Status of Empirical Methods for the Prediction of Protein Backbone Topography†

Frederick R. Maxfield‡ and Harold A. Scheraga*

ABSTRACT: An empirical prediction algorithm which uses information on the short-range (intraresidue) and medium-range (up to four neighbors on either side) interactions in 20 proteins to assign every residue in a protein to one of five conformational states is described. The conformational states are defined in terms of the backbone dihedral angles of the residue so that the prediction algorithm can be used to generate starting conformations for subsequent energy-minimization procedures, which would be necessary to predict the three-dimensional structure of a protein. An estimate is made of the statistical error in the determination of the parameters describing the effects of short-range and medium-range interactions in proteins, and it is shown that this statistical error plays a large role in limiting the accuracy of all prediction methods which rely on data from proteins of known structure. Using the method described in this paper, 56% of the residues in 20 proteins were assigned correctly to one of five conformational states. It seems unlikely that any prediction method can significantly improve on this accuracy for assigning residues to specific backbone conformations unless the size of the data base is increased greatly.

The determination of the amino acid sequences of many proteins and the elucidation of the three-dimensional structure of a few of these proteins have been followed by the development of several empirical algorithms which attempt to predict the backbone topography of a protein from a knowledge of its amino acid sequence. A review of many of the earlier prediction methods has been given in a paper which also reported the development of a new algorithm (Burgess et al., 1974). New prediction methods have recently been reported by Lim (1974b) and by Robson and Pain (1974a). One of the motivations for the development of these methods has been the suggestion that prediction algorithms might lead to starting conformations (topographical structures) from which energy minimization would lead to the correct native structure of the protein, thus avoiding the multiple-minimum problem (Scheraga, 1974). Some other workers (Chou and Fasman, 1974) have claimed that, even without energy minimization, prediction algorithms can be used to obtain information about the correlation between protein conformation and biological activity. However, it has been shown that even a perfect algorithm, which correctly assigned every residue in bovine pancreatic trypsin inhibitor to one of five conformational states, would not lead to the native structure without the introduction of additional information such as the formation of disulfide bonds and energy minimization (Burgess and Scheraga, 1975). This study indicated that small errors in the value of each dihedral angle in trypsin inhibitor resulted in the generation of a structure which was grossly different from the native structure. Since current prediction algorithms are still far from being able to predict the conformation of every residue in a protein, it should be clear that they cannot, by themselves, explain biological properties which depend on the precise three-dimensional arrangement of the atoms in a protein molecule. These limitations should be recognized to avoid overestimating the power of these algorithms.

In order to be useful for the generation of starting conformations, an algorithm should assign *each* residue in a protein to as small a region of conformational space as possible. It is necessary to assign each residue to a conformational state with specific dihedral angles since assignments such as "coil" or "bend" leave considerable ambiguity as to the values of the dihedral angles in a starting conformation. Since present prediction algorithms do not assign every residue correctly, it is useful to have some estimate of the reliability of the prediction for each residue. In an energy-minimization procedure, it might be possible to vary the dihedral angles of residues with less reliable predictions to a greater extent than the dihedral angles of residues with highly reliable predictions.

In this paper, a new prediction algorithm is described which assigns each residue to one of five conformational states. The assignment is based on the influences of both short-range (intraresidue) and medium-range (four neighbors on either side) interactions on the conformational preferences of residues. The medium-range interactions were carried out to the fourth nearest neighbor since it has been shown previously that the information contained in a nine-residue segment is frequently sufficient to determine the conformation of the central residue (Ponnuswamy et al., 1973; Burgess and Scheraga, 1975).

The estimates of the short-range and medium-range influences on the conformation of a residue were obtained from an analysis of the sequences and three-dimensional structures of 20 proteins with a total of 3681 residues. While the number of residues in the data set is sufficient for an analysis of the conformational preference of each amino acid based on intraresidue information, the consideration of medium-range interactions is statistically less reliable. Regarding a pairwise interaction as one that influences the conformation of an amino acid residue at position i when residue i interacts with another residue j units away (i.e., $i + j$, up to $|j| = 4$), and assuming that this interaction is independent of interactions between i and residues other than the $(i + j)$th, then the residues in the

† From the Department of Chemistry, Cornell University, Ithaca, New York 14853. *Received February 19, 1976.* This work was supported by research grants from the National Institute of General Medical Sciences, United States Public Health Service (GM-14312), and from the National Science Foundation (BMS75-08691).

‡ National Science Foundation Predoctoral Trainee, 1971-1974; National Institutes of Health Predoctoral Trainee, 1974-1976.

data set are distributed among 400 possible pairs so that there are an average of about nine observations per pairwise interaction. For triplets (i.e., the influence of a pair of residues, j and j' units away from i, up to $|j|, |j'| = 4$ on the conformation of the residue at position i), there would be an average of about 0.5 observation per triplet. For this reason, the method described in this paper is restricted to a consideration of intraresidue and pairwise interactions only.

Also, the method is designed so that pairwise interactions can be included in a manner which weights the pairs with many observations more heavily than those with only a few observations. An estimate is made of the statistical error in each prediction due to the size of the data set, and the importance of this error is discussed. Finally, the results obtained by this method and other recent prediction algorithms are discussed, and their suitability for determining starting conformations for energy minimization, and for other purposes, is evaluated.

Methods

Backbone dihedral angles were determined from the x-ray coordinates of the following proteins: myoglobin (Watson, 1969), thermolysin (Matthews et al., 1974), cytochrome b_5 (Mathews et al., 1972), tosyl elastase (Shotton and Watson, 1970; Shotton and Hartley, 1973), A, B, and C chains of tosyl-α-chymotrypsin (Birktoft and Blow, 1972), carboxypeptidase A (Quiocho and Lipscomb, 1971), papain (Drenth et al., 1971), bovine pancreatic trypsin inhibitor (Deisenhofer and Steigemann, 1975), concanavalin A (Edelman et al., 1972), high-potential iron protein (Carter et al., 1974; Freer et al., 1975), hen-egg-white lysozyme (Diamond, 1974), carp myogen (Moews and Kretsinger, 1975), subtilisin BPN′ (Alden et al., 1971), ribonuclease S (Fletterick and Wyckoff, 1975), staphylococcal nuclease (Arnone et al., 1971), cytochrome c_2 (Salemme et al., 1973a,b), D-glyceraldehyde-3-phosphate dehydrogenase (Buehner et al., 1974a,b), clostridial flavodoxin (Burnett et al., 1974), A and B chains of human deoxyhemoglobin (Fermi, 1975), and sea lamprey hemoglobin (Hendrickson and Love, 1971; Hendrickson et al., 1973). The coordinates for all of these proteins, except tosyl elastase and trypsin inhibitor, were obtained from the Protein Data Bank, Brookhaven National Laboratories, Upton, N.Y. For several proteins, the sequences used in the crystallographic papers have been revised to agree with more recent sequence determinations. For five residues, we were unable to resolve the difference between the sequence reported by the crystallographer and that determined by chemical sequencing, and these five residues were not used in our data set. For some of the proteins, the residues near the N or C terminus did not have any coordinates reported, and these residues were, therefore, excluded from the data set. Since the value of one of the backbone dihedral angles could not be determined for the first and last residue of every protein chain, these residues were not included in the data set. The complete list of residues which were used is part of the data available in the supplementary material (see paragraph at the end of this paper concerning supplementary material).

Calculations were performed on an IBM 370/168 computer system. The programs used to make the predictions, a program user's guide, and the complete set of data (including the sequences and backbone conformational states of the proteins used in this study) are available in the supplementary material.

Definitions of Conformational States. Each residue is assigned to one of five conformational states. For all amino acids except glycine, the conformational states were defined in the following way:

ϵ: $-180° \le \phi \le 0°$ and $-180° \le \psi < -120°$, or
$-180° \le \phi \le 0°$ and $100° \le \psi < 180°$, or
$140° \le \phi < 180°$ and $-180° \le \psi < -120°$, or
$140° \le \phi < 180°$ and $10° \le \psi < 180°$;
α_R: $-180° \le \phi \le 0°$ and $-120° \le \psi \le 10°$, or
$140° \le \phi < 180°$ and $-120° \le \psi < 10°$;
ζ_R: $-180° \le \phi \le 0°$ and $10° < \psi < 100°$;
α_L: $0° < \phi < 140°$ and $-10° \le \psi \le 110°$;
ζ_L: $0° < \phi < 140°$ and $-180° \le \psi < -10°$, or
$0° < \phi < 140°$ and $110° < \psi < 180°$

These definitions are similar to those of Burgess et al. (1974), except that they cover the whole ϕ, ψ conformational space.[1] The boundaries between these states were drawn through regions of conformational space which had a low frequency of occurrence for all of the amino acids except glycine. In order to account for the flexibility and lack of an asymmetric carbon in glycine, the conformational states were defined slightly differently for this residue, viz., as follows:

ϵ: $-180° \le \phi < 180°$ and $110° \le \psi < 180°$, or
$-180° \le \phi < 180°$ and $-180° \le \psi \le -110°$;
α_R: $-180° \le \phi \le 0°$ and $-110° < \psi \le 0°$;
ζ_R: $-180° \le \phi \le 0°$ and $0° < \psi < 110°$;
α_L: $0° < \phi < 180°$ and $0° \le \psi < 110°$;
ζ_L: $0° < \phi < 180°$ and $-110° < \psi < 0°$

In addition, if four or more consecutive residues have dihedral angles, $-130° \le \phi \le -10°$ and $-90° \le \psi \le -10°$, these residues are considered to be part of a right-handed α helix (Burgess et al., 1974) and are designated α_h. These residues are considered separately from other α_R residues to allow for differences in the effects of medium-range interactions on residues in α helices compared with isolated α_R residues.

The conformational states used in this analysis are shown in Figure 1 along with the frequencies of occurrence of the values of ϕ, ψ within a 10° grid for alanine and glycine.

Classification of Data on Short-Range and Medium-Range Interactions. Throughout this paper, we will be concerned with the conformation of a residue at position i in a protein. The conformational state of residue i will be designated by the index k, and the type of amino acid residue at position i will be indicated by the index m. When the effects of neighboring residues on the conformation of residue i are considered, the index l will be used to indicate the type of amino acid residue at position $i + j$. The use of these indices is described in the following sections.

Intraresidue Information. The number of occurrences, n_{km}, of amino acid m ($1 \le m \le 20$) in each of the conformational states k (k = 1 to 6 designating states ϵ, α_h, ζ_R, α_L, ζ_L, and α_R, respectively) was determined, and the relative frequencies of occurrence in each state were used as an estimate of the short-range conformational preference of amino acid m.

Nonspecific Medium-Range Interactions. The number of occurrences, n_{jkl}, of *any* amino acid residue in a conformational state k at position i in a protein chain when the residue at position $i + j$ is a specific amino acid l ($1 \le l \le 20$) was tabulated for values of j from -4 to 4 (excluding zero, which would be equivalent to intraresidue information). The relative

[1] The abbreviations and symbols for the description of the conformation of polypeptide chains conform to the rules adopted by an IUPAC-IUB Commission on Biochemical Nomenclature (1970), *Biochemistry 9*, 3471.

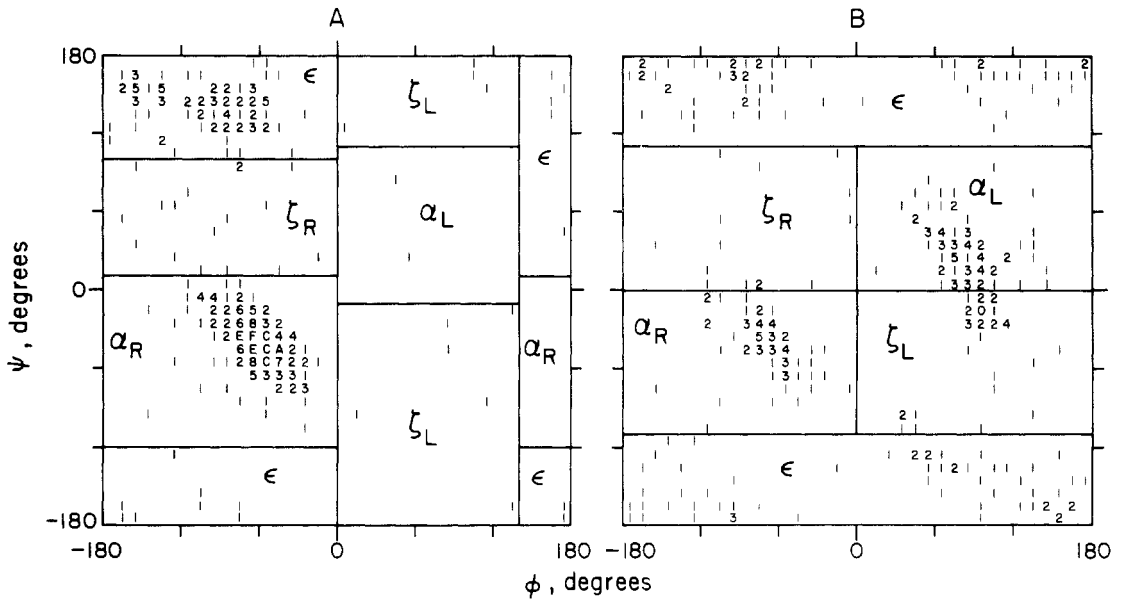

FIGURE 1: Distribution of the occurrences of ϕ, ψ conformations for 20 proteins. (A) Occurrence of Ala residues and the conformations used for all residues except Gly. (B) Occurrrences and definitions of conformations for Gly. The symbols A through F within the diagram represent 10 through 15 occurrences, respectively.

frequencies of occurrence in each state, for given values of j and l, are an estimate of the conformational preference of residue i when residue $i + j$ is amino acid l, *independent* of the identity of the amino acid at position i. These interactions are designated "nonspecific" to distinguish them from specific medium-range interactions, described below, which do take into account the identity of the amino acid at position i (as well as its conformation). For cases where $i + j < 1$ or $i + j > N$, where N is the number of residues in the protein chain, l is assigned the value 21, and values of n_{jkl} were also tabulated for $l = 21$. For negative values of j with $l = 21$, the relative frequencies of occurrence in the six states are an estimate of the conformational preference for any amino acid i when it is within $|j| - 1$ residues of the N-terminal residue. Similarly, for $j > 0$ and $l = 21$, the relative frequencies are an estimate of the conformational preference for any amino acid when it is within $j - 1$ residues of the C-terminal residue. This computational device enables one to detect the influence of the ends of the chain on residues near the ends.

Specific Medium-Range Interactions. The number of occurrences, n_{jklm}, of an amino acid residue m with conformation k at position i in a protein chain, when the residue at position $i + j$ is amino acid l, was tabulated for values of j from -4 to 4 (again excluding zero). As with the nonspecific medium-range interactions, when $i + j < 1$ or $i + j > N$, l is assigned the value 21. The relative frequencies of occurrence in each state, for fixed j, l, and m, are an estimate of the conformational preference of amino acid m at position i in a protein when residue $i + j$ is amino acid l.

Contingency Tables. An $r \times s$ contingency table is obtained when a number of items are classified into the exhaustive and exclusive classes A_1 through A_r, and, at the same time, the items are classified into the exhaustive and exclusive classes B_1 through B_s. The probability $\Theta_{\alpha\beta}$ for an item to be classified as both A_α and B_β is assumed to be the same for each item. This assumption means that there are no special characteristics among the items classified as A_α which would alter the probability for those items to be classified as B_β. The numbers, $O_{\alpha\beta}$, are the number of items classified as both A_α and B_β. As an example of a 2×2 contingency table, A_1 could be used to designate all the Ala residues in the data set, and A_2 would then represent all other amino acids. B_1 might designate the ϵ conformation, with B_2 representing all other conformations. The number O_{11} would then be the number of Ala residues found in the ϵ conformation in the data set. If the probability of being in B_1 or B_2 were independent of the A_α classification, one would expect O_{11}/O_{12} to be nearly the same as O_{21}/O_{22}. There are several methods available to measure the significance of deviations from equality of O_{11}/O_{12} and O_{21}/O_{22}. One such method (Fisher, 1958) has been described previously (Maxfield and Scheraga, 1975), and it will be used in this paper to measure the probability, P, that deviations of O_{11}/O_{12} from O_{21}/O_{22} as large or larger than those actually observed could occur by chance. This test (which is independent of the Bayesian method described below) is used to determine whether a specific medium-range interaction is significant.

Prediction Method Based on a Bayesian Analysis of Contingency Tables. Bayesian methods have been applied previously in the development of a prediction algorithm (Robson and Pain, 1971, 1974a; Robson, 1974), and the theory describing the application of these methods has been discussed (Robson, 1974). While the methods described in this paper are similar to those developed by Robson and Pain, there are also some important differences. One major addition is that, in the method described in this paper, an estimate is made of the statistical error due to the size of the data set. Also, a much greater emphasis is placed on weighting the various terms used in making the prediction, based on their statistical significance.

While the prediction method in its final form is easy to apply, the description of the method is quite involved. In order to make the important features of the method clear, an outline of the prediction procedure is presented in this section, and the details of the method are described in the Appendix. The quantity used in this study to measure the tendency for a residue to be in a particular conformation is the natural logarithm of the odds in favor of finding the residue in that conformation. If Θ_k is the probability for a residue to be in a conformation, k, then $\ln [\Theta_k/(1 - \Theta_k)]$ is the logarithm of the odds in favor of the residue being in conformation k. The prediction proce-

dure makes an estimate of the logarithm of the odds for a residue to be in each of the six conformations, based on short-range and medium-range interactions. The predicted conformation for a residue is that conformation which is estimated to have the largest value of $\ln [\Theta_k/(1 - \Theta_k)]$.

The effects of short-range interactions are estimated using the values of n_{km} shown in Table I. An estimate of the logarithm of the odds based entirely on short-range interactions can then be obtained from eq A-8 of the Appendix. The logarithm of the odds for an amino acid residue to be in a conformation, k, is approximately equal to the logarithm of the ratio of the number of occurrences in conformation k to the number in all other conformations. The medium-range interactions are then used to modify this estimate, using the assumption that each of the eight medium-range interactions ($-4 \leq j \leq 4, j \neq 0$) involving residue i influence the conformation of residue i independently.

The nonspecific medium-range interactions are incorporated using the values of n_{jkl} (see Table II). The logarithm of the odds favoring one conformation over all others, based on short-range and nonspecific medium-range interactions, is estimated using the right-hand side of eq A-12 of the Appendix. Each nonspecific medium-range interaction contributes one term to the summation in eq A-12.

The effects of specific short-range interactions should be estimated using the values of n_{jklm} (see Table III), but in general there are not enough observations for each n_{jklm} to make a good estimate. Therefore, two procedures were adopted to reduce the errors caused by the small number of observations. First, the data on the specific medium-range interactions were examined to determine the level of confidence with which one could say that the specific interaction influenced the conformational preference of the residue being predicted. This level of confidence, which we call P_{jklm}, was calculated using the statistical test for 2×2 contingency tables. The method used to calculate P_{jklm} is given in the Appendix. A small value of P_{jklm} indicates that it is likely that the interaction being studied does influence the conformational preference of residue i. A large value of P_{jklm} indicates either that the interaction does not affect the conformational preference of residue i or that there are simply not enough data to say with confidence that the interaction is important; that is, a high value of P_{jklm} does not necessarily mean that the interaction does not influence the conformational preference of the residue. The significance test has been used to limit the specific medium-range interactions which are used in the prediction by considering only the interactions which had a value of P_{jklm} below some value, P_{max}, which was used as an adjustable parameter. If a pair of residues had $P_{jklm} > P_{max}$ for every conformation, the effect of that interaction was estimated using nonspecific medium-range interactions exclusively. For specific medium-range interactions which were judged to be significant (i.e., $P_{jklm} < P_{max}$ for at least one value of k), the term representing the effect of nonspecific medium-range interactions in the estimate of the logarithm of the odds was replaced by a term based on specific medium-range interactions. For $P_{max} = 0$, only nonspecific medium-range interactions are considered; when $P_{max} = 1$, all of the specific medium-range interactions are used.

The second method used to decrease the statistical error in the estimate of the parameters describing specific medium-range interactions was to bias the estimates of these parameters using the more reliable intraresidue information and nonspecific medium-range interactions. This bias was introduced by adding a fixed number of observations, S, to each row in the specific medium-range interaction contingency tables, with the added observations distributed among the conformations in the proportions that would be expected on the basis of intraresidue information and nonspecific interactions. For example, if residue i is Ala, and residue $i - 4$ is Asp, the specific medium-range interaction for $j = -4$ would be estimated using the second row of Table III with S additional observations distributed among the entries in this row. The distribution of these additional observations would be determined by the intraresidue information for Ala (row 1 of Table I) and the nonspecific medium-range interactions for Asp at position $i - 4$ (row 2 of Table II). This information is used in a manner described in the Appendix (eq A-15). The number S was treated as an adjustable parameter.

This procedure allows for a compromise between the statistically reliable nonspecific medium-range interactions and the less well-known (but potentially more important) specific medium-range interactions. The relative importance of the two types of medium-range interactions in the prediction method depends on the value of S. When S is small, the specific medium-range interactions will predominate, and, as S gets larger, the nonspecific medium-range interactions become more important. This particular method for combining the two types of medium-range interactions was chosen so that, as S approaches infinity, the prediction will become exactly equivalent to one based entirely on nonspecific medium-range interactions. Since the same number is added to each pairwise interaction, those interactions with a large number of occurrences in the data set (e.g., Table III) will be affected less by this procedure than the interactions with only a few observations. In this way, the specific interactions for which there are many data are weighted more heavily than those with only a small amount of data. By varying the two adjustable parameters, P_{max} and S, it is possible to select only those specific medium-range interactions which are significant at a chosen

TABLE I: Short-Range Interaction Contingency Table.[a]

	Conformational State[b] (k)						
Amino Acid (m)	ϵ	α_h	ζ_R	α_L	ζ_L	α_R[c]	Total
Ala	109	168	17	2	10	56	362
Asp	52	51	34	10	5	56	208
Cys	30	19	2	2	0	14	67
Glu	40	85	12	3	3	28	171
Phe	53	46	18	1	1	16	135
Gly	125	50	14	81	32	19	321
His	25	31	18	3	0	16	93
Ile	82	63	5	2	1	21	174
Lys	72	90	16	15	9	47	249
Leu	103	103	12	2	3	38	261
Met	23	21	4	0	0	5	53
Asn	55	37	38	22	5	25	182
Pro	59	23	6	1	0	43	132
Gln	51	36	3	5	2	23	120
Arg	46	31	7	5	0	23	112
Ser	129	62	21	7	6	77	302
Thr	109	50	16	2	6	56	239
Val	145	97	12	2	5	27	288
Trp	24	20	7	0	1	10	62
Tyr	76	29	12	7	2	24	150
Total	1408	1112	274	172	91	624	3681

[a] The entries in this table are the n_{km} defined in the text. [b] Definitions of conformational states are given in the text. [c] Excluding α_h residues.

TABLE II: Nonspecific Medium-Range Interaction Contingency Table for Any Residue at Position i and for $j = -4$.[a]

	Conformational State[b] of Residue i (k)						
Residue $i-4$ (l)	ϵ	α_h	ζ_R	α_L	ζ_l	α_R[c]	Total
Ala	112	125	28	20	5	63	353
Asp	75	67	18	6	3	36	205
Cys	27	14	5	4	2	14	66
Glu	45	82	10	6	5	23	171
Phe	41	42	8	9	7	26	133
Gly	143	83	12	13	8	55	314
His	28	37	3	6	2	16	92
Ile	75	55	12	6	2	24	174
Lys	91	79	22	8	5	38	243
Leu	89	81	21	10	10	42	253
Met	21	18	4	4	1	6	54
Asn	75	51	14	10	1	28	179
Pro	56	38	11	4	4	17	130
Gln	47	30	9	8	6	17	117
Arg	41	25	12	9	2	19	108
Ser	119	84	23	10	6	56	298
Thr	96	63	17	10	9	37	232
Val	108	82	16	11	8	60	285
Trp	21	20	4	5	3	8	61
Tyr	54	31	21	9	1	30	146
End[d]	44	5	5	4	0	10	68
Total[e]	1408	1112	275	172	90	625	3682

[a] The entries in this table are the n_{jkl} for $j = -4$, as an example, as defined in the text. Data for other values of j can be generated from the computer program. The index k designates ϵ, α_h, etc., and the index l designates the amino acid residue in column 1. [b] Definitions of the conformational states are given in the text. [c] Excluding α_h residues. [d] For residues where $i - 4 < 1$. [e] The totals in this row are not always the same as those at the bottom of Table I. This is due to the five residues of uncertain identity whose backbone atom coordinates are known. In order for a residue at position i to be used in Table I, the conformation and identity of residue i must be known. For an interaction to be included in Table II, the conformation of residue i and the identity of residue $i - 4$ must be known.

level of confidence and then to bias the estimates of the parameters for those interactions as described above.

It may be useful to state explicitly which parts of this method followed directly from a Bayesian analysis of contingency tables and which parts contain assumptions and special procedures needed for the problem at hand. The method used to estimate the effects of short-range interactions is a straightforward application of Bayesian theory (Lindley, 1964). When medium-range interactions are included, it has been assumed that each medium-range interaction is independent of the other medium-range interactions. This assumption, which has been discussed in a similar context by Robson (1974), is made necessary by the paucity of data. The nonspecific medium-range interactions are incorporated in the prediction method by a procedure which is based on the assumption that, to a first approximation, the effect of a neighboring residue on the conformation of residue i is independent of the identity of residue i. The method for estimating the effect of a single nonspecific medium-range interaction (eq A-10) follows directly from the Bayesian theory (Lindley, 1964). Once the assumptions mentioned above have been made, the use of nonspecific medium-range interactions in eq A-12 is straightforward.

The method used to estimate the effect of a single specific medium-range interaction in eq A-14 follows directly from the Bayesian theory (Lindley, 1964). The distribution of a number, S, of additional observations to each row of the specific medium-range interaction contingency tables is permitted by Bayesian theory, but the particular method for distributing these observations among the various conformations and the value of S are not specified by the Bayesian theory. The method for distributing the S observations used in the present paper seems reasonable, and it will be shown that some improvement in the accuracy of the prediction results from its use. The use of a significance test to determine whether the effect of a medium-range interaction will be estimated using data from specific or nonspecific medium-range interactions does not have a Bayesian justification. However, it will be shown that the use of this procedure can increase the accuracy of the predictions by reducing the errors caused by the small number of observations for many of the specific medium-range interactions.

In addition to the logarithm of the odds favoring one conformation over *all* other conformations, the logarithm of the odds favoring one conformation over another *particular* conformation can be estimated (see eq A-20 of the Appendix). Also, the variance in the estimates of the logarithm of the odds can be determined (eq A-18 and A-22 of the Appendix) A useful quantity for determining the reliability of a prediction is the number of standard deviations by which the predicted conformation is favored over each of the other possible conformations individually (eq A-20 and A-22 of the Appendix).

Results

The contingency table that classifies residues according to conformation and amino acid is shown in Table I. These are the data used to estimate the intraresidue information. One of the contingency tables used to estimate nonspecific medium-range interactions is shown in Table II, and one of the tables used to estimate specific medium-range interactions is shown in Table III. The computer programs and data (used

TABLE III: Specific Medium-Range Interaction Contingency Table for Alanine at Position i and $j = -4$.[a]

	Conformational State[b] of Ala at Position i (k)						
Residue $i-4$ (l)	ϵ	α_h	ζ_R	α_L	ζ_L	α_R[c]	Total
Ala	13	19	4	0	0	6	42
Asp	9	12	0	0	1	5	27
Cys	0	1	0	0	0	2	3
Glu	2	12	2	0	0	1	17
Phe	3	4	0	0	1	3	11
Gly	17	12	0	0	2	3	34
His	6	3	0	0	0	0	9
Ile	1	8	0	1	0	0	10
Lys	7	9	1	0	1	4	22
Leu	8	11	0	0	1	3	23
Met	3	1	0	0	0	1	5
Asn	6	8	0	0	0	2	16
Pro	3	8	1	0	2	1	15
Gln	3	4	1	0	0	2	10
Arg	2	5	1	1	0	1	10
Ser	8	19	4	0	0	3	34
Thr	5	12	2	0	0	4	23
Val	7	11	1	0	2	8	29
Trp	3	3	0	0	0	1	7
Tyr	1	4	0	0	0	4	9
End[d]	2	2	0	0	0	2	6
Total	109	168	17	2	10	56	362

[a] The entries in this table are the n_{jklm} for $j = -4$ and m representing Ala, as an example, as defined in the text. Data for other j and m can be generated from the computer program: The index k designates ϵ, α_h, etc., and the index l designates the amino acid residue in column 1. [b] Definitions of the conformational states are given in the text. [c] Excluding α_h residues. [d] For residues where $i - 4 < 1$.

to generate the full set of 8 nonspecific medium-range interactions and 160 specific medium-range interaction tables) are presented in the microfilm edition (see paragraph at the end of this paper concerning supplementary material). There are an average of $3681/20 \simeq 184$ observations in each row of the intraresidue information contingency table, $3682/21 \simeq 175$ in each row of the nonspecific medium-range interaction contingency tables, and $3681/420 \simeq 9$ observations in each row of the specific medium-range interaction contingency tables.

The data on short- and medium-range interactions were used in eq A-16 to estimate the logarithm of the odds in favor of each possible conformation. The predicted conformation for each residue was the conformation that had the highest value of the logarithm of the odds. As a test of the method, predictions were made for each of the proteins used in creating the data set. The prediction for each residue was made after that residue had been removed from the data set; hence, the accuracy found in this test should be roughly the same as the accuracy of the predictions for proteins not used in this study. The predicted conformations were compared with the observed conformations, and Table IV shows the number of residues predicted correctly out of 3681 residues for several values of the parameters, P_{max} and S. When determining the number of residues predicted correctly, residues which were predicted as α_R and observed to be α_h, and vice versa, were considered to be correct since either assignment would put the residue in the same region of ϕ, ψ conformational space. No requirements for any predicted conformations to occur in runs were considered. Thus, a single residue could be predicted to be α_h even though it would require at least four α_h residues in a row to form an α helix. Even without a requirement for α_h residues to occur in runs, 1129 of the 1388 residues predicted in the α_h conformation, with $P_{max} = 0.1$ and $S = 40$, occurred in runs of four or more. It can be seen that the prediction based on intraresidue information and specific medium-range interactions only (i.e., $P_{max} = 1$ and $S = 0$) gave relatively poor results. The best prediction, with $P_{max} = 0.1$ and $S = 40$ or 60, correctly assigned 56% of the residues to one of five conformational states. A wide range of values of the adjustable parameters gave results nearly as good as the best prediction. The prediction based on intraresidue information and nonspecific medium-range interactions ($P_{max} = 0$ or $S = \infty$) has an accuracy nearly identical with that of the best prediction. This arises because of the small amount of data available on specific medium-range interactions; presumably, if more such data were available, they would improve the accuracy of the predictions.

TABLE IV: Number of Residues Predicted Correctly for Various Values of the Adjustable Parameters.[a]

	P_{max}			
S	0.0	0.02	0.1	1.0
0	2024	1960	1905	1742
20	2024	2022	2014	1974
40	2024	2038	2046	2010
60	2024	2040	2046	2031
100	2024	2040	2038	2037
∞	2024	2024	2024	2024

[a] From a sample of 3681 residues in 20 proteins.

The effect of the adjustable parameters on the average value of the variance in the estimate of the logarithm of the odds is shown in Table V. The values shown in that table are the average value of the variance, calculated using eq A-18 and A-19, for all the residues in the data set. As expected, low values of

TABLE V: Average Values for the Minimum Variance in the Logarithm of the Odds and the Maximum Number of Standard Deviations between the First and Second Choice of Predicted Conformation.

	P_{max}:	0	0.02	0.1	1.0	1.0
	S:	c	40	40	40	0
Average variance in the logarithm of the odds[a]		0.17	0.17	0.18	0.19	1.81
Average no. of SD's between 1st and 2nd choice[b]		0.75	0.66	0.48	0.29	0.56

[a] Calculated using eq A-18 and A-19 in the Appendix. [b] Calculated using eq A-20 to A-23 in the Appendix. The first and second choices are those with the highest and next highest values of the logarithm of the odds. [c] S is not used in the prediction when P_{max} is zero.

P_{max} and large values of S reduce the variance. The prediction with $P_{max} = 1$ and $S = 0$, which had relatively low accuracy, also has a very high variance, as shown in Table V. The average number of standard deviations separating the predicted conformation from the second choice, as determined by eq A-20 to A-23, is also shown in Table V. For all values of the adjustable parameters tested, the average number of standard deviations separating the first two choices is less than one, indicating that for most residues these methods do not produce a single clear choice for the predicted conformation.

The number of predicted occurrences of each conformation in the proteins studied is shown in Table VI, along with the number correctly predicted to occur in each conformation. For comparison, the number of residues which would be predicted correctly by a random assignment of conformations with the same number of residues in each conformation as were obtained by the prediction is also shown in Table VI. The number expected to be predicted correctly by this random method for any conformation, k, is given by

$$\text{no. corr by random assign.} = \binom{\text{no. of residues}}{\text{obsd in } k} \times \binom{\text{no. pred}}{\text{in } k} \Big/ \binom{\text{total}}{\text{residues}} \quad (1)$$

The randomness in this procedure occurs only in the determination of which residues will be assigned a given conformation. The total number of residues assigned to each conformation is not random since it is determined by the number of residues predicted to occur in each conformation. It can be seen from Tables I and VI that, for all values of the parameters, the number of residues predicted to be in the extended conformation is too large, and the numbers predicted in ζ_R, α_L, and ζ_L are all too small. The number of residues predicted in the $\alpha_R + \alpha_h$ conformation is approximately correct. For all values of P_{max} and S, the number of residues predicted correctly is considerably above the number expected for a random prediction.

The percentage of residues predicted correctly for each of the proteins used in this study is shown in Table VII. The value of the prediction index described by Burgess et al. (1974) is also shown. The value of the prediction index (PI) is defined by

$$\text{PI} = N_s Q - 1 \qquad \text{for } 0 \le Q \le 1/N_s$$

$$\text{PI} = (N_s Q - 1)/(N_s - 1) \qquad \text{for } 1/N_s \le Q \le 1 \quad (2)$$

where Q is the fraction of residues predicted correctly, and N_s is the number of conformational states possible for each residue. In the present study, $N_s = 5$.

In order to provide a comparison with previous methods, an analysis which assigned residues to one of two states (helix or not helix) was performed. In this two-state prediction, a residue was predicted to be part of an α helix if it occurred as part of a run of four or more residues which were predicted to be α_h by the five-state prediction method. All other residues (including isolated α_h residues) were predicted to be nonhelical. The percentage of residues predicted correctly and the value of the prediction index with $N_s = 2$ for each of the proteins used in this study are shown in Table VII.

Discussion

The prediction method described in this paper represents an attempt to use data from proteins of known three-dimensional structure to estimate parameters which describe the effect that an amino acid and its neighbors have on its backbone conformation and, then, to use the estimated parameters to predict the conformations of residues in proteins. With an average of about 184 occurrences for each amino acid, the effect that an amino acid has on its own conformation can be estimated fairly reliably. The estimates will vary somewhat depending on the proteins used in the data set, but the parameters describing intraresidue information are fairly well established. On the other hand, when one looks at the effect of an amino acid on the conformational preference of a neighboring amino acid (i.e., the specific medium-range interactions), the average of nine occurrences per pair of amino acids is not sufficiently large to estimate the conformational preference reliably. In order to obtain statistical reliability at the expense of detail, one can consider the effect of an amino acid on the conformational preference of a neighboring residue independent of the identity of the latter (i.e., nonspecific medium-range interactions). The estimates of the parameters describing these nonspecific interactions are nearly as reliable as the parameters describing intraresidue information.

Despite the small average number of occurrences per pair of residues (viz., 9), it is possible to show that some specific medium-range interactions do affect the conformational preference of residues in proteins significantly. For example, it has been shown that a charged residue at position i in a protein has a significantly enhanced likelihood to be in an α helix if the residue at either position $i + 4$ or $i - 4$ carries the opposite charge (Maxfield and Scheraga, 1975). Nonspecific medium-range interactions would not be able to take such an effect into account. Unfortunately, it requires less data to determine that an interaction, such as the electrostatic one mentioned above, is important than it does to estimate precisely what is the effect of that interaction. Although the present data are sufficient to determine that several specific medium-range interactions are significant, they are not sufficient to yield reliable predictions based on this type of interaction.

The results shown in Table IV for various values of the adjustable parameters indicate that, with the amount of data used in this study, one must rely heavily on nonspecific medium-range interactions. When the specific medium-range interactions were relied on heavily (i.e., for small S and large P_{max}), the accuracy of the predictions decreased because of the large statistical error, as shown in Table V for the case where $S = 0$ and $P_{max} = 1$. In fact, none of the predictions which included specific medium-range interactions was significantly better than the prediction based entirely on nonspecific interactions and intraresidue information ($S = \infty$, or $P_{max} = 0$). The current predictions, with about 56% of the residues assigned correctly to one of five conformational states (see Table VII), may be considered as a base against which future predictions

TABLE VI: Predicted Occurrences in Each Conformation for Several Values of the Adjustable Parameters.

P_{max}:[a]	0	0.02	0.1	1.0	1.0
S:[a]	b	40	40	40	0
ϵ (pred)[c]	1801[d]	1754	1766	1757	1595
ϵ (corr)[e]	909	903	911	899	739
ϵ (corr by random assignment)[f]	657	640	644	641	582
($\alpha_R + \alpha_h$) (pred)	1731[d]	1771	1760	1739	1686
($\alpha_R + \alpha_h$) (corr)	1079	1102	1097	1069	960
($\alpha_R + \alpha_h$) (corr by random assignment)	778	796	791	782	758
ζ_R (pred)	65[d]	64	67	84	180
ζ_R (corr)	16	14	15	14	18
ζ_R (corr by random assignment)	5	5	5	6	13
α_L (pred)	67[d]	73	74	83	153
α_L (corr)	19	18	23	26	23
α_L (correct by random assignment)	3	3	3	4	7
ζ_L (pred)	17[d]	19	14	18	67
ζ_L (corr)	1	1	0	2	2
ζ_L (correct by random assignment)	0	0	0	0	2

[a] The definitions of the adjustable parameters, P_{max} and S, are given in the text. [b] S is not used when P_{max} is zero. [c] The number of residues, out of 3681, predicted to occur in the indicated conformation. [d] The sum of these numbers is 3681 [e] The number of residues correctly predicted to occur in the indicated conformation. [f] Calculated using eq 1 in the text.

TABLE VII: Percentage of Residues Predicted Correctly and Value of a Prediction Index (PI) with $P_{max} = 0.1$ and $S = 40$.

Protein	% Corr (Five-State Prediction)	PI[a]	% Corr (Two-State Pred)[b]	PI[c]
Myoglobin	82	0.77	78	0.57
Thermolysin	49	0.36	84	0.68
Cytochrome b_5	72	0.65	78	0.55
Tosyl elastase	55	0.43	84	0.69
Tosyl-α-chymotrypsin A chain	57	0.46	100	1.00
B chain	50	0.38	75	0.50
C chain	52	0.40	74	0.47
Carboxypeptidase A	51	0.39	69	0.37
Papain	52	0.41	75	0.50
Bovine pancreatic trypsin inhibitor	55	0.44	95	0.90
Concanavalin A	47	0.34	87	0.74
High-potential iron protein	51	0.38	61	0.23
Lysozyme	59	0.49	79	0.57
Carp myogen	65	0.56	68	0.36
Subtilisin BPN′	52	0.40	81	0.61
Ribonuclease S	49	0.37	77	0.54
Staphylococcal nuclease	53	0.41	68	0.36
Cytochrome c_2	55	0.43	59	0.18
D-Glyceraldehyde-3-phosphate dehydrogenase	52	0.40	70	0.40
Clostridial flavodoxin	66	0.58	73	0.46
Human deoxyhemoglobin A chain	70	0.62	71	0.42
B chain	60	0.50	65	0.29
Sea lamprey hemoglobin	60	0.50	54	0.08
Average	56	0.44	76	0.52

[a] Calculated using eq 2 with $N_s = 5$. [b] The two states are: (1) runs of four or more α_h residues predicted by the five-state prediction and (2) all other conformations including isolated α_h residues. [c] Calculated using eq 2 with $N_s = 2$.

can be compared when the data on specific medium-range interactions rise above the current level of statistical noise.

Although the inclusion of many proteins in the data set would seem to be the most satisfactory way to increase the reliability of estimates of specific medium-range interactions, it may be worthwhile to explore other methods of improving these estimates pending the determination of a large number of protein crystal structures. One such method would involve the classification of amino acids into groups according to the nature of their side chains. For example, amino acids might be classified as nonpolar, positively charged, negatively charged, or polar uncharged. One could then, for example, replace a specific medium-range interaction describing the effect of an aspartic acid on an alanine residue with a medium-range interaction term for the effect of all negatively charged residues on alanine. The advantage of such a method would be the increased number of observations for each interaction. However, if different amino acids in each group did not have the same effect on all neighboring amino acids, the method would not be useful. It might prove quite difficult to

determine a suitable method for classifying the amino acids into these types of groups.

Another method that could be explored with data presently available involves a greater use of homologous proteins. Prediction methods that rely heavily on data from homologous proteins have already been proposed (Nagano, 1973, 1974; Gabel et al., 1976). The data set used in our predictions also includes a few homologous proteins. With the assumption that the structures of homologous proteins are nearly the same (Perutz et al., 1965; Endres et al., 1975; Gabel et al., 1976), it would be possible to include in the data set a large number of proteins whose crystal structures have not been determined but which have been shown to have sequences homologous to those of proteins of known three-dimensional structure. Since these homologous proteins contain a large number of amino acid replacements, many different interactions which are compatible with the same conformation could be included in the data (Gabel et al., 1976). There are, however, some drawbacks to the use of large numbers of homologous proteins in the data set. If the families of proteins used contain an unusually large percentage of one conformation, as in hemoglobin and myoglobin, the parameters for that conformation may become too large. Also, the method would not be able to take into account the possible differences in conformation of analogous residues in different proteins from the same family, if they exist, and this would result in some errors. Finally, it is not clear what the relative importance of invariant and variable residues in homologous proteins should be. It would not be correct to treat the repeated occurrences of invariant residues as a series of independent events. The inclusion of an invariant residue in the data set each time it occurred in a large series of homologous proteins might overestimate the preference for that amino acid to be in the conformation it had in the homologous family. Despite the difficulties involved, this method is likely to be useful in improving the estimates of parameters describing specific medium-range interactions in the interim until a large number of three-dimensional structures of proteins are determined.

Although the prediction method described in this paper would not be very useful, at the present, for the generation of starting conformations for protein energy minimization, we can consider its potential usefulness and limitations for that purpose. One of the advantages of the method over many of the previously published prediction methods is that *every* residue is assigned to a relatively small region of ϕ, ψ conformational space. This avoids the ambiguity in starting conformation for residues which are predicted to be in "bends," "coil," or "irregular" regions. In order to assign specific dihedral angles to each residue, one could use the mean values of ϕ and ψ for each amino acid in the predicted region, as described by Burgess and Scheraga (1975).

In addition to a prediction of the conformation which is most likely to be correct, the probability that the prediction is correct can be obtained from the logarithm of the odds in eq A-16 by the relationship:

$$\text{probability (conformation is } k) \propto [1 + \exp\{-(\log \text{ of odds in favor of } k)\}]^{-1} \quad (3)$$

This probability does not take into account statistical or experimental errors, but for predictions where the statistical error is small, it would be useful in assessing the reliability of the prediction.

The estimate of the statistical error in the predictions is useful for several reasons. First, it allows one to determine the error introduced by including parameters estimated with only a small amount of data. It emphasizes the importance of relying more heavily on parameters which can be estimated reliably, even if this means neglecting some interactions which are known to be important. Also, in a protein-folding and energy-minimization procedure, one might constrain residues with highly reliable predictions to remain near the predicted conformation while allowing residues with less reliable predictions to vary more freely. The statistical error also helps one to determine where the paucity of data leads to inaccuracies in the predictions. Even for predictions based entirely on nonspecific medium-range interactions and intraresidue information, the statistical error is large enough to contribute significantly to the inaccuracy of the predictions. For predictions based on specific medium-range interactions and intraresidue information, the statistical error is so large, at this time, that it is impossible to tell if these interactions would be sufficient for an accurate prediction. The inaccuracies of the prediction with $P_{max} = 1$ and $S = 0$ could be due largely to the statistical error, or it could be that pairwise specific medium-range interactions and intraresidue information are not sufficient to determine the conformational preference of many residues. The amount of data used in the present analysis is not sufficient to distinguish between these alternatives, but it should be possible to answer this question as more proteins are added to the data set. If sufficient data were available, it would be possible to include triplet interactions as well as pairwise interactions, but it is clear from the errors in pairwise interactions that it would take an enormous amount of data to estimate the effects of specific triplets on the conformational preference of residues accurately.

Another source of error in estimating the parameters used to predict conformation is the experimental error in the determination of the x-ray structure of the proteins in the data set. The errors in determining the dihedral angles are sufficient to cause many residues to be assigned to the wrong conformational states. These types of errors will be reduced as the resolution of the x-ray structures is improved and as new methods of refining x-ray structures are developed and applied.

One problem with the prediction algorithm described in this paper is that, for predictions based largely on nonspecific medium-range interactions, too many residues are predicted to be in the ϵ conformation, and too few are predicted in ζ_R, α_L, and ζ_L conformations. This reflects the fact that, for every amino acid, either the ϵ or $(\alpha_R + \alpha_h)$ conformation is the most likely (see Table I), and nonspecific interactions are apparently not often able to increase the probabilities for other conformations sufficiently to outweigh the intraresidue information. Inclusion of specific medium-range interactions leads to nearly the correct distribution of predicted residues among the five states, but most of the predictions of ζ_R, α_L, and ζ_L are incorrect. It is possible that the occurrence of the ζ_R, α_L, and ζ_L conformations is determined largely by long-range interactions, such as the requirement to form a compact globular structure, or it could be that, when sufficient data are obtained, the prediction of these conformations based on specific medium-range interactions and intraresidue information will improve greatly. It is encouraging, at least, to note that, of the residues predicted to be in ζ_R, α_L, and ζ_L conformations, the fraction correct is usually well above what would be expected from a random prediction.

The assumption that short-range and/or medium-range interactions are sufficient to determine the conformational state of most of the residues in proteins is fundamental to all empirical prediction algorithms. Changes in conformational

preference due to specific interactions with residues far away in the primary sequence or changes caused by prosthetic groups are not taken into account by empirical methods. Less specific types of long-range effects, such as the presence of nonpolar pairs in positions i and $i + 4$ of an α helix to allow favorable contacts with the nonpolar core of a protein, have been included explicitly in some prediction methods (Schiffer and Edmundson, 1967; Lim, 1974b) and are included implicitly in the present study and in some other methods (Robson and Pain, 1971; Nagano, 1973). As mentioned previously, the amount of data presently available is not sufficient to determine the percentage of residues which could be assigned correctly from short- and medium-range interactions, other than to say that at least 56% of all residues can be assigned correctly to one of five conformational states. Presumably, the percentage correct will increase as the data for specific medium-range interactions are improved. Whatever the final success of prediction algorithms is, the effects of long-range interactions and prosthetic groups in determining the conformational state of a residue will have to be accounted for by alternative procedures such as energy minimization.

The need for procedures such as energy minimization or Monte Carlo simulation of folding to predict the three-dimensional structure of proteins has recently been demonstrated in analyses of the problems involved in predicting the three-dimensional structure of bovine pancreatic trypsin inhibitor (Burgess and Scheraga, 1975; Tanaka and Scheraga, 1975). In one of those analyses, every residue in trypsin inhibitor was assigned to one of five conformational states (roughly equivalent to those used in the present paper) on the basis of the dihedral angles observed in the crystal structure. This is equivalent to a prediction algorithm with 100% accuracy in assigning residues to one of five conformational states. Each residue was then assigned dihedral angles which were the average value from a sample of eight proteins for the amino acid under consideration in the assigned conformation. Peptide bonds were held in the planar trans conformation, and bond lengths and bond angles were assigned values from a recent compilation (Momany et al., 1975). Using these mean values of ϕ and ψ for each amino acid in a conformational state and the molecular geometry for each residue, a backbone structure for trypsin inhibitor was generated. When this generated structure was compared with the experimentally observed structure, it was found that there were gross differences between the two. Some of the C^{α} atoms in the generated structure were separated by more than 50 Å while the maximum separation in the x-ray structure was about 15 Å. These differences arise even though every residue is assigned to the correct conformational state and are the result of the cumulative errors caused by small differences between the observed and assigned dihedral angles. Only after other procedures such as energy minimization and the formation of disulfide bonds (or Monte Carlo simulation of folding) were used did the predicted structure begin to resemble the observed structure. Since even a perfect five-state prediction algorithm gives such a grossly inaccurate three-dimensional structure, it is clear that empirical prediction algorithms can provide only the first step in predicting the backbone structure of a protein. Statistical mechanical procedures (Tanaka and Scheraga, 1976) may overcome some of these difficulties.

Comparison with Other Methods. Many of the earlier prediction methods have been reviewed previously (Burgess et al., 1974), and the usefulness of more recent prediction methods for generating starting conformations for energy minimization have also been discussed (Burgess and Scheraga, 1975). In this section, the methods and results of some of the recently published prediction methods will be discussed and compared with the methods proposed in this paper.

The prediction methods developed in this paper are most closely related to the information theory approach of Robson and Pain (1971, 1974a; Robson, 1974). There are, however, many differences in terminology and several substantial differences in the actual methods used. The "star" function used by Robson and Pain is the same as the logarithm of the odds in our nomenclature. The estimate of the information used by Robson and Pain consisted of sums and differences of "star" functions in much the same way as our final estimate of the logarithm of the odds was based on sums and differences of logarithms of the odds based on various types of interactions. The procedure used in eq A-14 to subtract out the intraresidue information is, for the most part, analogous to the "expected frequency" method of Robson and Pain. In addition, Robson and Pain have developed an alternative method for estimating the logarithm of the odds (their "star" function) from that given by our eq A-6. While their method is more accurate for small values of $a + n$ and $b + N - n$, the differences between the methods are not significant when these terms have values greater than 5. When these terms are less than 5, however, the uncertainty in the estimate of a "star" function or logarithm of the odds is so great that the differences in the two estimates are probably not important.

One of the important differences between the two methods is that, in the predictions described in this paper, the estimates for specific medium-range interactions which have a large statistical uncertainty can be biased by using estimates from better known, but less detailed, nonspecific medium-range interactions. The extent of this bias is determined by the parameter S. Also, the methods described in this paper allow for some medium-range interactions to be estimated by nonspecific interactions and others by specific interactions, depending on the level of significance of the effects caused by the specific interaction and on the level of P_{max}. Although Robson and Pain discussed and calculated the interactions described here as nonspecific medium-range interactions (Robson and Pain, 1974b), they never applied them in a prediction scheme. The method of Robson and Pain is, therefore, *roughly* equivalent to taking $P_{max} = 1$ and $S = 0$ (the difference being that their method for evaluating the value of their "star" function implies a *small* amount of weighting according to significance), whereas we allow for other values of P_{max} and S. It can be seen from Table IV that the accuracy of the prediction can be increased significantly by choosing other values for P_{max} and S. The choice of conformational states used by Robson and Pain also differs from that used in this paper. In an early paper (Robson and Pain, 1971), only α-helical runs of at least four residues were considered. The prediction of helical and nonhelical regions was the only case for which Robson and Pain included specific medium-range interactions. In a subsequent series of papers (Robson, 1974; Robson and Pain, 1974a–c), several sets of definitions of conformational states were investigated. While the method for including specific medium-range interactions was discussed in this later series, the predictions were based entirely on intraresidue information. In one of the multistate models considered (Robson and Pain, 1974a), 43% of the residues in nine proteins were assigned correctly to a 40° × 40° cell on the ϕ, ψ map.

In the prediction of helical and nonhelical regions (Robson and Pain, 1971), the data for each protein, including the residue being predicted, were left in the data set. We have found that failure to remove the residue being predicted from the data

TABLE VIII: Helix and β-Sheet Assignments used by Chou and Fasman (1974) for the C-Terminal Region of Trypsin Inhibitor.[a]

Residue No.	AA[c]	P_α	α-Helix Assign.	P_β	β-Sheet Assign.	Obsd Conformation[b,c]
44	Asn	0.73	b_α	0.65	b_β	C
45	Phe	1.12	h_α	1.28	h_β	C
46	Lys	1.07	I_α	0.74	b_β	H
47	Ser	0.79	i_α	0.72	b_β	H
48	Ala	1.45	H_α	0.97	I_β	H
49	Glu	1.53	H_α	0.26	B_β	H
50	Asp	0.98	i_α	0.80	i_β	H
51	Cys	0.77	i_α	1.30	h_β	H
52	Met	1.20	h_α	1.67	H_β	H
53	Arg	0.79	I_α	0.90	i_β	H
54	Thr	0.82	i_α	1.20	h_β	H
55	Cys	0.77	i_α	1.30	h_β	H
56	Gly	0.53	B_α	0.81	i_β	H
57	Gly	0.53	B_α	0.81	i_β	C
58	Ala	1.45	H_α	0.97	I_β	C

[a] The values of P_α and P_β as well as the α-helical and β-sheet assignments and the definitions of these terms are given by Chou and Fasman (1974). [b] According to the definitions of conformational states used by Chou and Fasman (1974). [c] AA, amino acid; C, coil; H, helix.

set can have striking effects. Therefore, we always removed the residue being predicted from the data set. For our prediction method with $P_{max} = 1$ and $S = 0$ (i.e., the closest equivalent to the method of Robson and Pain), the percentage of residues assigned correctly to one of the five states increases from 47 to 88% when the residue being predicted is left in the data set. this increase is due largely to the effect that a single observation can have on the estimates of the parameters describing specific medium-range interactions. It is likely that much of the success of these early predictions was due to this effect. Robson and Pain also noted a decrease in the precision of their predictions for lysozyme when that protein was removed from the data base. In the later predictions (Robson and Pain, 1974a), each protein was removed from the data set before the conformation of its residues was predicted.

Perhaps the most important difference between the method described in this paper and those of Robson and Pain is the determination of a minimum value for the statistical error in each prediction. This allows one to understand the factors which presently limit the accuracy of prediction methods.

Another recent prediction method which relies on data obtained from proteins of known conformation has been developed by Chou and Fasman (1974). The method is based primarily on short-range interactions, although medium-range effects are included by averaging the parameters describing α-helix and β-sheet formation over several adjacent residues. These authors report that they are able to assign residues to the helical, β-sheet, or coil states with an accuracy of 80%. However, we have attempted to reproduce these results and were unable to do so with the prediction rules as presently described. [See Tanaka and Scheraga (1976) for further discussion of the Chou–Fasman procedure.] The predictive analysis of trypsin inhibitor, which is used by Chou and Fasman as an example to illustrate their method, also serves to illustrate some of the difficulties encountered in applying the method. For this discussion we will adopt the nomenclature used by Chou and Fasman. The C-terminal region from residue 44 through 58 has the amino acid sequence shown in Table VIII. The residues are also classified according to their helical and β-sheet potentials. The first condition for the location of helical regions (A-1) states: "Locate clusters of four helical residues (h_α or H_α) out of six residues along the polypeptide chain. Weak helical residues (I_α) count as 0.5 h_α (i.e., three h_α and two I_α residues out of six could also nucleate a helix). Helix formation is unfavorable if the segment contains one-third or more helix breakers (b_α or B_α), or less than one-half helix formers." Although there are no segments of six residues in this section of trypsin inhibitor which satisfy this condition, the segment from residue 45 to 54 was predicted to be helical. The rationale for this prediction, in apparent violation of condition A-1, is not clear to us. Furthermore, the region from residue 51 to 55 with $\langle P_\beta \rangle = 1.27$ and $\langle P_\alpha \rangle = 0.87$ appears to satisfy the conditions for β-sheet formation. The possible objection to this assignment is the rather vague condition (B-4), which states: "Charged residues occur rarely at the N-terminal β-sheet end and infrequently at the inner β region and C-terminal β end." However, this condition did not prevent two arginine residues from being included in the β-sheet region predicted from residue 16 to residue 23 in trypsin inhibitor. The C-terminal region illustrates yet another difficulty in applying these rules. Ignoring the difficulty in predicting the nucleation of a helix in the region 45 through 54, the authors point out that $\langle P_\alpha \rangle = 1.05$ for this region while $\langle P_\beta \rangle$ for the same region is only 0.99, thus indicating a preference to be helical since $\langle P_\alpha \rangle$ is greater than $\langle P_\beta \rangle$. On the other hand, for the region 51 through 55, $\langle P_\alpha \rangle = 0.87$ and $\langle P_\beta \rangle = 1.27$, indicating to us that this region should be predicted as β. No criterion for determining the prediction for the overlap region 51 through 54 is given. In addition to the ambiguities pointed out already, there are others, such as condition B-3: "Glu occurs rarely in the β region. Pro occurs rarely in the inner β region." Further difficulties arise when these authors include the prediction of β turns in their analysis. The "preliminary guidelines for locating β turns" include the rule that $\langle P_\alpha \rangle$ should be less than 0.90, and $\langle P_t \rangle$ should be greater than 0.5×10^{-4} for a tetrapeptide to be predicted as a β turn. Nevertheless, in their analysis of trypsin inhibitor, tetrapeptide 23–26 with $\langle P_\alpha \rangle = 0.97$ and $\langle P_t \rangle = 0.3 \times 10^{-4}$ was predicted to be in a reverse turn. Despite these apparent problems, Chou and Fasman were able to do relatively well in predicting the helix, β-sheet, β-turn, and coil regions in a protein whose three-dimensional structure was not known beforehand (Schulz et al., 1974). It seems that these rules leave much to the intuition of the individual investigator, and it is, perhaps, not surprising that we have been unable to reproduce them. In this context, it should be noted that two papers have recently appeared which attempt to predict the

backbone conformation of the *lac* repressor (or a segment of it) using the Chou–Fasman rules (Patel, 1975; Chou et al., 1975). In many cases, the predictions reported in the two papers do not agree.

Another algorithm which uses data from proteins of known three-dimensional structure to predict the conformational preference of residues on the basis of short-range and medium-range interactions has been developed by Burgess et al. (1974). This algorithm assigned residues to the α-helix, extended, β-bend, and coil structures. The algorithm provides an unambiguous assignment of each residue to one of these four states. The method was designed to avoid assigning residues incorrectly to one of the three regular conformations since it was felt that energy minimization or other procedures could more easily correct a wrong assignment to the coil state than an incorrect assignment to one of the regular structures. However, the large number of coil residues predicted by this method, as well as the residues predicted to be in β bends, leaves a large percentage of the residues in a protein without any predicted value for the dihedral angles. Some method for assigning dihedral angles to residues in the coil and β-bend structures (see, e.g., Gabel et al., 1976) would have to be developed for this method, or any other method that predicts coil and β-bend structures, to be useful in generating starting conformations for energy minimization.

An alternative approach to the use of data from proteins of known three-dimensional structure has been described by Lim (1974a–c). The method is designed to take into account the principles governing the packing of polypeptide chains in globular proteins. On the basis of studies with space-filling models, Lim has attempted to determine the types of sequences which would be compatible with the formation of α helix and β structure. Particular attention was paid to the shielding of nonpolar side chains from water, the solvation of polar groups, and the tight packing of nonpolar side chains in the core of the protein. Consideration of these stereochemical principles led to a set of rules for the prediction of α-helical and β-sheet regions in proteins. The method was reported to have an accuracy of 70% in assigning residues in 25 proteins to the α-helical, β-sheet, and coil states. Although this method avoids the errors due to the small amount of data available on the three-dimensional structure of proteins, it is possible that important factors governing folding have been overlooked in formulating the prediction rules. Again, it should be pointed out that 100% accuracy in assigning residues to these states would leave roughly half the residues in a protein with unspecified dihedral angles.

Conclusions

The prediction method described in this paper assigns each residue in a protein to a specific region of conformational space with an accuracy of 56%. While there are other prediction methods which report an accuracy higher than this, many of the residues which are considered to be predicted correctly in these methods cannot be assigned specific dihedral angles. When the number of residues correctly assigned by these other methods to regions with specific dihedral angles is considered, it is less than 56% of the residues in the proteins studied. Since the correct three-dimensional structure of a protein cannot be generated even from a five-state prediction algorithm with 100% accuracy, it is clear that current prediction methods are not useful, by themselves, for understanding the structure of proteins. A more reasonable expectation is that prediction algorithms more accurate than those presently available will provide starting conformations for protein folding methods which will take into account the short-range and long-range interactions responsible for the three-dimensional structure of proteins.

There are several factors which limit the accuracy of current prediction methods. One limitation which is common to all methods is the neglect of long-range interactions. It is not yet known what limit this will place on the accuracy of empirical prediction methods based on short-range and medium-range interactions. Another limitation which is common to all of the methods that use data from the sequence and structure of proteins is the quality and quantity of those data. It is possible that the dihedral angles for a residue could be in error by as much as 30°, and this error could affect the parameters used in the prediction methods. As we have shown in this paper, the quantity of data presently available limits the inclusion of medium-range interactions almost exclusively to what we have called nonspecific medium-range interactions. This is unfortunate since it has been shown that specific medium-range interactions between pairs of residues can affect the conformational preference of these residues in proteins. We have also shown that even the predictions based entirely on intraresidue information and nonspecific medium-range interactions will contain many errors due to the small size of the data base currently available. We are, thus, unable to evaluate the *potential* accuracy of the method described in this paper since statistical errors play a large role in limiting the accuracy. Since any method employing data from proteins of known structure will be subject to the same statistical limitations, it seems unlikely that changes in the *method* for using those data will result in significantly increased accuracy until the quantity of data is also increased.

The success of empirical prediction algorithms has certainly established that there is a relationship between the local amino acid sequence and backbone conformation in proteins. However, the exact nature of that relationship is, at present, far from established. With the determination of the structure of additional proteins and the development of new methods which make better use of the available data, the nature of the relationship between sequence and structure may be understood more clearly, and it should be possible to use prediction algorithms to generate reasonable starting conformations for subsequent protein-folding procedures.

Acknowledgment

We acknowledge Dr. S. Tanaka and Mr. Z. Hodes for helpful discussions and Mrs. S. Rumsey for assistance with the protein coordinate data files. We are also indebted to Dr. T. F. Koetzle for providing us with the x-ray coordinates from the Protein Data Bank, Brookhaven National Laboratories, Upton, N.Y.

Appendix: Prediction Procedure

Bayesian Methods (*Lindley, 1965*). Consider a series of N independent trials, each of which has the same probability for success (Θ); $1 - \Theta$ is the probability for failure. For example, the N trials might consist of the observation of the conformational state of all the alanine residues in the data set. Success might be defined as the observation of an alanine in, say, the ϵ conformation, and failure would then be the observation of any other conformation. The statement that each of the trials has the same probability for success would then be equivalent to the assumption that, before any observations are made, all of the alanine residues have an equal probability to be observed in the ϵ conformation. We seek an estimate of the value of Θ. If there is some knowledge about the value of Θ before per-

forming the trials, this may be denoted by a probability density, $\pi(\theta|K)$, which represents our degree of belief that a number, θ, is the true value of Θ, given our prior knowledge, represented by K. The probability density $\pi(\theta|K)$ has all the usual properties of a probability density function, i.e.

$$\pi(\theta|K) \geq 0 \tag{A-1}$$

$$\int_{-\infty}^{\infty} \pi(\theta|K)\mathrm{d}\theta = 1 \tag{A-2}$$

$$\int_{c}^{d} \pi(\theta|K)\mathrm{d}\theta = \text{probability}\{c < \Theta < d\} \tag{A-3}$$

where $c < d$, and probability $\{c < \Theta < d\}$ is the estimate of the probability that Θ is between c and d. This probability density, $\pi(\theta|K)$, is called the *prior* density of θ. The *likelihood* that N trials with probability for success, θ, will result in an observed sequence of n successes and $N - n$ failures is $\theta^n(1 - \theta)^{N-n}$ The *posterior* probability density $\pi(\theta|R,K)$ represents the degree of belief in a particular value of θ, given both our prior knowledge, K, and the results of the N trials, represented by R. According to Bayes' theorem:

$$\pi(\theta|R,K) \propto \theta^n(1 - \theta)^{N-n}\pi(\theta|K) \tag{A-4}$$

where $\pi(\theta|K)$ is a measure of our prior knowledge and $\theta^n(1 - \theta)^{N-n}$ (which implicitly includes R) is a measure of our additional knowledge after N trials. The constant of proportionality is just a normalization constant so that

$$\int_{0}^{1} \pi(\theta|R,K)\mathrm{d}\theta = 1$$

In words, eq A-4 states that the posterior probability for any θ to be the true value of Θ is proportional to the prior probability of that θ multiplied by the likelihood of obtaining the observed results (after N trials) with that value of θ. The form of the posterior probability density will depend on the form of the prior probability density. For

$$\pi(\theta|K) = \theta^a(1 - \theta)^b \Big/ \int_{-\infty}^{\infty} \theta^a(1 - \theta)^b\mathrm{d}\theta$$

where a and b are constants, the posterior probability density will be:

$$\pi(\theta|R,K) \propto \theta^{n+a}(1 - \theta)^{N-n+b} \tag{A-5}$$

The constants, a and b, must be greater than -1 in order for the integral

$$\int_{-\infty}^{\infty} \theta^a(1 - \theta)^b\mathrm{d}\theta$$

to converge. The $\theta^a(1 - \theta)^b$ form of the prior probability is the one that will be used throughout this paper. The prior knowledge may then be thought of as equivalent to the observation of a successes and b failures. It follows that small values of a and b are equivalent to little prior knowledge (Lindley, 1964).

For our purposes, it is convenient to consider the natural logarithm of the odds in favor of success, $\ln [\theta/(1 - \theta)]$. It has been shown (Lindley, 1964) that the posterior distribution of $\ln [\theta/(1 - \theta)]$ is approximately a normal distribution, with the mean of the distribution given by:

$$\langle\{\ln [\theta/(1 - \theta)]\}|R,K\rangle = \ln [(a + n)/(N - n + b)] \tag{A-6}$$

i.e., by the natural logarithm of the ratio of the number of successes to failures, and the variance by:

$$\sigma^2 = (a + n + 1)^{-1} + (b + N - n + 1)^{-1} \tag{A-7}$$

These approximations are good for $a + n$ and $b + N - n$ greater than five and are probably adequate for our purposes for values greater than 3. The variance of the distribution is a measure of the statistical error due to the size of the sample. We use the log of the odds, rather than the probability, θ, because of the convenience of working with the normal distribution, and because of the additive properties of the logarithm of the odds for independent events.

Application of Bayesian Methods to Intraresidue Information. The first step in the prediction of the conformation of a residue is to estimate the logarithm of the odds in favor of each conformation, k, given only the identity of the residue, and no information about its neighbors. This is evaluated using the values of n_{km}, shown in Table I, with 0.1 added to each entry to avoid mathematical difficulties in evaluating $\log n_{km}$ when $n_{km} = 0$. (The entries in Table I do not contain the 0.1; when the 0.1 is added, it contributes a trivial error.) If Θ_k is the probability for the residue to be in conformation k and R_m represents the observations of amino acid m in the data set, then, using eq A-6 and A-7, the posterior probability density for the logarithm of the odds in favor of conformation k is approximately normal, with the mean given by:

$$\langle\{\ln [\theta_k/(1 - \theta_k)]\}|R_m, K_m\rangle = \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] \tag{A-8}$$

and variance:

$$\sigma_{km}^2 = (n_{km} + 1.1)^{-1} + (n_{\cdot m} - n_{km} + 1.5)^{-1} \tag{A-9}$$

where[2] $n_{\cdot m} = \Sigma_k n_{km}$. The addition of 0.1 to each value of n_{km} is formally equivalent to taking $a = 0.1$ and $b = 0.5$ in eq A-6 and A-7. The value of b is 0.5 because 0.1 was added to the entries for each of the five other conformations as well as conformation k.

Incorporation of Nonspecific Medium-Range Interactions. The influence of neighboring residues on the conformation of a residue at position i, independent of the identity of the residue at position i, is evaluated using the values of n_{jkl} described previously. If R_{jl} is used to represent the observations of the conformation of residue i when residue $i + j$ is amino acid l, then the posterior probability density for the logarithm of the odds in favor of conformation k, based on a single nonspecific medium-range interaction, is approximately normal, with mean given by:

$$\langle\{\ln [\theta_k/(1 - \theta_k)]\}|R_{jl}, K_{jl}\rangle = \ln [(n_{jkl} + 0.1)/(n_{j\cdot l} - n_{jkl} + 0.5)] \tag{A-10}$$

and variance:

$$\sigma_{jkl}^2 = (n_{jkl} + 1.1)^{-1} + (n_{j\cdot l} - n_{jkl} + 1.5)^{-1} \tag{A-11}$$

The logarithm of the odds in favor of a conformational state, based on both short-range and nonspecific medium-range interactions, may now be estimated as:

$$\begin{aligned}\langle\{\ln [\theta_k/(1 - \theta_k)]\}|R_m, R_{-4,l} \ldots R_{4,l}, K\rangle &= \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] \\ &+ \sum_{\substack{j=-4 \\ j\neq 0}}^{4} \{\ln[(n_{jkl} + 0.1)/(n_{j\cdot l} - n_{jkl} + 0.5)] \\ &- \ln [(n_{jk\cdot} + 2.1)/(n_{j\cdot\cdot} - n_{jk\cdot} + 10.5)]\}\end{aligned} \tag{A-12}$$

The small numbers added to the actual observations in eq

[2] A dot will be used throughout the Appendix to indicate summation over the index which has been replaced.

A-10 through A-13 result from the addition of 0.1 to each entry in the nonspecific medium-range interaction tables. Multiples of 0.1 result from the summation over many terms to which 0.1 has been added. The second term in the summation over j is the estimate of the logarithm of the odds in favor of conformation k, independent of the identity of the residue and its neighbors. The variance in this term is given by:

$$\sigma_{jk}^2 = (n_{jk\cdot} + 3.1)^{-1} + (n_{j\cdot\cdot} - n_{jk\cdot} + 11.5)^{-1} \quad \text{(A-13)}$$

The whole summation over j in eq A-12 is an estimate of the information transmitted by amino acid l at position $i + j$ minus the average of the information transmitted by all residues at position $i + j$. Since the average information transmitted by all residues at position $i + j$ is implicitly included in the term for short-range interactions in eq A-12, it would be incorrect to include it again in the terms for each of the nonspecific medium-range interactions; hence, it is subtracted out in eq A-12.

Test of Significance of Specific Medium-Range Interactions. Before incorporating the effect of specific medium-range interactions, we test to see if they are significant enough to include. The effect of specific medium-range interactions is estimated using the values of n_{jklm}. A 2×2 contingency table may be used to test the level of significance of a specific medium-range interaction. The entries in the contingency table are as follows: $O_{11} = n_{jklm}$, $O_{12} = n_{j\cdot lm} - n_{jklm}$, $O_{21} = n_{jk\cdot m} - n_{jklm}$, $O_{22} = n_{j\cdot\cdot m} - n_{j\cdot lm} - n_{jk\cdot m} + n_{jklm}$. The significance test referred to in the section on contingency tables may then be used to evaluate the probability, P_{jklm}, that deviations of O_{11}/O_{12} from O_{21}/O_{22} as large or larger than those actually observed could occur by chance [see eq 3 of Maxfield and Scheraga (1975) for computation of P_{jklm}]. A small value of P_{jklm} would indicate that the specific interaction under consideration does affect the preference for conformation k. If the value of P_{jklm} for all values of k is greater than a selected value, $P_{\max}$, the specific interaction is considered to be not significant.

The value of $P_{\max}$ was an adjustable parameter in the prediction method. When $P_{\max} = 0$, only nonspecific medium-range interactions are considered, as described previously. When $P_{\max} = 1$, only specific medium-range interactions are considered, as described in the following section. For intermediate values, a mixture of specific and nonspecific medium-range interactions are considered.

Incorporation of Specific Medium-Range Interactions. If the specific medium-range interaction under consideration is not statistically significant (i.e., if $P_{jklm} > P_{\max}$ for all conformational states), then the medium-range interactions for that value of j are estimated by the nonspecific medium range interactions, as in eq A-12. For each statistically significant interaction, the term in the summation of eq A-12 for that value of j is replaced by:

$$\ln [(n_{jklm} + a_{jklm} + 0.1)/(n_{j\cdot lm} - n_{jklm} + b_{jklm} + 0.5)] - \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] \quad \text{(A-14)}$$

and R_{jl} for that value of j is replaced by R_{jlm}, where R_{jlm} represents the observations of the conformations of amino acid m at position i when residue $i + j$ is amino acid l; a_{jklm} and b_{jklm} represent the prior knowledge about the interaction and are determined by a procedure described below. The term $\ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)]$, which is the estimate of the intraresidue information, is subtracted out since the quantity of interest here is the difference from the short-range conformational preference which is caused by the specific medium-range interaction. The intraresidue information is included explicitly in the first term of eq A-12, and it is also included implicitly in the values of n_{jklm} and, thus, in the first term of eq A-14. To avoid including the intraresidue information twice, it is subtracted out in eq A-14.

The values of a_{jklm} and b_{jklm} are evaluated using the intraresidue information and the nonspecific medium-range interactions to obtain a prior estimate of the logarithm of the odds favoring a particular conformation based on specific medium-range interactions. The ratio a_{jklm}/b_{jklm} is obtained, by analogy to eq A-12, from the equation:

$$\ln (a_{jklm}/b_{jklm}) = \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] + \ln [(n_{jkl} + 0.1)/(n_{j\cdot l} - n_{jkl} + 0.5)] - \ln [(n_{jk\cdot} + 2.1)/(n_{j\cdot\cdot} - n_{jk\cdot} + 10.5)] \quad \text{(A-15)}$$

The sum, $S = a_{jklm} + b_{jklm}$, was the same for all j, k, l, and m and was used as an adjustable parameter. The values of a_{jklm} and b_{jklm} are uniquely determined when a value for S has been chosen, and the ratio a_{jklm}/b_{jklm} has been evaluated from eq A-15. This procedure is equivalent to adding a fixed number of observations, S, to a row of the specific medium-range interaction table (e.g., Table III), with the observations distributed among the various conformational states, k, as would be expected from intraresidue information and the nonspecific medium-range interactions. As an example of this procedure, consider the case when $j = -4$, $k = \epsilon$, $l =$ Cys, and $m =$ Ala. For this example, we will take $S = 10$. It can be seen from Table III (first column, third row) that there were no observations of this particular combination in the proteins used for our data set. Reading across row 3 of Table III, we see that there were only three cases in which residue i was Ala and residue $i - 4$ was Cys. There are obviously not enough data for a good estimate of the effects of a Cys at position $i - 4$ on the preference for the Ala at position i to be in the ϵ conformation. There are, however, several observations for $m =$ Ala (row 1 of Table I), and there are also several observations for $j = -4$ and $l =$ Cys (row 3 of Table II). It is, therefore, possible to estimate the logarithm of the odds favoring the ϵ conformation for $j = -4$, $l =$ Cys, and $m =$ Ala on the basis of intraresidue and nonspecific medium-range interactions. This logarithm of the odds is given by eq A-15. Using the values of the terms on the right-hand side from Tables I and II, $\ln (a_{-4,\epsilon,\text{Cys,Ala}}/b_{-4,\epsilon,\text{Cys,Ala}}) = -0.741$. Since $a_{jklm} + b_{jklm} = S$ (= 10 in this example), we obtain $a_{-4,\epsilon,\text{Cys,Ala}} = 3.23$, and $b_{-4,\epsilon,\text{Cys,Ala}} = 6.77$. When these values of a_{jklm} and b_{jklm} are used in eq A-14, it is as if 10 observations had been added to row 3 of Table III. Of these 10 observations, 3.23 were in the ϵ conformation, and the remainder were divided among the other conformations. In this particular case, the addition of 10 observations will significantly alter the logarithm of the odds calculated in the first term of eq A-14, so that it will be almost the same as the logarithm of the odds obtained on the basis of intraresidue and nonspecific medium-range interactions. The a_{jklm} and b_{jklm} dominate the other terms in the logarithm of the odds, for this case. If, instead, we look at the case where $j = -4$, $k = \epsilon$, $l =$ Gly, and $m =$ Ala (row 6 of Table III), the addition of 10 observations to this specific medium-range interaction would not have as large an effect on the logarithm of the odds. The a_{jklm} and b_{jklm} for this case would be small compared with other terms in eq A-14. Thus, the specific medium-range interactions with only a few observations are more strongly affected by this procedure than are those with many observations. Of course, when $S = 0$, specific medium-range interactions always predominate and, for very large values of S, the intraresidue and nonspecific medium-range interactions will predominate. This

method can be used to reduce the statistical errors due to the small number of observations for each specific medium-range interaction. The final estimate of the logarithm of the odds in favor of a conformation is given by:

$$\langle\{\ln [\theta_k/(1-\theta_k)]|R, K\}\rangle = \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] + \sum_{\substack{j=-4 \\ j\neq 0}}^{4} I_j \quad \text{(A-16)}$$

where I_j is the medium-range interaction term given by:

$$I_j = \ln [(n_{jkl} + 0.1)/(n_{j\cdot l} - n_{jkl} + 0.5)] - \ln [(n_{jk\cdot} + 2.1)/(n_{j\cdot\cdot} - n_{jk\cdot} + 10.5)] \quad \text{(A-17a)}$$

when $P_{jklm} > P_{\max}$ for all k

$$I_j = \ln [(n_{jklm} + a_{jklm} + 0.1)/(n_{j\cdot lm} - n_{jklm} + b_{jklm} + 0.5)] - \ln [(n_{km} + 0.1)/(n_{\cdot m} - n_{km} + 0.5)] \quad \text{(A-17b)}$$

when $P_{jklm} \leq P_{\max}$ for at least one k

Estimation of Statistical Errors. The estimate of the logarithm of the odds in favor of a conformation given by eq A-16 is a sum of terms, each of which has a variance determined by eq A-7. With the assumption that the medium-range interaction for each value of j is independent of the short-range interactions and the other medium-range interactions, the variance of the logarithm of the odds is given by:

$$\sigma^2 = \sigma_{km}{}^2 + \sum_{\substack{j=-4 \\ j\neq 0}}^{4} \sigma_j{}^2 \quad \text{(A-18)}$$

since the variance of a sum is equal to the sum of the variances, where $\sigma_{km}{}^2$ is given by eq A-9, and $\sigma_j{}^2$ is the variance in the term representing the medium-range interaction due to residue $i + j$. The form of $\sigma_j{}^2$ depends on whether specific or nonspecific medium-range interactions are used for that value of j in eq A-16. For nonspecific medium-range interactions, the variance in I_j of eq A-17a is given by

$$\sigma_j{}^2 = \sigma_{jkl}{}^2 + \sigma_{jk}{}^2 - 2\rho_{jkl}\sigma_{jkl}\sigma_{jk} \quad \text{(A-19a)}$$

where σ_{jkl} is given by eq A-11, σ_{jk} is given by eq A-13, and ρ_{jkl} is a correlation coefficient with $|\rho_{jkl}| \leq 1$ (Feller, 1968). For specific medium-range interactions, the variance in I_j of eq A-17b is given by

$$\sigma_j{}^2 = \sigma_{jklm}{}^2 + \sigma_{km}{}^2 - 2\rho_{jklm}\sigma_{jklm}\sigma_{km} \quad \text{(A-19b)}$$

where $\sigma_{jklm}{}^2 = (n_{jklm} + a_{jklm} + 1.1)^{-1} + (n_{j\cdot lm} - n_{jklm} + b_{jklm} + 1.5)^{-1}$ and ρ_{jklm} is a correlation coefficient, again with $|\rho_{jklm}| \leq 1$. The values of a_{jklm} and b_{jklm} are determined by eq A-15 and the value of the parameter S. In order to determine the *minimum* value for $\sigma_j{}^2$, the values of ρ_{jkl} and ρ_{jklm} were set at unity in eq A-19a and A-19b, respectively. It should be noted that the estimate of $\sigma_j{}^2$, which depends primarily on the size of the sample, does not include experimental errors in the crystallographic data or errors in the prediction method, such as the assumption of independence of pairwise interactions and the negelct of longer range interactions, which we have no way of estimating.

Reliability of Prediction. In addition to the logarithm of the odds in favor of one conformation over all other conformations, as calculated by eq A-16, another quantity of interest is the logarithm of the odds favoring one conformation over another *single* conformation. In particular, the logarithm of the odds favoring the predicted conformation (that with the highest logarithm of the odds) over each other conformation is a useful quantity. The estimate of the logarithm of the odds favoring the predicted conformation, k', over another conformation k is given by:

$$\langle\{\ln (\theta_{k'}/\theta_k)|R, K\}\rangle = \ln [(n_{k'm} + 0.1)/(n_{km} + 0.1)] + \sum_{\substack{j=-4 \\ j\neq 0}}^{4} I_j' \quad \text{(A-20)}$$

where I_j' is the medium-range interaction term:

$$I_j' = \ln [(n_{jk'l} + 0.1)/(n_{jkl} + 0.1)] - \ln [(n_{k'\cdot} + 0.1)/(n_{k\cdot} + 0.1)] \quad \text{(A-21a)}$$

when $P_{jklm} > P_{\max}$ for all k, and $P_{jk'lm} > P_{\max}$

$$I_j' = \ln [(n_{jk'lm} + a_{jk'lm} + 0.1)/(n_{jklm} + a_{jklm} + 0.1)] - \ln [(n_{k'm} + 0.1)/(n_{km} + 0.1)] \quad \text{(A-21b)}$$

when $P_{jk'lm} \leq P_{\max}$ or $P_{jklm} \leq P_{\max}$ for at least one k.

Equations A-20 and A-21 are equivalent to eq A-16 and A-17, except that, in eq A-20, the logarithm of the odds favoring one conformation over another single conformation (e.g., ϵ vs. α_h) is estimated, while in eq A-16 the odds favoring one conformation against all other conformations (e.g., ϵ vs. α_h, ζ_R, α_L, ζ_L, and α_R) are estimated. The a_{jklm}'s used in eq A-21b are the same as those used in eq A-17b.

The minimum value for the variance in the logarithm of the odds given by eq A-20 can be estimated by:

$$\sigma_{k':k}{}^2 = (n_{k'm} + 1.1)^{-1} + (n_{km} + 1.1)^{-1} + \sum_{\substack{j=-4 \\ j\neq 0}}^{4} \sigma_{j,k':k}{}^2 \quad \text{(A-22)}$$

where

$$\sigma_{j,k':k}{}^2 = (n_{jk'l} + 1.1)^{-1} + (n_{jkl} + 1.1)^{-1} + (n_{k'\cdot} + 1.1)^{-1} + (n_{k\cdot} + 1.1)^{-1} - 2[(n_{jk'l} + 1.1)^{-1} + (n_{jkl} + 1.1)^{-1}]^{1/2}[(n_{k'\cdot} + 1.1)^{-1} + (n_{k\cdot} + 1.1)^{-1}]^{1/2} \quad \text{(A-23a)}$$

for nonspecific medium-range interactions, and

$$\sigma_{j,k':k}{}^2 = (n_{jk'lm} + a_{jk'lm} + 1.1)^{-1} + (n_{jklm} + a_{jklm} + 1.1)^{-1} + (n_{k'm} + 1.1)^{-1} + (n_{km} + 1.1)^{-1} - 2[(n_{jk'lm} + a_{jk'lm} + 1.1)^{-1} + (n_{jklm} + a_{jklm} + 1.1)^{-1}]^{1/2}[(n_{k'm} + 1.1)^{-1} + (n_{km} + 1.1)^{-1}]^{1/2} \quad \text{(A-23b)}$$

for specific medium-range interactions. Equations A-22 and A-23 are equivalent to eq A-18 and A-19, except that eq A-22 gives the variance in the logarithm of the odds favoring one conformation over another single conformation.

Using eq A-20 through A-22, the number of standard deviations by which the predicted conformation is favored over each other possible conformation can be evaluated. The number of standard deviations separating the predicted conformation from the other conformations can be used as an estimate of the reliability of the prediction. For example, in the prediction of trypsin inhibitor, with $P_{\max} = 0.1$ and $S = 40$, the third residue is predicted to be in the ϵ conformation. The conformation with the second highest logarithm of the odds is α_R. When eq A-20 and A-21 are used with $k' = \epsilon$ and $k = \alpha_R$, it is found that $\langle[\ln (\theta_\epsilon/\theta_{\alpha_R})|R, K]\rangle = 0.9$. However, the standard deviation in the estimate of this logarithm of the odds (the square root of the variance from eq A-22) is 1.5. Thus the logarithm of the odds favoring the ϵ conformation over α_R is only 0.6 of a standard deviation greater than 0. A value less

than zero would indicate that the α_R conformation was favored over the ϵ conformation. There is, therefore, a significant probability that the α_R conformation should actually be favored but that the data presently available do not show this. A separation of 0.6 standard deviation is certainly not sufficient to determine with high confidence that the ϵ conformation is

Supplementary Material Available

Additional data and computer programs used in calculations and predictions as noted in the text (37 pages). Ordering information is given on any current masthead page.

References

Alden, R. A., Birktoft, J. J., Kraut, J., Robertus, J. D., and Wright, C. S. (1971), *Biochem. Biophys. Res. Commun. 45,* 337.

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E., Jr., Richardson, D. C., Richardson, J. S., and, in part, Yonath, A. (1971), *J. Biol. Chem. 246,* 2302.

Birktoft, J. J., and Blow, D. M. (1972), *J. Mol. Biol. 68,* 187.

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1974a), *J. Mol. Biol. 82,* 563.

Buehner, M., Ford, G. C., Moras, D., Olsen, K. W., and Rossmann, M. G. (1974b), *J. Mol. Biol. 90,* 25.

Burgess, A. W., Ponnuswamy, P. K., and Scheraga, H. A. (1974), *Isr. J. Chem. 12,* 239.

Burgess, A. W., and Scheraga, H. A. (1975), *Proc. Natl. Acad. Sci., U.S.A. 72,* 1221.

Burnett, R. M., Darling, G. D., Kendall, D. S., LeQuesne, M. E., Mayhew, S. G., Smith, W. W., and Ludwig, M. L. (1974), *J. Biol. Chem. 249,* 4383.

Carter, C. W., Jr., Kraut, J., Freer, S. T., Xuong, N. H., Alden, R. A., and Bartsch, R. G. (1974), *J. Biol. Chem. 249,* 4212.

Chou, P. Y., Adler, A. J., and Fasman, G. D. (1975), *J. Mol. Biol. 96,* 29.

Chou, P. Y., and Fasman, G. D. (1974), *Biochemistry 13,* 222.

Deisenhofer, J., and Steigemann, W. (1975), *Acta Crystallogr., Sect. B 31,* 238.

Diamond, R. (1974), *J. Mol. Biol. 82,* 371.

Drenth, J., Jansonius, J. N., Koekoek, R., and Wolthers, B. G. (1971), *Adv. Protein Chem. 25,* 79.

Edelman, G. M., Cunningham, B. A., Reeke, G. N., Jr., Becker, J. W., Waxdal, M. J., and Wang, J. L. (1972), *Proc. Natl. Acad. Sci. U.S.A. 69,* 2580.

Endres, G. F., Swenson, M. K., and Scheraga, H. A. (1975), *Arch. Biochem. Biophys. 168,* 180.

Feller, W. (1968), An Introduction to Probability Theory and Its Applications, Vol. 1, New York, N.Y., Wiley, pp 230–237.

Fermi, G. (1975), *J. Mol. Biol. 97,* 237.

Fisher, R. A. (1958), Statistical Methods for Research Workers, London, Oliver and Boyd, p 96.

Fletterick, R. J., and Wyckoff, H. W. (1975), *Acta Crystallogr., Sect. A, 31,* 698.

Freer, S. T., Alden, R. A., Carter, C. W., Jr., and Kraut, J. (1975), *J. Biol. Chem. 250,* 46.

Gabel, D., Rasse, D., and Scheraga, H. A. (1976), *Int. J. Peptide Protein Res. 8,* 237.

Hendrickson, W. A., and Love, W. E. (1971), *Nature (London), New Biol. 232,* 197.

Hendrickson, W. A., Love, W. E., and Karle, J. (1973), *J. Mol. Biol. 74,* 331.

Lim, V. I. (1974a), *J. Mol. Biol. 88,* 857.

Lim, V. I. (1974b), *J. Mol. Biol. 88,* 873.

Lim, V. I. (1974c), *Biofizika 19,* 366.

Lindley, D. V. (1964), *Ann. Math. Stat. 35,* 1622.

Lindley, D. V. (1965), Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2, Inference, London, Cambridge University Press.

Mathews, F. S., Levine, M., and Argos, P. (1972), *J. Mol. Biol. 64,* 449.

Matthews, B. W., Weaver, L. H., and Kester, W. R. (1974), *J. Biol. Chem. 249,* 8030.

Maxfield, F. R., and Scheraga, H. A. (1975), *Macromolecules 8,* 491.

Moews, P. C., and Kretsinger, R. H. (1975), *J. Mol. Biol. 91,* 201.

Momany, F. A., McGuire, R., Burgess, A. W., and Scheraga, H. A. (1975), *J. Phys. Chem. 79,* 2361.

Nagano, K. (1973), *J. Mol. Biol. 75,* 401.

Nagano, K. (1974), *J. Mol. Biol. 84,* 337.

Patel, D. J. (1975), *Biochemistry 14,* 1057.

Perutz, M. F., Kendrew, J. C., and Watson, H. C. (1965), *J. Mol. Biol. 13,* 669.

Ponnuswamy, P. K., Warme, P. K., and Scheraga, H. A. (1973), *Proc. Natl. Acad. Sci. U.S.A. 70,* 830.

Quiocho, F. A., and Lipscomb, W. N. (1971), *Adv. Protein Chem. 25,* 1.

Robson, B. (1974), *Biochem. J. 141,* 853.

Robson, B., and Pain, R. H. (1971), *J. Mol. Biol. 58,* 237.

Robson, B., and Pain, R. H. (1974a), *Biochem. J. 141,* 869.

Robson, B., and Pain, R. H. (1974b), *Biochem. J. 141,* 883.

Robson, B., and Pain, R. H. (1974c), *Biochem. J. 141,* 899.

Salemme, F. R., Freer, S. T., Alden, R. A., and Kraut, J. (1973b), *Biochem. Biophys. Res. Commun. 54,* 47.

Salemme, F. R., Freer, S. T., Xuong, N. H., Alden, R. A., and Kraut, J. (1973a), *J. Biol. Chem. 248,* 3910.

Scheraga, H. A. (1974), *Curr. Top. Biochem., 1973,* 1.

Schiffer, M., and Edmundson, A. B. (1967), *Biophys. J. 7,* 121.

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., and Nagano, K. (1974), *Nature (London) 250,* 140.

Shotton, D. M., and Hartley, B. S. (1973), *Biochem. J. 131,* 643.

Shotton, D. M., and Watson, H. C. (1970), *Nature (London) 225,* 811.

Tanaka, S., and Scheraga, H. A. (1975), *Proc. Natl. Acad. Sci. U.S.A. 72,* 3802.

Tanaka, S., and Scheraga, H. A. (1976), *Macromolecules 9,* 142, 168.

Watson, H. C. (1969), *Prog. Stereochem. 4,* 299.