)]^2 = \left[\sum_{k=0}^{l-1} h_k \cos(k\omega)\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin(k\omega)\right]^2. \qquad \text{(A2)}$$

Equation (A2) is equation (1) of the main text. Several authors examine $P(\omega)$ (or its square-root, $I(\omega)$, called intensity) for the value of $\omega = \hat{\omega}$ that maximizes $P(\omega)$, and interpret $\hat{\omega}$ as the characteristic frequency of $\{h_k\}_{k=0}^{l-1}$ and $2\pi/\hat{\omega}$ as its characteristic period. The underlying reason, briefly explained below, is that $A(\hat{\omega})\cos(k\hat{\omega}) + B(\hat{\omega})\sin(k\hat{\omega}) \doteq h_k$ often is a good approximation and always is the harmonic approximation that has the highest (non-normalized) correlation with $\{h_k\}_{k=0}^{l-1}$, thus identifying $\hat{\omega}$ as a dominant frequency.

From elementary trigonometric identities, it follows that $P(\omega+2\pi) = P(\omega)$ and $P(\pi+\omega) = P(\pi-\omega)$ so that all of the information about $P(\omega)$ is expressed on $0 \le \omega \le \pi$. Therefore, in all of our graphs only the frequencies $0 \le \omega \le \pi$ are plotted, and for clarity we mark the frequencies in degrees, which we use interchangeably with radian measure.

† Because
$A\cos(k\omega) + B\sin(k\omega) = R(A/R\cos(k\omega) + B/R\sin(k\omega)) = R\cos(k\omega+\phi)$ when $R^2 = A^2 + B^2$ and $\phi = \arctan(-B/A)$, $A\cos(k\omega) + B\sin(k\omega)$ and $R\cos(k\omega+\phi)$ are equivalent expressions for harmonic sequences and we use them interchangeably.

The rationale for accepting $\hat{\omega}$ as the characteristic frequency of $\{h_k\}_{k=0}^{l-1}$ is the following. Using trigonometric identities†, it can be shown that:

$$h_k = \frac{1}{l}\sum_{j=0}^{l-1} [A(2\pi j/l)\cos(k\,2\pi j/l) + B(2\pi j/l)\sin(k\,2\pi j/l)]. \qquad \text{(A3)}$$

Thus, the original sequence can be written as a sum of clearly periodic sine and cosine terms that have frequencies $2\pi j/l$ and periods l/j, $j = 1, l-1$. The "power" (square of the amplitude) of the jth term is $P(2\pi j/l)$. When for a certain $j = \hat{j}$, the power $P(2\pi\hat{j}/l)$ is "large" when compared to the powers of all the other terms, the corresponding frequency, $2\pi\hat{j}/l$, and period, $l/\hat{j}$, are taken to be the characteristic frequency and period of the original sequence. The frequencies $\omega_j = 2\pi j/l$ at which $P(\omega)$ is evaluated thus are very important in the usual discrete Fourier analysis, and are referred to as the Fourier frequencies. Because $\omega_{l-j} - \pi = \pi - \omega_j$, $P(\omega_{l-j}) = P(\omega_j)$, so that the only relevant frequencies are $0 \le \omega_j \le \pi$. The number, l, of terms in the original sequence of hydrophobicity values determines the number of Fourier frequencies. In most uses of the discrete Fourier transform, l will be quite large, the Fourier frequencies will be thickly distributed in $[0, \pi]$, and $\hat{\omega}$ will be near some Fourier frequency, and usually near the Fourier frequency, $2\pi\hat{j}/l$. However, for sequence of length seven residues, for example, the only Fourier frequencies are (converted to degrees) 0°, 51·4°, 102·9° and 154·3°, and $\hat{\omega}$ may not be near any of them.

Often a specific frequency is important that may or may not be a Fourier frequency (100° per residue frequency for α-helices, for example). An alternative to equation (A3) is, for any ψ, the identity:

$$h_k = \frac{1}{l}\sum_{j=0}^{l-1} [A(2\pi j/l+\psi)\cos(k(2\pi j/l+\psi)) + B(2\pi j/l+\psi)\sin(k(2\pi j/l+\psi))]. \qquad \text{(A3')}$$

Any special frequency of interest can thereby be included in the reconstruction of $\{h_k\}_{k=0}^{l-1}$. For example, if the segment length were eight residues and 100° frequency were important, the Fourier frequencies are 0°, 45°, 90°, 135° and 180°, but by

† The critical step depends on the identity:

$$\frac{1}{l}\sum_{j=0}^{l-1}\left[\cos\left(m\frac{2\pi j}{l}\right)\cos\left(k\frac{2\pi j}{l}\right) + \sin\left(m\frac{2\pi j}{l}\right)\sin\left(k\frac{2\pi j}{l}\right)\right] = \begin{cases} 1 & \text{for } k = m \\ 0 & \text{for } k \ne m \end{cases}$$

choosing $\psi = 10°$ the relevant frequencies become 10°, 55°, 100°, 145°, 190° ≃ 170°, 235° ≃ 125°, 270° ≃ 80° and 315° ≃ 45°. The frequencies greater than π "fold back" on to frequencies less than π, but not on to frequencies already present in the sum.

It can be shown that an arbitrary phase angle ϕ can be added to each of the angle terms in equation (A2) without affecting $P(\omega)$. Thus:

$$P(\omega) = \left[\sum_{k=0}^{l-1} h_k \cos(k\omega)\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin(k\omega)\right]^2 = \left[\sum_{k=0}^{l-1} h_k \cos(k\omega+\phi)\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin(k\omega+\phi)\right]^2. \quad \text{(A2')}$$

Equation (A2') yields two useful results. The first is revealed if one selects $\phi = (m+p)\omega$, where m is an integer and $0 \leq p < 1$, and computes:

$$P(\omega) = \left[\sum_{k=0}^{l-1} h_k \cos(k\omega+(m+p)\omega)\right]^2 + \left[\sum_{k=0}^{l-1} h_k \sin(k\omega+(m+p)\omega)\right]^2 = \left[\sum_{k=m}^{l+m-1} h_{k-m} \cos((k+p)\omega)\right]^2 + \left[\sum_{k=m}^{l+m-1} h_{k-m} \sin((k+p)\omega)\right]^2. \quad \text{(A2'')}$$

Thus, the summation index in the expression for $P(\omega)$ may be translated as needed. For comparison with the least-squares computation, below, it is convenient to choose $m = -(l-1)/2$ and $p = 0$ for odd values of l, and $m = -l/2$ and $p = 1/2$ for even values of l.

The second result from equation (A2') follows from the observation that the algebraic sign of:

$$\sum_{k=0}^{l-1} h_k \sin(k\omega+\phi),$$

for $\phi = 0$ is opposite to that for $\phi = \pi$, so there is a value $\hat{\phi}$ for which:

$$\sum_{k=0}^{l-1} h_k \sin(k\omega+\hat{\phi}) = 0.$$

Because for $\phi = \hat{\phi}$ the "sine" terms of equation A(2') is zero, the "cosine" term is at its maximum value. We obtain the interpretation that $P(\omega)$ is the square of the largest possible (non-normalized) correlation:

$$\sum_{k=0}^{l-1} h_k \cos(k\omega+\phi),$$

of $\{h_k\}_{k=0}^{l-1}$ with the harmonic sequence $\{\cos(k\omega+\phi)\}_{k=0}^{l-1}$, for any value of ϕ, and $P(\hat{\omega})$ is the largest such correlation for any value of ω and of ϕ. This is the basis for the earlier statement that $A(\hat{\omega})\cos(k\hat{\omega}) + B(\hat{\omega})\sin(k\hat{\omega}) \doteq h_k$ is the harmonic approximation that has the highest (non-normalized) correlation with $\{h_k\}_{k=0}^{l-1}$.

(b) *Least-squares*

An alternative to the Fourier analysis, and the correlation just shown, to estimate the frequency and period of a sequence $\{h_k\}_{k=0}^{l-1}$ is, for each value of ω, to select D, R and ϕ that give a best least-squares fit of $\{D+R\cos(k\omega+\phi)\}_{k=0}^{l-1}$ to $\{h_k\}_{k=0}^{l-1}$ or, equivalently, to select D, A and B to give a best least-squares fit of $\{D+A\cos(k\omega)+B\sin(k\omega)\}_{k=0}^{l-1}$ to $\{h_k\}_{k=0}^{l-1}$. The analog to $P(\omega)$ is $Q(\omega)$, the sum of squares accounted for, and the value $\tilde{\omega}$ of ω for which $Q(\omega)$ is maximum is taken to be the frequency of $\{h_k\}_{k=0}^{l-1}$. For large values of l, $\tilde{\omega}$ and $\hat{\omega}$ are usually equal, but for small values of l they differ, and in some sense $\tilde{\omega}$ may be considered the more accurate of the two.

The price of greater accuracy for short sequences is some increase in algebraic complexity. As in equation (A2''), we may shift the index of summation and fit the sequence $\{D+A\cos((k+p)\omega)+B\sin((k+p)\omega)\}_{k=-l_1}^{l_2}$ to $\{h_k\}_{k=0}^{l-1}$, where:

$$l_1 = (l-1)/2, \quad l_2 = l_1, \quad \text{and } p = 0 \text{ if } l \text{ is odd, and}$$
$$l_1 = l/2, \quad l_2 = l_1 - 1, \quad \text{and } p = 0{\cdot}5 \text{ if } l \text{ is even.} \quad \text{(A4)}$$

We denote $k+p$ by k'. The parameters l_1, l_2 and k' above have been selected so that the angles $k'\omega$ for $k = -l_1$ to l_2 are evenly distributed around zero, yielding:

$$\sum_{k=-l_1}^{l_2} \sin k'\omega = \sum_{k=-l_1}^{l_2} \sin k'\omega \cos k'\omega = 0.$$

The procedure, then, is to find $\tilde{A}$, $\tilde{B}$ and $\tilde{D}$ that minimize:

$$S(A, B, D) = \sum_{k=-l_1}^{l_2} [h_{k+l_1} - (D + A\cos k'\omega + B\sin k'\omega)]^2,$$

and let:

$$Q(\omega) = \sum_{k=0}^{l-1} h_k^2 - S(\tilde{A}, \tilde{B}, \tilde{D}).$$

This is a standard linear regression. After several algebraic steps we obtain:

$$Q(\omega) = l\bar{h}^2 + \frac{\left[\sum_{k=-l_1}^{l_2} (h_{k+l_1} - \bar{h})\cos k'\omega\right]^2}{\sum_{k=-l_1}^{l_2} \cos^2 k'\omega - \frac{1}{l}\left[\sum_{k=-l_1}^{l_2} \cos k'\omega\right]^2} + \frac{\left[\sum_{k=-l_1}^{l_2} (h_{k+l_1} - \bar{h})\sin k'\omega\right]^2}{\sum_{k=-l_1}^{l_2} \sin^2 k'\omega}. \quad \text{(A5)}$$

In the special case that $\bar{h} = 0$, there is a marked similarity between the expression (A2) for $P(\omega)$ and the expression (A4) for $Q(\omega)$. This is particularly true for the Fourier frequencies, ω_j, for:

$$\sum_{k=-l_1}^{l_2} \cos k'\omega_j = 0,$$

and:

$$\sum_{k=-l_1}^{l_2} \cos^2 k'\omega_j = \sum_{k=-l_1}^{l_2} \sin^2 k'\omega_j = \frac{l}{2}.$$

Then (when $\bar{h} = 0$): $Q(\omega_j) = \frac{2}{l} P(\omega_j)$.

Consequently, when comparing the two graphically, we always compare $Q(\omega)$ with $(2/l)P(\omega)$. We call $Q(\omega)$ the least-squares power spectrum of $\{h_k\}_{k=0}^{l-1}$.

(c) *Subtraction of* $\bar{h}$

When computing the Fourier transform for a sequence $\{h_k\}_{k=0}^{l-1}$, we have followed the practice of first subtracting

$$\bar{h} = \left(\sum_{k=0}^{l-1} h_k\right) \Big/ l \text{ from } h_k,\ k = 0, l-1,$$

and actually computing the power spectrum of $\{h_k - \bar{h}\}_{k=0}^{l-1}$ as:

$$P_m(\omega) = \left[\sum_{k=0}^{l-1} (h_k - \bar{h}) \cos(k\omega)\right]^2 + \left[\sum_{k=0}^{l-1} (h_k - \bar{h}) \sin(k\omega)\right]^2. \qquad (A2''')$$

A comparison between $P_m(\omega)$ with $P(\omega)$ shows that:

(1) for $\omega = 0$: because $\cos k\ 0 = 1$ and $\sin k\ 0 = 0$, $P_m(0) = 0$ and $P(0) = (l\bar{h})^2$;

(2) for $\omega = \omega_j$, a positive Fourier frequency: because

$$\sum_{k=0}^{l-1} \begin{matrix}\cos \\ \sin\end{matrix}(k\omega_j) = 0,$$

$$\sum_{k=0}^{l-1} (h_k - \bar{h}) \begin{matrix}\cos \\ \sin\end{matrix}(k\omega_j) = \sum_{k=0}^{l-1} h_k \begin{matrix}\cos \\ \sin\end{matrix}(k\omega_j),$$

so that $P_m(\omega_j) = P(\omega_j)$, and;

(3) $P(\omega) - P_m(\omega)$ is alternately positive and negative for ω in the successive intervals determined by the Fourier frequencies.

Thus, subtraction of $\bar{h}$ makes no change at the non-zero Fourier frequencies, and a possible peak at zero in $P(\omega)$ is deleted in $P_m(\omega)$ ($P(0) = (l\bar{h})^2$ may be quite large). The difference is particularly significant for certain hydrophobicity scales that, for example, have all positive values, and for which $P(0)$ shows up as unduly significant for all protein segments, whereas $P_m(\omega)$, with $P_m(0) = 0$, avoids that difficulty. Subtraction of $\bar{h}$ also tends to stabilize certain computations; for example, the values of $\hat{\omega}_m$ shown in the last column of Table 6 and explained in Results, section (b), have less variation than the corresponding values of $\hat{\omega}$ (not shown).

A subscript m always indicates that $\bar{h}$ was subtracted in the computation so that, for example, $\hat{\omega}_m$ is the value of ω that maximizes $P_m(\omega)$ and $Q_m(\omega)$ denotes the least-squares power spectrum of $\{h_k - \bar{h}\}_{k=0}^{l-1}$ (but note from eqn (A5) that $Q(\omega) - Q_m(\omega)$ is constant and equal to $l\bar{h}^2$).

The computation of $Q(\omega)$ by equation (A5) is a bit awkward for $\omega = 0$, because both denominators are zero, and for $\omega = \pi$ because one of the denominators is zero, depending on whether l is odd or even. $Q(0)$ and $Q(\pi)$ are evaluated using L'Hospital's rule (possibly 4 times!). The following FORTRAN computer program computes both $Q(\omega)$ and $P(\omega)$ at 0·50, 1° or 5° intervals, as desired. With the program, the two power spectra may be used with equal ease.

A FORTRAN PROGRAM TO COMPUTE POWER SPECTRA

```
C     COMPUTES THE FOURIER POWER, (2/L)P(W), & THE LEAST SQUARES SS ACCOUNTED FOR, Q(W),
C     FOR ANY INPUT SEQUENCE LENGTH, L, AND INPUT SEQUENCE  H(K), K=0 , ..., L-1.

      REAL H(0:30),P(0:360),Q(0:360)
      REAL*8 CC,SS,C1,CS,SN,PI,W,PK,H1,HM1,HK1,HKM1,HK2,SK2,SK4
C     INPUT L; H(K): K=0,...,L-1; MESH = 1, 2, OR 10 (FOR 5, 1, OR 0.5 DEGREE INTERVALS)
      PI=4.0D00*DATAN(1.0D00)
C     HBAR=0.0                                          ! INCLUDE THESE
C     DO 20 K=0,L-1                                     ! SIX LINES
C 20          HBAR=HBAR+H(K)                          ! IF THE
C     HBAR=HBAR/FLOAT(L)                                ! AVERAGE
C     DO 30 K=0,L-1                                     ! IS TO BE
C 30          H(K)=H(K)-HBAR                          ! SUBTRACTED
      FL=FLOAT(L)
      L1=L/2                                            ! L1 IS THE INTEGER
      L2=L1                                             ! PART OF L/2.
      PK=0.0D00                                         ! IF L IS ODD, THEN
      IF (2*L1.EQ.L) THEN                               ! L2=L1 AND PK = 0.0.
            L2=L1-1                                     ! IF L IS EVEN, THEN
            PK=0.5D00                                   ! L2=L1-1 AND PK=0.5.
      END IF
      H1=0.0D00
      HM1=0.0D00
      HK1=0.0
```

```
      HKM1=0.0
      HK2=0.0
      SK2=0.0
      SK4=0.0
      DO 40 K=-L1,L2
           H1=H1+H(K+L1)
           HM1=HM1+H(K+L1)*(-1.0)**K
           HK1=HK1+H(K+L1)*(K+PK)
           HKM1=HKM1+H(K+L1)*(K+PK)*(-1.0)**K
           HK2=HK2+H(K+L1)*(K+PK)**2
           SK2=SK2+(K+PK)**2
40         SK4=SK4+(K+PK)**4
      P(0)=(2.0/FL)*(H1*H1)                                    ! COMPUTATIONS FOR
      AVE=H1/FL                                                ! FREQUENCY = 0
      D=AVE*H1                                                 ! DEGREES
      A=(HK2-AVE*SK2)*(HK2-AVE*SK2)/(SK4-SK2*SK2/FL)           !
      B=HK1*HK1/SK2                                            !
      Q(0)=D+A+B                                               !
      DO 60 J=1,36*MESH-1
           W=J*PI/(36.0D00*MESH)
           C1=0.0D00
           HC=0.0
           HS=0.0
           CC=0.0D00
           SS=0.0D00
           DO 50 K=-L1,L2
                CS=DCOS((K+PK)*W)
                SN=DSIN((K+PK)*W)
                C1=C1+CS
                HC=HC+H(K+L1)*CS
                HS=HS+H(K+L1)*SN
                CC=CC+CS*CS
50              SS=SS+SN*SN
           P(J)=(2.0/FL)*(HC*HC+HS*HS)                         ! COMPUTATIONS FOR
           A=(HC-AVE*C1)*(HC-AVE*C1)/(CC-C1*C1/FL)             ! FREQUENCIES
           B=HS*HS/SS                                          ! BETWEEN  0 AND
           Q(J)=D+A+B                                          ! 180 DEGREES
60    CONTINUE
      P(36*MESH)=(2.0/FL)*HM1*HM1                              ! COMPUTATIONS FOR
      A=(HM1-AVE*(-1)**L1)*(HM1-AVE*(-1)**L1)/(FL-1.0/FL)      ! FREQUENCY =
      B=HKM1*HKM1/SK2                                          ! 180 DEGREES
      IF(2*L1.EQ.L) THEN                                       !
           A=((HKM1-AVE*L1*(-1)**(L1+1))**2)/(SK2-L1*L1/FL)    !
           B=HM1*HM1/FL                                        !
      END IF                                                   !
      Q(36*MESH)=D+A+B                                         !
C     OUTPUT ((J*5.0/MESH,P(J),Q(J)),J=0,36*MESH)
      STOP
```

APPENDIX II

Index to the Hydrophobicity Scales

ZIMMR Zimmerman (1968) published a table of properties of amino acids in which there is a hydrophobicity index based on solubility data given by Tanford (1962). The value 0·10 for glycine is much less than the values for aspartic acid and glutamic acid, and the center chosen for computing the normalized scale in Table 4 is 1·0.

N TAN Nozaki & Tanford (1971) Calculate the free energies of transfer of amino acid side-chains and backbone peptide units from water to ethanol and dioxane from the solubilities of the amino acids in those media. The free energies of transfer of the hydrophobic side-chains to 100% ethanol and dioxane are essentially identical, and are the basis for their hydrophobicity scale.

NTANR Rose modification of the Nozaki–Tanford scale. The scale shown appears in Rose *et al.* (1985*b*) and was first used by Rose (1978).

JONES Jones (1975) Adjusts the Zimmerman scale according to the experimental data given by Nozaki

& Tanford (1971). We chose the center to be 1·0, as in ZIMMR.

LEVIT Levitt (1976) Uses the Nozaki–Tanford values supplemented with estimates for the residues not characterized by Nozaki & Tanford (1971) based on the relationship of accessible surface area given by Lee & Richards (1971) and the hydrophobicity given by Chothia (1974).

HOPPW Hopp & Wood (1981) Adjust the Levitt scale so that the hydrophobicity profile would more successfully identify antigenic determinants in 12 proteins.

YUNGD Yunger & Cramer (1981) Measure quite accurately the octanol/water distribution coefficients of 12 amino acids. We have used their log *D* (logarithm of distribution ratio experimentally determined) and not their log *P* (logarithm of partition coefficient, calculated assuming the species has no net charge). The center of the scale for computing the normalized scale was shifted slightly to −2·80 from the glycine value of −3·11.

FAUPL Fauchère & Pliška (1983) Extend the octanol/water distribution measurements to all 20 amino acids. There are special computations for cystine and proline. We used the value of 1·54 given for cysteine rather than the 0·98 of half-cystine.

ZASLZ Zaslavsky *et al.* (1981) In order to extend the Nozaki–Tanford type measurements to polar side-chains, they measure the partition of solutes in an aqueous two-phase Ficoll-dextran system.

WOLF Wolfenden *et al.* (1981) Measure the distribution of amino acid side-chains between dilute aqueous solutions and the vapor phase. Proline is omitted (not strictly a side-chain). The side-chain of glycine (hydrogen) gives a value, 2·39, for glycine, larger than the value for any other side-chain. The center for this scale was chosen to be −8·0.

KUNTZ Kuntz (1971) Hydration. Measures the amount of water that does not freeze when an aqueous macromolecular solution is rapidly frozen and then equilibrated at −20 to −40°C. Addresses the question: can the hydration of a globular protein be described in terms of the hydration of its constituent amino acids? The value, 1·0, for glycine is the same as for cystine, valine, isoleucine, leucine and methionine; the center for the scale was chosen to be 1·6.

ABODR Aboderin (1971) Measures the mobilities of the amino acids on chromatography paper. The hydrophobicity scale is linearly proportional to R_F values on Whatman no. 3 paper, using the monophasic apolar solvent system ethyl acetate/pyridine/water (8 : 2 : 1, by vol.).

MEEK Meek (1980) Measures the retention times of 25 peptides in high-pressure liquid chromatography and computes the "retention coefficients" for the amino acids. The retention coefficients are selected to obtain a maximum correlation between the observed retention times and the retention times computed as the sum of the retention coefficients for each residue of the peptide.

BULDG Bull & Breese (1973) Measure the effect of each amino acid on the surface tension of water, and consider the slope of the surface tension relative to the concentration of the amino acid. They suggest that these slopes constitute a hydrophobicity scale, and convert that scale to the free energy of transfer of the amino acid from solution to the surface.

EISEN Eisenberg *et al.* (1982*b*). The scale is an average of five other scales: those of Nozaki & Tanford (1971), Wolfenden *et al.* (1981), Chothia (1976), Janin (1979), and von Heijne & Blomberg (1979).

KYTDO Kyte & Doolittle (1982). Combine the Wolfenden scale, the Chothia scale, and estimates based on the constituent parts of the side-chains.

CHOTH Chothia (1976) For each amino acid X, computes the proportion of all X-residues in a certain set of six proteins that are 95% buried in the native structure of the protein.

WERSC Wertz & Scheraga (1978) The residues in 20 proteins are classified as being inside or outside. The scale value, P_{in}, for amino acid X is the number of interior X-residues divided by the total number of X-residues.

JANIN Janin (1979) The scale is the ratio, *f*, of buried to accessible molar fractions of each amino acid, as measured in 22 proteins.

OLSEN Olsen (1980) Computes the average internal preference from the data given by Chothia (1975).

MEIRO Meirovitch *et al.* (1980) Compute the average normalized distance of the alpha-carbon of amino acid X from the center of the protein, for 19 proteins. The normalized distance is the actual distance from the center divided by the radius of gyration of the protein.

PONNU Ponnuswamy *et al.* (1980) Compute the "surrounding hydrophobicity" of each residue in 23 proteins, then, for each amino acid X, compute the average "surrounding hydrophobicity" for all of the occurrences of X. The surrounding hydrophobicity of residue *k* in a protein is the sum of the Jones hydrophobicities of all of the other residues in the protein whose alpha-carbon atoms are within 8 Å of the alpha-carbon of residue *k*. Also see Results, section (e) of this paper.

NNIEG A dominant eigenvector of the nearest neighbor matrix of Results, section (e) of this paper.

ROBOS Robson & Osguthorpe (1978) Is an information theory (or Bayesian) evaluation of −log (fraction of X buried/fraction of X not buried) +log (fraction all residues buried/fraction of all residues not buried). Their data base is the 25 proteins studied by Tanaka & Scheraga (1976), and their scale is proportional to the transfer free energy.

CHDLG WSDLG JADLG Chothia delta *G*, Wertz & Scheraga delta *G*, and Janin delta G are transfer energies computed from the respective scales. For example, in Janin it is simply $RT \ln f$, where *f* is the ratio of buried to accessible molar fractions. In Wertz & Scheraga, delta G is $-RT \ln P_{in}/(1-P_{in})$, and is computed by Guy (1985).

GUY Guy (1985) Is a transfer free energy

computation based on data given by Prabhakaran & Ponnuswamy (1980) classifying residues of 19 proteins as lying in one of six layers, from the surface to the center of the protein.

GUY M Guy mean is the average of four scales, his own, Ponnuswamy *et al.*, Meirovitch, and Wertz & Scheraga delta G.

KRIDG AND GRIGK Krigbaum & Komoriya (1979) Begin with the assumption that the free energy of transfer of a side-chain of type j from ethanol to water can be approximated by: delta G = volume$_j$ times $(A - Bx_j)$, where A and B are constants characteristic of water and ethanol, and x_j is an interaction parameter characteristic of the side-chain. Using solubility data given by Cohn & Edsel (1943) for 15 amino acids, they obtain first approximations of x_j for those side-chains. They estimate x_j for the remaining side-chains, and also estimate a maximum error in x_j for all 20 side-chains. Then, using crystallographic data from 23 proteins and a hypothesis that there should be a linear relation between the interaction parameter of a residue and the average of the interaction parameters of its nearest neighbors, they compute a least-squares best estimate of all the interaction parameters, subject to the constraint that they do not depart from the first approximate values by more than the estimated errors. KRIGK is the list of interaction parameters, and KRIDG values are the transfer free energies computed using the relationship above.

NIOII Nishikawa & Ooi (1980) The contact number of a residue is the number of α-carbon atoms within 8 Å of the α-carbon atom of the residue, omitting the α-carbon atoms of the residue and the two residues adjacent to it in the protein sequence. The contact number is a measure to show the location of the residue, on the surface or in the interior of the protein, and therefore is related to hydrophobicity. They compute a 19 by 20 matrix M so that a predicted contact number of the middle residue of a 19-residue block is the sum for $k = -9$ to 9 of $m_{k,j(k)}$, where $j(k)$ is the type of residue k. The scale shown is the middle row of M, giving the amount that a residue of type j contributes to its own contact number estimate.

MIJER Miyazawa & Jernigan (1985) Begin with crystallographic data for 42 proteins, count the number of residue–residue and residue–solvent contacts for different types of residues, and using methods of statistical mechanics, estimate from this information effective contact energies $e_{i,j}$ between residues of types i and j. An average contact energy e_i of the type i residue is computed from the $e_{i,j}$, and the average number of contacts, q_i, for a residue of type i is computed from the crystallographic data. The scale shown is $-0{\cdot}6\ q_i\ e_i/2$.

ROSEF Rose *et al.* (1985*a*) Compute the mean fractional area loss, f, for each amino acid as it appears in 23 proteins. $f = 1 - \langle A \rangle / A_0$, where A_0 is the solvent accessible surface area of the amino acid, X, in a standard state, and $\langle A \rangle$ is the average solvent accessible surface area of X in the 23 proteins.

SWEET Sweet & Eisenberg (1983) Compute "optimal matching hydrophobicity" (OMH) on the hypothesis that the hydrophobicity of each amino acid should be the average of the amino acids observed to substitute for it in point mutations, weighted according to the frequency of substitution. The mutation frequencies are given by Dayhoff *et al.* (1978). See Results, section (e).

SWEIG The component of a second dominant eigenvector of the normalized Dayhoff *et al.* (1978) matrix that is orthogonal to the vector with each entry = 1. See Results, section (e).

REKKR Rekker (1977) For each of 29 potential substituent parts (submolecule) of an organic molecule, Rekker computes a hydrophobic fragmental constant, the lipophilicity contribution of that constituent part to the total lipophilicity of the structure. The hydrophobicity of an amino acid is the sum of the fragmental constants of its constituent parts, adjusted slightly for separation of electronegative groups. The fragmental constants are computed from measured partition values for a large number of organic molecules and linear regression against the 29 potential substituent parts.

VHEBL von Heijne & Blomberg (1979) Estimate the free energy of transfer of a single residue in a polypeptide from a random coil conformation in an aqueous phase to the helix conformation in the nonpolar environment of a membrane interior. The scale shown is the sum of an hydrophobic contribution (computed from the accessible surface area values given by Chothia (1976)), a hydrogen bond contribution (each polar atom is assumed to form one hydrogen bond in water that is broken upon transfer into the membrane) and a charge contribution (each charged group is neutralized by adding or removing a proton). Interestingly, it was observed that the sum of the first two contributions yielded a scale (not shown) for which all of the parameters of Table 6 were higher.

FROMM Frömmel (1984) The apolar accessible surface area of a side-chain is the total accessible surface area of the side-chain minus a constant, f, times the sum of the products of the absolute value of the partial charge on each atom of the side-chain and the accessible area of the atom in the side-chain. The constant f is chosen to give a good correlation of apolar surface area with free energies of transfer, computed as the average of several experimental scales. This scale is computed from data in the paper, using $f = 3{\cdot}4$.

EIMCL Eisenberg & MacLachlan (1986) Compute the solvation energy of each amino acid as a sum of energies of the constituent atomic groups. The contribution from each atom is the product of its solvent-accessible area as determined by Lee & Richards (1971) and its atomic solvation parameter. Solvation parameters for five classes of atoms (carbon, neutral oxygen and nitrogen, charged oxygen, charged nitrogen and sulphur) are computed to give the best fit of the computed energies of transfer to the scale of Fauchère & Pliška (1983).

PRIFT Is the scale that maximizes the amphipathic index of the Fourier transform (FT) composite power spectrum for the primary (PRI) set of helices.
PRI, Primary set of helices. ALT, Alternative set of helices. TOT, Total set of helices, the union of the primary and alternative sets of helices.
LS, Least-squares power spectrum. FT, Fourier transform power spectrum.

References

Aboderin, A. A. (1971). *Int. J. Biochem.* **2**, 537–544.
Ben-Naim, A. (1980). *Hydrophobic Interactions*, Plenum Press, New York.
Bloomfield, P. (1976). *Fourier Analysis of Time Series: An Introduction*, John Wiley, New York.
Bull, H. B. & Breese, K. (1973). *Arch. Biochem. Biophys.* **161**, 665–670.
Charton, M. & Charton, B. I. (1982). *J. Theor. Biol.* **99**, 629–644.
Chothia, C. (1974). *Nature (London)*, **248**, 338–339.
Chothia, C. (1975). *Nature (London)*, **254**, 304–308.
Chothia, C. (1976). *J. Mol. Biol.* **105**, 1–14.
Cohn, E. J. & Edsel, J. T. (1943). *Proteins, Amino Acids, and Peptides as Ions and Dipolar Ions*, Reinhold, New York.
Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1978). In *Atlas of Protein Sequence and Structure 1978*, vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation, Silver Spring, MD.
Delisi, C. & Berzofsky, J. A. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 7048–7052.
Dunnill, P. (1968). *Biophys. J.* **8**, 865–875.
Eisenberg, D. & McLachlan, A. D. (1986). *Nature (London)*, **319**, 199–203.
Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1982*a*). *Nature (London)*, **299**, 371–374.
Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. (1982*b*). *Faraday Symp. Chem. Soc.* **17**, 109–120.
Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 140–144.
Eventhoff, W., Rossman, M. G., Taylor, S. S., Hans-Joachim, T., Meyer, H., Keil, W. & Hans-Hermann, K. (1977). *Proc. Nat. Acad. Sci., U.S.A.* **74**, 2677–2681.
Faddeev, D. K. & Faddeev, V. N. (1963). *Computational Methods of Linear Algebra*, W. H. Freeman & Co., San Francisco.
Fauchère, J. & Pliška, V. (1983). *Eur. J. Med. Chem.* **18**, 369–375.
Finer-Moore, J. & Stroud, R. M. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 155–159.
Frömmel, C. (1984). *J. Theor. Biol.* **111**, 247–260.
Guy, H. R. (1985). *Biophys. J.* **47**, 61–70.
Hopp, T. P. (1986). *J. Immunol. Methods*, **88**, 1–18.
Hopp, T. P. & Wood, K. R. (1981). *Proc. Nat. Acad. Sci., U.S.A.* **78**, 3824–3828.
IMSL, Inc. (1984). The IMSL Library, IMSL, Inc., Houston.
Janin, J. (1979). *Nature (London)*, **277**, 491–492.
Jones, D. D. (1975). *J. Theor. Biol.* **50**, 167–183.
Kabsch, W. & Sander, C. (1983). *Biopolymers*, **22**, 2577–2637.
Kaiser, E. T. & Kézdy, F. J. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 1137–1143.
Klein, P. & DeLisi, C. (1986). *Biopolymers*, **25**, 1659–1672.
Krigbaum, W. R. & Komoriya, A. (1979). *Biochim. Biophys. Acta*, **576**, 204–228.
Kuntz, I. D. (1971). *J. Amer. Chem. Soc.* **93**, 514–516.
Kyte, J. & Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105–132.
Lee, B. & Richards, F. M. (1971). *J. Mol. Biol.* **55**, 379–400.
Levitt, M. (1976). *J. Mol. Biol.* **104**, 59–107.
Lim, V. I. (1974*a*). *J. Mol. Biol.* **88**, 857–872.
Lim, V. I. (1974*b*). *J. Mol. Biol.* **88**, 873–894.
McLachlan, A. D. & Stewart, M. (1976). *J. Mol. Biol.* **103**, 271–298.
Margalit, H., Spouge, J. L., Cornette, J. L., Cease, K. B., DeLisi, C. & Berzofsky, J. A. (1987). *J. Immunol.* **138**, 2213–2229.
Meek, J. L. (1980). *Proc. Nat. Acad. Sci., U.S.A.* **77**, 1632–1636.
Meirovitch, H., Rackovsky, S. & Scheraga, H. A. (1980). *Macromolecules*, **13**, 1398–1405.
Miyazawa, S. & Jernigan, R. L. (1985). *Macromolecules*, **18**, 534–552.
Nishikawa, K. & Ooi, T. (1980). *Int. J. Pept. Protein Res.* **16**, 19–32.
Nozaki, Y. & Tanford, C. (1971). *J. Biol. Chem.* **246**, 2211–2217.
Olsen, K. W. (1980). *Biochim. Biophys. Acta*, **622**, 259–267.
Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965). *J. Mol. Biol.* **13**, 669–678.
Ponnuswamy, P. K., Prabhakaran, M. & Manavalan, P. (1980). *Biochim. Biophys. Acta*, **623**, 301–316.
Prabhakaran, M. & Ponnuswamy, P. K. (1980). *J. Theor. Biol.* **87**, 623–637.
Rekker, R. F. (1977). *The Hydrophobic Fragmental Constant*, Elsevier Scientific Publishing Company, Amsterdam.
Robson, B. & Osguthorpe, D. J. (1978). *J. Mol. Biol.* **132**, 19–51.
Rose, G. D. (1978). *Nature (London)*, **272**, 586–590.
Rose, G. D., Geselowitz, A. R., Glenn, J. L., Lee, R. H. & Zehfus, M. H. (1985*a*). *Science*, **229**, 834–838.
Rose, G. D., Gierasch, L. M. & Smith, J. A. (1985*b*). *Advan. Protein Chem.* **37**, 1–109.
Schiffer, M. & Edmundson, A. B. (1967). *Biophys. J.* **7**, 121–135.
Shrake, A. & Rupley, J. A. (1973). *J. Mol. Biol.* **79**, 351–371.
Sweet, R. M. & Eisenberg, D. (1983). *J. Mol. Biol.* **171**, 479–488.
Tanaka, S. & Scheraga, H. A. (1976). *Macromolecules*, **9**, 945–950.
Tanford, C. (1962). *J. Amer. Chem. Soc.* **84**, 4240–4247.
Wertz, D. H. & Scheraga, H. A. (1978). *Macromolecules*, **11**, 9–15.
Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. (1981). *Biochemistry*, **20**, 849–855.
von Heijne, G. & Blomberg, C. (1979). *Eur. J. Biochem.* **97**, 175–181.
Yunger, L. M. & Cramer, R. D. (1981). *Mol. Pharmacol.* **20**, 602–608.
Zaslavsky, B. Y., Mestechkina, L. M. M. & Rogozhin, S. V. (1981). *J. Chromatogr.* **240**, 21–28.
Zimmerman, J. M. (1968). *J. Theor. Biol.* **21**, 170–201.

Edited by A. Klug