Statistical Mechanical Treatment of Protein Conformation. III. Prediction of Protein Conformation Based on a Three-State Model

Seiji Tanaka[2] and Harold A. Scheraga*

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received August 19, 1975

ABSTRACT: The method proposed for the evaluation of statistical weights in paper I, and the three-state model [α-helical (α), extended (ϵ), and other (c) states] formulated in paper II, have been used to develop a procedure to predict the backbone conformations of proteins, based on the concept of the predominant role played by short-range interactions in determining protein conformation. Conformational probability profiles, in which the probabilities of formation of three consecutive α-helical conformations (triad) and of four consecutive extended conformations (tetrad) have been defined relative to their average values over the whole molecule, are calculated for 19 proteins, of which 16 had been used in paper I to evaluate the set of statistical weights of the 20 naturally occurring amino acids. By comparing these conformational probability profiles to experimental x-ray observations, the following results have been obtained: 80% of the α-helical *regions* and 72% of the extended conformational *regions* have been predicted correctly for the 19 proteins. The percentage of residues predicted correctly is in the range of 53 to 90% for the α-helical conformation and in the range of 63 to 88% for the extended conformation for the 19 proteins in the two-state models [α-helical (α) and other (c) states, and extended (ϵ) and other (c) states]. In the three-state model, the percentage of residues predicted correctly is in the range of 47% to 77 for 19 proteins. These results suggest that the assumption of the dominance of short-range interactions, on which the predictive scheme is based, is a reasonable one. The present predictive method is compared with that of other authors.

In the accompanying papers, we presented a method[3] for evaluating *empirical* statistical weights of various conformational states of amino acids, based on conformations of native proteins determined by x-ray crystallography, and formulated a three-state model[4] for specific-sequence polypeptides that included α-helical (α), extended (ϵ), and other (c) states. These will be referred to here as papers I[3] and II,[4] with equations designated as I-1, II-1, etc.

In the present paper, we use the statistical weights of paper I and the theory of paper II to predict the occurrence of α, ϵ, and c states in proteins, based on the concept (discussed in paper I) that short-range interactions dominate in determining the conformational preferences of the amino acid residues in proteins.[5,6] As stated in paper I, it is intended that such predictions be used to obtain starting conformations for subsequent minimization of the energy of the whole molecule. In section I, we will describe the model and parameters used in this prediction scheme. The methods to calculate conformational-sequence probabilities and conformational probability profiles of proteins are described in section II. The predictive scheme is outlined in section III, and the results of this scheme are presented in section IV and discussed in section V. The use of predictive schemes, augmented by an algorithm incorporating long-range interactions, to study protein folding will be discussed in a forthcoming paper.

I. Theoretical Model and Statistical Weights

A. Three-State Model. In paper II, we formulated a three-state model that is applicable to homopolymers and to specific-sequence copolymers of amino acids.[7] In order to calculate the partition function and molecular averages, a 4 × 4 statistical weight matrix for the ith residue was formulated to correlate the states of residues $i-1$, i, and $i+1$ of the polymer chain (see eq II-13). An approximate nearest-neighbor interaction model, in which the states of residues $i-1$ and i were correlated, was also introduced;[4] a 3 × 3 matrix was required in this model (see eq II-20). Both the 4 × 4 and 3 × 3 matrix formulations involved four parameters ($u_{c,i}$, $w_{h,i}$, $v_{\epsilon,i}$, and $v_{h,i}$). By choosing the c state as a standard state, the number of statistical weights was reduced to three (relative to that of the c state), viz., $w_{h,i}$*, $v_{\epsilon,i}$*, and $v_{h,i}$*. Hence, for example, the 3 × 3 matrix of eq II-20, for the nearest-neighbor interaction model, could be replaced by eq II-26, viz.,

$$\mathbf{W}_i = \begin{bmatrix} 1 & v_{\epsilon,j}^* & (v_{h,j}^*)^2/w_{h,j}^* \\ 1 & v_{\epsilon,j}^* & (v_{h,j}^*)^2/w_{h,j}^* \\ 1 & v_{\epsilon,j}^* & w_{h,j}^* \end{bmatrix}_i \tag{1}$$

where the subscript i is placed on the matrix instead of on the individual statistical weights, and j designates the species of amino acid in the ith residue of the protein.[8]

As mentioned in section III of paper II, the nearest-neighbor interaction model (3 × 3 matrix formulation) of eq II-20 and II-26 is a good approximation to the 4 × 4 matrix formulation of eq II-13. Therefore, in this paper, we use (only) the 3 × 3 matrix formulation of eq II-26 rather than the 4 × 4 matrix formulation of eq II-13 in order to save computer time.[9] For a specific-sequence copolymer such as a protein, the partition function, Z, of the polymer molecule of N residues can be calculated by substituting the statistical weights of the jth species of amino acid[8] (j = 1 to 20 for proteins) at the ith position of the chain for the elements of the matrix operator of eq II-26 by using eq II-21 and II-22 which are given by

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N^* \tag{2}$$

where

$$\mathbf{e}_1 = (1 \quad 0 \quad 0) \tag{3a}$$

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{3b}$$

and $\mathbf{W}_i$ is given in eq 1. A method for computing average molecular quantities will be described in section IIA.

B. Statistical Weights. The *empirical* relative statistical weights for the three-state model ($w_{h,j}^{(3)*}$, $v_{\epsilon,j}$*)[10] that appear in eq 1 were calculated in paper I[3] for the 20 naturally occurring amino acids, using x-ray crystallographic data from native proteins. These are summarized in the second and third columns of Table I (see also Table IV of paper I[3]).

Table I
Relative Statistical Weights $w_{h,j}$*, $v_{h,j}$*, and $v_{\epsilon,j}$* of the jth Amino Acid Used in the Present Prediction Scheme

Amino acid j	Relative statistical weights[a]		
	$w_{h,j}$*[b]	$v_{\epsilon,j}$*[b]	$v_{h,j}$*[c]
Ala	1.549	0.500	0.030
Arg	0.468	0.298	0.026
Asn	0.304	0.152	0.023
Asp	0.481	0.130	0.023
Cys	0.444	0.556	0.031
Gln	0.795	0.432	0.029
Glu	1.188	0.250	0.025
Gly	0.226	0.195	0.024
His	0.535	0.233	0.025
Ile	0.891	0.587	0.032
Leu	1.343	0.702	0.034
Lys	0.726	0.253	0.025
Met	1.000	0.667	0.033
Phe	0.727	0.318	0.026
Pro	0.315	0.333	0.027
Ser	0.336	0.234	0.025
Thr	0.488	0.464	0.029
Trp	1.105	0.421	0.028
Tyr	0.262	0.410	0.028
Val	1.028	0.789	0.036

[a] The statistical weights are given in terms of the three-state model relative to the c state (see ref. 3 and 4 for more details). These statistical weights are assigned tentatively, as described in paper I.[3] [b] Cited from Table IV of ref 3. [c] Assuming that $v_{h,j}^{*(2)} = 0.020$ for all residues in the two-state model, these values for the three-state model are calculated by using eq II-29.

The statistical weights $v_{h,j}$* in eq 1 may be calculated from x-ray data on proteins by either of the following two methods. (i) Using the values of $v_N^{(h)}$* and $v_C^{(h)}$* evaluated for the ends of helical sequences in section IIC of paper I[3] (see also Tables V and VI of paper I), $v_{h,j}$* may be calculated as

$$v_{h,j}^* = [v_N^{(h)*} v_C^{(h)*}]^{1/2} \quad (4)$$

The square root in eq 4 arises from the fact that $v_N^{(h)}$* pertains to the contribution of the ch state of residues $(i-1)$ and i in the chc triad of residues $(i-1)$, i, and $(i+1)$, as seen in

$$v_N^{(h)*} = \sigma_N^{(h)} w^* \quad (5)$$

(which is equivalent to eq I-33), and $v_C^{(h)}$* pertains to the contribution of the hc state of residues i and $(i+1)$ in the chc triad, as seen in

$$v_C^{(h)*} = \sigma_C^{(h)} w^* \quad (6)$$

Equation 4 may be rewritten, with the aid of eq 5 and 6, as

$$v_{h,j}^* = [\sigma_N^{(h)} \sigma_C^{(h)}]^{1/2} w^* \quad (7)$$

Equation 4 (or 7) may be used to calculate the relative statistical weight $v_{h,j}$* for both the two-state and three-state models. (ii) Using atomic coordinates from x-ray structures[11] for isolated h states, i.e., those that are *not* involved in hydrogen bonding (in contrast to the hydrogen-bonding description used to detect the ends of helical sequences in paper I), the statistical weights $v_{h,j}$* can be calculated by the method described in ref 46 of paper I.[3] In this method, the ends of helical sequences are not used to evaluate $v_{h,j}$*.

However, since both ends of a helical sequence are not structurally equivalent for a homopolymer, and even less so for a specific-sequence copolymer, the value of $v_{h,j}$* from method (i) does not correspond to a specific structure, but can be used effectively in a nearest-neighbor model. On the other hand, method (ii) does involve a direct relation between $v_{h,j}$* and molecular structure. The statistical weights $v_N^{(h)}$* and $v_C^{(h)}$*, and hence $v_{h,j}$* from method (i), involve a lower statistical reliability than $w_{h,j}$* and $v_{\epsilon,j}$* because there are fewer data for the ends of helical sequences within the limited x-ray data set used in paper I;[3] the x-ray observations from which the statistical weights of paper I were deduced did not identify isolated h states, but identified only helical sequences. In the future,[11] the values of $v_{h,j}$* will be calculated by method (ii), using more reliable x-ray coordinates than are available at the present time. Therefore, instead, in this paper, we make a tentative assignment of the values of $v_{h,j}^{(3)*}$ based on the value of $v_h^{(2)*}$ which was estimated from theoretical calculations and experimental observations. For this purpose, we first quote the theoretical values of $v_h^{(2)}$ for Ala and Gly, which were obtained with a molecular theory of the helix–coil transition of polyamino acids.[12] In brief, following ref 12, the value of $v_h^{(2)}$ can be calculated by integrating the Boltzmann factor of the potential energy of the amino acid residue over the small region of the α-helical state as defined in eq II-5. Similarly, integration over the whole space gives the statistical weight $u_c^{(2)}$. The values of $v_h^{(2)*}$ for the two-state model of the helix–coil transition of poly(L-alanine) and poly(glycine) (at 300°K), evaluated in ref 12, are given, together with the enthalpic and entropic contributions $\Delta H_{[v_h^{(2)*}]}$ and $\Delta S_{[v_h^{(2)*}]}$ to $v_h^{(2)*}$, in rows 1 and 2 of Table II. In the same table, experimental values[13] of $v_h^{(2)*}$, which corresponds to this parameter of the Lifson–Roig theory,[14] are also summarized. As seen in Table II, it seems that $v_h^{(2)*}$ does not vary drastically among the amino acids. Therefore, in this paper, the calculated value of $v_{h,j}^{(2)*}$ for alanine, viz., 0.02, which is also in reasonable agreement with experimental values, will be used, as a first approximation, for the values of $v_h^{(2)*}$ for the other 19 amino acids. Since the values of v_h* may be expected to vary among the 20 amino acids, as seen in Table XI of paper I,[3] this approximation should be considered as a tentative one until the values of v_h* for all 20 amino acids are evaluated either empirically[11] or theoretically.[12] Making this approximation, it is then necessary to convert the value of $v_{h,j}^{(2)*}$ of Ala for the two-state model into the value of $v_{h,j}^{(3)}$ of Ala for the three-state model. Similarly, $v_{h,j}^{(2)*}$ is converted to $v_{h,j}^{(3)*}$; i.e., the values of $v_h^{(3)*}$ (for the three-state model) given in the fourth column of Table I were converted from the values of $v_h^{(2)*}$ (for the two-state model), which were all assumed to be 0.02, by using eq I-18 or II-29. Thus, all the parameters, which are needed to construct the statistical weight matrix of eq 1, are now available.

II. Calculation of Conformational-Sequence Probabilities and of Conformational Probability Profiles of Proteins

A. Conformational-Sequence Probabilities. The probability, $P(i|n|\{\rho\})$, of finding a sequence of n amino acid residues in a certain specific conformational state $\{\rho\}$, starting at the ith residue of the chain, is a useful quantity for determining the most stable or probable conformations of protein molecules. The conformational-sequence probability, $P(i|n|\{\rho\})$, was introduced in section VI of paper II.[4] As demonstrated in paper II, the number of amino acid residues, n, can be any finite value (up to the chain length N), and the conformational-sequence $\{\rho\}$ can be any combination of any possible states η_i for a residue, e.g., the α-helical (α), extended conformational (ϵ), and other (c) states for the present scheme using a three-state model, as shown in eq II-38, i.e.,

$$\{\rho\} = \eta_i \eta_{i+1} \cdots \eta_{i+n-1} \quad (8)$$

According to the formulation of paper II,[4] the probability

Table II
Theoretical and Experimental Values of $v_h^{(2)*}$ and Related Quantities for Various Polyamino Acids

	Polyamino acid	$\Delta H_{[v_h^{(2)*}]}$, kcal/mol	$\Delta S_{[v_h^{(2)*}]}$, eu	$v_h^{(2)*}$
Theor calculation[a]	Alanine	1.35	−3.48	0.018[b]
	Glycine	0.71	−4.38	0.034[b]
Exptl[c] observation	Alanine			0.012[d]
	N^5-(3-Hydroxypropyl) L-glutamine			0.017
	Glutamic acid			0.05
	γ-Benzyl L-glutamate			0.014

[a] Cited from ref 12. [b] At 300°K. [c] The data are cited from Table III of ref 13. [d] Assumed to be independent of temperature in the range of 10 to 80°C.

$P(i|n|\{\rho\})$ may be calculated by using either eq II-41 or II-54.

For the present purpose, to predict the α-helical (α), extended conformation (ε), and other state (c) *regions* of proteins, we are concerned only with specific *regular* conformational sequences of n residues, such as $\{\rho\} = \alpha\alpha\alpha \ldots \alpha\alpha\alpha$, $\{\rho\} = \varepsilon\varepsilon\varepsilon \ldots \varepsilon\varepsilon\varepsilon$, and $\{\rho\} = \mathrm{ccc} \ldots \mathrm{ccc}$, but not with combined conformations of α, ε, and c. As will be discussed in section IIB, in order to detect the possible conformational states of proteins, we have to vary the position of the ith residue and the length of the sequence, n. In this respect, it was pointed out[4] that it is more advantageous to use eq II-54 in computing $P(i|n|\{\rho\})$ since, in eq II-41, the matrix multiplication has to be repeated every time that n and i are varied, even for a protein, which is an extremely time-consuming process for computer calculations. On the other hand, the use of eq II-54, i.e.,

$$P(i|n|\{\rho\}) = F_{i;\eta_i}\left[\prod_{k=i}^{i+n-1} P_{k+1;\eta_k\eta_{k+1}}\right]_{\{\rho\}} \tag{9}$$

where $F_{i;\eta_i}$ is the first-order a priori probability of finding the ith residue in conformational state η_i and $P_{k+1;\eta_k\eta_{k+1}}$ is the conditional probability that (given that the kth residue is in conformational state η_k) the $(k+1)$th residue is in conformational state η_{k+1}, can save considerable computer time; this arises because, after once computing the values of $F_{i;\eta_k}$ and $P_{k+1;\eta_k\eta_{k+1}}$ for all N residues of the protein, the computation of $P(i|n|\{\rho\})$ may be performed by simple multiplication of the values of P_{k+1} corresponding to the conformational state $\{\rho\}$. The values of $F_{i;\eta_i}$ and $P_{k+1;\eta_k\eta_{k+1}}$ can be calculated by using eq 1–3, together with eq II-42 and II-41, and together with eq II-63, II-43, and II-41, respectively, for any starting position i of any length of sequence n, in general (see section VI of paper II[4] for more details, and section IIB and III of this paper for more specific details for the present special case in which we are concerned only with regular sequences).

B. Calculation of Conformational Probability Profiles of Proteins. In order to obtain quantitative information about the tendencies for sequences of n residues to be in the α-helical (α), extended conformational (ε), and other (c) states, we compute the probabilities $P(i|n|\{\rho\})$ of finding a sequence of n residues in the conformational states

$$\{\rho\} = \alpha_i\alpha_{i+1}\alpha_{i+2} \ldots \alpha_{i+n-1} \tag{10a}$$

$$\{\rho\} = \varepsilon_i\varepsilon_{i+1}\varepsilon_{i+2} \ldots \varepsilon_{i+n-1} \tag{10b}$$

and

$$\{\rho\} = \mathrm{c}_i\mathrm{c}_{i+1}\mathrm{c}_{i+2} \ldots \mathrm{c}_{i+n-1} \tag{10c}$$

To detect α-helical and extended conformation sequences of various lengths, n, we would have to calculate $P(i|n|\{\alpha\})$ and $P(i|n|\{\varepsilon\})$ for all possible values of n, as a function of i. However, this would require an extensive amount of computer time. Therefore, to develop a convenient predictive method, we shall compute $P(i|n|\{\rho\})$ only for short sequences, and introduce empirical rules to detect α-helical and extended conformations in proteins, as described below and in section III.

To detect the α-helical and extended conformation regions in proteins, we shall first compute the conformational-sequence probabilities $P(i|n|\{\alpha\})$ and $P(i|n|\{\varepsilon\})$ for short sequences of α-helical and extended conformations. As a minimum length of such sequences, we choose those originally defined in paper I:[3] in section IA of paper I, for the minimum length of an α-helical sequence, we took a sequence of three amino acid residues, which can form one hydrogen bond if these three residues are in the α-helical state; this is also the conformational state corresponding to the statistical weight w_h or w_h^* for the ith residue, described in section I of paper II; the minimum length of an extended conformation is taken to be four residues, as described in section IB of paper I to calculate v_ε^* from x-ray data. Thus, we compute the probabilities $P(i|n|\{\alpha\})$ of finding a sequence of three ($n = 3$ in eq 10a) and of four residues ($n = 4$ in eq 10b) in the conformational states $\alpha\alpha\alpha$ and $\varepsilon\varepsilon\varepsilon\varepsilon$, respectively; this may be written as $P(i|3|\{\alpha\})$ and $P(i|4|\{\varepsilon\})$. The method for computing $P(i|3|\{\alpha\})$ and $P(i|4|\{\varepsilon\})$ has already been given in terms of a general expression in the last paragraph of section IIA; for computing $P(i|3|\{\alpha\})$, 3 and $\alpha\alpha\alpha$ are substituted for n and $\{\rho\}$, respectively, of eq 9, and for $P(i|4|\{\varepsilon\})$, 4 and $\varepsilon\varepsilon\varepsilon\varepsilon$ are substituted. Computations of $P(i|3|\{\alpha\})$ and $P(i|4|\{\varepsilon\})$ are performed for all residues i of the chain (i.e., for $1 \leq i \leq N - n + 1$) to obtain the probabilities of occurrence of all possible triads $\alpha\alpha\alpha$ and tetrads $\varepsilon\varepsilon\varepsilon\varepsilon$ of the chain.

In this connection, in order to predict the helical conformations of proteins, Lewis et al.[15] used an α-helical probability which corresponds to $P(i|1|\alpha)$ in eq 9 (which is also defined as a first-order a priori probability $F_{i;\alpha}$ in eq II-42 in paper II). It should be noted that $P(i|3|\{\alpha\})$ is not obtained by simply multiplying the first-order a priori probabilities $F_{i;\alpha}$, $F_{i+1;\alpha}$, and $F_{i+2;\alpha}$, as demonstrated in eq II-65.

In order to locate the three-residue α-helical segments and the four-residue extended conformational segments of proteins, we test whether the α-helical probability $P(i|3|\{\alpha\})$ for the three residues i, $i+1$, and $i+2$ or the extended conformation probability $P(i|4|\{\varepsilon\})$ for the four residues i, $i+1$, $i+2$, and $i+3$ equals or exceeds the mean α-helical probability of a triad (i.e., $\alpha\alpha\alpha$), $\theta_\alpha^{(3)}$, or the mean extended conformation probability of a tetrad (i.e., $\varepsilon\varepsilon\varepsilon\varepsilon$), $\theta_\varepsilon^{(4)}$, which can be calculated from

$$\theta_\alpha^{(3)} = \frac{1}{N-2}\sum_{i=1}^{N-2} P(i|3|\{\alpha\}) \tag{11}$$

and

$$\theta_\varepsilon^{(4)} = \frac{1}{N-3}\sum_{i=1}^{N-3} P(i|4|\{\varepsilon\}) \tag{12}$$

Therefore, we define relative probabilities $P^*(i|3|\{\alpha\})$ of

finding any triad of the chain in the $\alpha\alpha\alpha$ conformation and $P^*(i|4|\{\epsilon\})$ of finding any tetrad of the chain in the $\epsilon\epsilon\epsilon\epsilon$ conformation by the following equations:

$$P^*(i|3|\{\alpha\}) = P(i|3|\{\alpha\})/\theta_\alpha{}^{(3)} \quad (13)$$

and

$$P^*(i|4|\{\epsilon\}) = P(i|4|\{\epsilon\})/\theta_\epsilon{}^{(4)} \quad (14)$$

Then, if the value of $P^*(i|3|\{\alpha\})$ or $P^*(i|4|\{\epsilon\})$ equals to or exceeds unity, the triad of residues i, $i + 1$, and $i + 2$ may be considered as a *possible* α-helical triad of *relatively high* probability, or the tetrad of residues i, $i+1$, $i + 2$, and $i + 3$ as a *possible* extended conformation tetrad of *relatively high* probability in the protein chain [it should be noted that, at this stage, we have not yet made a choice as to whether residues i, $i + 1$, and $i + 2$ (and $i + 3$) are in α-helical or extended conformation states].

Thus, the conformational probability profiles (i.e., the relative probabilities $P^*(i|3|\{\alpha\})$ and $P^*(i|4|\{\epsilon\})$) have been calculated as a function of i. For comparison, we have also computed the relative probability $P^*(i|3|\{c\})$ and $P^*(i|4|\{c\})$ for the triad and tetrad of other (c) states because, if $P^*(i|3|\{c\})$ or $P^*(i|4|\{c\})$ equals or exceeds unity, the c states would then have to be regarded as *possible* conformational states for the triad or tetrad. The criterion for determining the most probable conformation out of the three states (α, ϵ, and c) is described in section III. It should be noted that $P^*(i|3|\{\alpha\})$ is always compared with $P^*(i|3|\{c\})$ and, independently, $P^*(i|4|\{\epsilon\})$ is always compared with $P^*(i|4|\{c\})$.

III. Predictive Scheme

In this section, we will describe the predictive method and, in section IV, the application of the predictive scheme to bovine pancreatic trypsin inhibitor will be described as an illustrative example.

In section II, we have computed the conformational probabilities $P^*(i|3|\{\alpha\})$ and $P^*(i|4|\{\epsilon\})$, as well as $P^*(i|3|\{c\})$ and $P^*(i|4|\{c\})$, for the same $\alpha\alpha\alpha$ triads and $\epsilon\epsilon\epsilon\epsilon$ tetrads, respectively, for all values of i. The relative stabilities of triads or tetrads in α, ϵ, and c states depends on whether these relative probabilities equal or exceed unity. We then examine longer runs of α, ϵ, or c states by combining the triads and tetrads into clusters of α, ϵ, or c states, respectively, without allowing an intervening triad or tetrad to have a relative probability smaller than unity. For example, if $P^*(i|3|\{\alpha\}) \geq 1$ for all three consecutive triads [$P^*(i|3|\{\alpha\})$ for $(i, i + 1, i + 2)$, $P^*(i + 1|3|\{\alpha\})$ for $(i + 1, i + 2, 1 + 3)$, and $P^*(i + 2|3|\{\alpha\})$ for $(i + 2, i + 3, i + 4)$], then the *five* residues from i to $i + 4$ can be regarded as a highly probable α-helical sequence. It should be noted that $P^*(i + 3|3|\{\alpha\})$ and $P^*(i + 4|3|\{\alpha\})$ do not have to be larger than unity, even though residues $i + 3$ and $i + 4$ are involved in the α-helical sequence from residue i to residue $i + 4$. If any given sequence can thus be unambiguously assigned to α, ϵ, or c conformations, then the prediction for that sequence is finished. However, it sometimes happens that more than one of the relative probabilities of such a cluster of α, ϵ, or c states will be equal to or greater than unity, i.e., there is then a duplicate choice as to whether the cluster is in an α or c, or ϵ or c conformation; in such a case, we proceed as follows to resolve this ambiguity.

We begin by determining the *tendency* of *sequences* (or clusters) to adopt a given conformation. This conformational tendency of the sequence under consideration is characterized by the *largest* value of $P^*(i|3|\{\alpha\})$, $P^*(i|4|\{\epsilon\})$, $P^*(i|3|\{c\})$, or $P^*(i|4|\{c\})$ among the triads or tetrads found in the sequence. This sequence is said to have a *"weak tendency"* toward formation of the given conformation if $1 \leq P^*(i|3|\{\alpha\}) \leq 2$ [or $1 \leq P^*(i|3|\{c\}) \leq 2$] for *every* triad in the helical [or c state] sequence or $1 \leq P^*(i|4|\{\epsilon\}) \leq 2$ [or $1 \leq P^*(i|4|\{c\}) \leq 2$] for *every* tetrad in an extended [or c state] sequence. This sequence is said to have a *"strong tendency"* toward formation of the given conformation if at least one of $P^*(i|3|\{\alpha\})$ [or $P^*(i|3|\{c\})$] or $P^*(i|4|\{\epsilon\})$ [or $P^*(i|4|\{c\})$] is greater than 2 among the triads and tetrads, respectively.

We then adopt a three-stage process to assign the conformational states of a sequence *unambiguously*, in those cases where a duplicate choice was found to be possible.

A. Stage I. In stage I, we make the decision as to whether a given sequence has a chance (its "first possibility") to be in the α or ϵ regions (if it is found to be in neither of these, it is assigned to the c region). This decision is based on the following conditions.

1. Compare the sequence of values of $P^*(i|3|\{\alpha\})$ with corresponding values of $P^*(i|3|\{c\})$ in a cluster. If one of the values of $P^*(i|3|\{\alpha\})$ is larger than all values of $P^*(i|3|\{c\})$ in the cluster, then the "first possibility" is α, and vice versa if one of the values of $P^*(i|3|\{c\})$ is larger than all values of $P^*(i|3|\{\alpha\})$. This enables a choice to be made between α and c. A similar procedure is applied independently to $P^*(i|4|\{\epsilon\})$ and $P^*(i|4|\{c\})$, and an independent choice is made between ϵ and c. It has never happened, if c were chosen over α, that ϵ would be chosen independently over c, and vice versa. Thus, at this stage, the cluster is either c, or duplicately assigned as α or ϵ; all the c states have been determined.

2. If the computed values of the relative probabilities indicate that *both* sequences α and c (or independently, ϵ and c) are "weak tendency" ones (i.e., if the largest probabilities in the sequence lie in the range $1 \leq$ [*both* $P^*(i|3|\{\alpha\})$ and $P^*(i|3|\{c\})$] ≤ 2 or $1 \leq$ [*both* $P^*(i|4|\{\epsilon\})$ and $P^*(i|4|\{c\})$] ≤ 2), then conditions (a) and (b) are imposed in order to determine the boundary between α and c (and ϵ and c) regions. (a) An *isolated* α triad or ϵ tetrad with a "weak tendency" is ignored (but *isolated* ones with a "strong tendency" are retained). (b) An α sequence is terminated after the *second* residue of the last triad for which $1 \leq P^*(i|3|\{\alpha\}) \leq 2$ and $1 \leq P^*(i|3|\{c\}) \leq 2$; the c region can extend only one residue into the C-terminal portion of an α sequence. Similarly, an ϵ sequence is terminated after the *third* residue of the last tetrad for which $1 \leq P^*(i|4|\{\epsilon\}) \leq 2$ and $1 \leq P^*(i|4|\{c\}) \leq 2$; again, the c region can extend only one residue into the C-terminal portion of an ϵ sequence.

At the end of stage I, "first possibilities" have been assigned, i.e., the sequence is either c or (α or ϵ). If the region has *not* been duplicately assigned to α and ϵ, but only to one of these, then this region has its final assignment if condition II-3 is satisfied for α-helical conformations (and condition III for any conformation).

B. Stage II. In stage II, we select between duplicate α and ϵ assignments by the following criteria.

1. We select α or ϵ (consistent with conditions II-3 and III) depending on which sequence has the highest relative probability in *one* of its triads (or tetrads), for both "weak" and "strong" tendency conformations.

2. If a sequence was assigned as α in stage II-1, and if some portion of the ϵ sequence extends beyond the overlap region with the α sequence, this remaining portion of the ϵ region is assigned as ϵ if the remaining region involves ≥ 4 residues and satisfies condition III. Likewise, if a sequence was assigned as ϵ in stage II-1, and if some portion of the α sequence extends beyond the overlap region with the ϵ sequence, this remaining portion of the α region is assigned as α if the remaining region involves ≥ 3 residues and satisfies conditions II-3 and III.

3. We now introduce a criterion for α-helix formation that involves electrostatic interactions. In the α-helical

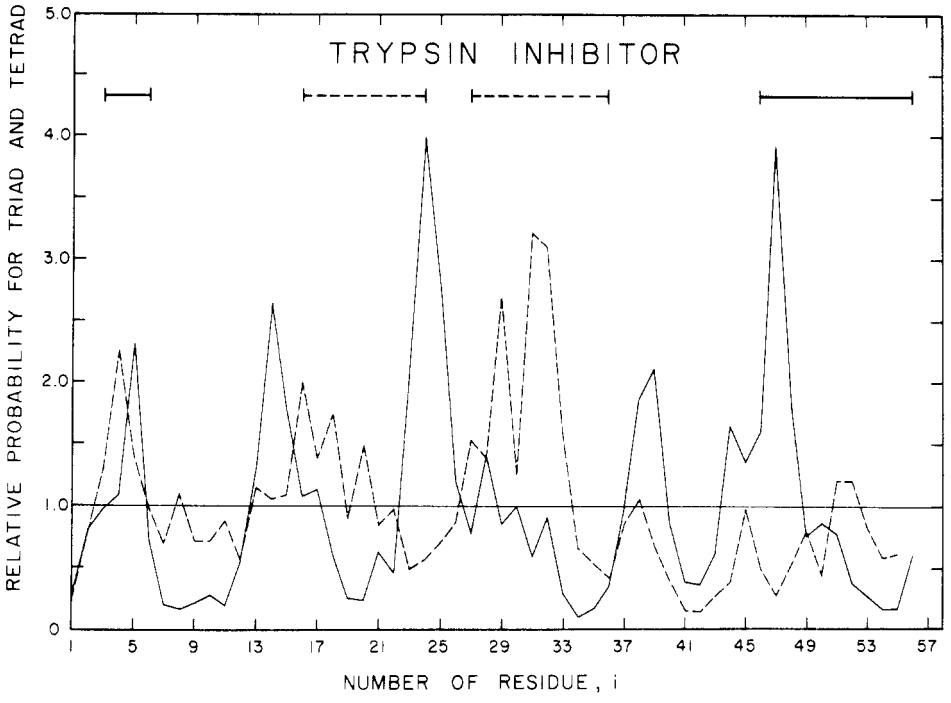

Figure 1. Conformational probability profiles of α-helical and extended conformations for bovine pancreatic trypsin inhibitor.[21] The relative probabilities $P^*(i|3|\{\alpha\})$ for the α-helical state and $P^*(i|4|\{\epsilon\})$ for the extended state are plotted as a function of i, the position of the initial residue of a triad of α-helical states or a tetrad of extended conformational states, which are designated by solid and dashed curves, respectively. The observed regions[21] of α-helical and extended conformations are denoted by solid and dashed lines in the upper part of the figure.

conformation, the side chains of residues i and $i + 3$ and i and $i + 4$ can approach each other closely. According to Zimm and Rice,[16] the distances between the charges of the COO^- groups of the ith side chain with side chains $i + 1$, $i + 2$, $i + 3$, and $i + 4$ of poly(L-glutamic acid) in the α-helical conformation are 10.1, 13.1, 7.9, and 7.5 Å, while those in the (extended) random coil are ca. 10, 10, 10, and 16 Å. This indicates that the electrostatic interactions between the side chains of residues i and $i + 3$, and of i and $i + 4$, act to stabilize the random coil over the α-helical conformation. If the side chains of residues i and $i + 3$, and i and $i + 4$, in protein molecules are in the same charged state, electrostatic interactions can destabilize the α-helical conformation. In principle, electrostatic interactions between charged side chains of protein molecules can be incorporated into our formulation of the three-state model described in paper II. However, the size of the statistical weight matrix whould have to be increased, and this would require considerably more computer time to calculate probability profiles of protein molecules.[17,18] Therefore, in this predictive method, we will take electrostatic interactions between side chains into consideration by a simple empirical rule in order to conserve computer time. We thus adopt the following rule; the α-helical conformation cannot form if the same state of charge is present at residues i and $i + 3$ and/or i and $i + 4$ without an intervening or neighboring oppositely charged amino acid. For this purpose, His was taken as a positively charged residue.

It should be noted that Maxfield and Scheraga[19] found evidence for such i to $i + 4$ electrostatic interactions in proteins, and that Lewis and Bradbury[20] took such electrostatic interactions into account in an empirical predictive scheme.

C. Stage III. As a final confirmation, the relative probabilities for *whole* sequences in the α and β conformations assigned in stages I and II are required to equal or exceed unity. For this purpose, the relative probability for a sequence of n residues is calculated in a similar manner to that shown in eq 11–14, by using eq II-53 and II-63, viz.,

$$P^*(i|n|\{\rho\}) = P(i|n|\{\rho\})/\theta_{|\rho|}{}^{(n)} \qquad (15)$$

where $\{\rho\}$ should be replaced by $\{\alpha\}$ or $\{\epsilon\}$ for the α or ϵ sequences, respectively, and n is the length of the sequence (not simply 3 or 4) being examined. If $P^*(i|n|\{\rho\}) < 1$, the sequence of n residues has to be reduced, residue by residue, until $P^*(i|n|\{\rho\}) \geq 1$. However, in practice throughout this work, we found that all assignments made up to this stage also satisfied the condition $P^*(i|n|\{\rho\}) \geq 1$ without altering the regions assigned. Therefore, the condition of stage III may be regarded only as a confirmatory one, but one which demonstrates that the empirical procedure introduced here provides a valid substitute for the more extensive computation of $P^*(i|n|\{\rho\})$ for *all* sequences.

IV. Results

The values of $P^*(i|3|\{\alpha\})$, $P^*(i|4|\{\epsilon\})$, $P^*(i|3|\{c\})$, and $P^*(i|4|\{c\})$ have been computed as a function of i for respective triads and tetrads in the proteins that had been used (in paper I) to obtain the statistical weights of Table I, and in three proteins[21-23] that had not been included in the original data set in paper I. The conformational probability profiles for the latter three proteins are shown in Figures 1 to 3. In these figures, a baseline is drawn at a relative probability of unity.

According to stage I in section III, the parts of the curves above the baseline are assigned as *possible* α-helical or extended conformations. Since only the first position of a triad or tetrad is plotted in Figures 1 to 3, the length of an α-helical or extended conformation sequence should be augmented by two and three residues, respectively, when interpreting these figures.

Using the probabilities $P^*(i|3|\{\alpha\})$ and $P^*(i|4|\{\epsilon\})$, together with $P^*(i|3|\{c\})$ and $P^*(i|4|\{c\})$ which are needed in stage I-2b, and following the predictive scheme of section III, we

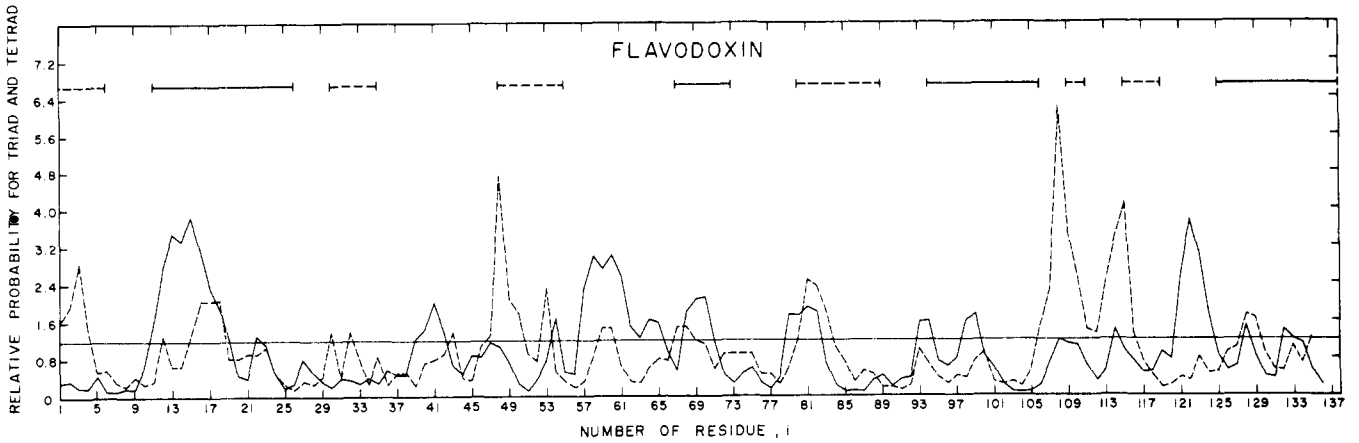

Figure 2. The same as Figure 1, but for clostridial flavodoxin.[22]

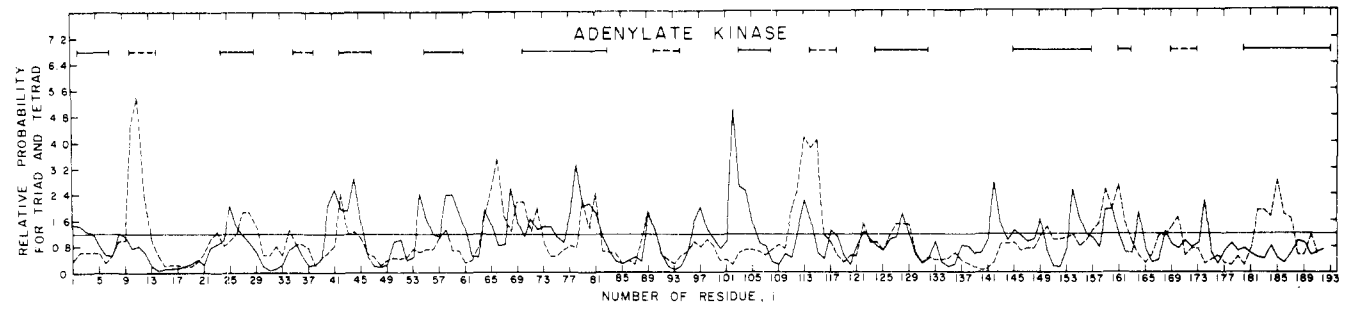

Figure 3. The same as Figure 1, but for porcine muscle adenylate kinase.[23]

have obtained the predictive results (summarized in Table III) for the proteins which were included in the original data set. Predictions were also made (Table IV) for the three proteins that were not included in the original data set. For purposes of illustration, we will describe the predictive process in detail for bovine pancreatic trypsin inhibitor since this protein contains only 58 residues, and the x-ray data on its native conformation have been reported.[21]

For bovine pancreatic trypsin inhibitor, the values of $P^*(i|3|\{\alpha\})$, $P^*(i|4|\{\epsilon\})$, $P^*(i|3|\{c\})$, and $P^*(i|4|\{c\})$ have been computed according to the procedure described in section II. The numerical values of these probabilities are given in columns 3, 5, 6, and 8 of Table V, and the conformational probability profiles for the α-helical and extended conformations are given in Figure 1. The probabilities $P^*(i|3|\{c\})$ and $P^*(i|4|\{c\})$ are not drawn in Figure 1 to avoid overcrowding. However, the numerical values of $P^*(i|3|\{c\})$ and $P^*(i|4|\{c\})$ for bovine pancreatic trypsin inhibitor are given in columns 5 and 8 of Table V.

First, all possible α-helical and extended conformation regions are chosen independently, according to stage I-1; these are shown by the solid vertical lines in columns 4 and 7 of Table V. The regions designated by broken vertical lines in columns 4 and 7 (i.e., residues 28–30 in the α-helix region, and residues 8–11, 20–23, and 38–41 in the extended conformation region) are the isolated triad and tetrads, which are ignored as described in stage I-2a. Therefore, the first candidates for α-helical regions are residues 4–7, 13–19, 23–28, 38–41, and 44–50, while those for extended regions are residues 3–8, 13–21, 27–36, and 51–55. According to stage I-2b (which is applied only to weak-tendency regions), a weak tendency region is terminated by a residue which has a value of $P^*(i|3|\{c\})$ or $P^*(i|4|\{c\})$ for the c state equal to or larger than unity. The only region in which this condition is encountered is the extended region of residues 51–55. Therefore, following stage I-2b, the extended region of residues 51–55 is shortened to 51–54.

The regions which are not duplicately assigned are the α-helical regions 38–41 and 44–50, and the extended region 51–54. However, among these three regions, the α-helical region 38–41 has to be discarded because of stage II-3, i.e., because of the effect of electrostatic interactions which are shown in column 2, where the + and − signs indicate positively charged and negatively charged amino acid residues.

We next advance to stage II. Since all of the other α and extended conformation regions are duplicately assigned, we will make a definite assignment according to stage II. The duplicate assignment of α-helical 4–7 and extended 3–8 is resolved by selecting the α-helical choice because the α-helical sequence 4–7 involves residue 5, for which $P^*(5|3|\{\alpha\}) = 2.32$; this is larger than $P^*(4|4|\{\epsilon\})$ for residue 4, the largest value of $P^*(i|4|\{\epsilon\})$ in the extended sequence 3–8. Since the remaining residues 3 and 8 cannot form the extended region (see stage II-2), they are not assigned to the extended region. In a similar manner, the duplicate assignment of α-helical region 13–19, and of extended region 13–21, is resolved by selecting the α-helical assignment. However, because of the effect of electrostatic interaction (stage II-3), the region 17–20 cannot form the α-helical conformation (see column 2 of Table V). Therefore, only the remaining region 13–16 has to be assigned as α helical. Furthermore, the remaining region 17–21 satisfies stage II-2 [i.e., the sequence 17–21 consists of more than four residues and $P^*(17|5|\{\epsilon\}) \geq 1.0$, the latter condition satisfying stage III], which means that the region 17–21 is assigned to be extended (see also the last column of Table V). Then, the duplication between the α-helical region 23–28 and the extended region 27–36 is resolved by selecting the α-helical conformation because the value of $P^*(24|3|\{\alpha\}) = 4.00$, which is larger than $P^*(31|4|\{\epsilon\}) = 3.23$, as seen in columns

Table III
Predicted and Experimentally Observed α-Helical and Extended Conformation Regions for Proteins Involved in the Original Data Set[a]

Protein[a]	No. of amino acid residues	α-Helical region			Extended-conformation region		
		Obsd[b]	Predicted[c]	$P^*(i\|n\|\{\alpha\})$[d]	Obsd[b]	Predicted[c]	$P^*(i\|n\|\{\epsilon\})$[d]
Myoglobin	153	4–17	13–20	3.61	None	6–12	1.84
		21–35				27–32	1.49
		37–41	37–41	1.73		65–72	5.90
		52–57	50–57	2.70		112–117	1.10
		59–76	73–77	2.19		127–130	1.83
		87–94	80–87	2.77			
			88–91	1.49			
		101–117	103–107	1.96			
			108–111	2.18			
			124–126	1.57			
		125–148	132–137	4.31			
			141–145	1.39			
			146–149	1.28			
Lysozyme	129	5–14	4–13	15.7	43–46	40–43	2.17[e]
		25–35	26–34	2.43		53–59	1.39
		81–84	80–84	1.48		88–92	2.72
		89–95	93–98	1.15		118–125	2.09
		110–113	104–111	5.44			
		121–123					
Ribonuclease S	124	4–12	1–10	8.10	43–47	45–48	2.23
			17–21	1.41	61–65	62–66	1.15
		25–33	27–30	2.05	77–88	78–82	1.40
		51–55	49–57	3.67	95–111	95–98	1.48
			99–103	1.56		104–111	2.88
						114–119	2.20
Oxyhemoglobin α chain	141	4–17	9–14	4.23	None	1–4	2.12
		21–35	24–32	9.42		65–72	2.03
		37–41[f]				103–114	5.90
		53–70	51–57	2.14		132–138	2.80
			59–63	2.07			
			78–80[e]	2.33			
		81–88	84–89	1.54			
		95–111	97–101	1.36			
		119–137	119–123	2.44			
Oxyhemoglobin β chain	146	5–17	5–17	28.9	None	29–40	2.84
		20–33	19–24	2.19		103–107	1.46
			25–28	2.75		108–117	3.33
		36–40[f]				130–135	2.31
		51–55	51–53[g]	1.14			
		57–75	58–62	1.20			
		86–93	84–89	2.39			
		100–116					
		124–142	124–128	2.13			
			136–142	1.99			
α-Chymotrypsin B chain	131	None	3–8	1.91	4–9		
			29–32	1.01	15–20	12–19	2.10
			(61–73)	(12.5)	27–31		
			94–98	1.45	36–40	35–43	2.03
			112–117	2.39		43–47	1.36
					49–53	49–54	1.23
					67–70	67–71	1.30
					88–93	88–94	2.66
					95–100	99–111	4.73
					103–110		
					119–125	119–125	2.19
α-Chymotrypsin C chain	97		4–9	1.91	6–16	10–16	1.60
		17–24	25–31	2.77	32–35	32–35	1.73
		88–96	82–96	17.30	41–44		
					49–55	49–53	1.53
					58–61	58–65	4.13
					76–82	78–81	1.13
Carboxypeptidase A	307	15–25	19–22	2.20	11–14	9–15	2.03
			29–32	1.20		23–28	1.52
		73–87	70–74	2.05	33–37	33–36	1.27
			79–85	6.83	45–53	45–51	2.23
		95–99	95–101	1.85	60–66	59–65	1.91
			(102–108)	(3.72)		75–78	1.84
		113–121			103–107	104–112	1.12
			(123–127)	(1.22)		129–133	1.27
			152–155	1.35		136–141	2.19
		174–186	171–177	2.66	191–198	198–206	3.20
			188–193	3.78	201–205		

Table III (Continued)

Protein[a]	No. of amino acid residues	α-Helical region			Extended-conformation region		
		Obsd[b]	Predicted[c]	$P^*(i\|n\|\|\alpha\|)$[d]	Obsd[b]	Predicted[c]	$P^*(i\|n\|\|\epsilon\|)$[d]
		216–230	214–220	1.40		208–212	1.10
			220–231	22.6	236–242	242–251	2.63
		255–259			266–271	266–271	2.97
		286–305	288–295	3.51		277–287	3.90
						296–302	1.55
Subtilisin BPN′	275	6–9	7–17	2.28	2–5	1–6	1.58
		15–19			26–30	26–32	2.42
			41–45	1.74		68–76	2.97
		65–72				81–87	2.02
		104–116	110–116	8.31	88–91	88–97	3.80
		133–144	131–144	51.1		121–124[e]	2.04
			193–200	1.61	147–152	145–154	4.03
		224–237	225–235	8.24	174–181	171–180	4.52
		243–251			190–193		
		270–274	269–275	3.88	205–210	205–209	1.93
					213–219		
Insulin							
A chain	21	3–5	1–5	3.69	None	6–9	1.05
		14–18	14–17	3.96		10–13	1.41
B chain	30	8–18	9–18	6.15	2–7		
					24–29	23–27	1.09
Elastase	240		4–8	3.79	14–22	15–21	1.36
			24–28	1.60	25–35	33–36	1.85
			38–45	6.31	38–44		
			46–50	2.12	53–58	51–59	3.31
			73–76	2.55	69–80	69–72	1.29
			85–90	3.12		77–80	2.27
			92–99	22.3	93–101	100–106	1.12
			118–122	2.87		107–114	2.48
			139–146	10.6	124–135	126–130	1.82
		156–159	155–157[e]	2.33	139–142		
			200–203	2.80	145–153	147–153	3.25
		234–239	235–238	2.31	171–180	171–175	1.44
					185–196	191–197	1.29
					199–209	206–212	1.47
					215–226	219–230	2.85
Staphylococcal nuclease	149		(4–8)	(1.25)	7–14	10–16	1.69
		55–66	55–61	1.04	22–27	22–27	2.25
			62–67	1.82	30–37	32–41	4.70
			71–75	2.51	39–43		
			96–99	1.41	71–77		
		100–105	100–104	3.58	88–94	87–94	2.15
		123–133	126–137	5.78	109–112	108–115	6.10
						121–125	1.31
Papain	212		(4–8)	(1.63)		27–35	4.91
		25–40	24–26	1.64		91–95	2.07
			36–40	1.77		110–114	2.66
		51–56	48–52	4.31		128–137	8.07
			(52–55)	(2.07)	162–167	155–166	2.20
		68–77	67–77	10.4	169–175		
			102–107	1.29	185–192	186–189	1.21
		117–125	118–122	1.29		199–203	
			123–126	1.37			
		139–142	140–143	1.30			
			(155–164)	(13.2)			
Ferricytochrome *c*	104		1–4	1.49	None	8–22	8.30
		10–12				44–49	1.50
		15–17				80–85	2.26
		50–53				93–96	1.27
		63–69	56–66	7.62			
			67–69[e]	2.46			
		72–74					
			77–79	1.05			
		92–100	(88–93)	(1.62)			
			97–101	3.81			
Cytochrome b_5	93	9–14	7–14	4.19	4–8	1–6	1.09
		34–37	31–38	4.50	21–24	20–30	5.36
		43–48	41–50	6.21	28–33		
		56–61	(52–57)	(2.26)	75–80	72–77	1.87
			(57–60)	(1.16)			
		65–73	65–70	1.95			

Table III (Continued)

Protein[a]	No. of amino acid residues	α-Helical region			Extended-conformation region		
		Obsd[b]	Predicted[c]	$P^*(i\|n\|\{\alpha\})$[d]	Obsd[b]	Predicted[c]	$P^*(i\|n\|\{\epsilon\})$[d]
		81–85	78–80	1.03			
Myogen	108	8–14	7–21	13.9		1–6	3.49
		27–32				28–36	4.18
		41–50	43–45	1.39		46–50	4.14
			56–58[h]	1.73	74–77	74–78	2.27
		68–70	69–73	1.13		83–87	1.45
		79–88				101–108	4.39
		103–106					
Sea lamprey hemoglobin	148	13–28	10–16	3.62	None	35–39	2.03
			21–24	2.77		125–128	1.51
		31–43				139–147	2.63
		(45–52)[i]	45–51	1.53			
		(62–66)[i]	60–65	1.74			
		68–87	71–76	1.82			
			82–86	1.84			
		(92–98)[i]					
		99–105	(102–106)	(1.34)			
		(111–114)[i]					
		115–126	114–124	9.93			
		133–144	132–137	3.58			
		(145–148)[i]					

[a] The original papers on the x-ray crystallographic studies of the proteins in this table, which were used to evaluate the statistical weights given in Table I, are cited in the footnotes of Table XII of our previous paper I.[3] [b] The definitions of α-helical and extended conformation regions were given in sections IA and IB of paper I.[3] [c] The regions in the parentheses are not predicted only because of the effect of electrostatic interaction described in stage II-3 in the text. [d] The relative probability of finding a sequence of n residues in a conformational sequence $\{\rho\}$ starting at the ith residue of the chain defined by eq 9, where $\{\rho\}$ is α and ε for the α-helical and extended conformation sequence, respectively, i is given by the number of the left-hand residues of the α-helical and extended conformation sequence (i_L) shown in columns 4 and 7, and the number of residues, n, for a sequence is given by $n = i_R - i_L + 1$ where i_R and i_L are the numbers of the residues given on the right- and left-hand sides, respectively, of columns 4 and 7. The values of $P^*(i|n|\{\rho\})$ are calculated from eq 15. [e] An isolated α-helical sequence of three residues, or an isolated extended conformation sequence of four residues, but having strong tendencies to form the α-helical or extended conformation sequences, respectively, with $P^*(i|3|\alpha) > 2.0$ and $P^*(i|4|\{\epsilon\}) > 2.0$ (see stage I-2a of the text). [f] A 3_{10} helical sequence. [g] A triad of α-helical states, or a tetrad of extended conformation states, remaining after the rest of the (originally duplicately assigned) sequence has been assigned to another region (e.g., by stage II). [h] Residues 56–65 were predicted to be α helical. However, since residues 59, 60, and 62 are positively charged, the portion from residues 59 to 65 could not be helical (see stage II-3 criterion). Thus, the helix is assigned only to residues 56–58. [i] α_{II} helical sequences (not counted as α helices, but as c).

3 and 6 of Table V (see also the last column of Table V). According to stage II-2, the remaining region 29–36 is assigned to the extended conformation, as seen in the final column of Table V. Thus, all of the α-helical and extended regions have been determined, as given in the last column of Table V.

Finally, the relative probabilities $P^*(i|n|\{\rho\})$ for these regions have been computed (by using eq 15) to satisfy stage III, the results of which are given in parentheses in the last column of Table V. It can be seen that all of the assignments satisfy stage III. The results of the prediction for bovine pancreatic trypsin inhibitor determined in this manner are summarized in Table IV, together with the values of $P^*(i|n|\{\rho\})$ where $\{\rho\} = \{\alpha\}$ or $\{\epsilon\}$.

In a similar manner, the predictions for clostridial flavodoxin and adenylate kinase, which were not involved in the original data set, have been carried out. The conformational probability profiles are shown in Figures 2 and 3, and the final results, determined in a manner similar to that described above for bovine pancreatic trypsin inhibitor, are summarized in Table IV.

In Table VI, we have summarized the number of regions of α-helical and extended conformations predicted theoretically and observed experimentally (see Tables III and IV). As another measure of the correctness of the prediction, one may employ the similar criterion used by Kotelchuck and Scheraga.[24] In the first place, in order to compare the results of the present prediction method with those carried out by other authors in terms of two-state models (all of them being for the α-helical and coil states), we will introduce the quantity $P_{(\eta)}^{(2)}$, for the two-state model, to measure the correctness of the prediction of state η, where

$$P_{(\eta)}^{(2)} = 100[N - n_{\mathrm{incor}(\eta)}]/N \tag{16}$$

where N is the number of residues in the protein, $n_{\mathrm{incor}(\eta)}$ is the number of residues in state η (where η is α or ε on the one hand, and c on the other), which are predicted incorrectly. The number of residues in state η that are predicted incorrectly is given by

$$n_{\mathrm{incor}(\eta)} = n_{\mathrm{o}(\eta)} + n_{\mathrm{m}(\eta)} \tag{17}$$

where $n_{\mathrm{o}(\eta)}$ is the number of residues whose conformations are over predicted (i.e., the number of residues predicted to be in the state η, which are not observed experimentally to be in state η), and $n_{\mathrm{m}(\eta)}$ is the number of residues predicted erroneously (i.e., the number of residues predicted to be in states other than the state η, which are observed experimentally to be in the state η). It should be noted that the predictions in Tables III and IV are *not* based on a two-state model (i.e., α and c, or ε and c), but on the three-state model. Hence, in order to compute $P_{(\eta)}^{(2)}$, we compute $n_{\mathrm{o}(\eta)}$ and $n_{\mathrm{m}(\eta)}$ by ignoring the results of predictions of states other than η in Tables III and IV.

The percentage of the total number of residues in the proteins predicted correctly (in the three-state scheme of α, ε, and c states) is measured by

Table IV
Predicted and Experimentally Observed α-Helical and Extended Conformation Regions for Proteins That Were *Not* Involved in the Original Data Set

Protein	No. of amino acid residues	α-Helical region			Extended-conformation region		
		Obsd[a]	Predicted[b]	$P^*(i\|n\|\{\alpha\})$[c]	Obsd[a]	Predicted[b]	$P^*(i\|n\|\{\epsilon\})$[c]
Bovine pancreatic trypsin inhibitor[d]	58	3–6	4–7	1.34	16–24	17–21	1.38
			13–16	1.86	27–36	29–36	4.02
			23–28	2.54		51–54	1.20
			(38–41)	(1.93)			
		46–56	44–50	4.20			
Clostridial flavodoxin[e]	138	11–26	11–21	8.50	1–6	1–7	1.52
			22–25	1.30	30–35		
			39–44	1.80	48–55	46–53	2.02
		67–73	(57–67)	(9.38)	80–89	80–87	1.31
			68–73	2.21	(109–111)[g]	106–119	10.4
		94–106	93–96	1.95	115–119		
			98–101	1.75			
		125–137	(121–126)	(3.54)			
			132–136	1.55			
Adenylate kinase[f]	194	2–7	1–6	2.32	10–14	10–15	4.46
		24–29	24–29	1.26	35–38		
		42–47	40–47	5.02		64–72	5.10
		55–61	54–63	5.25	90–94	89–92	1.87
		70–83	73–76	1.32	114–118	111–119	3.54
			77–84	4.09		126–130	1.71
		101–106	96–101	1.80		150–153	1.43
			102–107	5.96		157–165	
		124–132	122–125	1.68	169–173	167–173	1.22
			(127–131)	(1.20)		182–190	5.66
			(142–144)	(2.35)			
		145–157	145–149	1.39			
			154–156	2.53			
		161–163					
			174–176	2.20			
		180–193					

[a] See footnote *b* of Table III. [b] See footnote *c* of Table III. [c] See footnote *d* of Table III. [d] From ref 21. [e] From ref 22. [f] From ref 23. [g] This extended conformation region has only three residues according to the original authors, whereas our definition of an extended sequence is, at least, four residues. However, since the x-ray coordinates have not yet been made available to us, we tentatively assign this region to the extended conformation region.

$$P = 100[N - n_{\text{incor(t)}}]/N \quad (18)$$

where $n_{\text{incor}(t)}$ is the total number of residues predicted incorrectly in the protein and is given by

$$n_{\text{incor(t)}} = n_{o(\alpha)} + n_{m(\alpha)} + n_{o(\epsilon)} + n_{m(\epsilon)} - n_{\text{twice}} \quad (19)$$

where n_{twice} is the number of incorrectly predicted residues counted twice, in $n_{o(\alpha)}$ and $n_{m(\epsilon)}$ or in $n_{o(\epsilon)}$ and $n_{m(\alpha)}$. In Table VII, the results of $P_{(\alpha)}^{(2)}$, $P_{(\epsilon)}^{(2)}$, and P are summarized, together with N, $n_{(\alpha)}$, $n_{o(\alpha)}$, $n_{m(\alpha)}$, $n_{(\epsilon)}$, $n_{o(\epsilon)}$, $n_{m(\epsilon)}$, and $n_{\text{incor(t)}}$, where $n_{(\alpha)}$ and $n_{(\epsilon)}$ are the numbers of residues in α-helical and extended conformational states of the protein observed in x-ray experiments.

V. Discussion

A. Evaluation of Our Prediction Scheme. As can be seen in Figures 1 to 3, the peaks of the probability profiles are related closely to the α-helical and extended regions of proteins observed experimentally in the native state. Indeed, as can be seen for other proteins in Table VI, 80% of the α-helical regions (i.e., 84 regions out of 105 observed experimentally; Table VI) have been predicted correctly, with 33 regions being over-predicted as α helical (i.e., not observed experimentally), and 72% of the extended conformation regions (i.e., 58 regions out of 81), with 55 over-predicted regions. The results for the number of residues predicted correctly, converted into the two-state model, i.e., α and c states, and ε and c states, are in the range of 53 to 90% for the α-helical conformations of 20 proteins as seen in $P_{(\alpha)}^{(2)}$ in column 10 of Table VII, and in the range of 63 to 88% for the extended conformations of 19 proteins as seen in $P_{(\epsilon)}^{(2)}$ in column 11 of Table VII. The α-helical residues have been predicted slightly better than in the previous work by Lewis et al.,[15] in which the number of α-helical residues predicted correctly was in the range of 47 to 80% in the two-state model of α-helical and coil states. The total number of residues predicted correctly to be α-helical and extended, i.e., in the three-state model, is in the range of 47 to 77% for 19 proteins, as seen in the last column of Table VII.

B. Comparison with Other Prediction Schemes. In comparing our prediction scheme with those of others in this subsection, we limit our discussion to recent papers (those of Ptitsyn and Finkelshtein,[25] Robson and Pain,[26] Nagano,[27] Burgess et al.,[28] Chou and Fasman,[29] and Wu and Kabat[30]) and only to methodological aspects of these schemes. All of these schemes, as well as the one developed here, are based on the concept of the predominant role of short-range interactions in determining protein conformation.[3,6] In order to determine the probable conformations of proteins, all of these authors,[25,27–30] except Robson and Pain,[26] used, as a quantitative measure of the probability of occurrence of any conformation, the frequencies of occurrence of single amino acid residues (or pairs of amino acid residues) in the given conformation, as observed in x-ray crystal structures of proteins. Robson and Pain also used these frequencies, but converted them to a "measure of information translation" (defined in information theory).[26]

On the other hand, we have developed a statistical me-

Table V
Example of Assignments of α-Helical and Extended Conformation Regions for Pancreatic Trypsin Inhibitor

Amino acid *i*		Distribution of charged residues	α-Helical conformation			Extended conformation			*c*
			$P^*(i\|3\|\{\alpha\})$	First[a] possibility	$P^*(i\|3\|\{c\})$	$P^*(i\|4\|\{\epsilon\})$	First[a] possibility	$P^*(i\|4\|\{c\})$	
1	Arg	+	0.289		1.260	0.241		1.385	
2	Pro		0.809		1.324	0.783		1.320	
3	Asp	−	0.976		1.297	1.294		1.035	
4	Phe		1.087		0.862	2.252		0.937	
5	Cys		2.315		0.744	1.406		0.757	
6	Leu		0.727		0.868	0.982		0.884	α (1.34)
7	Glu	−	0.204		1.109	0.692		1.066	
8	Pro		0.174		0.982	1.097		0.911	
9	Pro		0.218		0.895	0.716		1.015	
10	Tyr		0.283		0.998	0.716		1.014	
11	Thr		0.199		1.055	0.879		0.918	
12	Gly		0.548		0.991	0.560		1.073	
13	Pro		1.308		0.945	1.146		0.856	
14	Cys		2.646		0.841	1.053		0.879	
15	Lys	+	1.766		1.008	1.091		0.863	α (1.86)
16	Ala		1.074		0.797	2.000		0.682	
17	Arg	+ [b]	1.134		0.754	1.378		0.789	
18	Ile		0.608		0.754	1.742		0.724	
19	Ile		0.261		0.847	0.899		0.931	ε (1.38)
20	Arg	+	0.238		1.089	1.480		1.159	
21	Tyr		0.623		1.108	0.847		1.301	
22	Phe		0.459		1.353	0.972		1.225	
23	Tyr		2.089		1.007	0.491		1.090	
24	Asn		4.004		1.134	0.564		1.027	
25	Ala		2.784		0.872	0.696		0.988	α (2.54)
26	Lys	+	1.195		1.093	0.863		0.872	
27	Ala		0.768		0.806	1.526		0.702	
28	Gly		1.415		0.777	1.383		0.736	
29	Leu		0.848		0.649	2.687		0.601	
30	Cys		0.996		0.755	1.244		0.829	
31	Gln		0.589		0.951	3.226		0.870	
32	Thr		0.908		0.918	3.107		0.883	ε (4.02)
33	Phe		0.294		0.953	1.598		1.082	
34	Val		0.098		0.817	0.651		0.926	
35	Tyr		0.174		1.220	0.527		1.063	
36	Gly		0.356		1.106	0.415		1.154	
37	Gly		0.978		1.017	0.850		0.921	
38	Cys		1.864		0.812	1.054		0.880	
39	Arg	+	2.122		1.009	0.677		1.055	
40	Ala		0.861		1.010	0.389		1.189	
41	Lys	+	0.387		1.315	0.154		1.546	
42	Arg	+	0.366		1.428	0.145		1.569	
43	Asn		0.617		1.501	0.268		1.733	
44	Asn		1.652		1.472	0.384		1.612	
45	Phe		1.353		1.371	0.972		1.241	
46	Lys	+	1.601		1.059	0.487		1.151	
47	Ser		3.937		1.063	0.277		1.277	
48	Ala		1.844		1.151	0.522		1.012	α (4.20)
49	Glu	−	0.751		1.119	0.794		0.750	
50	Asp	−	0.858		0.692	0.435		0.606	
51	Cys		0.767		0.505	1.199		0.468	
52	Met		0.372		0.536	1.199		0.468	
53	Arg	+	0.264		0.834	0.810		0.946	ε (1.20)
54	Thr		0.166		0.905	0.574		1.025	
55	Cys		0.173		1.106	0.605		1.002	
56	Gly		0.604		1.149				
57	Gly								
58	Ala								

[a] A region denoted by a broken line is an isolated α-helical sequence of three residues or an isolated extended sequence of four residues (see stage I-2a of section III in the text). [b] The vertical line shows the repulsive electrostatic interactions (*i* to *i* + 3 or *i* to *i* + 4), when the criterion of stage II-3 is applied. [c] Final assignment of α-helical and extended conformation regions. The numerals in parentheses are the conformational-sequence probabilities for the sequence of *n* residues calculated from eq 15. Broken lines in this column simply distinguish ε from α conformations.

chanical theory of polypeptide chains to treat the conformations of protein molecules.[4] To use this model, it is necessary to have a set of statistical weights for each of the 20 naturally occurring amino acids in the respective conformations. In principle, these statistical weights can be obtained by experiment (e.g., by computing the parameters *s* and *σ* of the Zimm–Bragg theory from experiments on the helix–coil transition in solution), or by conformational energy calculations.[5,6,12] At present, neither of these methods has been carried far enough to provide as complete a set of statistical weights as is required (e.g., see Table I). Thus, we have resorted to an alternative method to obtain the data of Table I, viz., the use of x-ray data on proteins, as described in paper I.[3] The assumption used in evaluating

Table VI
Predicted and Experimentally Observed α-Helical and Extended Conformation Regions

Protein[a]	α-Helical, No. of regions			Extended conformation, No. of regions		
	Obsd	Predicted correctly	Over-predicted[b]	Obsd	Predicted correctly	Over-predicted[b]
Myoglobin	8	7	0	0		4
Lysozyme	6	5	0	1	1	3
Ribonuclease S	3	3	2	4	4	1
Oxyhemoglobin						
α chain	7	6	1	0		4
β chain	8	6	0	0		4
α-Chymotrypsin						
B chain	0		4	10	8	1
C chain	2	2	1	6	5	0
Carboxypeptidase	8	6	3	9	9	7
Subtilisin BPN'	8	6	2	8	6	3
Insulin						
A chain	2	2	0	0		2
B chain	1	1	0	2	1	1
Elastase	2	2	10	13	11	1
Staphylococcal nuclease	3	3	2	7	6	1
Papain	5	5	1	3	2	5
Ferricytochrome *c*	6	2	2	0		4
Cytochrome b_5	6	5	0	4	4	0
Myogen	6	3	1	1	1	5
Sea lamprey hemoglobin	8	6	0	0		3
Bovine pancreatic trypsin inhibitor	2	2	2	2	2	1
Clostridial flavodoxin	4	4	1	6	5	0
Porcine muscle adenylate kinase	10	8	1	5	4	5
Total	105	84	33	81	58	55

[a] See footnotes in Table XII of paper I[3] for the original papers on the x-ray crystallographic studies of the proteins from myoglobin up through sea lamprey hemoglobin, which were part of the original data set; see footnotes *d–f* of Table IV for those from bovine pancreatic trypsin inhibitor to adenylate kinase, which had not been involved in the original data set. [b] Predicted to be α-helical or extended conformation, but not observed experimentally. "Region" refers to an α-helical or ε sequence.

Table VII
Correctness of Predictions

Protein[a]	N	$n_{(\alpha)}$	$n_{o(\alpha)}$	$n_{m(\alpha)}$	$n_{(\epsilon)}$	$n_{o(\epsilon)}$	$n_{m(\epsilon)}$	$n_{incor(t)}$[b]	$P_{(\alpha)}^{(2)c}$	$P_{(\epsilon)}^{(2)c}$	Pd
Myoglobin	153	107	15	57	0	30	0	78	52.9	80.4	49.0
Lysozyme	129	41	9	14	4	23	3	42	82.2	79.8	67.4
Ribonuclease S	124	28	21	4	39	8	15	43	79.8	81.5	65.3
Oxyhemoglobin α chain	141	96	6	56	0	27	0	68	56.0	80.9	51.8
Oxyhemoglobin β chain	146	99	3	53	0	33	0	59	61.6	77.4	59.6
Tosyl-α-chymotrypsin B chain	131	0	22	0	58	18	16	44	83.2	74.0	66.4
Tosyl-α-chymotrypsin C chain	97	17	19	8	37	4	13	43	72.2	82.5	55.7
Carboxypeptidase A	307	93	18	48	55	66	19	134	78.5	72.3	56.4
Subtilisin BPN'	275	70	21	29	44	41	13	97	81.8	80.4	64.7
Insulin	51	19	2	3	12	9	8	22	90.2	66.7	56.9
Tosyl elastase	240	14	59	4	124	25	63	123	73.7	63.3	48.7
Staphyloccal nuclease	149	28	15	4	43	13	11	35	87.2	83.9	76.5
Papain	212	45	16	15	21	37	12	71	85.4	76.9	66.5
Ferricytochrome *c*	104	29	14	18	0	31	0	56	69.2	70.2	46.2
Cytochrome b_5	93	37	13	14	21	10	8	36	71.0	80.6	61.3
Carp myogen	108	40	14	28	4	34	0	57	61.1	68.5	47.2
Sea lamprey hemoglobin	148	80	18	46	0	18	0	69	56.8	87.8	53.4
Bovine pancreatic trypsin inhibitor	58	15	13	7	19	4	6	23	65.5	82.8	60.3
Clostridial flavodoxin	138	49	17	15	38	9	10	50	76.8	86.2	63.8
Adenylate kinase	194	85	18	34	24	43	6	76	73.2	74.7	60.8

[a] See footnote *a* of Table VI. [b] From eq 19. [c] From eq 16. [d] From eq 18.

these statistical weights was that the conformations of each amino acid residue are distributed according to the Boltzmann factor of the free energy of the amino acid unit [including the free energy contribution from hydrogen bond formation, in the case of three α-helical states, i.e., in evaluating $w_{h,j}$ (see paper I[3])]. The (ϕ,ψ) map of the frequencies of occurrence of the conformational states of the amino acids obtained from x-ray data on proteins can be seen, for example, for lysozyme in Figure 3 of ref 31 and for each amino acid in Figures 5–8 ref 28. In other words, our assumption is that those frequency maps correspond to ones that can be calcualted from the free energy of a peptide unit or from the statistical weights (according to the method described in section IIA of paper I). Thus, it should be noted that our parameters calculated in paper I,[3] and given in Table I of this paper, should not be regarded as probabilities, but as statistical weights (see ref 60 of paper I for a discussion of the difference between a statistical

Table VIII
Illustrative Predictions of Insulin A Chain Based on the Chou and Fasman Scheme

	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21
Amino acid sequence	Gly Ile Val Glu Gln Cys Cys Thr Ser Ile Cys Ser Leu Tyr Gln Leu Glu Asn Tyr Cys Asn
$\langle P_\alpha \rangle$[a]	⊢—This region cannot nucleate the α helix—⊣ ⊢—$\langle P_\alpha \rangle$ = 1.12—⊣
Helical assignment[b]	B I h H h i i i i I i i H b h H H b b i b
β-sheet assignment[b]	i H H B h h h h b H h b h h h h B b h h b
$\langle P_\beta \rangle$[a]	⊢—$\langle P_\beta \rangle$ = 1.17—⊣⊢—$\langle P_\beta \rangle$ = 1.24—⊣
	⊢—$\langle P_\beta \rangle$ = 1.19—⊣
Predictive results (i)[c] by the present authors[d]	⊢—β-sheet—⊣
(ii)[c]	⊢—β-sheet—⊣⊢—α-helix—⊣
Chou–Fasman[e]	⊢—α-helix—⊣ ⊢—α-helix—⊣
Observed by x-ray[f]	⊢—α-helix—⊣ ⊢—α-helix—⊣

[a] These values are computed using the values of P_α and P_β of Table I of ref 29b, which involve some error.[35] [b] From Table I of ref 29b. [c] From the two possible assignments, (i) and (ii), described in the text. [d] Based on Chou and Fasman's method, and not on ours. [e] From ref 29b. [f] From ref 33 (see also ref 34).

weight and the probability of occurrence of a certain conformational state); we have also calculated the probability of finding a certain conformation by incorporating the statistical weights into a statistical mechanical method of averaging over a *whole molecule,* as has been discussed already in section VI of paper II[4] and briefly in section II of this paper.

Another difference between the present prediction scheme and those of previous authors[25-30] is that, in the latter studies, the most probable conformation of a protein is determined residue by residue, or for a segmental sequence of amino acids out of a whole protein molecule; thus, the short-range and long-range cooperativities of the chain molecule (within a one-dimensional Ising model) are neglected. Furthermore, in order to determine the conformation with the highest probability for a residue or for a partial sequence of the protein molecule, one has to take into account all possible conformational states for parts of the molecule other than the residue or the sequence under consideration. For example, if one excises a short sequence of amino acids out of a whole protein molecule, the short sequence cannot form a stable α-helical conformation (even though the sequence contains a number of residues sufficient to form the α-helical conformation). Thus, in the predictive methods proposed by those authors,[25-30] the connectivity of the chain molecule, hence the cooperativity, is ignored in determining the conformations of the chain molecule by assuming the independent occurrence of conformational states in individual amino acid residues of the proteins. Because of this neglect of chain connectivity, these other methods usually cannot discriminate between the conformational preference of ABCDE . . . from that of, say CDBAE . . . , where ABCDE . . is a given amino acid sequence.

C. Evaluation of Chou–Fasman Prediction Scheme. Chou and Fasman[29] (referred to here as CF) proposed a scheme to predict α-helical and ϵ conformations in proteins. However, as pointed out by Burgess et al.,[28] there are ambiguities in some of their prediction rules; also, their scheme seems to be incomplete in the sense that the ambiguities from duplicately assigned regions do not seem to be resolved adequately. Further, in ref 60 of paper I,[3] we pointed out some misconceptions in their mathematical treatment and in the physical meaning of their parameter P_α. Nevertheless, we will adopt their parameters[29a] P_α and P_β (despite the numerical errors in them, as pointed out in ref 60 of paper I[3]), and follow their rules,[29b] and will show two examples in which it is not possible to reproduce their claimed results.

As a first example, we consider residues 45 to 54 of bovine pancreatic trypsin inhibitor which they assigned as α helical (with $\langle P_\alpha \rangle$ = 1.05), and claimed agreement with the observed[21] helical region in residues 45–56 (see Figure 1[32] and Table IV of ref 29b). However, this assignment is impossible, if one follows their rules. Their rule 1 demands satisfaction of their condition A-1 which requires four of six residues to be helical (h_α or H_α, with I_α counting as 0.5 h_α). However, no four out of six residues[32] in region 45–54 satisfy their condition A-1, even though their *partial* criterion that $\langle P_\alpha \rangle \geq 1.03$ is satisfied in this sequence. In fact, if we follow their rule 2, we find that this sequence must be assigned as a β-sheet region, according to the following argument. (a) Their condition B-1 (three of five residues must be h_β or H_β) is satisfied in regions 50–54 (h_β, H_β, h_β), 51–55 (h_β, H_β, h_β, h_β), and 52–56 (H_β, h_β, h_β). (b) Their condition B-2 (extend β region in *both* directions until terminated by tetrapeptides with $\langle P_\beta \rangle < 1.00$) is satisfied for 54–57 (for which $\langle P_\beta \rangle$ = 1.03) and also for 50–55 ($\langle P_\beta \rangle$ = 1.20), 50–56 ($\langle P_\beta \rangle$ = 1.14), and 50–57 ($\langle P_\beta \rangle$ = 1.10). (Parenthetically, it should be noted that all the values of $\langle P_\beta \rangle$ = 1.20, 1.14, and 1.10 are greater than $\langle P_\alpha \rangle$ = 1.05 for the region 45–54.) Since $\langle P_\beta \rangle > 1.05$ (part of their rule 2), and $\langle P_\alpha \rangle$ = 0.80, for 50–57, it is clear that they should have assigned residues 50–57 as a β-sheet region (the sequence 46–49 can be assiged neither as an α-helical or β-sheet region). Thus, their scheme does *not* predict the existence of the α helix observed[21] experimentally.

As a second example (illustrated in Table VIII), consider the A chain of insulin to which CF assigned helices to residues 2–7 and 13–18, and claimed agreement with the observed[33] helical region in residues 2–8 and 13–19[34] (see Table II of ref 29b). However, according to their condition A-1, there are only 3.5, and not 4, helical residues among the six in the sequence 2 to 7. Thus, residues 2–7 *cannot* be helical, according to *their* rules. Further, the sequence 5 to 16 has many β-sheet residues, and many consecutive sets of 5 residues in this region satisfy condition B-1 of CF. However, there is an unresolved ambiguity in the assignment if we follow the CF rules. The following *two* possible assignments exist. (i) A β sheet can be assigned to residues 5 to 16 because $\langle P_\beta \rangle$ = 1.19 and $\langle P_\alpha \rangle$ = 0.95 (and the value of $\langle P_\beta \rangle$ = 1.19 for residues 5 to 16 is larger than $\langle P_\alpha \rangle$ = 1.12 for residues 13 to 18). As seen in Table VIII, this assignment cannot be altered even if region 5–16 is divided into two regions, i.e., one from 5–12 ($\langle P_\beta \rangle$ = 1.17) and the other from 13–16 ($\langle P_\beta \rangle$ = 1.24), because the latter value is greater than the value of $\langle P_\alpha \rangle$ = 1.12 for region 13–18. (ii) If one wanted to argue that the strong β-sheet region 5–16 is ter-

minated by the α-helical region 13–18, then region 5–12 would be assigned as β sheet and region 13–18 as α helix. The CF rules do not resolve the ambiguity between choices (i) and (ii), both of which differ from the original CF assignment.

These two examples raise doubts about the high percentage of correct predictions claimed by Chou and Fasman[29] (see also ref 60 of paper I for further criticism of the CF procedure[35]).

D. Evaluation of Froimowitz–Fasman Method. Froimowitz and Fasman (FF) recently proposed a two-state model for predicting α-helical regions in proteins.[36] While similar to the model of Lewis et al.,[15] that of FF involves some misconceptions which we point out here. First of all, as pointed out in ref 60 of paper I,[3] the parameters P_α and P_β of Chou and Fasman should not be identified with the statistical weights σ and s of the Zimm–Bragg theory of the helix–coil transition. Nevertheless, FF[36] used the probabilities of Chou and Fasman[29a] as statistical weights. Second, although they used a two-state model, they omitted the ϵ state from the c state. Thus, they underweighted the c state. Third, there is an error in the treatment in the matrix operator in their eq 2, which can be written in our notation as

$$W_i = \begin{bmatrix} w_h^* & v_C^* & 0 \\ 0 & 0 & 1 \\ v_N^* & v^* & 1 \end{bmatrix} \tag{20}$$

The element v^*, which is assigned to the middle residue of the triad in the conformational state chc, should have been written as

$$v^* = (v_N^* v_C^*)^{1/2} \tag{21}$$

(see eq 4) instead of

$$v^* = v_N^* v_C^* \tag{22}$$

as used by FF, because the statistical weight of an isolated helical state has to have contributions, one-half from the ch pair from residues $i - 1$ and i, and one-half from the hc pair from residues i and $i + 1$, when using a matrix that correlates three residues (see the discussion in section IB). They tried to rationalize this error in their ref 14 by stating that "this would not affect the computational results, since the contribution of isolated helical residues, with their relatively low statistical weights, would not be significant to the overall helical probabilities". It is true that v^* is small, if defined erroneously as in eq 22, instead of as in eq 21, since $v^* < 1$. However, if one uses the statistical weight of eq 21, the magnitude of v^* is by no means small in comparison with those of other statistical weights (the elements v_N^* and v_C^* given in eq 20). It turns out that the contribution of v^* to the final prediction is as large as those of v_N^* and v_C^*, which were introduced as their main improvement of the previous study.[15]

Addendum

The model developed in this series of papers, and applied to proteins, is based on short-range interactions (the one-dimensional Ising model). This model can be augmented by the introduction of medium- and long-range interactions, as shown elsewhere.[37]

One of the main points of this series is to test the adequacy of a short-range interaction model, and not to advocate its use for the prediction of protein conformation. While short-range interactions are the predominant ones, they are not the only ones. Thus, as shown here, a model based solely on short-range interactions cannot provide 100% accuracy in the prediction of protein conformation. Hopefully, the inclusion of medium- and long-range interactions[37] will improve the predictability.

The empirical rules based on P^* were introduced here to make the computed probabilities deterministic, i.e., to enable the backbone conformations to be determined from the probability P. For this purpose, a standard value (viz., θ_η) was introduced to convert P to P^*, and thereby incorporate the deterministic feature. As stated in ref 45 of paper I, the cooperative feature of the one-dimensional Ising model already appears in P. Our further efforts in this direction will not be to improve the empirical rules, but to eliminate the need for them.

References and Notes

(1) This work was supported by research grants from the National Institute of General Medical Sciences of the National Institutes of Health, U.S. Public Health Service (GM-14312), and from the National Science Foundation (BMS71-00872 A04).
(2) From Kyoto University, 1972–1975.
(3) S. Tanaka and H. A Scheraga, *Macromolecules*, paper I in this issue.
(4) S. Tanaka and H. A. Scheraga, *Macromolecules*, paper II in this issue.
(5) For recent reviews, see (a) H. A. Scheraga, *Chem. Rev.*, **71**, 195 (1971); (b) H. A. Scheraga, "Peptides, Polypeptides and Proteins", E. R. Blout, F. A. Bovey, M. Goodman, and N. Lotan, Ed., Wiley, New York, N.Y., 1974, p 49.
(6) H. A. Scheraga, *Pure Appl. Chem.*, **36**, 1 (1973).
(7) See the last paragraph of the introductory section of paper II[4] for the reasons why the eight-state and six-state models for α-helical and extended conformations, respectively, and the treatment of the asymmetric properties of helix nucleation, are not included in the theory used in this paper.
(8) For a specific-sequence copolymer, we use the subscript j to designate the species of amino acid, and the subscript i to designate the position of the amino acid in the polymer chain. However, the statistical weight of amino acid of type j in position i will be designated simply as w_{hj}^* (with i omitted), as in eq 1.
(9) In general, the execution time required for matrix multiplication on a computer is proportional to the cube of the order of a matrix.
(10) As mentioned in section IIB of paper I,[3] the statistical weight for the extended conformational state is not defined for the two-state model. Hence, the superscript (3), which designates the three-state model, is not needed in $v_{\epsilon j}^*$.
(11) S. Tanaka and H. A. Scheraga, manuscripts in preparation, multi-state models.
(12) S. Tanaka and H. A. Scheraga, manuscript in preparation, same as ref 40 of paper I.
(13) R. T. Ingwall, H. A. Scheraga, N. Lotan, A. Berger, and E. Katchalski, *Biopolymers*, **6**, 331 (1968).
(14) S. Lifson and A. Roig, *J. Chem. Phys.*, **34**, 1963 (1961).
(15) P. N. Lewis, N. Gō, M. Gō, D. Kotelchuck, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **65**, 810 (1970).
(16) B. H. Zimm and S. A. Rice, *Mol. Phys.*, **3**, 391 (1960).
(17) For a matrix treatment of such long-range interactions in α-helical structures, see ref 18.
(18) R. K. H. Liem, D. Poland, and H. A. Scheraga, *J. Am. Chem. Soc.*, **92**, 5717 (1970).
(19) F. R. Maxfield and H. A. Scheraga, *Macromolecules*, **8**, 491 (1975).
(20) P. N. Lewis and E. M. Bradbury, *Biochim. Biophys. Acta*, **336**, 153 (1974).
(21) R. Huber, D. Kukla, A. Ruehlmann, and W. Steigemann, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 141 (1972).
(22) (a) R. M. Burnett, G. D. Darling, D. S. Kendall, M. E. LeQuesne, S. G. Mayhew, W. W. Smith, and M. L. Ludwig, *J. Biol. Chem.*, **249**, 4383 (1974); (b) the amino acid sequence of clostridial flavodoxin was reported in M. Tanaka, M. Haniu, K. T. Yasunobu, and S. B. Mayhew, *ibid.*, **249**, 4393 (1974).
(23) G. E. Schulz, M. Elzinga, F. Marx, and R. S. Schirmer, *Nature (London)*, **250**, 120 (1974).
(24) D. Kotelchuck and H A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **62**, 14 (1969).
(25) O. B. Ptitsyn and A. V. Finkelshtein, *Biofizika*, **15** 757 (1970).
(26) (a) B. Robson and R. H. Pain, *J. Mol. Biol.*, **58**, 237 (1971); (b) B. Robson, *Biochem. J.*, **141**, 853 (1974); B. Robson and R. H. Pain, *ibid.*, **141**, 869, 883 (1974).
(27) K. Nagano, *J. Mol. Biol.*, **75**, 401 (1973); **84**, 337 (1974).
(28) A. W. Burgess, P. K. Ponnuswamy, and H. A. Scheraga, *Isr. J. Chem.*, **12**, 239 (1974).
(29) (a) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 211 (1974); (b) *ibid.*, **13**, 222 (1974).
(30) T. T. Wu and E. A. Kabat, *J. Mol. Biol.*, **75**, 13 (1973).
(31) C. M. Venkatachalam and G. N. Ramachandran, "Conformation of Biopolymers," No. 1, G. N. Ramachandran, Ed., Academic Press, New York, N.Y., 1967, p 83.
(32) In this figure, they assigned h_β to Glu-49. However, according to Table I of ref 29b, Glu should be B_β, and we retain the B_β assignment here. Since the values in Table IV[29b] of $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ for this sequence are

correct, the wrong assignment in Figure 1 of ref 29b is probably a typographical error.

(33) (a) M. J. Adams, T. L. Blundell, E. J. Dodson, G. G. Dodson, M. Vijayan, E. N. Baker, M. M. Harding, D. C. Hodgkin, B. Rimmer, and S. Sheat, *Nature (London)*, **224**, 491 (1969); (b) T. L. Blundell, J. F. Cutfield, S. M. Cutfield, E. J. Dodson, G. G. Dodson, D. C. Hodgkin, D. A. Mercola, and M. Vijayan, *ibid.*, **231**, 506 (1971).

(34) In ref 33a, it was reported that residues 2–6 and 13–19 of the A chain of insulin were α helical. This corresponds to 3–5 and 14–18 in our definition of an α-helical sequence (see section I of paper I[3]). In their earlier paper [*J. Mol. Biol.*, **74**, 263 (1973)], CF cited 2–6 and 13–19 as the helical residues. Thus, we do not know whether their report of the x-ray results as helical in residues 2–8 and 13–19 in Table II of ref 29b is simply a typographical error or not. However, in the last line of Table VIII of the present paper, we listed the helical region as that given by the x-ray crystallographers.[33a]

(35) As pointed out in ref 60 of paper I,[3] there are errors in the values of P_α and P_β in Table II of ref 29a, and hence in Table I of ref 29b, because of a mathematical error. However, in the illustrative examples in the present paper, we retain the numerical values of Chou and Fasman to compute $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$.

(36) M. Froimowitz and G. D. Fasman, *Macromolecules*, **7**, 583 (1974).

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

Comments on Exclusion of Polymer Chains from Small Pores and Its Relation to Gel Permeation Chromatography

Edward F. Casassa

Department of Chemistry, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.
Received August 15, 1975

ABSTRACT: Some criticisms of our theoretical treatment of the partial exclusion of flexible-chain polymers in solution from cavities of macromolecular size and its application to gel permeation chromatography are examined. In other discussion, it is confirmed by simple reasoning that the identification, explicit or implicit in various studies, of the mean projection of a polymer molecule onto a line as a characteristic dimension governing the extent of permeation of simple pores does not depend on specific molecular models. Our previous calculation of permeation by certain random-flight branched-chain species is shown to lead, incidentally, to the mean projection for these structures. From relations between the mean projection and the hydrodynamic volume of a molecule, it appears that the product of intrinsic viscosity and molecular weight is not a common calibration factor for elution of all molecular species from a gel chromatographic column, but theory and experience do support the validity of this correlation among solutes with similar molecular architecture.

Although it has been a subject of controversy over the last decade, it now seems generally agreed that peak separation in gel permeation chromatography (or exclusion chromatography, to use a better but less accepted term) is governed primarily by the equilibrium distribution of solute species between a macroscopic solution phase (the mobile phase in chromatography) and solution within pores of macromolecular size.[1-6] In advocating this point of view on chromatography of flexible-chain polymers, we calculated distribution coefficients for random flight chains in cavities of simple geometrical form with rigid impermeable, but otherwise noninteractive, walls.[1,3-5] For these idealized models, the problem is to obtain appropriate solutions of the equation

$$\partial P_n(\mathbf{r})/\partial n = (b^2/6)\nabla^2 P_n(\mathbf{r}) \quad (1)$$

where $P_n(\mathbf{r})$ is the probability density for finding the nth step of a random flight with root-mean-square length b at a point designated by a vector $\mathbf{r}$ drawn from the coordinate origin. In an infinite medium the familiar Gaussian form for $P_n(\mathbf{r})$ is obtained, but by imposing the boundary condition that all $P_n(\mathbf{r})$ vanish when $\mathbf{r}$ is on the surface S defining the walls of the cavity, the equation can also be solved for a random flight within a cavity of sufficiently simple geometry. The probability $P_n(\mathbf{r}|\mathbf{r}')d\mathbf{r}$, thus obtained, that a chain of n steps beginning at $\mathbf{r}'$ will terminate in the volume element $d\mathbf{r}$ without ever intersecting S is averaged over all $\mathbf{r}$ inside S to obtain the probability that a chain beginning at $\mathbf{r}'$ will not touch the boundary, and a final integration over $\mathbf{r}'$ gives the permeation (or distribution) coefficient K, the fraction of all unrestricted chain conformations beginning inside S that is still allowed in the presence of the impenetrable barrier or, equivalently, the ratio of polymer concentrations in solutions inside and outside cavities. Certain artifices may simplify calculation (e.g., imagining random flights to be generated simultaneously throughout the volume of the cavity), but results are equivalent to those obtained by enumerating conformations of a single chain.

Since our theoretical work was published, several authors have questioned some aspects of it or its application, and/or have suggested amplifications of it. Here we comment on three of these studies, and, at the end, discuss the relation between a characteristic molecular dimension governing the exclusion phenomenon and the question of a "universal" calibration in gel chromatography.

I. Confrontation of Theory and Experiment. Characterization of Pore Size and Shape

If K is identified with the effective fraction of pore volume available to solute in a chromatographic column, the theoretical calculations can be compared with elution data. Yau and Malone[6] have addressed themselves to the comparison of theory and experiment that we made in ref 1. Using the same elution and pore size data, together with newer results from their laboratory, they show how our analysis should be improved.

In ref 1, we plotted theoretical values of K against a dimensionless size parameter R/a, the ratio of the root-mean-square molecular radius of gyration, $R = (nb^2/6)^{1/2}$, of a linear random-flight chain to the pore size a (half the thickness of the slab-shaped cavity, or the radius of the cylinder or sphere), and compared the curves with the experimental dependence of K deduced from column elution