J. Mol. Biol. (1978) **120**, 97–120

Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins

J. GARNIER†

Laboratoire de Biochimie physique, I.N.R.A.
Bât. 433, Université de Paris Sud
91405 Orsay, France

D. J. OSGUTHORPE AND B. ROBSON‡

Department of Biochemistry
University of Manchester
Manchester M139PL, England

(Received 30 June 1977, and in revised form 15 December 1977)

(1) Co-operation between a laboratory interested in developing the theory for protein secondary structure prediction methods and a laboratory interested in applying and comparing such methods has led to the development of a simple predictive algorithm.

(2) Four-state predictions, in which each residue is *unambiguously* assigned one conformational state of α-helix, extended chain, reverse turn or coil, predict 49% of residue states correctly (in a sample of 26 proteins) when the overall helix and extended-chain content is not taken into account.

(3) When the relative abundances of helix, extended chain, reverse turn and coil observed by X-ray crystallography are taken into account, a single constant for each protein and type of conformation can be used to bias the prediction. When predictions are optimized in this way, 63% of all residue states are unambiguously and correctly assigned.

(4) By analysing the nature of the bias required, proteins can be classified into helix-rich types, pleated-sheet-rich types, and so on. It is shown that, if the type of protein can be determined even approximately by circular dichroism, 57% of residue states can be correctly predicted without taking into account the X-ray structure. Further, comparable predictions can be obtained if, instead of circular dichroism, preliminary predictions are made to assess the protein type.

(5) It is emphasized that the numbers quoted here depend on the method used to assess accuracy, and the algorithm is shown to be at least as good as, and usually superior to, the reported prediction methods assessed in the same way.

(6) Ways of further enhancing predictions by the use of additional information from hydrophobic triplets and homologous sequences are also explored. Hydrophobic triplet information does not significantly improve predictive power and it is concluded that this information is used by proteins in the next stage of folding. On the other hand, the use of homologous sequences appears to be very promising.

(7) The implication of these results in protein folding is discussed.

† From whom programs may be obtained.
‡ To whom requests for reprints should be submitted.

0022-2836/78/1201-9720$02.00/0

1. Introduction

Following experimental evidence that the conformation of a globular protein is determined by its amino acid sequence, there have been many attempts to predict protein secondary structure from the amino acid sequence alone (for a review, see Robson & Suzuki, 1976). As the accuracy of the predictions has developed, confidence in their potential use as a routine tool for biochemists has also grown. Unfortunately, there are many ways of describing the accuracy of secondary structure predictions, and if different authors were to arrive at exactly the same assignment of secondary structure, they would be likely to quote quite different numbers to express that accuracy. Further, this may have led to the impression that the current status of the art is considerably better than it actually is. This is unfortunate because, as discussed in detail by Burgess & Scheraga (1975), the ability to make a good prediction of secondary structure is likely to remain a factor of crucial importance whenever an attempt is made to predict the three-dimensional structure of a protein molecule.

Here we are concerned with describing and testing the *simplest possible* statistical procedures which can be used routinely for the prediction of secondary structure and, at the same time, yield good results. In this connection we have been particularly interested in the promising method of Chou & Fasman (1974), this method being attractive to those workers who do not have ready access to large digital computers and who must carry out the calculations by hand or by desk-top calculator. Unfortunately, some of their rules are qualitative rather than quantitative, and being open to interpretation they have not always yielded such promising results in the hands of other workers (Burgess & Scheraga, 1975; Garnier *et al.*, 1976). Recently, Rerat & Rerat (personal communication) have written a computer program for the Chou & Fasman method and applied it to 21 proteins of known sequence and conformation. Although the degree of success was significant, the method developed by us in this paper has overall superiority when the accuracy is compared by identical measures. This improvement is of interest to us because it helps to pinpoint the relative importance of different mechanisms for secondary structure formation *in vivo* which are emphasized to different extents in the Chou & Fasman method and the procedures described here. In this context it may be noted that the basic conformational parameters used by Chou & Fasman, as opposed to the algorithms for their application, are very similar when compared through the use of equation (13) of Robson & Suzuki (1976).

Finally, some attention in this paper is given to ways of circumnavigating certain kinds of error which may arise in making predictions. Two kinds of predictive procedure are developed, one that makes use of additional information from circular dichroism concerning the amount of helix or pleated sheet, and one that yields reasonable results in the absence of this information. Preliminary observations are also made on the advantage of using homologous proteins.

2. Theory

The relative complexity of the theory (Robson, 1974), which is necessary to justify the formal correctness of the method for handling both large and small statistical samples and the additivity of the different terms, may have obscured the simplicity of the predictive methods derived from it. As we shall demonstrate, the predictive

method recommended here is simpler in concept than that proposed by Chou & Fasman (1974), which has enjoyed some popularity because of its apparent simplicity.

Only theoretical aspects which relate directly to the predictive procedure will be considered here. A prediction of the secondary structure of a protein can be considered as a prediction of the conformational state of each of the residues in that sequence. Let S_j be the conformational state, say α-helix, of the jth residue in the sequence, and let R_k be the type of residue, say alanine, at the kth position in the sequence. *A priori*, a good prediction of S_j, for example one that correctly assigns it as α-helix, could depend not on the type of one or even several residues, but on the whole amino acid sequence of the protein. The most general statement concerning the information we have for the conformation of the jth residue is thus

$$I(S_j; R_1, R_2, \ldots R_{last}),$$

which reads as the "information which the residues at the first, second, and so on up to the last position carry about the conformation of the jth residue". If we define the conformational states of the residues in such a way that there are, say, four possible states A, B, C and D for each residue, then a predictive procedure leads to four values for information associated with each residue:

$$I(S_j = A; R_1, R_2, \ldots R_{last})$$

$$I(S_j = B; R_1, R_2, \ldots R_{last})$$

$$I(S_j = C; R_1, R_2, \ldots R_{last})$$

$$I(S_j = D; R_1, R_2, \ldots R_{last}).$$

Whichever of these values is highest defines the conformational state predicted. For example, if there was most information for state B, this is the state which is predicted. Units used here are based on the nat (Robson, 1974).

Each of these four information values may be estimated by inspection of the amino acid sequence. The actual procedure for carrying out this inspection and assigning the values depends on how we expand $I(S_j; R_1, R_2, \ldots R_{last})$ as a series of simpler terms. The expansion could be made exact, but this would ultimately mean assigning certain parameters for the effect of all possible amino acid sequences of the same length as the protein whose structure is to be predicted. Statistical analysis of a finite sample of, at the present time, about 30 different proteins, can only provide parameters which are dependent on one, two or at best three residues. Parameters depending on larger combinations can, of course, be derived but the necessary and desirable consequence of the information theory approach used by us (Robson, 1974) means that the values so obtained for these parameters would be zero or close to zero. Hence the expansion is performed to produce a summed series of simple (depending on few residues) and complex (depending on many residues) terms, and the complex terms can then be neglected, since the values obtained by our statistical analysis would necessarily be close to zero because they imply the limited information available to the observer.

Previous studies (Robson & Pain, 1974*b*; Robson & Suzuki, 1976) have shown that the effect of one residue type on the conformation of residues up to eight residues

distant plays a predominating role, while choosing shorter separation distances neglects significant information.

A particularly simple approximation is thus:

$$I(S_j; R_1, R_2, \ldots R_{last}) \simeq \sum_{m=-8}^{m=+8} I(S_j; R_{j+m}), \quad (1)$$

in which terms in the expansion containing more than one R_j parameter have been neglected and interactions between residues separated by more than eight positions in the amino acid sequence are neglected.

The measure $I(S_j; R_{j+m})$ represents the information which the type of residue at $j+m$ carries about the conformational state of the jth residue. Since there are 20 types of residue R and, say, four conformational states, there are 20×4 different parameters for each separation m. Since we are taking $-8 \leqslant m \leqslant +8$, including $m = 0$, there are effectively 17 separations. If m is negative, the information concerns a residue on the N-terminal side of R_j. If m is positive, the information concerns a residue on the C-terminal side of R_j. If m is zero, the information is $I(S_j; R_j)$, which is the information the residue carries about its own conformation. There is thus a total of $20\times4\times17$ parameters provided in this paper, the conformational states being defined in Methods.

3. Methods

(a) *Parameters*

Parameters $I(S_j; R_{j+m})$ are given in Tables 1 to 4 for 4 conformational states: $S_j = H$, the α-helical state; $S_j = E$, the extended-chain state; $S_j = T$, the reverse-turn state, and $S_j = C$, the coil (or aperiodic) state, which is defined as any state not H, E or T. The detailed definitions of these states have been given by Robson & Suzuki (1976), although the state T sometimes corresponds to the state T_m described by Robson & Suzuki (1976). This is because the 2 middle residues of a 4-residue reverse turn are the only residues with conformations relevant to the definition of a reverse turn (see, e.g., Robson & Pain, 1974*c*). Hence, a prediction that, say, residues 26, 27, 28 and 29 constitute a reverse turn makes no statement concerning the conformations of residues 26 and 29, and is thus of limited interest in the assignment of starting conformations for the simulation of protein folding. It is interesting to note, however, that information concerning T (the turn as 4 residues) and T_m (the turn as 2 residues), is not significantly different except in the instances tryptophan and proline. Further, E information is similar to that for β sheet.

These parameters are obtained from the directional information plots of $I(S_j; R_{j+m})$ *versus m* given by Robson & Suzuki (1976). These were based on the statistical (information theory) analysis of 25 proteins of known sequence and conformation. However, on the rationale that a scatter of the parameters of up to about 2 decinats is purely due to the finite size of the sample base (Robson & Pain, 1974*b*; Robson & Suzuki, 1976), smooth curves have been fitted to these points and values (Tables 1 to 3) read from these curves in order to correct the parameters for sampling noise. Directional values for the coil state (Table 4) have not previously been reported.

In order to take into account clustering of hydrophobic residues, particularly on the surface of an α-helix, the following information function was elucidated:

$$I_{clus}(R_j) = \ln \frac{P(S_j = H; R_j | R_{j-m'} = \Phi, R_{j+m'} = \Phi)}{P(S_j = H; R_j)}, \quad (2)$$

where the numerator probability on the right-hand side is the probability that R_j will be α-helical, when $R_{j-m'}$ (the side-chain at R_{j-3} or R_{j-4}) and $R_{j+m'}$, (the side-chain at R_{j+3} or R_{j+4}) is hydrophobic (Φ), as determined from the data base of Robson & Suzuki

TABLE 1

Directional information measure for the α-helical conformation†

Amino acid residue	Residue position‡ (centinats)																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	−5	−10	−15	−20	−30	−40	−50	−60	−86	−60	−50	−40	−30	−20	−15	−10	−5
Ala	5	10	15	20	30	40	50	60	65	60	50	40	30	20	15	10	5
Val	0	0	0	0	0	0	5	10	14	10	5	0	0	0	0	0	0
Leu	0	5	10	15	20	25	28	30	32	30	28	25	20	15	10	5	0
Ile	5	10	15	20	25	20	15	10	6	0	−10	−15	−20	−25	−20	−10	−5
Ser	0	−5	−10	−15	−20	−25	−30	−35	−39	−35	−30	−25	−20	−15	−10	−5	0
Thr	0	0	0	−5	−10	−15	−20	−25	−26	−25	−20	−15	−10	−5	0	0	0
Asp	0	−5	−10	−15	−20	−15	−10	0	5	10	15	20	20	20	15	10	5
Glu	0	0	0	0	10	20	60	70	78	78	78	78	78	70	60	40	20
Asn	0	0	0	0	−10	−20	−30	−40	−51	−40	−30	−20	−10	0	0	0	0
Gln	0	0	0	0	5	10	20	20	10	−10	−20	−20	−10	−5	0	0	0
Lys	20	40	50	55	60	60	50	30	23	10	5	0	0	0	0	0	0
His	10	20	30	40	50	50	50	30	12	−20	−10	0	0	0	0	0	0
Arg	0	0	0	0	0	0	0	0	−9	−15	−20	−30	−40	−50	−50	−30	−10
Phe	0	0	0	0	0	5	10	15	16	15	10	5	0	0	0	0	0
Tyr	−5	−10	−15	−20	−25	−30	−35	−40	−45	−40	−35	−30	−25	−20	−15	−10	−5
Trp	−10	−20	−40	−50	−50	−10	0	10	12	10	0	−10	−50	−50	−40	−20	−10
Cys	0	0	0	0	0	0	−5	−10	−13	−10	−5	0	0	0	0	0	0
Met	10	20	25	30	35	40	45	50	53	50	45	40	35	30	25	20	10
Pro	−10	−20	−40	−60	−80	−100	−120	−140	−77	−60	−30	−20	−10	0	0	0	0

† The data for Tables 1 to 4 are obtained from 25 proteins by Robson & Suzuki (1976), but the values quoted here are read from curves fitted through the directional plots. The coil values come from the same source but have not previously been quoted. Values are in centinats (nats × 100).

‡ For example, the information at position $j-6$ is the information which the residue j carries about the conformation of any residue 6 away in the N-terminal direction and at position $j+6$ about any residue 6 away in the C-terminal direction (see Robson & Suzuki, 1976). At position j, it is the information carried by the residue itself to be in the given conformation (single-residue information).

TABLE 2

Directional information measure for the extended conformation

Amino acid residue	Residue position																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	10	20	30	40	40	20	0	−20	−42	−20	0	20	40	40	30	20	−10
Ala	0	0	0	0	−5	−10	−15	−20	−23	−20	−15	−10	−5	0	0	0	0
Val	0	0	−10	−20	0	20	40	60	68	60	40	20	0	−20	−10	0	0
Leu	0	0	0	0	0	5	10	20	23	20	10	5	0	0	0	0	0
Ile	0	−10	−20	−10	0	20	40	60	67	60	40	20	0	−10	−20	−10	0
Ser	0	10	20	10	0	−5	−10	−15	−17	−15	−10	−5	0	10	20	10	0
Thr	5	10	15	20	15	15	10	10	13	10	10	15	15	20	15	10	5
Asp	0	5	10	15	20	0	−20	−30	−44	−30	−20	0	0	0	0	0	0
Glu	−10	−15	−20	−25	−30	−35	−40	−45	−50	−55	−60	−60	−50	−40	−30	−20	−10
Asn	10	30	50	30	20	0	−15	−30	−41	−30	−15	0	20	30	50	30	10
Gln	0	0	0	0	0	−5	−10	0	12	20	30	40	50	50	40	30	15
Lys	−5	−10	−15	−20	−30	−40	−50	−40	−33	−20	−10	0	10	10	0	0	0
His	−10	−20	−40	−20	−10	0	−10	−20	−25	−35	−30	−25	−20	−15	−10	−5	0
Arg	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0
Phe	0	0	0	0	0	5	10	20	26	10	−10	−30	−60	−65	−60	−40	−20
Tyr	0	5	10	15	20	25	30	35	40	35	30	25	20	15	10	5	0
Trp	0	0	0	0	0	−10	−10	−10	−10	−10	−10	−15	−20	−25	−30	−20	−10
Cys	0	0	0	0	0	10	20	30	44	30	20	10	0	0	0	0	0
Met	−10	−20	−30	−40	−40	−30	0	10	23	10	0	−30	−40	−40	−30	−20	−10
Pro	10	20	30	30	20	10	0	−10	−18	−20	−10	10	30	40	30	20	10

TABLE 3

Directional information measure for turns†

Amino acid residue	Residue position																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	0	0	0	0	10	30	55	55	57	40	0	0	0	0	0	0	0
Ala	0	0	0	−10	−20	−30	−40	−50	−50	−40	−30	−20	−10	0	0	0	0
Val	0	0	0	0	−10	−20	−30	−40	−60	−40	−30	−20	−10	0	0	0	0
Leu	0	0	0	−10	−20	−30	−40	−50	−56	−20	−10	0	0	0	0	0	0
Ile	0	0	0	0	0	−10	−20	−30	−46	−40	−10	0	0	20	30	20	10
Ser	0	−10	−20	−20	10	15	20	25	26	25	20	15	10	0	0	0	0
Thr	0	10	20	20	20	15	18	5	3	5	10	15	20	20	20	10	0
Asp	0	0	0	0	0	0	5	10	31	10	5	0	0	0	0	0	0
Glu	0	−5	−10	−15	−20	−30	−40	−45	−47	−20	0	10	5	0	0	0	0
Asn	0	0	0	10	20	30	35	40	42	40	35	30	20	10	5	0	0
Gln	10	20	30	25	20	15	10	5	4	20	30	40	50	60	50	40	20
Lys	−10	−20	−30	−40	−25	−10	0	10	10	10	0	−20	−30	−20	−10	−5	0
His	0	0	0	0	0	0	0	0	−3	0	10	20	30	20	10	0	0
Arg	0	0	0	0	0	0	0	10	21	30	40	30	20	10	0	0	0
Phe	0	0	0	0	0	−5	−10	−15	−18	−15	0	15	30	25	20	10	0
Tyr	0	0	0	5	15	15	20	25	29	25	20	15	15	5	0	0	0
Trp	0	0	0	10	20	30	40	80	36	−30	30	40	50	60	70	40	20
Cys	20	40	50	60	60	55	50	45	44	40	35	30	25	20	15	10	5
Met	−5	−15	−20	−25	−30	−35	−40	−45	−48	−45	−40	−35	−30	−25	−20	−15	−5
Pro	10	20	30	40	50	70	10	−90	36	90	10	0	0	0	0	0	0

† Defined as 2 residues.

TABLE 4

Directional information measure for coil

Amino acid residue	Residue position																
	$j-8$		$j-6$		$j-4$		$j-2$		j		$j+2$		$j+4$		$j+6$		$j+8$
Gly	0	0	0	0	10	30	40	45	49	45	40	30	10	0	0	0	0
Ala	0	0	0	0	−5	−10	−20	−25	−25	−25	−20	−15	−10	−5	0	0	0
Val	0	0	0	0	−10	−20	−25	−30	−35	−30	−25	−20	−10	0	0	0	0
Leu	0	0	0	−10	−20	−30	−40	−30	−20	−20	−10	0	0	0	0	0	0
Ile	0	0	0	0	0	−10	−20	−30	−33	−30	−10	0	10	20	30	20	0
Ser	0	−10	−20	−20	10	15	20	25	50	25	20	15	10	0	0	0	0
Thr	0	10	20	30	20	15	10	15	17	15	10	15	20	30	20	10	0
Asp	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Glu	0	0	10	20	40	20	0	−10	−44	−40	−20	−10	0	0	0	0	0
Asn	0	0	0	10	20	30	35	40	46	40	35	30	20	10	0	0	0
Gln	10	20	30	25	20	15	10	0	−5	20	30	40	50	60	50	40	20
Lys	−10	−20	−30	−40	−25	−20	−10	−8	−8	0	0	−20	−30	−20	−10	−5	0
His	0	0	0	0	0	0	0	10	16	15	10	10	10	10	5	0	0
Arg	0	0	0	0	0	0	0	0	−12	0	20	30	20	10	0	0	0
Phe	0	0	0	0	0	−5	−10	−20	−41	−20	0	15	30	25	20	10	0
Tyr	0	0	0	0	0	0	0	0	−6	0	0	0	0	0	0	0	0
Trp	0	0	0	10	20	30	40	20	12	20	30	40	50	60	70	40	20
Cys	0	0	0	0	0	0	−10	−30	−47	−30	−10	0	0	0	0	0	0
Met	0	−5	−10	−15	−20	−25	−30	−40	−41	−40	−30	−25	−20	−15	−10	−5	0
Pro	0	0	10	20	30	40	50	55	58	50	10	0	0	0	0	0	0

(1976). The definition of a hydrophobic side-chain is discussed below in Results. Clustering was only taken into account in a limited series of studies (see below).

(b) *The basic method (directional method)*

As described in Theory, the prediction of the conformational state of each residue j involves evaluation of eqn (1) for each conformational state, and then choosing the conformation with the highest information content. This procedure for the conformation of every residue in a long sequence is, of course, considerably facilitated by a simple computer program.

From the information content of eqn (1) and for each conformational state, a constant value or decision constant (DC_S) can be subtracted before comparing the information contents. DC_S is an adjustable parameter which is chosen with the aim of producing optimal predictions. It is a function only of S_j and if there are N different conformational states S, the prediction to be made is a function of eqn (1) and of $N - 1$ values of DC_S. The term $N - 1$ arises from the fact that only relative values of DC_S affect the prediction so that one of the DC_S, usually for the coil state, can be fixed at an arbitrary value of zero.

(c) *The single-residue information method*

In this method we include only information $I(S_j; R_j)$, i.e. the information that each residue carries about its own conformation. It thus corresponds to the use of eqn (1) neglecting all values of m except $m = 0$. This study was undertaken in order to estimate the importance of neighbour interactions by comparing the accuracy of the predictions. It was not undertaken to provide a simpler method, for the following reasons. Neglect of all separations except $m = 0$ loses all information concerning the co-operativity between residues. For example, there is now no tendency for a weakly helical residue to be pulled into the helical conformation by its neighbours, and this leads to poor predictions. However, this co-operativity may be accounted for by an algorithm by which neighbouring residues share their net information content. This can be done in the absence of directional information by the use of an algorithm which would include the averaged single-residue information measures of any run of n_S residues, the number n_S being referred to as the run constant (Robson, 1974). From this a decision constant DC_S for each conformational state can be subtracted as in the basic or directional method. The overall consequence of using single-residue information only is thus an increased complexity of the algorithm, which makes it no easier to use than the directional method. The predictive algorithm resulting from these considerations is:

$$I(S_j; R_1, R_2, \ldots R_{last}) \simeq \mathbf{MAX}_{k=j-(n_S-1)}^{k=j} \left[\frac{1}{n_S} \sum_{i=k}^{i=k+n_S-1} I(S_i; R_i) - DC_S \right]. \quad (3)$$

Here **MAX** indicates that for a specified conformational state the maximum value of the bracketed function is to be sought in the interval of k bounded by the subscript and superscript.

(d) *Optimization of the accuracy of predictions as a function of decision constants* DC_S *and run constants* n_S

In making a prediction, the decision constant DC_S and run constant n_S for each conformation are considered to be constants for the protein being used. The accuracy index or fraction of residues correct (see Appendix) may be optimized as a function of the decision and run constants, and this optimization can be carried out using single proteins or groups of proteins. The use of the accuracy index was preferred for optimization of individual proteins because it gives equal weighting to each conformation irrespective of its abundance in each protein. In studies on groups of proteins, however, we preferred the use of fraction of residues correct because (1) similar values of DC_S and n_S are usually obtained to those using the accuracy index, (2) the fraction of residues correct is arguably the most useful statistic to optimize for predictions to be used as starting conformations in simulating protein folding, and (3) this measure is widely used in the literature and is therefore most suitable for comparison with other work.

4. Results

(a) *Co-operativity effect (run constants)*

The value of the run constant n_S is a measure of the co-operativity between residues in state S. This has already been found to be essential for improving α-helix predictions from single-residue information (Robson & Pain, 1971).

Different values of n_S have been proposed for α-helix and β-sheet (Kotelchuck & Scheraga, 1969; Robson & Pain, 1971; Chou & Fasman, 1974). The optimized values of n_S for the single-residue information program using equation (2) were found in this work to be 6 for α-helix, 5 for extended, 4 for reverse turn and 3 for coil. When the decision constants are set to zero, this group of run constants lead to the correct prediction of 45·3% of all residues, for the four conformations and for 26 proteins, compared with 36·4% when all run constants are made equal to unity (see Table 5). It is emphasized that these are true four-state predictions assessed on a four-state basis. They thus compare very favourably with earlier results (e.g. Chou & Fasman, 1974) quoted on a two-state basis. In a two-state prediction (e.g., of helix and non-helix), 50% of residues would be correctly assigned even by a random prediction, in a four-state prediction, only 25% would be correctly assigned by a random prediction.

Unless otherwise stated, the simple use of equation (1), rather than equation (3) constitutes the program used here. No run constant as in equation (3) is required; this is shown as follows. When a run constant was introduced into equation (1) as in equation (3), it was observed that the accuracy of the predictions was considerably less sensitive to the choice of a run constant. Uniform run constants of 1, 2, 3 and 4 yield 49·0%, 48·1%, 48·0% and 47·4%, respectively, of correctly predicted residues for the four conformations and the 26 proteins. The run constants produced by optimization of the single-residue information program produced 48·6% residues correct when applied to the directional information program, and no combination of run constants tested gave better than 49%. There is thus no evidence whatsoever for introducing a run constant other than unity, providing accurate directional information is used. We therefore assumed that the directional information measures already include the effects of co-operativity which are relevant to secondary structure predictions.

It may be noted (Table 5) that the directional information program with all run constants equal to unity gave marginally better predictions than the single-residue information program with optimized run constants. This improvement is even more pronounced when no decision constant is included (i.e. all decision constants are zero). However, the directional information program is recommended not on the basis of its marginal advantage in terms of predictive power, but because of its simplicity of use which results when taking all run constants as unity.

(b) *Optimization of the decision constants*

The predictions may be greatly improved by introducing different decision constants for all four states. Although an important aim of this paper is to enable good predictions to be made from sequence and circular dichroism data alone, it is interesting to determine the optimal values of the decision constants for each protein, knowing its secondary structure in advance. Indeed, as described below and in section (c), such investigations provide the information necessary for choosing decision constants from circular dichroism data when the X-ray secondary structure is not known.

TABLE 5

Effect of decision and run constants on the percentage of correctly predicted residues for the four conformations H, E, T and C

Proteins	Predictive programs: Single-residue			Directional			
	$DC = 0$		Optimized DC	$DC = 0$		Optimized DC	DC‡
	$n = 1$	n opt.	n† opt.	$n = 1$	n† opt.	$n = 1$	$n = 1$
Rubredoxin	35·8	15·1	47	32	32·1	52·8	47·2
Concanavalin A	34·4	37	53	42·4	41·6	56·3	49·2
Prealbumin	34·6	40·2	—	36·2	33·9	60·6	55·1
Chymotrypsin	32·7	36·7	57	42·4	42·4	55·5	52·4
Elastase	32·5	40	59	51·2	48·7	59·6	59·2
Carbonic anhydrase C	29·3	30·9	51	29·7	27	47·5	37·5
Trypsin inhibitor	39·7	55·2	—	56·9	50	58·6	62·1
Bovine ribonuclease	31·4	33·9	60	41·1	35·5	60·5	58·1
Papain	35·8	35·4	53	41	39·2	52·8	49·1
Staphylococcal nuclease (foggi strain)	31·7	53·5	63	50	51·4	59·2	46·5
Subtilisin DPN′	33·8	42·9	50	40·7	39·3	46·2	43·3
Bovine carboxypeptidase	31·9	42·7	51	41·7	42	53·4	43·3
Porcine insulin, B chain	36·7	50	76	46·7	56·7	70	53·3
Thermolysin	33·2	44·6	54	42·7	42·4	55·4	47·8
Dogfish apo-lactate dehydrogenase	32·8	43·2	54	48·3	45·9	54·4	49·2
Hen egg-white lysozyme	40·3	51·9	61	48·1	42·6	67·4	56·6
Horse cytochrome *c*	38·4	40·4	55	51·9	54·8	56·7	54·8
Cytochrome *b*5	34·4	57	65	55·9	53·8	62·4	60·2
Carp myogen	37	63	63	59·2	59·3	64·8	63
Adenylate kinase	60·3	56·7	68	63·9	56·7	73·7	71·1
Cytochrome *c*2	43·7	56·2	67	58·9	56·2	65·2	58
Lamprey haemoglobin	32·4	49·3	72	52·8	56·1	68·9	61·5
Horse oxyhaemoglobin α	36·2	48·9	77	58·9	60·3	79·4	73
Horse oxyhaemoglobin β	35·6	44·5	74	52·7	58·2	83	71·9
Sperm whale myoglobin	47·1	58·1	81	71·2	73·2	83·1	81
Erythrocruorin	33·8	50·7	81	59·6	63·2	83·1	75·7
Averaged %	36·4	45·3	62·2	49	48·6	62·7	56·9

† Run constants of 6, 5, 4 and 3 for α-helix, extended, turn and coil conformations, respectively.

‡ Decision constants as in Table 6. The proteins are in order of increasing helix content: 1 to 6, 0 to 20% helix; 7 to 17, 20 to 50%; 17 to 26, >50%.

A rapid optimization program, capable of making tens of thousands of predictions for various combinations of decision constants in only a few minutes of Central Processor Unit (CPU) time was developed. This program was designed to trace the highest percentage of correctly predicted residues for the four conformations by varying the decision constants independently for α-helix, extended chain, and reverse turns. Recalling that only relative values of the decision constant are important, the decision constant for coil was used as the reference and set to zero throughout. The reverse-turn decision constant was allowed to vary between −300 and +300 centinats, and there was no evidence of an optimal value lying outside this range.

An optimization surface can be constructed for each protein (or group of proteins), in which a measure of accuracy of the prediction is determined at each point in a three-dimensional space with axes DC_H, DC_E and DC_T. In practice, we produced Tables of percentage residues correct, each row corresponding to a different DC_H value and each column to a different DC_E value. At each point the percentage of residues correct corresponded to the maximum value of the percentage of residues correct in the DC_T dimension, that is the turn decision constant was optimized at each point in DC_H–DC_E space. Contours of iso-percentage can then be drawn through equivalent points in DC_H–DC_E space, as shown in Figure 1. As a check, similar plots were frequently constructed using the accuracy index; this was useful in resolving ambiguities in a few cases when the percentage of residues correct was an insensitive index in certain regions of the DC_H–DC_E surface. Although all such plots showed that the difference between the decision constants of helix and extended chain was a major factor determining the accuracy of a prediction, there are a number of differences in detail from protein to protein. Figures 1 to 3 exemplify a single, well-defined optimum, an optimum covering a large range of decision constants, and a double optimum, respectively. In these Figures, the numbers enclosed in boxes are the

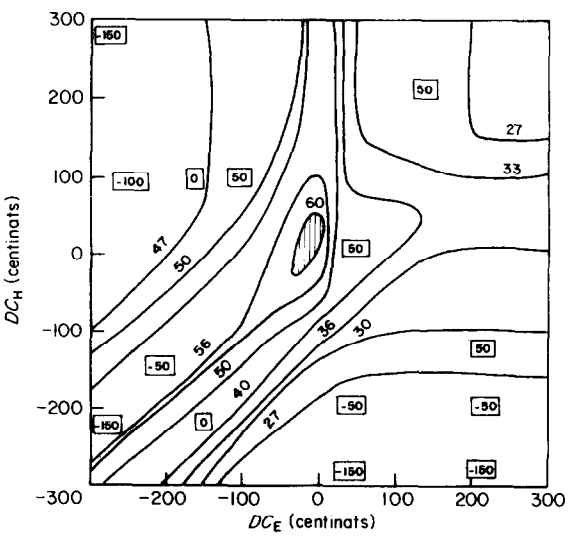

FIG. 1. Curves of iso-percentage of correctly predicted residues in the 4 conformations H, E, T and C *versus* the decision constants in centinats for α-helix (DC_H) and β-sheet (DC_E) with the use of the directional program. The protein is bovine ribonuclease. The optimized values of the decision constant for reverse turn (T) are in boxes. The hatched area corresponds to >60% residues correct.

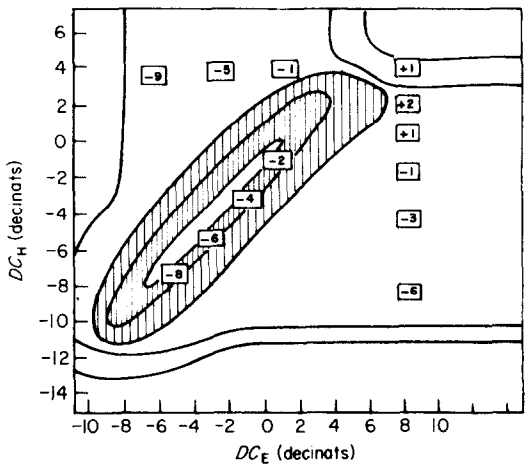

Figure 2.

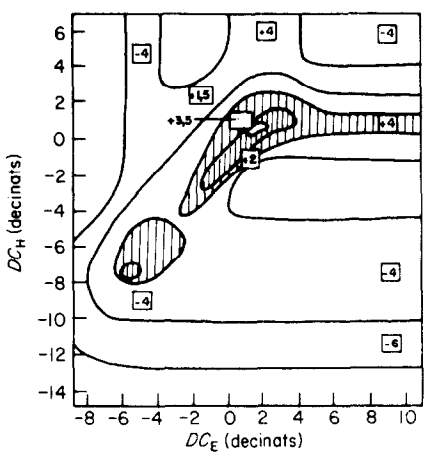

Figure 3.

FIGS 2 and 3. Iso-accuracy index curves for the 4 conformations H, E, T and C with the use of the single-residue information program (decinat units). Optimum accuracy >2·0. Dense hatching, 1·5 to 2·0; sparse hatching, 1·0 to 1·5; contours, at 0·5 intervals down to 0·0. Fig. 2, Adenylate kinase; Fig. 3, Lysozyme.

optimal values of DC_T found at various points. Such variations from plot to plot justify the use of adjustable decision constants in order to introduce further essential information relating to factors which change from protein to protein.

Using these plots, the maximum percentage of residues predicted correct which can be obtained for a protein by varying DC_H, DC_E and DC_T was recorded, and this was done for all 26 proteins and for programs using equations (1) and (3). The maximum percentages of correctly predicted residues for the four conformations for each protein are listed in Table 5. The use of optimized decision constants considerably improved the predictions: the averaged percentage correct of the 26 proteins is increased from 45·3% or 49% up to 62·2% or 62·7%, depending on the predictive program. Both predictive programs yield about the same overall accuracy of predic-

tion with a slight advantage in favour of the directional program. For this reason and for the simplification brought by the use of a run constant of unity as seen above, the directional program based on equation (1) was used for further investigations.

Plots of the percentage of a given observed conformation H, E, T or C *versus* the individually optimized decision constants are presented in Figures 4 to 6. Although the points are rather scattered, the higher the content of a given conformation the more negative the optimal value of decision constant. Since this negative value is subtracted from the information values at each residue, this means that α-helix-rich proteins have additional information in their sequence leading to further helix formation and, inversely, helix-poor proteins carry negative information to make less helices. The same seems to hold for the extended conformation. It may be noted that the accuracy index gives equal weighting to conformations irrespective of their abundance so that, for example, predictions for helix-rich proteins are not improved simply because of an over-prediction of α-helix.

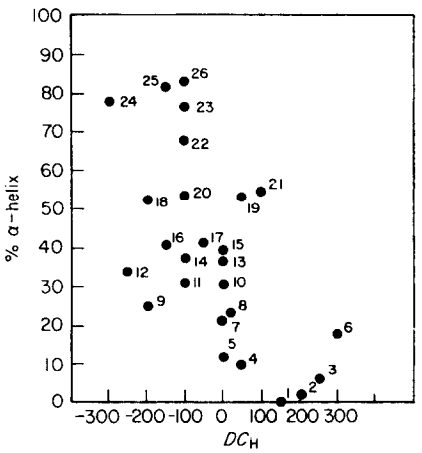

Figure 4.

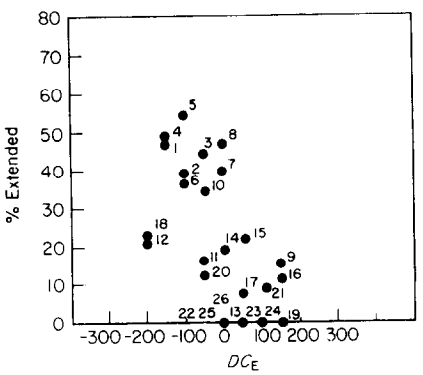

Figure 5.

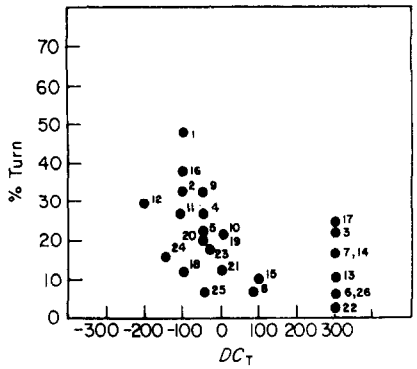

Figure 6.

FIG. 4 to 6. Observed percentages of α-helix (H), extended (E) and reverse turns (T) for 26 proteins *versus* the optimized decision constants for α-helix (DC_H), extended (DC_E) and reverse turns (DC_T), respectively, with the directional program. The numbers refer to the proteins as listed in Table 5. Decision constants in centinats.

On the other hand, one may notice that there is a kind of balance between the two periodic structures, α-helix and extended (Fig. 7). The α-helix content varies from zero to 83% in the 26 proteins and the content of extended structure from zero to 54%, but the sum of the contents of the two structures varies only from 40% to 83% (60% on average for these 26 proteins). If one includes the third regular conformation, the reverse turn, which itself varies from 3% to 47% according to the protein, the sum of the three conformation contents varies even less from 73% (except carbonic anhydrase with 61%) to 94% (80% on average for 26 proteins). The fourth conformation, coil, is quantitatively a minor conformation, 5% to 30%, with an average of only 20% for 26 proteins.

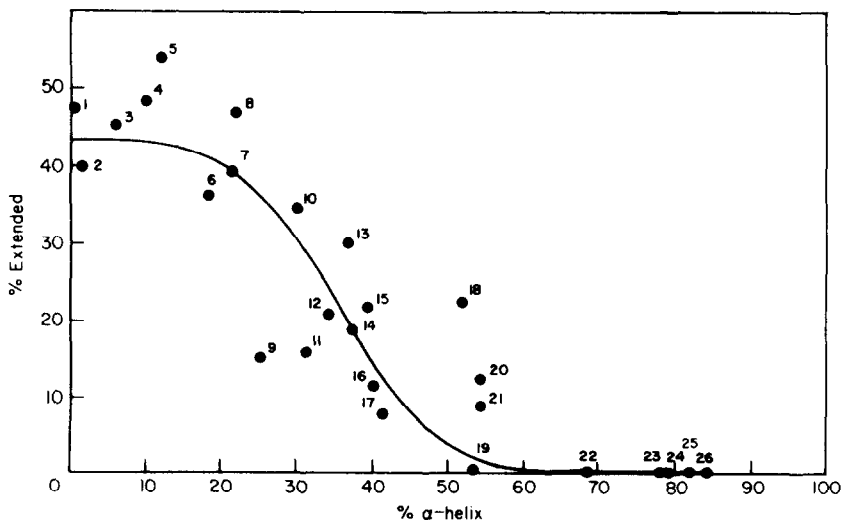

FIG. 7. Observed percentages of extended *versus* α-helix conformation contents for 26 proteins. The numbers refer to the proteins as listed in Table 5.

(c) *Choice of decision constants for the prediction of secondary structure from an amino acid sequence*

From the above results the choice of a correct set of decision constants is a definitive step towards improvement of a predictive program. Whatever its physical significance (see Discussion and Conclusions), the improvement is such that this choice is critical. Since our results suggested some relation between the content of a given conformation and the value of the corresponding decision constant, we recalculated for each protein the decision constants for the three conformations H, E and T from the linear regression line of the percentage of the observed conformation *versus* the optimized decision constants (Figs 4 to 6). The use of this new set of decision constants allowed us to predict correctly 54·8% of the residues for the four conformations of the 26 proteins, which is an improvement compared to a prediction with all the decision constants equal to zero (49%, see Table 5). However, the linear correlation is poor and suggests that a more complex relation exists. Because of this, and because this procedure is not recommended, we have not included the regression line in Figures 4 to 6.

Several attempts were therefore made to choose only a small number of decision constants for each conformational state depending on the range of content of that state rather than to use individual decision constants from the regression line. From inspection of Figures 4 to 6 and after different trials, three values of decision constant for the α-helix conformation and two values of decision constant for the extended conformation were retained (see Table 6). They were calculated as the average of the optimal decision constants of the proteins having an α-helix content less than 20%, between 20% and 50% and over 50%, i.e. $DC_H = 158$, $DC_H = -75$ and $DC_H = -100$, respectively. For extended conformation content below 20% the average decision constant was $DC_E = 50$ and above 20%, $DC_E = -87{\cdot}5$. With $DC_T = 0$ and $DC_C = 0$, this new set of decision constants applied to the prediction of the 26 proteins according to their observed content by X-ray of α-helix and extended conformations yielded an overall percentage of correctly predicted residues of 56·9% (Table 5). This constitutes a significant improvement compared to the use of the regression line. Other sets of decision constants according to different ranges of secondary structure content failed to yield better predictions. Attempts to define a more precise decision constant for reverse turn other than zero failed also to improve the prediction. This is certainly related in part to the greater scattering of the optimized decision constants for turns especially for a significant number of proteins requiring a high decision constant irrespective of their content in reverse turns (Fig. 6).

The choice of a decision constant requires an independent measurement of α-helix and β-sheet content. This can be achieved easily by circular dichroism or optical rotatory measurements. Although this method is not so accurate as was expected (Garnier *et al.*, 1976), it is sufficient to determine the range of secondary structure contents needed to choose the decision constants from Table 5 (see Discussion and Conclusions).

As an example of the usefulness of such a procedure, the predictions were made on 17 proteins with a choice of decision constants based on Table 6 and secondary structure contents from circular dichroism data as compiled by Chou & Fasman (1974). This yielded 56·1% of residues correct for the four states, which is slightly

TABLE 6

Best set of decision constants related to the secondary structure content of the protein

% of secondary structure (α-helix or extended)	Decision constants† α-Helix DC_H (centinats)	Extended DC_S (centinats)
I. less than 20%	158	50
II. between 20% and 50%	−75	−87·5
III. over 50%	−100	−87·5

† These constants have to be subtracted from the information measure at each residue, DC_T and DC_C being equal to zero (see text).

less than the value obtained with the use of more accurate secondary structure contents from X-ray results (57·5%). On the same 17 proteins, when all the decision constants are made equal to zero, only 50% of the residues were predicted correctly.

When experimental measurements of secondary structure contents are not available, the choice of the decision constants DC_H and DC_E can still be made from the α-helix and β-sheet contents obtained from a preliminary prediction with all decision constants equal to zero. For 26 proteins this procedure yielded 52% of correctly predicted residues for the four states, instead of 49% with all decision constants equal to zero. This should be recommended when no circular dichroism data are available, even though this is most effective for the α-helix-rich proteins (myoglobin, haemoglobins and erythrocruorin).

The predictive program (directional) that we recommend allows the prediction of four conformations including the coil conformation. From our data it is certainly the latter that is least well-predicted. For instance, with optimized and directional predictions made on the 26 proteins included in Table 5, 66·1% on average of the residues observed as α-helix were correctly predicted as helix, 55·2% as extended, 42·1% as reverse turns but only 28·6% as coil. This might arise from the rich variety of aperiodic structure implied by the "coil" conformation in the 26 proteins analyzed. From our data we observed that the lowest values of correctly predicted residues in a given protein for a given conformation often occurred in cases where the content in that conformation was low.

(d) *Comparison with other prediction methods using proteins which have been extensively studied as test cases*

For two proteins, one of which, bacteriophage T4 lysozyme, was not included in our statistics for information measures, the comparison was made between our predictions and those made by several authors (Schulz *et al.*, 1974; Matthews, 1975) at a time when the X-ray results were not yet known (Tables 7 and 8). In this very limited example, our predictions, with different sets of decision constants, were better for α-helix and somewhat similar for extended and reverse turns individually taken. Based on the predictions of two conformations H and E, they were equal to

TABLE 7

Conformation prediction for T4 phage lysozyme†

Method of prediction	Number of residues correctly predicted H	E	T	C	Percentage of correctly predicted residues %HETC	%HET‡	%HE‡
Directional with optimized *DC*	67	5	0	27	60·4	61	68
Directional with *DC* from Table 6	71	4	3	9	53	66·1	70·7
Schellman	67	8					70·7
Ptitsyn & Finkelstein	67	0	4			60	63·2
Prothero	58						
Lim	56	0					52·8
Nagano & Hasegewa	51	2	5			49	50
Chou & Fasman	50	5	4			50	51·9
Barry & Friedman	88	6					41·5
Burgess *et al.*	38	0	6			37·3	35·8
Observed	94	12	12	46			

† The predictions of the different authors are taken from Matthews (1975).

‡ Evaluated only at H, E or T regions which are the only available data from the published paper.

TABLE 8

Conformation prediction for adenylate kinase†

Method of prediction	Number of residues correctly predicted H	E	T	C	Percentage of correctly predicted residues %HETC	%HET‡	%HE‡
Directional with optimized *DC*	92	19	24	6	73·7	80·4	86
Directional with *DC* from Table 6	102	16	11	9	71·1	76·8	91·5
Lim	82	13					73·6
Finkelstein & Ptitsyn	79	13					71·3
Chou & Fasman	70	20	28			70·2	69·8
Nagano	61	22	38			72·0	64·3
Barry & Friedman	56						
Burgess & Scheraga	46		33				
Levitt & Robson	42						
Observed	105	24	39	26			

† The predictions are taken from Schulz *et al.* (1974).

‡ Evaluated only at H, E or T regions which are the only available data from the published paper.

one prediction for T4 phage lysozyme, and better for all the others. When we considered the joint predictions for the three conformations H, E and T, whenever it was possible (i.e. when they were made by the same authors), our predictions were always better.

(e) *Taking account of the clustering of non-polar residues*

Any region of the protein chain will usually be in contact with a non-polar region of the protein interior on at least one side. The stereochemistry of the α-helix thereby

implies that if a hydrophobic residue occurs at position j in the sequence, then the residues at position $j \pm 1$, $j \pm 3$ and $j \pm 4$ are likely to be hydrophobic. This will tend to generate a hydrophobic cluster of residues on the α-helix surface, since these separations bring the side-chains together in space. Preliminary studies suggest that consideration of hydrophobic residues at $j \pm 1$ does not generally enhance predictions, because strongly helix-forming residues already provide strong information that their neighbours will be helical. Such considerations have been used to make predictions (Schiffer & Edmundson, 1967; Lim, 1974*a*,*b*).

Robson & Pain (1971) first noted that including predictions of helical regions by the "helical wheel" approach of Schiffer & Edmundson (1967) usually simply confirmed helical regions already predicted by single and pairwise information statistics, and in those occasional regions where improvement resulted, this was compensated by the introduction of errors elsewhere. This was taken to imply that information in hydrophobic clustering is *degenerate*, and is involved not in the calling down of information for defining helical regions, but in the next stage of protein folding by providing hydrophobic facets on helices on which other chain conformations are laid down.

Identical results have been obtained by Nagano (1977), and his interpretation is basically the same, although expressed in the more recent terms of "supersecondary structure". Of course, the fact that the clustering of hydrophobic residues does not determine helical regions is quite distinct from saying that it does not strengthen them, and statistical (Nagano, 1977), statistical mechanical (Suzuki & Robson, 1976) and experimental (Robson & Pain, 1976) studies suggest that helices are stabilized in this way.

Because of the obvious tendency of helical residues to form such clusters, and the current evidence that paradoxically this seems to provide information only for tertiary interactions, we have tested this phenomenon in some detail. Figure 8 shows the information for residues at position j in a helix to have a hydrophobic neighbour at $j - 3$ or $j - 4$, and at $j + 3$ or $j + 4$, that is to have at least two hydrophobic neighbours on the helix surface, one of each in each direction. This information is plotted as a function of the hydrophobicity of the side-chain of j, measured as the free energy of transfer of this side-chain from water to ethanol (Nozaki & Tanford, 1970). Similar plots are also obtained from the tendency of a helical residue to form non-polar contacts with at least two hydrophobic residues not on the helix surface but brought close to j by tertiary folding (a statistical study of hydrophobic clustering of residues produced by tertiary packing will be submitted shortly). Reiterative studies of this type were also used to ensure that our classification of hydrophobic and hydrophilic side-chains was reasonable: only alanine, valine, leucine, isoleucine, phenylalanine, tyrosine, tryptophan and methionine consistently gave optimal information values of 40 ± 20 centinats. Repeating the above studies by, for example, including cysteine or cystine as hydrophobic, or treating tyrosine, with some polar character, as a member of the hydrophilic group, gave plots similar to Figure 8 but in which the higher values did not reach 40 centinats.

Predictions were then made in which 40 centinats were added to the jth residue when residues $j - 3$ or $j - 4$, and $j + 3$ or $j + 4$ and j itself were hydrophobic. Since this produced no improvement, 40 centinats were then added to the information carried in residues including and between the first and last hydrophobic residues in the cluster. Other investigations were made in which 40 centinats were also added

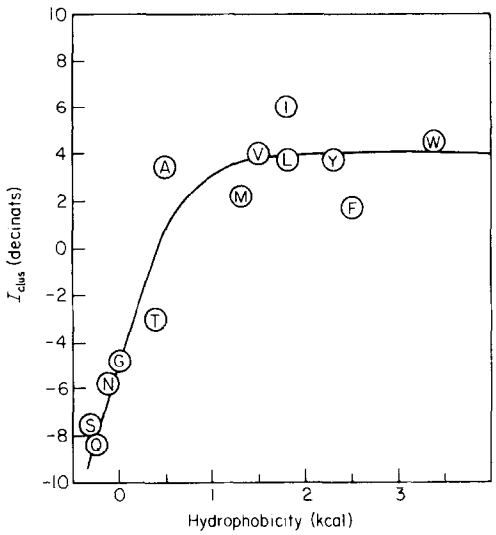

FIG. 8. Information (eqn (2)) for forming clusters, I_{clus}, *versus* the hydrophobicity of the amino acid side-chain R_j(IUPAC 1-letter code). The hydrophobicity scale is from Nozaki & Tanford (1970). This shows that if helix residues are hydrophobic, then their neighbours on the α-helix surface tend to be more hydrophobic. It also quantifies the effect, and reveals that 3 hydrophobic residues at R_j, R_{j+3}, R_{j+4} contribute about 4 decinats. This information is, however, usually degenerate (see text).

for polar clusters, subtracted for periodicities characteristic of β-pleated sheet, and so on. Some 20 different ways of including helical wheel effects have been tested, and most of these were also tested with values other than 40 centinats ranging up to 200 centinats. None of these significantly enhanced the predictions.

5. Discussion and Conclusions

(a) *Suitability of these methods as a standard predictive tool*

The methods presented here compare favourably with those of other authors when comparison is made using the same accuracy statistics. Apart from general performance of the algorithms, our most objective evidence for this statement is a direct comparison with the predictions of Rerat & Rerat (personal communication) using a computer program of the Chou & Fasman (1974) method on 21 proteins, and the direct comparison for T4 phage lysozyme and adenylate kinase (Tables 7 to 8) for which many authors have quoted detailed results. We have not used other prediction procedures (except for that of Chou & Fasman), because writing computer programs for these methods is invariably less satisfactory in the hands of other workers. This is because of occasional ambiguities in the description of the method, and in this respect we concede that the Chou & Fasman method may yield somewhat improved results in their hands. This emphasizes that a predictive method, however apparently simple it may be, is best presented as an algorithm (e.g. eqns (1) and (3)). Only in this way can ambiguities be resolved and objective predictions guaranteed.

As we have shown the decision constant may be altered to further improve the prediction of β-sheet-rich proteins or α-helix-rich proteins when an independent rough estimate of β-sheet and α-helix content is available. This can easily be achieved from

an analysis of the optical activity, mainly the circular dichroism spectra. Altogether we estimate that the relative experimental variation from circular dichroism will be within a range of $\pm 15\%$ of the content of α-helix or β-pleated sheet, and will result in predictions with overall accuracies of about 60% residues correct (4-state prediction).

(b) *Further improving the accuracy of predictions*

A fraction of about 60% residues correct for a four-state prediction of helix, β-pleated sheet, reverse turns and coil, appears to be the maximum predictive power available with the procedures described here. This value is, of course, obtained by predicting a large set of proteins: larger or smaller values will be obtained for individual proteins. Nevertheless, the idea of a limit for the accuracy of predicting secondary structures has a clear theoretical basis: tertiary interactions between residues far apart in the amino acid sequence must override the intrinsic conformational tendencies of many residues in order to achieve a compact, globular structure.

A way of overcoming this problem and improving predictions is to take into account the amino acid sequences of homologous proteins which are also believed to have the same or similar secondary and tertiary structure. In scanning through the homologous sequences at the locus of the jth residue, the information provided by each residue of the homologue is simply added and the sum divided by the number of homologues. With the single-residue information program, this averaging is done prior to the subtraction of the decision constants and prior to taking into account the run constants. We recommend this procedure whenever homologues are available as, on the basis of preliminary work, improvements of 5 to 10% residues correct are probable. The option to use it has been built into the program, which is available on request from J. Garnier.

(c) *Theoretical implications*

The above findings establish directional information (Tables 1 to 4) as a very important determinant of secondary structure in globular proteins. Although directional information implies interactions between, at most, two residues, only one side-chain need be known and this has been interpreted as due to an interaction between a side-chain and the backbone of residues further along the polypeptide chain in both directions (Robson & Suzuki, 1976).

Directional information also includes information concerning the co-operativity effect, so that no algorithm (e.g. eqn (3)) to represent this effect need be included.

As well as interactions between a side-chain and the backbone of residues up to eight residues distant, "long-range" interactions (between residues far apart in the amino acid sequence) are also relevant to good predictions, as revealed by consideration of protein classes. In conjunction with Figure 7, the results suggest that the simplest assignment of classes is into (1) α-helix-rich proteins, (2) β-pleated-sheet-rich proteins, and (3) proteins which do not contain large amounts of either type of structure.

Long-range interactions which differ between these classes appear to be particularly significant because there is a fine balance between the α-helical and extended-chain conformations of many regions. This is because many residues have strong tendencies to be both α-helix and extended-chain formers, even though in some cases certain crucial residues (such as the very strong helix-former glutamate) play a major part in determining the final conformation of the neighbouring, less partial residues.

Evidence for a fine balance is found in Figures 1 to 3, which show that optimal predictions run along a diagonal from the bottom left to the approximate origin ($DC_E = 0$, $DC_H = 0$) of the diagrams, so that a large negative value of DC_H (increasing the amount of helix predicted) also requires a large negative value of DC_E (increasing the amount of extended chain predicted) to maintain the high accuracy of predictions. A relatively small change in DC_E or DC_H alone, which may be interpreted as additional information concerning extended chain or α-helix content, will swing a prediction in favour of α-helix or extended chain. From the point of view of improving predictions in other proteins, this may be information from experimental studies such as circular dichroism measurements of helix or pleated-sheet content. From the point of view of protein folding dictated by the amino acid sequence, this information is coded as the long-range interactions between regions which have innate tendencies to be either helix or extended chain.

A mechanism for co-operativity between extended-chain regions is easily envisaged as arising from hydrogen bonding between regions of the backbone lying side by side in parallel or anti-parallel alignment to form continuous pleated-sheet regions. This co-operativity would arise from (1) the free energy required to initiate pleated sheet by placing two backbone regions side by side with the appropriate backbone conformation, (2) the probability of encountering other extended regions and the free energy gain in adding successive extended regions to the pleated sheet. Co-operativity is also possible between different α-helices (see Robson & Suzuki, 1976) by means of contacts between hydrophobic regions of the helix surfaces, although hydrophobic contacts presumably may also stabilize helix to pleated-sheet contacts. However, the results given in this paper could be adequately explained by considering only a simple model in which only the interaction between extended chains is co-operative. In proteins which have many regions with a slightly greater tendency to form extended chain than α-helix, co-operativity between these and other, less decisive regions will pull the latter into the extended-chain conformation. Hence regions which might have otherwise been α-helical then appear as components of β-pleated sheet. Conversely, in proteins with few regions with a marginal preference for the extended conformation, all marginal preferences for α-helix formation may be more readily expressed.

APPENDIX

Assessment of the Accuracy of Predictions

(a) *The accuracy index* (α)

The accuracy index used here is that of Robson (1974) extended to the case of more than two conformational states:

$$\alpha = \sum_S \frac{F_S^+}{F_S} - 1, \tag{4}$$

where F_S is the number of residues observed in state S and F_S^+ is the number of residues which are both observed in state S and correctly predicted to be in that state. The value of the information provided by a prediction algorithm is then:

$$I_a = \ln(1 + \alpha). \tag{5}$$

Note the following properties of α and I_a:

(1) If all residues are correctly predicted, and there are N different states

$$\alpha = N - 1 \tag{6}$$

and

$$I_a = \ln N. \tag{7}$$

(2) If all predicted states are assigned at random then for large F_S values, as the probability of correctly assigning any one conformational state is $1/N$:

$$\alpha = \sum_S \frac{(1/N)F_S}{F_S} - 1 = N(1/N) - 1 = 0 \tag{8}$$

and $I_a = 0$.

(3) If N_+ states can be predicted with certainty and the rest are assigned entirely at random:

$$\alpha = N_+ + (N - N_+) \cdot \frac{(1/N) \cdot F_S}{F_S} - 1 = N_+ - \frac{N_+}{N}$$

$$(= N_+, \text{ when } N \gg N_+) \tag{9}$$

and

$$\ln(N_+) \leqslant I_a \leqslant \ln(N_+ + 1). \tag{10}$$

Thus, to a first approximation, the accuracy index can be interpreted as the number of types of conformational state which may be predicted perfectly by an algorithm carrying the same amount of information. The information I_a provided by an algorithm is increased by increasing both its accuracy and its resolution (i.e. the number of types of conformational state considered).

(b) *Fraction of residues correct*

A commonly used index of predictive power is the fraction of residues correctly assigned, f, where

$$f = \frac{\sum_S F_S^+}{\sum_S F_S}, \tag{11}$$

which, for consistency with other work, is conveniently expressed on a percentage basis ($f \times 100$).

If there are N possible states S, then the value of f expected on a purely random assignment, is

$$f = \frac{1}{N}. \tag{12}$$

Of interest in this work are predictions based on two-state (set S = helix, non-helix), three-state (set S = helix, pleated-sheet, not-helix nor pleated-sheet) and four-state (set S = helix, pleated-sheet, reverse turn, coil).

We are thankful to Professor R. Pain for critically reading the manuscript.

REFERENCES

Burgess, A. W. & Scheraga, H. A. (1975). *Proc. Nat. Acad. Sci., U.S.A.* **72**, 1221–1225.
Chou, P. Y. & Fasman, G. D. (1974). *Biochemistry*, **13**, 222–245.
Garnier, J., Salesse, R., Rerat, B., Rerat, C. & Blake, C. (1976). *J. Chim. Phys.* **73**, 1018–1023.

Kotelchuck, D. & Scheraga, H. A. (1969). *Proc. Nat. Acad. Sci., U.S.A.* **62**, 14–21.
Lim, V. I. (1974*a*). *J. Mol. Biol.* **88**, 857–872.
Lim, V. I. (1974*b*). *J. Mol. Biol.* **88**, 873–894.
Mathews, B. W. (1975). *Biochim. Biophys. Acta*, **405**, 442–451.
Nagano, K. (1977). *J. Mol. Biol.* **109**, 251–274.
Nozaki, Y. & Tanford, C. (1970). *J. Biol. Chem.* **246**, 2211–2217.
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. (1974*a*). *Biochem. J.* **141**, 869–882.
Robson, B. & Pain, R. H. (1974*b*). *Biochem. J.* **141**, 883–897.
Robson, B. & Pain, R. H. (1974*c*). *Biochem. J.* **141**, 899–904.
Robson, B. & Pain, R. H. (1976). *Biochem. J.* **155**, 331.
Robson, B. & Suzuki, E. (1976). *J. Mol. Biol.* **107**, 327–356.
Schiffer, M. & Edmundson, A. B. (1967). *Biophys. J.* **7**, 121–135.
Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, B. & Nagano, K. (1974). *Nature (London)*, **250**, 140–142.
Suzuki, E. & Robson, B. (1976). *J. Mol. Biol.* **107**, 357–362.

Note added in proof: We thought it worth mentioning that, in connection with equation (9) (Appendix), a rational person would not assign states "entirely at random" but would make use of the fact that the actual state cannot be one of the N_+ states which he is helped to predict with certainty. Hence $1/N$ would then be replaced by $1/(N-N_+)$ in equation (9), and $\alpha = N_+$. The computer program would not, of course, make use of this extra information provided by logic. Either way, the accuracy is seen to be a good estimate of the number of states which could be predicted perfectly with quantitatively (though not qualitatively) similar information.