J. Mol. Biol. (1983) **171**, 479–488

Correlation of Sequence Hydrophobicities Measures Similarity in Three-dimensional Protein Structure

ROBERT M. SWEET† AND DAVID EISENBERG

Molecular Biology Institute and Department of Chemistry and Biochemistry
University of California, Los Angeles
Los Angeles, Cal. 90024, U.S.A.

(Received 12 May 1983, and in revised form 26 July 1983)

The degree of similarity in the three-dimensional structures of two proteins can be examined by comparing the patterns of hydrophobicity found in their amino acid sequences. Each type of amino acid residue is assigned a numerical hydrophobicity, and the correlation coefficient r_H is computed between all pairs of residues in the two sequences.

In tests on sequences from two properly aligned proteins of similar three-dimensional structure, r_H is found in the range 0·3 to 0·7. Improperly aligned sequences or unrelated sequences give r_H near zero.

By considering the observed frequency of amino acid replacements among related structures, a set of optimal matching hydrophobicities (OMHs) was derived. With this set of OMHs, significant correlation coefficients are calculated for similar three-dimensional structures, even though the two sequences contain few identical residues. An example is the two similar folding domains of rhodanese ($r_H = 0·5$).

Predictions are made of similar three-dimensional structures for the alpha and beta chains of the various phycobiliproteins, and for delta hemolysin and melittin.

1. Introduction

Fitch (1966), McLachlan (1971), Barker & Dayhoff (1972), Doolittle (1981), Jue *et al.* (1980) and others have developed statistical methods to detect if two amino acid sequences are related by divergent evolution. We address a related question: whether or not two amino acid sequences are likely to fold into closely similar three-dimensional structures. Of course, these questions are related because protein primary structure determines three-dimensional folding: methods that detect evolutionary relationships among sequences should detect similarity in three-dimensional structure. But what of cases of weakly related sequences? A method that compares sequences in structural rather than statistical terms might detect structural similarities more effectively. This could be especially useful for

† Current address: Biology Department, Brookhaven National Laboratories, Upton, Long Island, N.Y. 11973, U.S.A.

0022-2836/83/360479-10 $03.00/0

synthetic analogues to proteins for which there is no evolutionary relationship and where there may even be no identical residues in the two structures.

A property of an amino acid residue that is related to three-dimensional protein structure is the hydrophobicity. Among the observed relationships between protein structure and hydrophobicity are the following. (1) Alpha helices that lie at protein surfaces tend to have one face projecting mainly hydrophobic residues and an opposite face projecting mainly hydrophilic residues (Perutz *et al.*, 1965; Schiffer & Edmundson, 1967; Eisenberg *et al.*, 1982*a*). (2) Beta sheets, which frequently occur in the interior of globular proteins (Richardson, 1981), tend to be particularly rich in hydrophobic residues (Chou & Fasman, 1978). (3) Beta turns and other abrupt bends that reverse chain directions at the surface of globular proteins tend to be especially hydrophilic (Kuntz, 1972; Rose, 1978). These relationships are sufficiently strong that they suggest that two amino acid sequences folding into similar three-dimensional structures are likely to have highly correlated hydrophobicities. Section 2 describes quantitative tests of this hypothesis using globins, cytochromes, rhodanese and melittin.

The hydrophobicity method of comparing sequences can be combined with statistical methods. In McLachlan's statistical method for determining relationships between two protein sequences, differences between the sequences are compared to a matrix of known substitution frequencies, determined from a large population of proteins from many different families. Moreover, comparative sequence studies (Margoliash, 1963; Dayhoff *et al.*, 1979; McLachlan, 1971) have established that amino acids that substitute for each other in related three-dimensional proteins have similar chemical properties. This leads to the method of section 3 in which we combine McLachlan's statistically based matrix method with hydrophobicity correlation. From substitution data, we determine a new refined set of amino acid properties, called OMHs†, which fit best to the substitution data. In section 4, predictions of structural similarity are made with these OMHs.

2. Hydrophobicity Correlation

Numerical hydrophobicities were taken from the consensus scale of Eisenberg *et al.* (1982*b*), in which hydrophobicity is roughly proportional to the free energy required to move an amino acid residue from the interior to the surface of a hydrated protein, except that the hydrophobicities were first normalized (to a mean of zero and a standard deviation from the mean of 1·0). Then the correlation between the two sequences is calculated from:

$$r_{\mathbf{H}} = \frac{\sum_i H_{i1} H_{i2}}{\left(\sum_i H_{i1}^2 \sum H_{i2}^2\right)^{\frac{1}{2}}}, \tag{1}$$

where the sums are over the residues in the matched sequences; H_{i1} and H_{i2} are hydrophobicities of the ith residue in sequences 1 and 2, respectively. When

† Abbreviation used: OMH, optimal matching hydrophobicity.

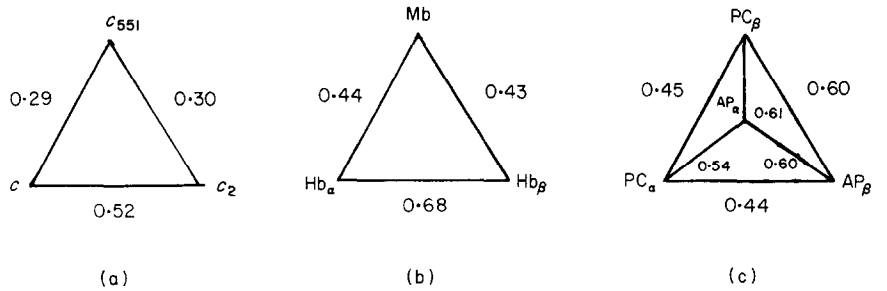

FIG. 1. Hydrophobicity correlation coefficients (r_H from equation (1)) among pairs of aligned sequences. (a) Cytochrome c_2 from *Rhodospirillum rubrum*, cytochrome c from tuna, and cytochrome c_{551} from *Pseudomonas aeruginosa*. (b) Sperm whale myoglobin and the α and β chains of horse hemoglobin. (c) α and β chains from allophycocyanin (AP) and phycocyanin (PC) from *Mastigocladus laminosus*.

alignment of the sequences requires that gaps appear in one, the residues of the other make no contribution to r_H. For two identical sequences $r_H = 1$, for two random sequences $r_H = 0$, and for two sequences of anti-correlated hydrophobicities, $r_H = -1$. Simple computer programs in FORTRAN, available from the first author, are used to perform these calculations.

In actual comparisons, r_H falls between 0·99 and about 0 and one must have a basis for deciding what value indicates similarity of structure. One indication of similarity is the range of values found for r_H for pairs of similar structures. Values of $r_H = 0{\cdot}3$ are found (sections 3 and 4 below) for similar molecules of 100 to 200 residues with 20% identical residues (or $r_H = 0{\cdot}1$ if identical residues are exluded from the comparison) with the consensus hydrophobicities. Using the OMH scale, we find $r_H = 0{\cdot}35$ (0·20 if identical residues are excluded).

A quantitative measure of the significance of r_H can be estimated from Student's t test. The relationship between r_H and t is $t = [r^2 v/(1-r^2)]^{1/2}$, in which v is the number of residues less two, and the significance of t is tabulated. In the examples below, two cytochromes (Fig. 1(a)) (100 residues) which yield $r_H = 0{\cdot}3$ can be taken as similar to the 99·9% level of confidence, and two melittin-like molecules (26 residues) which yield $r_H = 0{\cdot}71$ can be taken to be similar well beyond the 99·9% level of confidence.

In an alternative test of significance, one can compare r_H with r_H values calculated after the sequence of one of the peptides has been shuffled in a random way. The mean r_H after many such shufflings is close to zero, and the ratio of the observed r_H to the standard deviation of this mean can be used in a t-test for significance. This test is virtually identical to that performed by McLachlan (1971) and it gives results that are virtually the same as the more direct method above. In the comparison of human α-hemoglobin to whale myoglobin, McLachlan (1971) finds that the similarity can be rejected with a probability of 10^{-9}. Using the method of the previous paragraph, we find the similarity can be rejected with a probability of 10^{-13}, and using the method of this paragraph, we find it can be rejected with a probability of 10^{-9}.

Both these tests for significance may overestimate the confidence level, as noted

by Walter Fitch (personal communication). The reason is that the sample sequences have been rationally aligned, with gaps where most appropriate, to maximize the correlation, but no comparable procedure has been carried out with the randomized sequences. Consequently, we avoid reliance on the t-tests, and use them only to indicate a general statistical threshold for significance. Instead we use comparison with r_H values from pairs of structures known to be similar to set meaningful criteria for r_H.

3. Tests with Cytochromes, Globins, Rhodanese and Melittin

Values of r_H were determined for pairs of proteins of the same three-dimensional structure. Known structures were used to assure that residues compared in the correlation play as precisely as possible the same structural role in the two molecules.

In the first test, all three pairwise comparisons were made between cytochrome c_2 from *Rhodospirillum rubrum*, cytochrome c from tuna, and cytochrome c_{551} from *Pseudomonas aeruginosa*. X-ray analyses have shown that these molecules have the same basic folding pattern, and they differ mainly in the addition or deletion of surface loops of polypeptide chain (Dickerson, 1980). Cytochrome c_2, with enlarged loops to the left and right, is classified as a "large" cytochrome, tuna cytochrome c is termed "medium", and cytochrome c_{551}, with a large deletion at the bottom, is termed "small". Alignment of the amino acid sequences is derived mainly from the three-dimensional structures (Dickerson, 1980). Figure 1(a) shows that the pairwise correlation coefficients between these three proteins is 0·29 or greater.

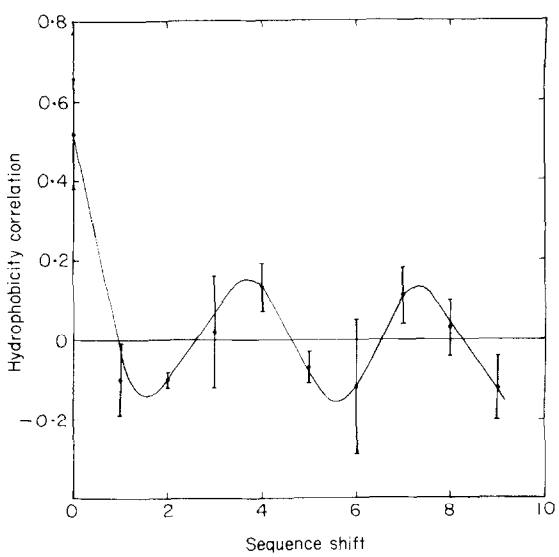

FIG. 2. Mean values of r_H for comparisons among the 3 globin chains used in Fig. 1(b), plotted against a misalignment shift applied to one of the chains before the correlation was calculated. Error bars are standard deviations of the 3 calculated values; a free-hand curve is cast through the points.

A second test was carried out with sperm whale myoglobin and the α and β chains of horse hemoglobin. Once again the sequences were aligned by reference to three-dimensional structures (Dickerson & Geis, 1982; Lesk & Chothia, 1980). Figure 1(b) shows that the pairwise correlations in hydrophobicities are 0·43 or greater.

The method is sensitive in detecting proper alignment of two sequences, as is shown for several globin sequences in Figure 2. A misalignment by exactly one residue produces a drastic drop in r_H. This suggests that the three-dimensional positioning of hydrophobic elements in a stable structure must be precise. When sequences are misaligned by more than one position, small maxima result, with a periodicity of three to four positions. These arise when the polar and non-polar sides of the globin helices come back into proper registration, but the r_H values are too small to be confused with the dominant maximum at the correct alignment.

A final test with globins was designed to discover whether identical residues dominate r_H or whether a significant value of r_H can result from non-identical residues alone. Fifteen diverse globin sequences were aligned (Dickerson & Geis, 1982; Lesk & Chothia, 1980) and compared using equation (1). However this time all of the residue pairs that are identical *were excluded from the summation.* Figure 3 displays the relationship between r_H values and the fraction of residues that are identical for all 105 possible pairs of peptides. The great bulk of the r_H

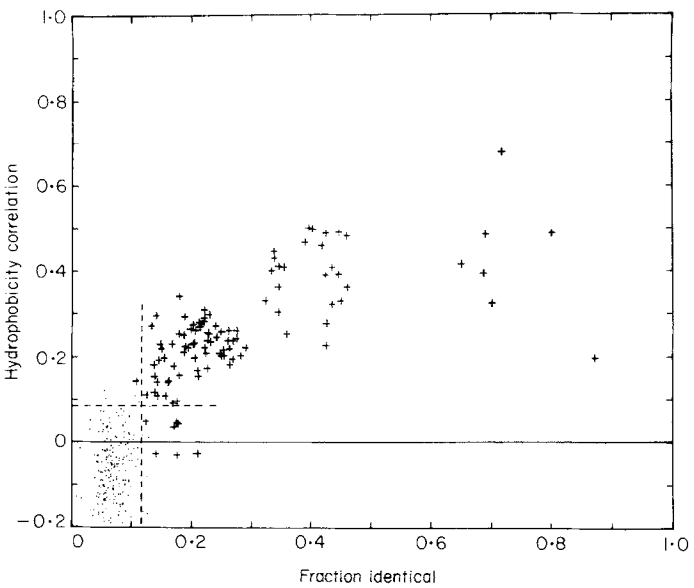

FIG. 3. The hydrophobicity correlation coefficient, r_H, for each possible comparison of 15 globin chains, only for residue pairs *that are not identical,* plotted against fraction of residues found to be identical (+). Also plotted as a control is r_H *versus* fraction identical, after 2 sequence randomizations each, of both chains used in the pairwise comparisons (·). Threshold values of r_H and fraction identical are shown as broken lines at the 95% level of significance. The mean value of r_H is negative in this distribution because of the systematic exclusion of the occasional occurrence of identical residues in the 2 randomized sequences.

values lie above the threshold of 95% significance, as indicated by Student's *t*-test. Therefore we conclude that the non-identical residues alone suggest structural similarity because the patterns of hydrophobicity they display are significantly similar.

Tests were carried out with two pairs of sequences having less strong relationships. One was on the sequence of the enzyme rhodanese (Ploegman *et al.*, 1978), which contains two domains folded in similar manner but with only 13 identical residues of the 98 in the common fold. The hydrophobicity correlation, r_H between the two domains is 0·26 (0·21 with identical residues excluded). The final test was for melittin (a tetramer of helical peptides) (Terwilliger & Eisenberg, 1982) from bee venom, compared to a synthetic analogue (DeGrado *et al.*, 1981), believed also to be a tetramer of helical peptides. Here there are 11 identical residues out of 26, and $r_H = 0{\cdot}91$ (0·71 with identical residues excluded).

In summary, properly aligned sequences from similar structures produce r_H values greater than about 0·25 with hydrophobicities of the consensus scale. In the following, we show that there is another scale which produces larger r_H values for similar structures.

4. A Hydrophobicity Scale to Optimize the Match between Similar Sequences

Dayhoff *et al.* (1979) and McLachlan (1971) have used observed frequencies of amino acid replacement within several families of proteins to estimate the similarities between pairs of sequences. Starting with similar data (1572 amino acid replacements among closely related proteins, compiled by Dayhoff *et al.* (1979)), we have "refined" hydrophobicities to values which best represent the observed substitutions.

To achieve refinement, the hydrophobicity H_i of each residue type was taken to be the average hydrophobicity of all residues found to replace it in the compared sequences. To achieve self-consistency, an iterative refinement was performed. At each cycle the new hydrophobicity was averaged with that from the last cycle, permitting gradual convergence. After each cycle the scale was normalized to a mean of zero and a standard deviation of 1·0. This process has the effect of producing a scale that will give the maximum possible value of r_H for the sequences being compared.

The final H_i values are largely independent of the starting hydrophobicities. This was found from refinements from three different starting scales. These included the consensus scale, and the trivial scale in which Leu had a value of 1·0, Lys had a value of $-1{\cdot}0$, and all others were 0. The refined values are indistinguishable; therefore, the refinement is robust.

The refined OMH scale is given in Table 1, where it is compared to other hydrophobicity scales. The agreement of the scales can also be expressed as a correlation coefficient: if equation (1) is applied to the OMH scale and to each of the other scales, r_H is found to be 0·72 with the consensus scale, 0·54 with the Wolfenden scale and 0·58 with the Janin scale. (With a sample of 20, $r = 0{\cdot}54$ is significant at the 99% level of confidence.) Thus, the OMH scale, although derived

TABLE 1

Hydrophobicity scales

Residue	OMH scale (from Dayhoff data)	Experimental scales		
		Consensus	Wolfenden	Janin
Phe	1·92	1·19	0·67	0·87
Tyr	1·67	0·26	−0·23	−0·40
Ile	1·25	1·38	1·16	1·16
Leu	1·22	1·06	1·18	0·87
Met	1·02	0·64	0·55	0·73
Val	0·91	1·08	1·13	1·02
Trp	0·50	0·81	−0·19	0·59
Cys	0·17	0·29	0·59	1·44
Thr	−0·28	−0·05	−0·02	−0·12
Ala	−0·40	0·62	1·12	0·59
Pro	−0·49	0·12	0·54	−0·26
Ser	−0·55	−0·18	−0·05	0·02
Arg	−0·59	−2·53	−2·55	−1·82
His	−0·64	−0·40	−0·93	0·02
Gly	−0·67	0·48	1·20	0·59
Lys	−0·67	−1·50	−0·80	−2·39
Gln	−0·91	−0·85	−0·78	−0·83
Asn	−0·92	−0·78	−0·83	−0·55
Glu	−1·22	−0·74	−0·92	−0·83
Asp	−1·31	−0·90	−0·83	−0·69

The OMH scale was refined from Dayhoff and co-workers' replacement data (1979) on 1572 amino acid substitutions among closely related proteins. The consensus scale (Eisenberg *et al.*, 1983) is an average of 5 other scales, including those of Wolfenden *et al.* (1979,1981) and Janin (1979). These were derived, respectively, from vapor pressures of side-chain analogues and counts of buried and exposed residues in globular proteins. All scales have been normalized with a mean of 0 and a standard deviation of 1·0.

from data on amino acid substitutions, reasonably represents the hydrophobicities of the residues.

The effectiveness of the OMH scale in correlating similar sequences is shown in Table 2. In all cases, except the unusual melittin-like peptides, r_H is greater for this scale than for the others. Notice that r_H for the weakly related domains of rhodanese more than doubles in using the OMH scale in place of the consensus.

5. Predictions

As illustrations of the possible usefulness of the present method, we can predict the extent of similarity of two pairs of proteins, based in both cases on sequence alignments proposed by others. The first prediction is on phycobiliproteins. For these proteins, work in progress on X-ray diffraction studies (Sweet *et al.*, 1977; Fisher *et al.*, 1980) would benefit from a firm estimate of the structural similarity of their α and β subunits for both allophycocyanin and phycocyanin. Sidler *et al.* (1981) aligned the sequences of the alpha and beta chains of allophycocyanin and phycocyanin from the cyanobacterium *Mastigocladus laminosus*. Then by applying the Chou & Fasman (1978) rules for secondary structure prediction, they

TABLE 2

Hydrophobicity correlation coefficients

	Sequences compared			
	Globins and cytochromes	Biliproteins	Rhodanese	Melittins
Number of residues considered	13000	950	98	26
Number of non-identical residues	8520	630	85	15
Hydrophobicity scales	r_H values:		*All residues* Non-identical residues	
OMH	0·61/0·38	0·67/0·47	0·53/0·46	0·77/0·60
Consensus	0·48/0·19	0·54/0·30	0·26/0·21	0·91/0·71
Janin	0·43/0·10	0·53/0·29	0·23/0·19	0·85/0·57
Wolfenden	0·44/0·14	0·51/0·24	0·21/0·13	0·90/0·70

Hydrophobicity correlation coefficients (r_H values) for comparisons of known protein structures with 4 hydrophobicity scales. Globins and cytochromes represent all 66 possible comparisons among 12 globin sequences (Dickerson & Geis, 1982) and all 36 possible comparisons among 9 *c*-type cytochrome sequences (Dickerson, 1980). The biliproteins are all possible combinations represented in Figure 1(c). Rhodanese is the 98 pairs of residues from the 2 domains of rhodanese that can be closely superimposed (Ploegman *et al.*, 1978). Melittins are the bee venom peptide melittin and its synthetic analogue (DeGrado *et al.*, 1981). The hydrophobicity scales are those shown in Table 1.

concluded that the four peptides have different folding patterns. In contrast, the present method of hydrophobicity correlation suggests that the four chains are similar to each other. The results of correlations of the sequences as aligned by Sidler *et al.* (1981) are given in Table 2 and Figure 1(c). The fact that these correlations, for peptide chains of about the same length as the globins, are all greater than those for the average of the cytochromes/globins, suggests that all four of these phycobiliproteins will be found to be as similar in detail as myoglobin is to either of the haemoglobin chains.

A second prediction is for delta hemolysin, a toxic peptide from *S. aureus*, sequenced by Fitton *et al.* (1980). Like bee melittin, it is a 26-residue peptide. Fitton *et al.* (1980) determined the amino acid sequence of delta hemolysin and proposed that it can be aligned residue-for-residue with melittin. Accepting this alignment and using the consensus hydrophobicity scale, we calculate values of r_H of 0·55 and 0·52 for all residues and non-identical residues, respectively (only one residue is common). With the OMH scale, these r_H values become 0·51 and 0·47. These values are smaller than those relating melittin to the synthetic peptide known to have a structure similar to melittin (see section 3). Nevertheless, the r_H values are probably large enough to indicate a similar structure also for delta hemolysin.

6. Discussion

The hydrophobicity correlation method of assessing similarity in protein structures is presented here as an alternative to the methods based on amino acid

identity (reviewed by Doolittle, 1981), or substitution frequency (Dayhoff *et al.*, 1979; McLachlan, 1971). Compared to methods relying only on amino acid identities, the present method is probably more useful for establishing structural similarities between very distantly related or unrelated proteins. Examples in this paper of distant relationships detected by hydrophobicity correlation are the two domains of rhodanese, and the similarities of melittin to its synthetic analogue and to delta hemolysin. Compared to the substitution frequency method of McLachlan, the present method uses a one-dimensional property, the hydrophobicity, to establish a relationship, whereas McLachlan's method uses a two-dimensional substitution matrix. The tests on globins described above indicate that the two methods yield similar measures of the relationship between pairs of proteins.

However, the hydrophobicity correlation method may yield additional biochemical information. For example, regions along the polypeptides of high average hydrophobicity may be membrane-penetrating sequences (Segrest & Feldman, 1977; Kyte & Doolittle, 1982), or regions of high hydrophobic moment may lie at the surfaces of membranes (Eisenberg *et al.*, 1982*a*) or at the surfaces of globular proteins (Eisenberg *et al.*, 1982*b*). These additional insights are evident immediately upon establishing the similarities of the two polypeptides. One such application of the present method of hydrophobicity correlation to achieving such insights has already been made with several chains of acetylcholine receptor (Stroud, personal communication).

For hydrophobicity correlation to become a stronger predictive tool, it will be necessary to add an alignment algorithm to the method, including a penalty for introducing gaps in each sequence during alignment. Also there may be various optimal "hydrophobicity" scales for different applications. For example, the OMH scale provides a useful indication of similarity among globular proteins while scales based on experimental hydrophobicity are useful for the melittin-like peptides (Table 2). Conceivably the method can be extended to recognition of supersecondary structures in proteins by their patterns of hydrophobicity. The method may also be effective for the design of synthetic peptides that mimic parts of known structures.

We thank Drs T. C. Terwilliger, J. A. Lake, C. Paul and Z. F. Burton for discussions and the National Science Foundation and National Institutes of Health (GM27076, GM16925) for support.

REFERENCES

Barker, W. C. & Dayhoff, M. Q. (1972). *Atlas of Protein Sequence and Structure,* vol. 5, pp. 101–110, National Biomedical Research Foundation, Washington, D.C.

Chou, P. Y. & Fasman, G. D. (1978). *Annu. Rev. Biochem.* **47**, 251–276.

Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. (1979). In *Atlas of Protein Sequence and Structure 1978,* vol. 5, suppl. 3, pp. 345–352, National Biomedical Research Foundation, Silver Spring, Md.

Dickerson, R. E. (1980). *Sci. Amer.* **242**, 136–153.

Dickerson, R. E. & Geis, I. (1982). *Hemoglobin,* Benjamin/Cummings, Menlo Park, Ca.

DeGrado, W. F., Itezdy, F. J. & Kaiser, E. T. (1981). *J. Amer. Chem. Soc.* **103**, 679–681.

Doolittle, R. F. (1981). *Science,* **214**, 149–159.
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.
Fisher, R. G., Woods, N. E., Fuchs, H. E. & Sweet, R. M. (1980). *J. Biol. Chem.* **255**, 5082–5089.
Fitch, W. W. (1966). *J. Mol. Biol.* **16**, 9–16.
Fitton, J. E., Del, A. & Shaw, W. V. (1980). *FEBS Letters,* **115**, 209–212.
Janin, J. (1979). *Nature (London),* **277**, 491–492.
Jue, R. A., Woodbury, N. W. & Doolittle, R. F. (1980). *J. Mol. Evol.* **15**, 129–148.
Kuntz, I. D. (1972). *J. Amer. Chem. Soc.* **94**, 4009–4012.
Kyte, J. & Doolittle, R. F. (1982). *J. Mol. Biol.* **157**, 105–132.
Lesk & Chothia (1980). *J. Mol. Biol.* **136**, 225–270.
Margoliash, E. (1963). *Proc. Nat. Acad. Sci., U.S.A.* **50**, 672–679.
McLachlan, A. D. (1971). *J. Mol. Biol.* **61**, 409–424.
McLachlan, A. D. (1972). *J. Mol. Biol.* **64**, 417–437.
Ploegman, J. H., Drenth, G., Kalk, K. H. & Hol, W. G. J. (1978). *J. Mol. Biol.* **123**, 557–594.
Perutz, M. F., Kendrew, J. C. & Watson, N. C. (1965). *J. Mol. Biol.* **13**, 669–678.
Richardson, J. S. (1981). *Advan. Protein Chem.* **34**, 167–339.
Rose, G. D. (1978). *Nature (London),* **272**, 586–590.
Schiffer, M. & Edmundson, A. B. (1967). *Biophys. J.* **7**, 121–135.
Segrest, J. P. & Feldman, R. J. (1977). *Biopolymers,* **16**, 2053–2065.
Sidler, W., Fuglistaller, P., Gysi, J., Isker, E. & Zuber, H. (1981). In *Photosynthesis III. Structure and Molecular Organization of the Photosynthetic Apparatus* (Akoyunoglou, G., ed.), pp. 583–594, Balaban International Science Services, Philadelphia.
Sweet, R. M., Fuchs, H. E., Fisher, R. G. & Glazer, A. N. (1977). *J. Biol. Chem.* **252**, 8258–8260.
Terwilliger, T. C. & Eisenberg, D. (1982). *J. Biol. Chem.* **257**, 6010–6015, 6016–6022.
Wolfenden, R. V., Cullis, P. M. & Southgate, C. C. F. (1979). *Science,* **206**, 575–577.
Wolfenden, R., Andersson, L., Cullis, P. M. & Southgate, C. C. B. (1981). *Biochemistry,* **20**, 849–855.

Edited by M. Gellert