Analysis of the Code Relating Sequence to Conformation in Globular Proteins

DEVELOPMENT OF A STEREOCHEMICAL ALPHABET ON THE BASIS OF INTRA-RESIDUE INFORMATION

By BARRY ROBSON and ROGER H. PAIN

*Department of Biochemistry, University of Newcastle upon Tyne,
Newcastle upon Tyne NE1 7RU, U.K.*

(Received 6 August 1973)

1. The relation of primary sequence to all residue backbone conformations was explored to test out starting conformations for protein folding. 2. Information theory was used to obtain measures of information which quantitate the role of each residue in determining its own conformation; i.e. intra-residue information. 3. The information measures are plotted as a function of ϕ , ψ peptide-backbone angles and ϕ , ψ contour maps obtained for each of the 20 amino acids. These show characteristic differences between residues. 4. To find practical ways of relating sequence to ϕ , ψ angles, several types of stereochemical alphabet were investigated. The value of these was tested by using them to predict the ϕ , ψ angles of nine different proteins. 5. A difference plot was constructed to show regions of the sequence that require little or no information extra to the intra-residue information in order to predict a correct conformation. These regions are suggested to be candidates for nucleating sites in the protein.

The prediction of the native structure of an entire protein molecule would involve the use of a computer to find the conformation of lowest free energy that is kinetically accessible to the protein plus solvent system within the lifetime of the protein molecule as a biologically active entity. This requires the search of a theoretically calculated free-energy surface which is large, multi-dimensional and occupied by many minima (Scheraga, 1971). Any means of limiting the search or of directing the search would clearly be essential. This problem is being approached (Robson & Pain, 1973) by attempting to carry out the energy minimization under the control of a heuristic computer program. In this context, a heuristic program is one that makes use of information other than that from the theoretical free-energy surface itself to make the search more efficient. The most readily accessible information for this purpose is that concerning the conformation of amino acid residues arising from the local interactions of the side chain and associated polypeptide backbone. Studies of these interactions can lead to the discovery of conformations of such high energy that they may be eliminated from the search and, further, of local energy minima some of which may approximate to the local conformation in the native protein. This latter has been expressed in the concept of a stereochemical alphabet (Liquori, 1969). In the present context, each letter of a stereochemical alphabet is identified with the domain of ϕ , ψ space* defined by a chosen range of angles. It would be an added advantage to be able to place

these energy minima in order, in a way such that the most likely letters or conformations are tried first by the computer in the overall energy-minimization procedure.

The setting up and weighting of a stereochemical alphabet may be approached either by calculating energy functions for specific local conformations or by observation of known conformations and relating them to the amino acid sequence. For the present purpose of providing heuristic information the latter approach was chosen and, specifically, the type of analysis involving the use of information theory. This analysis has significant advantages in terms of the way in which the available observations are weighted and assessed and of the potential that it offers for analysing into its component contributions the information contained by a particular sequence concerning its conformation.

Information theory has already proved to be a useful tool in the search for the factors responsible for the formation of the right-hand α -helix conformation in globular proteins. The validity of the various measures obtained is shown by their ability to account accurately for observed helical regions (Pain & Robson, 1970; Robson & Pain, 1971), or for the results of theoretical energy calculations (Pain & Robson, 1970), and for the distribution of different types of side chain within both helical and non-helical regions (Robson & Pain, 1972). There is now a sufficient number of proteins whose sequence and three-dimensional conformation have been determined to allow the application of information theory to be extended to other backbone conformations. It is possible to discover preferred conforma-

*The angle ϕ denotes rotation about the N-C α bond, and the angle ψ that around the C α -C bond.

tions and to assign them information measures in much the same way as energy estimates can be assigned to the conformational energy minima of the residue as calculated by using the C_α -CO-NH-CHR-CO-NH- C_α unit; hence conformations can be ordered so that the most likely are tried as a first resort.

The information approach also has advantages over the theoretical energy calculation approach, in that it is an empirical approach in which no assumptions are made concerning the physical nature, magnitude, or origin of the interactions involved. Assignments are made concerning the way in which the units behave within a protein environment. If there are any consistent interactions between units, they will not be overlooked by this approach and it may be possible to identify these constant factors by comparing the results of theoretical energy and information-theory approaches. The measures obtained are essentially teleological and take advantage of information available in the X-ray data concerning the final solution of the minimization problem; the native conformation of a protein molecule. Finally, the measures obtained are in a convenient form for statistical decision-theory processes which would direct the trajectory of the search of the energy surface.

It is intended, by using the information-theory approach, to study the dependence of protein conformation on side chain-own backbone interactions, side chain-neighbouring backbone interactions and side chain-side chain interactions. The present paper is concerned with the first of these types of interaction. The purpose is to assess the validity of the general notion of a stereochemical alphabet, to explore its properties, to assess the relative importance of the information carried within residues and that carried by interaction between residues, and finally to obtain measures which will be used to predict starting conformations for the search of the free-energy surface.

It is shown here that measures obtained from side chain-own backbone interactions alone enable sufficiently good orderings to be made to justify the expectation that measures from all local interactions will be able significantly to decrease the extent of the search.

Wu & Kabat (1973) have recently made predictions of the actual angles of a protein backbone by reference to the influence of residues R_{j-1} and R_{j+1} on the native conformation of R_j . They made use of residue triplets, but were not concerned with quantitative evaluation of the contributions of single residues or pairs of residues.

Theory

Let S_j be the conformation of the j th residue in the primary sequence, and R_{j+m} the $(j+m)$ th residue in the primary sequence; then the information trans-

ferred from R_{j+m} to S_j is described by $I(S_j; R_{j+m})$. There is no need to consider the numerical value of j since only the separation m is of interest. Since in this paper we are interested only in the interactions between a side chain and the attached segment of backbone and the effect such interactions have on the conformation of the attached segment of backbone, m is taken as zero.

For the sake of simplicity, but with no loss of generality, it is desirable to study the information transferred as if residue conformations were binary events (i.e. specified as the occurrence or non-occurrence of a defined conformation). Neglecting distortions of the NH- C_α -CO angle and of the planar and *trans* peptide bond (Ramachandran & Sasisekharan, 1968; Robson, 1972), then $S_j = X$ is a joint event in that X is defined by the bond rotation angles ϕ and ψ which lie in the domain defined by the limits $\phi_{\min.}$ and $\phi_{\max.}$, $\psi_{\min.}$ and $\psi_{\max.}$ respectively. If either ϕ or ψ lies outside the domain limits, the conformation is designated as $S_j = \bar{X}$. Thus, we are interested in the binary outcome $S_j = X:\bar{X}$ and in the information $I(S_j = X:\bar{X}; R_j)$, subsequently called the intra-residue information about R_j .

This intra-residue information for a binary outcome is expanded as follows:

$$I(S_j = X:\bar{X}; R_j) = *(S_j = X:\bar{X}; R_j) - *(S_j = X:\bar{X}) \quad (1)$$

$*(S_j = X:\bar{X}; R_j)$ represents the relative abundance of conformation X for each different type of residue and $*(S_j = X:\bar{X})$ the distribution of X irrespective of residue type.

The 'star' functions $*$ () are derived from the frequencies of the events $S_j = X$, $S_j = \bar{X}$ and R_j . They are determined by analysis of primary sequence and X-ray data (see the Methods section). Both of the star functions in eqn. (1) are of interest. A plot of $*(S_j = X:\bar{X})$ versus X would resemble the empirical energy map of Pohl (1971). This statistic, however, says nothing about the involvement of different side chains which determine the characteristic conformation of a residue. The effects of the different side chains R may be investigated by the statistic $*(S_j = X:\bar{X}, R_j)$. Both these functions $*$ () are analogous to the natural logarithm of a 'K-statistic' (Jeffreys, 1948), and may be recognized as a measure of the preference for conformation X as opposed to $\bar{X}$. More exactly, we can deal directly only with statistical estimates of functions $*$ () (Robson & Pain, 1971) which represent the available information about the preference for conformation X as opposed to $\bar{X}$ (Robson, 1974).

Methods

The evaluation of the terms of eqn. (1) and all associated studies were carried out by using pro-

cedures described in the preceding paper (Robson, 1974).

Information-theory analysis

Conformational data were taken from the following proteins: hen egg-white lysozyme; carboxypeptidase A (Quiocho & Lipscomb, 1971); *Staphylococcal* nuclease; sperm whale myoglobin (Watson, 1969); chymotrypsin; ribonuclease; elastase; trypsin inhibitor protein; cytochrome *c* (in Wu & Kabat, 1973) (see also the acknowledgments).

To our knowledge none of the conformational data used in this work has been refined with the use of energy functions; it is therefore wholly empirical. Most of the data was, however, processed by its authors by using the refinement programme of Diamond (1971), or some development of it. The frequencies of events ($S_j = X, R_j$) and ($S_j = \bar{X}, R_j$) were counted in this data, except that trypsin inhibitor was excluded so that its conformation could be predicted by using only information from other proteins.

To obtain estimates of functions $^*(S_j = X: \bar{X})$ and $^*(S_j = X: \bar{X}; R_j)$ it is first necessary to define the conformation X . This involves defining a set of non-overlapping domains on the ϕ - ψ surface. The simplest definition is based on the division of the ϕ - ψ diagram into square cells. Initially 20° square cells, i.e. 324 cells, are used to examine the distribution of conformations. However, the observed distribution of conformations may then suggest the use of non-square and even non-rectangular domains. For example, the cells enclosing a dense group of conformations as revealed by a region of high information may be pooled if it is assumed that the important feature of a conformation is that it belongs to that group as a whole and not to any particular position in the group. The practical advantage of such a grouping is that it saves computer memory when information measures are being derived and applied to the problem of choosing a suitable starting conformation. The saving may be trivial when intra-residue information only is being used, but for pairs, triplets, and higher combinations of residues it will be considerable.

Information measures are obtained by counting the number of conformations in each domain, this being done for the 20 kinds of amino acid residue in the case of intra-residue information for the 20×20 kinds of amino acid pair in the case of pair information (Robson & Pain, 1971) and so on. In the case of the 20° cells initially used, 18×18 different conformations were assigned information measures thus generating $18 \times 18 \times 20$ measures overall for the intra-residue information.

In estimating the 'star' functions $^*()$ from observed frequencies of residues and conformations

the procedure of Robson (1974) was used; in particular, eqns. (47a) and (47b) of that paper. These equations are then estimated as:

$$E[^*(S_j = X: \bar{X}; R_j) | D] = \# [f(X, R_j)] - \# [f(\bar{X}, R_j)] \quad (2a)$$

and:

$$E[^*(S_j = X: \bar{X}) | D] = \# [f(X)] - \# [f(\bar{X})] \quad (2b)$$

with $\#(0) = 1$, $\#(1) = 1$, $\#(2) = 1$, $\#(3) = 1 + 1/2$, $\#(4) = 1 + 1/2 + 1/3, \dots$, $\#(n) = 1 + 1/2 + 1/3 + \dots + 1/(n-1)$.

$E[^*(S_j = X: \bar{X}; R_j) | D]$ is a function of the relative abundance of conformation X for each different type of residue conditional upon, i.e. found in, the data pool D . $f(X, R_j)$ is the numerical frequency of occurrence of residue R_j in conformation X in that pool; $f(X)$ is similarly the frequency of occurrence of conformation X . $\#(n)$ is an information number, representing the information contained in n observations in each of which a particular conformation S_j has been observed.

Results

The term $^*(S_j = X: \bar{X})$

The term $^*(S_j = X: \bar{X})$ of eqn. (1) is enumerated for different values of ϕ and ψ in Fig. (1). Minimally stressed hand-sketched contours are fitted to the Figure to emphasize salient features. The outermost

Fig. 1. Estimates of the function $^*(S_j = X: \bar{X})$ calculated by using data from the nine proteins excluding trypsin inhibitor protein and plotted as a function of angles ϕ and ψ

Contours are at intervals of 1 nat above floor level.

contour encloses values more than 1.0nat (the information measures observed as a natural logarithm) above the global minimum. Three principal information maxima are obtained. The highest (-2.3 nats) occurs at $\phi = -50^\circ \pm 10$, $\psi = -50^\circ \pm 10$, which is in the region of the right-hand α -helix conformation and is subsequently referred to as the H maximum. The second highest (-3.4 nats) occurs at $\phi = -90^\circ \pm 10$, $\psi = 130^\circ \pm 10$, which is in the region of the pleated-sheet conformations and is referred to as the P maximum. The third highest (-4.6 nats) occurs at $\phi = 70^\circ \pm 10$, $\psi = 30^\circ \pm 10$, the region of the left-hand α -helical conformation, and is referred to as the L maximum. Although no conformation is so favoured as to have a positive measure, most of the observed conformations occur close to these maxima. This suggests that a stereochemical alphabet of preferred conformations on the basis of stereochemical symbols H, P and L is a reasonable proposition for the prediction of the backbone conformation of individual residues as a preliminary to energy minimization.

The term $^*(S_j = X:\bar{X}; R_j)$

The possibility that a stereochemical alphabet may be dependent on the type of side chain has not previously been investigated by using empirical data from proteins of known conformation. To explore this possibility, plots of $^*(S_j = X:\bar{X}, R_j)$ were

generated for the 20 amino acids. Five representative plots (for alanine, valine, glycine, lysine and threonine) are shown in Figs. 2–6. All 20 plots are summarized by Table 1 and Fig. 7, which record all maxima higher than 0.3nat above the global minimum of each plot, this being the lowest value for which a significant structural response is expected (Robson & Pain, 1972). The positions of the maxima vary widely, although of course always remaining within the outermost contour of Fig. 1.

Several features of these plots are worthy of comment. Alanine, as expected, shows preference for the classical α_1 conformation coupled, however, with a drive toward the α_{II} conformation (Nemethy *et al.*, 1967; Robson & Pain, 1971; Pain & Robson, 1971). This contrasts with valine which, surprisingly, strongly prefers the α_1 conformation, showing no tendency for the α_{II} conformation, but prefers the pleated-sheet conformation much more than alanine. Threonine prefers the α_{II} conformation to α_1 and, overall, prefers the pleated sheet. Glycine is notable in preferring the left-hand α -helix conformation.

Clearly, a more accurate description of the preferred conformation of a residue requires consideration of the type of side chain. Although an HPL alphabet can be used to define the preferred conformation of a residue within $\pm 70^\circ$, an alphabet dependent on side-chain type (H_{Ala} , P_{Val} , etc.) would be required to define the preferred conformation within $\pm 10^\circ$.

Fig. 2. Estimates of $^*(S_j = X:\bar{X}, R_j)$ for alanine calculated by using data from the nine proteins excluding trypsin inhibitor protein and plotted as a function of angles ϕ and ψ .

Contours are at intervals of 1 nat above floor level.

Fig. 3. Values of $^*(S_j = X:\bar{X}, R_j)$ for valine as a function of angles ϕ and ψ .

Details are as for Fig. 2.

Fig. 4. Values of $*(S_j = X: \bar{X}; R_j)$ for glycine as a function of angles ϕ and ψ

Details are as for Fig. 2.

Fig. 6. Values of $*(S_j = X: \bar{X}; R_j)$ for threonine as a function of angles ϕ and ψ

Details are as for Fig. 2.

Fig. 5. Values of $*(S_j = X: \bar{X}; R_j)$ for lysine as a function of angles ϕ and ψ

Details are as for Fig. 2.

When the positions of the $*(S_j = X: \bar{X}; R_j)$ maxima are collated as in Fig. 7 it is seen that they form clusters. This grouping of $*(S_j = X: \bar{X}; R_j)$ maxima suggests an alternative alphabet in which a stereochemical letter can be assigned to each cluster. This

has the double advantage of making it independent of the type of amino acid side chain and thereby decreasing the size of the alphabet to be used. These clusters lie in the regions of the H, P and L maxima of the $*(S_j = X: \bar{X})$ plot and generally more than one such cluster is associated with the H, P and L maxima. This suggests a notation H_1, H_2, P_1, P_2 etc., in which the letter indicates to which $*(S_j = X: \bar{X})$ maximum the $*(S_j = X: \bar{X}; R_j)$ cluster belongs. It is useful to refer to these clusters as representing the fine structure of the $*(S_j = X: \bar{X})$ maxima though this does not mean that any such fine structure can be inferred from the $*(S_j = X: \bar{X})$ plot alone. This is the basis of a fine-structure HPL alphabet.

Table 1 shows that each type of residue, R_j , frequently contributes a $*(S_j = X: \bar{X}; R_j)$ maximum to more than one cluster of maxima notated with the same alphabetic character (e.g. to H_1, H_2 and H_3), sometimes to only one cluster with that alphabetic character and occasionally two or more maxima to the same cluster. However, the clusters are sufficiently well defined and narrow in distribution to merit exploring their usefulness in the practical application of stereochemical alphabets.

Tests of the practical value of different stereochemical alphabets

The practical value of a stereochemical alphabet in determining a starting conformation can be assessed by measuring the agreement between the predicted

Table 1. Preferred conformations of amino acid residues

Information values were obtained from nine proteins.

Residue	ϕ	ψ	$*(S_j = X:\bar{X}, R_j)$	Residue	ϕ	ψ	$*(S_j = X:X, R_j)$
Gly (floor level, -53)	+ 70	+ 30	-26	Asn (floor level, -49)	- 50	- 50	-28
	- 90	+170	-29		- 90	+ 50	-30
	+ 90	- 30	-31		- 90	+130	-33
	- 50	- 50	-34		+ 70	+ 30	-34
	-170	+170	-38	Gln (floor level, -45)	-130	+150	-21
	- 70	- 10	-43		- 70	- 30	-26
	+ 70	-150	-43		- 60	+130	-29
	+ 90	+150	-43	Lys (floor level, -52)	- 60	- 50	-27
Ala (floor level, -53)	- 50	- 50	-15		-110	+130	-29
	-150	+150	-34		-150	+170	-42
	- 90	+140	-34		- 70	-130	-42
	-110	+ 10	-38		- 10	+ 90	-42
	-180	-170	-43	His (floor level, -42)	- 50	- 50	-18
	-130	+110	-43		-110	- 30	-27
	- 50	+130	-43		-110	+ 30	-32
Val (floor level, -51)	- 50	- 50	-19		- 70	+130	-32
	-110	+150	-24	Arg (floor level, -45)	- 50	- 50	-21
	-170	+170	-41		- 90	+120	-23
Leu (floor level, -52)	- 50	- 50	-22		-130	+150	-29
	- 90	+140	-30	Phe (floor level, -43)	- 60	- 40	-27
	- 90	+ 70	-41		-110	+130	-27
Ile (floor level, -48)	-130	+130	-22		- 50	+130	-33
	- 50	- 50	-25	Tyr (floor level, -45)	-130	+150	-26
Ser (floor level, -53)	- 70	+150	-26		- 50	- 50	-30
	- 50	- 50	-29		- 90	- 10	-30
	-110	- 10	-34		-130	-10	-35
	-130	+130	-37		- 50	+130	-35
	+150	- 10	-42	Trp (floor level, -40)	-110	- 10	-20
Thr (floor level, -52)	-130	+150	-29		- 50	- 70	-24
	- 90	- 30	-31		-110	+150	-29
	-110	+ 10	-33	Cys (floor level, -41)	- 70	- 50	-25
	- 70	+170	-37		-130	+130	-25
Asp (floor level, -46)	- 50	- 50	-24		- 70	+130	-25
	-110	+130	-30		- 80	- 10	-31
	- 90	- 30	-30		-150	+170	-31
	- 90	+ 30	-35	Met (floor level, -35)	- 50	- 50	-12
	- 80	+ 80	-35		-130	+170	-24
	- 70	+170	-35	Pro (floor level, -43)	- 70	+150	-16
Glu (floor level, -48)	- 50	- 50	-13		- 50	+110	-28
	- 70	+150	-29		- 30	- 50	-28
	-110	+150	-37		- 80	0	-33
	-110	+110	-37				
	- 90	- 10	-37				

and observed conformations for an amino acid sequence. The conformation of a residue predicted on the basis of intra-residue information is the domain with the highest valued measure $*(S_j = X:\bar{X}, R_j)$. The initial studies in this paper used a 20° square cell, but the localized distribution of information in these 18×18 cells suggests that a three-letter HPL alphabet could be just as effective. The advantage of the HPL alphabet is that it requires $3/(18 \times 18)$ times less computer memory. The disadvantage is that it is

much less precise and only allows indication of a very broad region to which a suitable starting conformation is likely to belong. A possible compromise is to use the fine-structure HPL alphabet which still takes advantage of a natural distribution of conformations but affords greater predictive accuracy. Alternatively, one could make a simpler compromise between computer storage and precision by using an alphabet of 40° square cells which has 9×9 letters.

Fig. 7. Configuration of sub-domains comprising the fine-structure HPL alphabet

Numbers indicate the number of information maxima which occur in each 20° cell (see the text).

Development of a fine-structure HPL alphabet

An example of domain definition is shown in Fig. 7. To compare the merits of the HPL alphabet and the fine-structure HPL alphabet, it is convenient to assign as sub-domains H_1 , H_2 , P_1 etc. and to define domains H, P and L by the union of sub-domains with the same prefix. Hence sub-domains H_1 and H_2 together make up domain H. In specifying the areas of the ϕ - ψ diagram for the fine-structure HPL alphabet it is clear that the placing of boundaries is to some extent rather arbitrary and any specification can only be justified by the accuracy of predicted starting conformations. We have attempted to optimize the specification of sub-domains with the help of an interactive programme and visual display screen. The effect was observed of altering sub-domain size and shape on the fraction of residues correctly predicted in trypsin inhibitor protein. Initially, sub-domains were specified to cover the ϕ - ψ diagram. The alphabet chosen could then be used to describe not only the preferred but any possible conformation including the observed conformation, a property already possessed by the 20° and 40° square-cell alphabets.

In attempting to optimize the configuration of sub-domains the constraints were imposed that each sub-domain should enclose each well-defined cluster of $^*(S_j = X:X; R_j)$ maxima, that boundaries between sub-domains should correspond closely

to saddle points in the distribution of these maxima and that boundaries should be composed of as long vertical and horizontal lines as possible. H, P and L regions could be resolved into seven sub-domains, H_1 , H_2 , P_1 , P_2 , P_3 , L_1 and L_2 . The suffix in this notation is lower, the higher the population density of the sub-domain (population of sub-domain/size of sub-domain). Thus sub-domain H_1 has a higher population density than H_2 . There is justification for defining a sub-domain to enclose the extended-chain conformation ($\phi = +180^\circ = -180^\circ$, $\psi = +180^\circ = -180^\circ$), but this region was included in the P_2 sub-domain because its population was low and because the population of conformations between the extended-chain conformation and the P_2 cluster is high. Before optimization the boundaries of the above sub-domains not in contact with another sub-domain were positioned so as to enclose approx. 90% of the tail from the $^*(S_j = X:X)$ maxima, the exact position being chosen so that the boundaries were conveniently vertical or horizontal lines on the ϕ - ψ surface. The regions not yet belonging to any defined sub-domain were called the G domain because they are populated principally by glycine residues, and this was split into two sub-domains G_1 and G_2 . Proline was treated as a special case because the covalent link between the side chain and backbone constrains the backbone geometry to two small regions of the ϕ - ψ diagram. Its conformation is defined as Pro^+ ($\psi \geq 0^\circ$) or Pro^- ($\psi < 0^\circ$).

Since the sub-domains are by definition completely surrounded and non-overlapping on the ϕ - ψ surface (which is mathematically the surface of a sphere), there are sufficient overall constraints to make a reasonably comprehensive evaluation of the possible sub-domain configurations achievable by using a partially interactive rather than completely automatic optimization programme. The optimization was further helped by the possibility of breaking down the overall problem of optimization into sub-problems. For example, the boundaries separating P_1 , P_2 and P_3 were adjusted so that the prediction of P sub-domains was improved without reference to other domains. This was also done with the H sub-domains and, in the absence of adequate information for the L region, boundaries for the latter were defined so as to preserve symmetry with the H domain. When it was necessary to handle more than two variables at a time, such as in the case of the position of the lower boundary of the P_1 and P_2 domains and of the boundary between them, subjective conclusions were checked against results obtained by positioning domain boundaries to give optimal predictions, by using an adaptation of Ghani's optimization procedure CMPX (Ghani, 1972). More extensive use of CMPX procedure was limited by the time required to carry out a prediction for each sub-domain conformation.

Comparison of alphabets

The best configuration found for sub-domains on the basis of the fine-structure HPL code is that shown in Fig. 7. The information measures for the sub-domains in this configuration are given in Table 2 and the prediction based on these measures is shown in Table 3 (column 3), together with an indication of the approximation to success of this prediction [columns (4) and (5)]. It may be compared with predictions based on the 20° and 40° square cells (Table 4). For the 40° square-cell alphabet, a residue was considered as correctly predicted if the observed conformation belonged to the predicted cell or to one of the cells adjacent to, and in contact with, it.

For the 20° alphabet a comparable criterion of success was chosen and success was claimed if the angle distance on the ϕ, ψ plot was $\leq 68^\circ$. Adoption of an angle distance of 60° as the criterion of success and subsequent correction of the degree of success on an area basis yielded essentially the same results.

A summary of the degree of success of predictions made by using the 40° cell alphabet for a total of nine proteins is given in Table 5, showing an overall 43% residues correctly predicted solely on the basis of intra-residue information. Also included in Table 3 (column 6) is the prediction that results when the preferred sub-domain is chosen, not on the basis of the measures in Table 2, but on the basis of those in Table 1, where the measure assigned to each sub-domain is the highest $\ast(S_j = X:\bar{X}, R_j)$ measure of all the 20° cells enclosed by that sub-domain. Columns (7) and (8) of Table 3 indicate the approxi-

mation to success of the predictions for sub-domains and domains.

In assigning starting conformations no attempt was made to account for co-operativity between residues. For example, it is possible that the cysteine residue 51 is pulled into the helical conformation by the five other residues around it which favour the helical conformation, particularly since the H_1 conformation is the second choice for cysteine. Co-operativity is essentially an effect arising between residues and should not be used in a prediction intended to study the effect that a residue has on its own conformation. However, it would enhance the predictive powers of the measures available and it should be possible to produce a generalized formulation of the qualitative procedure used earlier (Robson & Pain, 1971) for taking into account co-operativity in helix formation.

Role of single residue information as opposed to information arising from interactions between residues

It is well established that interactions between residues play a part in determining backbone conformation (Periti *et al.*, 1967; Schiffer & Edmundson, 1967; Robson & Pain, 1971). It is possible, however, that the relative importance of the information in single residues and that in interactions between residues shows a regional distribution along the primary sequence. This would be consistent with one view of nucleation in which the information in single residues is responsible for the formation of a

Table 2. Information measures for fine-structure HPL sub-domains

Residue	Sub-domain	...	H ₁	H ₂	P ₁	P ₂	P ₃	L ₁	L ₂	G ₁	G ₂
Gly			-20	-25	-21	-17	-43	-21	-9	-38	-21
Ala			-2	-17	-18	-13	-38	-53	-43	-53	-53
Val			-8	-21	-16	-4	-51	-51	-41	-51	-51
Leu			-7	-16	-12	-12	-36	-51	-52	-41	-52
Ile			-6	-15	-20	-7	-47	-48	-47	-47	-48
Ser			-10	-13	-10	-15	-34	-42	-42	-37	-53
Thr			-12	-11	-13	-10	-37	-52	-42	-52	-52
Asp			-7	-14	-24	-15	-17	-45	-46	-46	-35
Glu			0	-19	-20	-17	-32	-48	-47	-47	-48
Asn			-14	-14	-16	-21	-16	-21	-34	-39	-49
Gln			-9	-16	-12	-8	-45	-35	-45	-45	-45
Lys			-3	-18	-17	-16	-33	-33	-42	-52	-52
His			-6	-13	-16	-23	-27	-42	-42	-32	-32
Arg			-9	-14	-10	-13	-45	-44	-44	-45	-45
Phe			-7	-21	-15	-10	-33	-43	-43	-33	-43
Tyr			-15	-15	-18	-6	-30	-45	-30	-35	-45
Trp			-8	-8	-20	-17	-24	-40	-40	-40	-40
Cys			-13	-15	-22	-2	-41	-41	-41	-41	-41
Met			-3	-24	-15	-9	-35	-35	-35	-35	-35
Pro			-10	-17	0	-43	-43	-43	-43	-43	-43
Average*			-8	-16	-15	-12	-31	-36	-32	-41	-43

*Information from all residues without regard to side-chain type.

nucleating region around which a protein folds (see, for example, Lewis *et al.*, 1970). The possibility was explored by the construction of a smoothed difference plot measuring the difference between $*(S_j = X_p; \bar{X}_p, R_j)$ for the predicted conformations, X_p , and $*(S_j = X_o; \bar{X}_o, R_j)$ for the observed conformations, X_o , for each residue j in trypsin inhibitor protein. This corresponds to the minimal amount of information required from sources other than the residue itself, which will produce the observed conformation. Such a plot is shown in Fig. 8 which is smoothed by a factor of 5 (i.e. each point is an average of five contiguous residues). The regions requiring least information from other sources are the α -helical structure 48–54, followed by runs of residues in pleated-sheet structures and in pleated-sheet structure distorted in the direction of the extended chain conformation. These structures are therefore potential candidates for the role of an initial nucleating device.

Discussion

Intra-residue information

The information concerning whether the conformation of residue number j is X or not X can be written as $I(S_j = X; \bar{X}; R_1, R_2 \dots R_L)$ in a protein of length L residues. This can be expanded to represent the sum of contributions from various sources (cf. Robson & Pain, 1973) as follows:

$$\begin{aligned} I(S_j = X; \bar{X}; R_1, R_2 \dots R_L) \\ = I(S_j = X; \bar{X}, R_j) + \sum_{m \neq 0} I(S_j = X; \bar{X}; R_{j+m} | R_j) \\ + \sum_{m \neq 0, n \neq 0} I(S_j = X; \bar{X}; R_{j+n} | R_j, R_{j+m}) \\ + \dots, -M \leq m \leq +M, -M \leq n \leq +M, \dots \quad (3) \end{aligned}$$

The right-hand-side terms represent the information concerning whether the conformation S_j is X or $\bar{X}$; first, independent of residue side chains; secondly, dependent on residues taken in pairs; thirdly, on residues taken in triplets, and so on.

The term $I(S_j = X; \bar{X}, R_j)$ represents the intra-residue information, which is a function only of the backbone torsion angles of residue R_j . This does not imply that the interactions between atoms that are responsible for this information occur only between atoms confined to the residue structure NH-CHR-CO (where R is the side-chain type determining the residue). On the contrary, assuming the peptide unit to be planar and *trans* and the dielectric constant to be invariant, the larger unit C_α -CO-NH-CHR-CO-NH- C_α is the group of atoms whose interactions are dependent on the rotation angles of a single residue. These interactions will therefore all make a constant contribution to $I(S_j = X; \bar{X}, R_j)$. Averaged interactions can also arise between atoms inside and outside C_α -CO-NH-CHR-CO-NH- C_α , although as

atoms further away from the unit are considered there are so many intervening torsion angles that a very large number of interactions are possible within different proteins and different regions of the same protein. In this case the constant contribution to $I(S_j = X; \bar{X}, R_j)$ becomes zero. Hence, the intra-residue information concerns a variety of interactions which arise essentially between atoms within the C_α -CO-NH-CHR-CO-NH- C_α unit and is identical with 'single residue information' as used previously (Robson & Pain, 1971). The term $*(S_j = X; \bar{X})$ of eqn. (1), on the other hand, arises from those interactions that are independent of the type of side chain and these are essentially interactions between backbone atoms. Although plots for $*(S_j = X; \bar{X}; R_j)$ and $*(S_j = X; \bar{X})$ are not identical, their overall similarity suggests the importance of backbone interactions. The $*(S_j = X; \bar{X})$ plot of Fig. 1 corresponds broadly to the analogous plot of Pohl (1971). The additional amount of refined data used in producing the $*(S_j = X; \bar{X})$ plot results in greater detail, notably, the narrowness of the neck linking the H and P maxima.

The information plot for all residues regardless of side chain (Fig. 1) is similar to the energy plot for alanine (Scott & Scheraga, 1966). This confirms the importance of interactions between the β -carbon atom and the attached segment of backbone. However, despite the importance of van der Waals repulsive forces in such interactions, the information plot reflects the occurrence of conformations with fairly severe atomic overlaps in terms of the classical Pauling-Corey van der Waals radii (Pauling & Corey, 1951). Thus softer repulsive potentials or distortion of the α -carbon valence angle or peptide-bond angle, ω , must be involved. More precisely, conformations are allowed which are not confined within the Ramachandran outer limit (Ramachandran *et al.*, 1963), but which represent atomic contact distances as short as 70% of the sum of the minimal van der Waals radii (Leach *et al.*, 1966). This includes conformations that classically represent a clash between β -carbon atom and backbone. Two other discrepancies between information and energy plots are also apparent from this analysis. The first is differences between the observed population and the populations expected on the basis of theoretical energy calculations. Pohl (1971) has suggested that the relative abundance around $\phi = -90^\circ$, $\psi = -60^\circ$ is due to excluded volume effects dependent on interactions with atoms of other residues, owing principally to van der Waals repulsion forces. These interactions occur primarily as a consequence of the backbone chain tending to take up conformations which are variants of right-hand α -helices.

A further difference is the increased value of the information along the line $\psi = 0^\circ$ whenever there are no very close atomic overlaps. This reflects the

Table 3. *Predicted conformations of trypsin inhibitor protein on the basis of domain assignments*

Column (2): from W. Steigemann, personal communication. Columns (4) and (5): e.g. the observed sub-domain for residue 6, Leu, occurs fourth in descending order of information in Table 2. Column (3): information measures from Table 2. Column (6): information measures from Table 1. Columns (7) and (8): e.g. the observed sub-domain for residue 6, Leu, occurs fourth in descending order of information in Table 1.

Residue (1)	Observed sub-domain (2)	Sub-domain predicted from HPL sub-domains (Table 2) (3)	Correct sub-domain: choice number (4)	Correct domain: choice number (5)	Sub-domain predicted from 20° cell maxima (Table 1) (6)	Correct sub-domain: choice number (7)	Correct domain: choice number (8)
1 Arg	P ₁	H ₁	2	2	H ₁	2	2
2 Pro	Pro ⁺	Pro ⁺	1	1	Pro ⁺	1	1
3 Asp	H ₁	H ₁	1	1	H ₁	1	1
4 Phe	H ₂	H ₁	4	1	H ₁	4	1
5 Cys	H ₂	P ₂	3	2	H ₁	4	1
6 Leu	H ₂	H ₁	4	1	H ₁	4	1
7 Glu	P ₁	H ₁	4	2	H ₁	2	2
8 Pro	Pro ⁺	Pro ⁺	1	1	Pro ⁺	1	1
9 Pro	Pro ⁺	Pro ⁺	1	1	Pro ⁺	1	1
10 Tyr	P ₂	P ₂	1	1	P ₂	1	1
11 Thr	H ₁	P ₂	3	2	P ₂	5	2
12 Gly	G ₂	L ₂	4	4	L ₁	7	4
13 Pro	Pro ⁻	Pro ⁺	2	2	Pro ⁺	2	2
14 Cys	P ₁	P ₂	4	1	H ₁	3	2
15 Lys	P ₃	H ₁	5	2	H ₁	5	2
16 Ala	P ₁	H ₁	4	2	H ₁	3	2
17 Arg	P ₃	H ₁	6	2	H ₁	4	2
18 Ile	P ₂	H ₁	2	2	P ₂	1	1
19 Ile	P ₁	H ₁	4	2	P ₂	3	1
20 Arg	P ₂	H ₁	3	2	H ₁	3	2
21 Tyr	P ₂	P ₂	1	1	P ₂	1	1
22 Phe	P ₂	H ₁	2	2	H ₁	2	2
23 Tyr	P ₁	P ₂	4	1	P ₂	5	1
24 Asn	P ₂	H ₂ /H ₁	5	2	H ₁	5	2
25 Ala	H ₁	H ₁	1	1	H ₁	1	1
26 Lys	H ₁	H ₁	1	1	H ₁	1	1
27 Ala	H ₂	H ₁	3	1	H ₁	4	1
28 Gly	L ₂	L ₂	1	1	L ₁	3	1
29 Leu	P ₂	H ₁	2	2	H ₁	4	2
30 Cys	P ₁	P ₂	4	1	H ₁	3	2
31 Gln	P ₂	P ₂	1	1	P ₂	1	1
32 Thr	P ₁	P ₂	4	1	P ₂	4	1
33 Phe	P ₂	H ₁	2	2	H ₁	2	2
34 Val	P ₁	P ₂	3	1	H ₁	4	2
35 Tyr	P ₁	P ₂	4	1	P ₂	5	1
36 Gly	H ₂	L ₂	6	3	L ₁	6	3
37 Gly	L ₂	L ₂	1	1	L ₁	3	1
38 Cys	P ₂	P ₂	1	1	H ₁	2	2
39 Arg	L ₁	H ₁	5	3	H ₁	4	4
40 Ala	P ₁	H ₁	4	2	H ₁	3	2
41 Lys	P ₂	H ₁	2	2	H ₁	2	2
42 Arg	H ₂	H ₁	4	1	H ₁	4	1
43 Asn	P ₃	H ₂ /H ₁	3	2	H ₁	2	2
44 Asn	P ₂	H ₂ /H ₁	5	2	H ₁	5	2
45 Phe	P ₂	H ₁	2	2	H ₁	2	2
46 Lys	H ₂	H ₁	4	1	H ₁	6	1
47 Ser	P ₂	P ₁ /H ₁	4	1	P ₁	4	1
48 Ala	H ₁	H ₁	1	1	H ₁	1	1
49 Glu	H ₁	H ₁	1	1	H ₁	1	1
50 Asp	H ₁	H ₁	1	1	H ₁	1	1
51 Cys	H ₁	P ₂	2	2	H ₁	1	1

Table 3. *continued*

Residue (1)	Observed sub-domain (2)	Sub-domain predicted from HPL sub-domains (Table 2) (3)	Correct sub-domain: choice number (4)	Correct domain: choice number (5)	Sub-domain predicted from 20° cell maxima (Table 1) (6)	Correct sub-domain: choice number (7)	Correct domain: choice number (8)
52 Met	H ₁	H ₁	1	1	H ₁	1	1
53 Arg	H ₁	H ₁	1	1	H ₁	1	1
54 Thr	H ₁	P ₂	2	2	P ₂	5	2
55 Cys	H ₂	P ₂	3	2	H ₁	4	1
56 Gly	P ₁	L ₂	4	2	L ₁	2	2
57 Gly	G ₂	L ₂	4	4	L ₁	7	4
58 Ala	P ₂	H ₁	2	2	H ₁	2	2
Residues successfully predicted on first choice ...			31%	52%		28%	54%

possibility of 1–3 hydrogen bonding or amide to peptide-bond π -orbital hydrogen bonding in the vicinity of this line. In the region $\phi = -90^\circ$, $\psi = -20^\circ$, there is also an abundance of 1–(3, 4) bifurcate hydrogen bonding within helical regions. Indeed, the 1–4, 1–(3, 4), π and 1–3 hydrogen bonds represent a continuum of hydrogen-bonding arrangements between which there is no clear demarcation in the X-ray crystallographic data and which lead to increased population of the diagonal $\phi + \psi = 0^\circ \pm 30^\circ$.

It is possible to recognize physical analogues of the stereochemical alphabet based on the sub-domains of Fig. 7. Sub-domain H₁ corresponds to the classical α_1 -helix and H₂ to the α_{11} -helix discussed above. Sub-domain P₂ corresponds to parallel and anti-parallel pleated-sheet structures, whereas sub-domain P₁ is largely populated by markedly twisted pleated-sheet structures distorted in the direction of a right-hand screw (see e.g. Chothia, 1973; Fig. 6). Sub-domains, L₁ and L₂ represent conformations of helices analogous to those of the H₂ and H₁ regions respectively but of opposite sense. Thus the stereochemical alphabet relates closely to the conformations favoured by homopolymer systems. This is not because of an exclusive opportunity for hydrogen bonding in these structures for, as noted by Watson (1969), most groups that are capable of being involved in strong hydrogen bonds are indeed involved in hydrogen bonding inside myoglobin or between myoglobin and the solvent; this also appears to be true of the other proteins. The conformations must be favoured by interactions other than hydrogen bonding, and this would explain the population of the P₁, L₁ and L₂ sub-domains. In particular, left-hand α -helices are not observed in the conformational data used and L conformations are therefore not stabilized by helical hydrogen bonds.

Distribution of residue conformations within domains

Pohl (1971) has produced empirical energy maps based on the assumption that the conformations of

residue backbones in protein data correspond to a Boltzmann distribution. The assumption of this distribution in the absence of further information is certainly consistent with the Jaynes hypothesis which is not dependent on thermodynamic argument (Jaynes, 1957). Nevertheless, the fact that Pohl (1971) obtained good agreement with theoretical energy calculations shows that these interactions have an effective temperature in the range of the thermal temperature at the time of folding and this has implications for the folding of proteins. In particular, it is of value to the minimization of the free energy by computer, since the probability that starting and intermediate conformations will survive to the end of the folding can be estimated and taken into account by the heuristic program.

Practical applications of intra-residue information measures

The results of this work show that each type of side chain has a marked effect on the conformation of the backbone to which it is attached. The 'difference plot' (Fig. 8) shows that in approximately half the residues in trypsin inhibitor this effect predominates over interactions between residues, including those 'tertiary' interactions between residues brought together by the tertiary folding of the molecule. To some extent the regions associated with the H, P and L maxima enable individual residues to be associated with certain broad conformational types, but the exact conformation within each of these regions is also strongly determined by interactions within each residue and even the fine-structure HPL alphabet lacks sufficient resolution to take full advantage of the available intra-residue information.

The results shown in Tables 3 and 4 enable a comparison to be drawn between different ways of relating intra-residue information to ϕ , ψ conformations. In general, it is clear that the system of grouping, or alphabet, involving uniform cells in the ϕ , ψ plot is superior to the use of an alphabet based on natural

Table 4. *Predicted conformations of trypsin inhibitor protein on the basis of 20° and 40° cell information measures*

Column (2): from W. Steigemann, personal communication. Column (3): angle distance = $[(\phi_{\text{obs.}} - \phi_{\text{pred.}})^2 + (\psi_{\text{obs.}} - \psi_{\text{pred.}})^2]^{\frac{1}{2}}$; choices in descending value of information (see Table 1). Column (4): 1 means predicted angles fall in 40° cell identical with or adjacent to that observed. 0 means not 1.

Residue (1)	Observed		20° cells			40° cells, success of first choice (4)
	ϕ	ψ	Angle distances, observed to predicted, for choice:			
			1	2 (3)	3	
1 Arg	-62	146	196	38	284	1
2 Pro	-4	153	66	63	205	1
3 Asp	-62	-33	21	170	28	1
4 Phe	-68	-3	38	139	134	1
5 Cys	-75	-20	30	160	150	1
6 Leu	-86	-8	55	148	78	1
7 Glu	-80	148	200	10	30	0
8 Pro	-17	156	53	57	206	1
9 Pro	3	142	73	62	195	1
10 Tyr	-121	106	45	171	120	0
11 Thr	-67	-42	202	26	67	1
12 Gly	104	175	149	194	205	0
13 Pro	-32	-14	168	125	36	0
14 Cys	-80	157	207	57	29	0
15 Lys	-117	43	109	87	131	0
16 Ala	-90	177	230	66	37	0
17 Arg	-127	81	152	131	69	1
18 Ile	-107	123	24	182	—	1
19 Ile	-83	117	49	170	—	1
20 Arg	-131	172	236	66	22	1
21 Tyr	-111	151	19	210	162	1
22 Phe	-143	149	206	38	95	0
23 Tyr	-88	130	47	184	140	1
24 Asn	-104	106	219	58	28	0
25 Ala	-60	-30	22	201	173	1
26 Lys	-73	-36	19	170	220	1
27 Ala	-86	-21	46	183	161	1
28 Gly	78	15	17	228	47	1
29 Leu	-153	174	247	72	122	0
30 Cys	-94	146	197	39	29	0
31 Gln	-132	166	16	206	80	1
32 Thr	-92	157	39	187	148	0
33 Phe	-156	161	223	55	110	0
34 Val	-93	119	174	35	92	0
35 Tyr	-94	135	39	190	145	1
36 Gly	-75	-11	151	182	166	0
37 Gly	107	-12	56	268	25	1
38 Cys	-144	152	215	26	77	0
39 Arg	57	41	140	167	216	0
40 Ala	-60	156	206	90	34	0
41 Lys	-103	179	233	49	48	0
42 Arg	-80	-16	45	158	173	0
43 Asn	-91	75	48	125	205	0
44 Asn	-158	110	193	91	71	0
45 Phe	-127	157	208	32	112	0
46 Lys	-80	-10	45	146	193	1
47 Ser	-157	162	88	237	178	0
48 Ala	-70	-35	186	201	249	1
49 Glu	-63	-42	15	192	198	1
50 Asp	-69	-38	22	173	22	1
51 Cys	-64	-47	7	189	177	1

Table 4. *continued*

Residue (1)	Observed		20° cells			40° cells, success of first choice (4)
	ϕ	ψ	Angle distances, observed to predicted, for choice:			
			1	2	3	
			(2)	(3)		
52 Met	-73	-29	31	207	—	1
53 Arg	-65	-44	16	166	205	0
54 Thr	-80	-42	198	16	60	1
55 Cys	-101	7	65	126	127	1
56 Gly	-56	140	167	45	224	0
57 Gly	60	179	149	212	211	0
58 Ala	174	-151	246	442	393	0
Residues successfully predicted* ...			48%	33%	26%	52%

* For column (3), success denoted by angle differences $\leq 68^\circ$; for column (4), success = 1.

Table 5. *Assessment of the effectiveness of intra-residue information based on the 40° alphabet in predicting conformation*

Note: these predictions were made with the respective proteins omitted from the initial data from which the measures were obtained.

Protein	% of residues correct
Hen egg-white lysozyme	55
Carboxypeptidase	45
<i>Staphylococcus</i> nuclease	36
Myoglobin	69
Ribonuclease	28
Elastase	33
Cytochrome c	50
Trypsin inhibitor protein	54
Chymotrypsin	31
Overall average	43

Fig. 8. Difference plot of $\Delta^* = *(S_j = X_p; \bar{X}_p, R_j) - *(S_j = X_0; \bar{X}_0, R_j)$ for each residue j in trypsin inhibitor protein

Values of Δ^* are smoothed, each point being an average of five contiguous residues. The lower the values of Δ^* the less information required from sources other than the intra-residue information to give a correct prediction. X_p is the predicted and X_0 the observed conformation. ----, 40° Square-cells alphabet; —, fine-structure HPL alphabet; ▨, pleated sheet; ▩, α -helix; □, distorted helix.

groupings of information. The latter approach may be used by making prediction of domains according to the highest information maximum or according to the population density of a given domain or sub-domain. These two methods, compared in Table 3, show little difference in the accuracy of prediction. Of the 'cell' approaches that utilizing 20° cells is not significantly better than that involving 40° cells when the former is assessed by a limiting angle distance of 68° (see footnote to Table 4, and the Results section).

It is concluded that the best practical method of choosing starting conformation is that using 40° cells, despite the fact that this alphabet uses considerably more computer memory when extended to a consideration of interactions between residues. The alphabet, which exploits the natural distribution of residue information, is economical of computer memory and has been found useful in analysing

certain conformations involving residue pairs, in particular the ' β -bend' (Robson & Pain, 1974b).

The percentage of residues correctly predicted for a variety of proteins (Table 5) derives from a consideration of intra-residue information only, taking no account of theoretical co-operative effects or of any empirical rules for deciding whether a certain type of secondary structure is present or not. (For example, such a rule might be that a helix will be present if any four out of six consecutive residues prefer helix, provided that the other two residues which prefer another conformation do not occur together.)

Hence these results provide a baseline for the development of such rules and for the addition of further information. The results show that about 43% of all residues exist in a conformation which is determined purely by their type of side chain and not by the type or conformation of any other residue.

Helical (H) regions are predicted well with all methods, as would be expected on the basis of the information approach directed solely towards the prediction of helical and non-helical regions (Robson & Pain, 1971). Pleated-sheet (P) regions are again well predicted by using the 40° and 20° cell approach, but not by using the HPL alphabet. It is clear that single residues carry a significant amount of information concerning their occurrence or non-occurrence in pleated-sheet conformations, provided the appropriate analytical method is used to display it. It should be noted in passing that the pleated-sheet conformation (P) is here described in terms of ϕ , ψ backbone angles and does not necessarily imply a classical hydrogen-bonded β -pleated sheet.

The overall goodness of fit is not high (about 50% for the cell alphabet). It should be emphasized, however, that these results are being obtained on the basis of the ability of the single residue to determine its own conformation, i.e. on intra-residue information with no account of co-operativity. This result provides an important baseline for assessing predictions utilizing co-operativity, specific interactions between pairs of residues, between triplets, and so on. In particular, it shows the importance of co-operativity since, at least in the case of α -helix, the very simplest treatment of co-operativity leads to fairly accurate predictions (Robson & Pain, 1971). The importance of neighbouring residues in both co-operative and more specific interactions is the subject of a further communication (Robson & Pain, 1974a), and supports the conclusion that the grouping of helix formers and pleated-sheet formers implied by these interactions must be taken into account.

This work was supported by a grant from the Medical Research Council. The expert assistance of Miss Moira Bain is gratefully acknowledged. The authors are indebted to Dr. W. Steigemann for the use of his crystallographic results on trypsin inhibitor protein before publication and for ϕ , ψ angles from Dr. F. A. Cotton for *Staphylococcus* nuclease, from Dr. F. M. Richards for ribonuclease S, from Dr. L. Sawyer for a refined structure of elastase, from

Dr. R. Diamond for a refined structure of hen egg-white lysozyme and from Dr. D. M. Blow for chymotrypsin.

References

- Chothia, C. (1973) *J. Mol. Biol.* **75**, 295–302
 Diamond, R. (1971) *Acta Crystallogr.* **A27**, 426–452
 Ghani, S. N. (1972) *Comput. Aided Design* **4**, 71–78
 Jaynes, E. T. (1957) *Physical Rev.* **106**, 620–630
 Jeffreys, H. (1948) *Theory of Probability*, 1st edn., p. 194, Oxford University Press, Oxford
 Leach, S. J., Nemethy, G. & Scheraga, H. A. (1966) *Biopolymers* **4**, 369–407
 Lewis, P. N., Gö, N., Kotelchuck, D. & Scheraga, H. A. (1970) *Proc. Nat. Acad. Sci. U.S.* **65**, 810–815
 Liquori, A. M. (1969) *Quart. Rev. Biophys.* **2**, 65–92
 Nemethy, G., Phillips, D. C., Leach, S. J. & Scheraga, H. A. (1967) *Nature (London)* **214**, 363–365
 Pain, R. H. & Robson, B. (1970) *Nature (London)* **227**, 62–63
 Pain, R. H. & Robson, B. (1971) *Proc. Eur. Congr. Biophysics 1st* **1**, 33–37
 Pauling, L. & Corey, R. B. (1951) *Proc. Nat. Acad. Sci. U.S.* **37**, 729–740
 Periti, P. F., Quagliarotti, G. & Liquori, A. M. (1967) *J. Mol. Biol.* **24**, 313–322
 Pohl, F. M. (1971) *Nature (London) New Biol.* **234**, 277–281
 Quiocho, F. A. & Lipscomb, W. N. (1971) *Advan. Protein Chem.* **25**, 1–78
 Ramachandran, G. N. & Sasisekharan, V. (1968) *Advan. Protein Chem.* **23**, 283–437
 Ramachandran, G. N., Ramakrishnan, C. & Sasisekharan, V. (1963) *J. Mol. Biol.* **7**, 95–99
 Robson, B. (1972) *Spec. Per. Rep. Chem. Soc.: Amino Acids, Peptides and Proteins* **4**, 224–229
 Robson, B. (1974) *Biochem. J.* **141**, 853–867
 Robson, B. & Pain, R. H. (1971) *J. Mol. Biol.* **58**, 237–259
 Robson, B. & Pain, R. H. (1972) *Nature (London) New Biol.* **238**, 107–108
 Robson, B. & Pain, R. H. (1973) in *Conformation of Biological Molecules and Polymers, 5th Jerusalem Symp.* (Bergman, E. D. & Pullman, A., ed.), pp. 161–172, Academic Press, London and New York
 Robson, B. & Pain, R. H. (1974a) *Biochem. J.* **141**, 883–897
 Robson, B. & Pain, R. H. (1974b) *Biochem. J.* **141**, 899–904
 Scheraga, H. A. (1971) *Chem. Rev.* **71**, 195–217
 Schiffer, N. & Edmundson, A. B. (1967) *Biophys. J.* **7**, 121–135
 Scott, R. A. & Scheraga, H. A. (1966) *J. Chem. Phys.* **45**, 2091–2101
 Watson, H. C. (1969) *Progr. Stereochem.* **4**, 299–333
 Wu, T. T. & Kabat, E. A. (1973) *J. Mol. Biol.* **75**, 13–31