Statistical Mechanical Treatment of Protein Conformation. I. Conformational Properties of Amino Acids in Proteins[1]

Seiji Tanaka[2] and Harold A. Scheraga*

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

ABSTRACT: A statistical mechanical (one-dimensional Ising model) treatment, based on the dominance of short-range interactions, is developed in this series of papers; it is intended as an improvement over empirical prediction schemes for obtaining approximate initial conformations of proteins (to be used to try to deduce the native conformation by subsequent energy minimization). In the present paper, the statistical weights for a two-state model (α-helical and other conformations) and for a three-state model (α-helical, extended, and other conformations) are evaluated from x-ray data on 16 native proteins. The method for evaluating the statistical weights is presented. Asymmetric α-helical nucleation parameters are also evaluated for the 20 naturally occurring amino acids. On the basis of these statistical weights, the conformational properties of the twenty naturally occurring amino acids are discussed. The statistical weights evaluated from x-ray data are also discussed in comparison with experimental results on the helix–coil transition in polyamino acids in solution. The predominant role of short-range interactions, and some possible long-range effects in determining the statistical weights, are discussed in conjunction with the mechanism of protein folding.

The prediction of the three-dimensional structure of a protein from its amino acid sequence has not yet been achieved because of the existence of multiple minima in the many-dimensional conformational energy surface of a protein.[3] To circumvent this problem, empirical prediction schemes are being developed to obtain a starting conformation from which energy minimization could lead to the native structure.[4] In this connection, it is important to realize that energy minimization (which includes long-range interactions and solvent effects) is necessary in order to determine the precise dihedral angles of native proteins. The use of the *average* conformational state of each amino acid residue cannot lead to the three-dimensional structure of a native protein;[5] i.e., it was shown[5] that, in order to predict protein conformation, it was necessary but not sufficient[5] to determine the *average* conformational state of each residue, but that this had to be followed by a determination of the exact dihedral angles, rather than the average ones.[5]

The success of empirical prediction algorithms, to varying extents,[4] is based on the demonstrated[6] dominance of short-range interactions (those between the side chain of an amino acid and its own backbone) in determining the average conformational preferences (α-helical, extended, coil, etc.) of amino acid residues in polypeptides and proteins.[7] The dominance of short-range interactions also manifests itself in the applicability[8] of the one-dimensional Ising model (based on nearest-neighbor interactions) to the determination of the parameters σ and s of the Zimm–Bragg theory[9] of the helix–coil transition, and to the use of such parameters for the calculation of helix-probability profiles in proteins;[10] the correlation between regions of high helix probability in the denatured protein and those where α-helices are found in the native protein[10,11] attests to the approximate validity of the one-dimensional Ising model despite its omission of long-range interactions. Further justification of the use of the one-dimensional Ising model is provided in section IIIE.

Therefore, the purpose of this set of three papers is to formulate a statistical mechanical one-dimensional Ising model treatment of protein conformation, in which three states (α-helix, extended, and other) are allowed for each amino acid residue, in order to develop a basis for an empirical prediction scheme (to be used subsequently in an energy minimization algorithm[5]). The statistical mechanical treatment avoids the ambiguities arising from the use of arbitrary rules in current empirical prediction schemes. In the present paper, we deduce the statistical weights from x-ray data on protein structures. In paper II,[12] we formulate the model and show how to compute the probabilities of occurrence of helical and extended conformations, respectively. In paper III,[13] we compute these probabilities for specific proteins, using the theory of paper II and the statistical weights of paper I. An extension of this treatment to include more than three states (e.g., chain reversal, bridge-region, left-handed helical conformations, etc.) will be described elsewhere.[14]

In section I of this paper, we analyze the conformations of native proteins obtained from x-ray crystallographic studies. In section II, we describe the methods for computing the statistical weights for the α-helical and extended conformations, based on both a two-state and a three-state model. The numerical results are presented and discussed in section III, and the conclusions are summarized in section IV.

I. Statistical Analysis of Protein Data

Sixteen proteins, whose amino acid sequences and x-ray crystal structures have been reported,[15–36] are surveyed in the present analysis. The α-helical and extended conformational regions (designated by h and ϵ, respectively) found by x-ray observations are given in Table XII, in which the descriptions of the h and ϵ regions are different from those found in the original papers[15–36] because of the different definitions for these states, as described in sections IA and IB.

A. The Helical Conformation. In order to describe a helical sequence, eight states are assigned to the ith residue, depending on its conformation, i.e., its dihedral angles (ϕ_i,ψ_i),[37a] and on whether or not its constituent NH and CO groups are hydrogen bonded; this model is the one that Gō et al.[38] applied to homopolymers and specific-sequence copolymers. The eight states are specified by three independent factors:[39,40] (a) the range of the dihedral angles (ϕ_i,ψ_i), i.e., whether they lie in the helical region, h, or whether they lie outside of it (and hence designated as "c"); (b) the presence or absence of a hydrogen bond between the CO group of the ith residue in question and the NH group of the $(i + 4)$th residue, when the three intervening residues, $(i + 1)$, $(i + 2)$, and $(i + 3)$, are all in h states; (c) the presence or absence of a hydrogen bond be-

Table I
Definition of the Eight Helical States of the ith Residue

Conformational state (ϕ_i, ψ_i)[a]	Hydrogen bond[b] on C_iO_i	Hydrogen bond[b] on N_iH_i	State
c	No	No	1
c	No	Yes	2
c	Yes	No	3
c	Yes	Yes	4
h	No	No	5
h	No	Yes	6
h	Yes	No	7
h	Yes	Yes	8

[a] The symbols c and h denote the conformational state of residue i, i.e., whether the residue is allowed to occupy the whole (ϕ,ψ) space, and whether it is restricted to the small region of the (ϕ,ψ) space characteristic of the right-handed α-helix, respectively. [b] The presence or absence of a hydrogen bond on the CO and NH groups of the ith residue is designated in terms of Yes or No, respectively.

tween the NH group of the ith residue in question and the CO group of the $(i - 4)$th residue, when the three intervening residues, $(i - 3)$, $(i - 2)$, and $(i - 1)$, are all in h states. The possible combinations of these three factors and the corresponding states are listed in Table I.

If a hydrogen bond map is available for the α-helical regions of a native protein,[41] then the conformational states of the residues can be assigned easily, according to the states given in Table I and illustrated, for several examples, in Figure 1. Even without such maps, helical states can be assigned, according to Table I, by assuming that all possible hydrogen bonds are present in the α-helical sequences reported by the x-ray crystallographers. For example, a reported helical sequence of six amino acid residues would be assigned the states shown in Figure 1b. Unless they specify otherwise, the crystallographers usually designate a helical sequence as beginning with the residue whose CO group is hydrogen bonded and ending at the residue whose NH group is hydrogen bonded. Thus, one of eight states can be assigned to every residue of a protein, ignoring all hydrogen bonds[42-44] that are not involved in α-helix formation. In the present paper, the helical sequences listed in Table XII are those consisting of states 5, 6, 7, and 8; those in states 2, 3, and 4 are not counted as helical even though they are designated as helical in the original papers.[15-35] Thus, every residue in all 16 proteins given in Table XII was assigned to one of states 1 to 8. It should be noted that, in this section, the analysis is based on a two-state scheme, i.e., h and c; therefore, the c state includes the extended (ε) conformation, which is defined in section IB.

In Appendix I, we have summarized the number, $N_{h,j}^{(i)}$, of the jth amino acid (where j = 1 to 20) in conformational state i (where i = 1 to 8), found in the 16 proteins given in Table XII. The fraction, $f_{h,j}^{(i)}$, of the jth amino acid in state i is also tabulated, where $f_{h,j}^{(i)} = N_{h,j}^{(i)}/N_j$, with N_j being the total number of occurrences of the jth amino acid in the 16 proteins (the values of N_j are given later in Table III). The amino acids are listed in the order recommended by an IUPAC-IUB commission.[37b]

B. The Extended Conformation. We use the definition of Burgess et al.[4] for the extended conformation, viz., as a sequence of four or more consecutive residues whose dihedral angles are in the range (designated by ε): $-180° \leq \phi \leq -45°$ and $100° \leq \psi \leq 180°$ or $-180° \leq \psi \leq -140°$; and $140° \leq \phi \leq 180°$ and $100° \leq \psi \leq 180°$ or $-180° \leq \psi \leq -140°$. It should be noted that this definition specifies an extended structure regardless of whether it is or is not hydrogen bonded to another part of the chain (in a parallel or antiparallel manner).

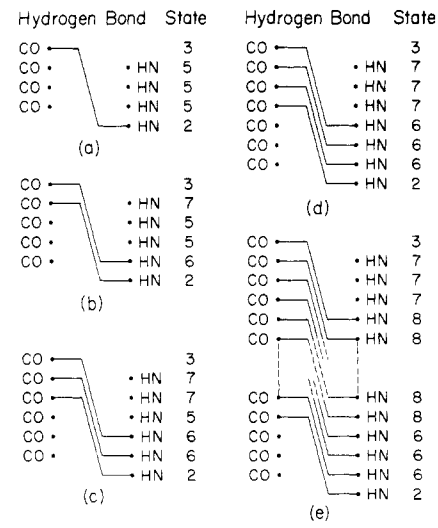

Figure 1. Schematic illustration of hydrogen-bond formation in α-helical sequences, and the assignments of conformational states to the residues in the α-helical sequences, according to the definitions given in Table I. (a), (b), (c), (d), and (e) illustrate the hydrogen bonds and the corresponding states of the residues in α-helical sequences consisting of five, six, seven, eight, and nine or more residues.

To describe the extended conformation (and its immediate environment), we assign six possible states to an ith residue. State 1 is assigned to the ith residue when it is in the interior of a sequence of ε conformations. State 2 is assigned when the ith residue is the N terminal one of a sequence of ε conformations (following a sequence of c conformations). Similarly, state 3 is assigned to the C-terminal residue of a sequence of ε conformations (followed by a sequence of c conformations). State 4 is assigned when the ith residue is in a c conformation preceding a residue in state 2. State 5 is assigned when the ith residue is in a c conformation following a residue in state 3. State 6 is assigned when the ith residue is in the interior of a run of c states. In all of these assignments, the symbol c includes the h conformation, but not the ε conformation. In section IA, the symbol c included the ε conformation. In the three-state model (section IIA), c does not include either the h or ε conformations.

The extended conformation region is specified in terms of states i = 1 to 6 in Appendix II, based on the proteins listed in Table XII. The data are tabulated as $N_{\epsilon,j}^{(i)}$ and $f_{\epsilon,j}^{(i)}$, the number and fraction, respectively, of the jth residue in extended state i (where i = 1 to 6), with $f_{\epsilon,j}^{(i)} = N_{\epsilon,j}^{(i)}/N_j$, where N_j is again the total number of occurrences of the jth amino acid in the 16 proteins.

II. Computation of Statistical Weights

A. Statistical Weights of Three-State Model. In a three-state model, we divide the conformational space (ϕ_i,ψ_i) of the ith residue into three regions, one corresponding to the α-helix (h state), one to the ε conformation, and the remainder of the space designated as a c state. We define statistical weights v_h, v_ϵ, and u for a residue in an h, ε, or c state, respectively. We define an additional statistical weight w_h for the ith residue when it is in an h state and there is a hydrogen bond between the CO group of the $(i - 2)$th residue and the NH group of the $(i + 2)$th residue;[39] v_h corresponds to an h state without a hydrogen bond. The physical meaning of the statistical weights w_h, v_h, v_ϵ, and u is described in detail in section I of paper II.[12] In brief, these statistical weights can be calculated by integrating the Boltzmann factor, $e^{-E/kT}$, over the conformational

space assigned to the h, ϵ, and c states, respectively, where E is the energy of the amino acid residue (including the hydrogen bond energy, and related interaction energies, when computing w_h). In computing w_h, the energy of formation of the hydrogen bond is assigned to the statistical weight of the ith residue.[12]

The residue partition function, z, may then be written as

$$z = u + v_\epsilon + w_h + v_h \tag{1}$$

On the basis of the concept that short-range interactions are dominant,[6] the conformational states of an amino acid residue in the (ϕ,ψ) space observed in proteins obey a Boltzmann distribution;[45,46] i.e., the x-ray data on proteins can then be used to obtain the statistical weights. Then the observed fractions of c, ϵ, and h states, f_c, f_ϵ, and f_h, are given by

$$f_c = u/z \tag{2}$$

$$f_\epsilon = v_\epsilon/z \tag{3}$$

$$f_h = (w_h + v_h)/z \tag{4}$$

where z is given by eq 1.[46] In order to specify that the parameters pertain to the jth species of amino acid, the subscript j is added to all quantities appearing in eq 1–4, viz., z_j, u_j, $v_{\epsilon,j}$, $v_{h,j}$, $w_{h,j}$, $f_{c,j}$, $f_{\epsilon,j}$, and $f_{h,j}$. Based on the statistical analysis carried out in section I, the fractions of the jth amino acid in the ith conformational state can be calculated by

$$f_{c,j} = n_{c,j}/N_j \tag{5}$$

$$f_{\epsilon,j} = n_{\epsilon,j}/N_j \tag{6}$$

and

$$f_{h,j} = n_{h,j}/N_j \tag{7}$$

where $n_{h,j}$, $n_{\epsilon,j}$, $n_{c,j}$ are given by

$$n_{h,j} = N_{h,j}^{(5)} + N_{h,j}^{(6)} + N_{h,j}^{(7)} + N_{h,j}^{(8)} \tag{8}$$

$$n_{\epsilon,j} = N_{\epsilon,j}^{(1)} + N_{\epsilon,j}^{(2)} + N_{\epsilon,j}^{(3)} \tag{9}$$

and

$$n_{c,j} = N_j - [n_{h,j} + n_{\epsilon,j}] \tag{10}$$

where $N_{h,j}^{(i)}$, $N_{\epsilon,j}^{(i)}$, and N_j were defined in section I. It should be noted, parenthetically, that the symbol c used in eq 2 and 5 does not include either the h or ϵ state; in contrast, the c conformation defined in the two-state schemes (h and c in section IA, and ϵ and c in section IB) includes ϵ and h in sections IA (state 1) and IB (states 4, 5, 6), respectively.

Generally speaking, in theoretical treatments of conformations of polymer chains, only the relative statistical weights are the interesting quantities. In other words, one may arbitrarily choose one of the various states as the reference state. Indeed, in the conformational theory based on a three-state model of a polyamino acid,[12] the c state was chosen as the reference state, as was also done in the helix–coil transition theories of polyamino acids, e.g., those of Zimm and Bragg[9] and Lifson and Roig.[47] In the Tanaka–Scheraga theory[12] of the three-state model of a polyamino acid, the following relative statistical weights were defined:

$$w_h^* = w_h/u \tag{11}$$

$$v_\epsilon^* = v_\epsilon/u \tag{12}$$

and

$$v_h^* = v_h/u \tag{13}$$

Inserting the subscript j to specify the jth species of amino acid, and using eq 2–4 and 11–13, $v_{\epsilon,j}^*$ can be calculated by the following relations:

$$v_{\epsilon,j}^* = v_{\epsilon,j}/u_j = (v_{\epsilon,j}/z_j)/(u_j/z_j) = f_{\epsilon,j}/f_{c,j} \tag{14}$$

Using eq 2 and 4, the following relation can be obtained

$$w_h/u = (w_h/z)/(u/z) = f_h/f_c \tag{15}$$

in which v_h is omitted (see reference 46). Equations 11 and 15 lead to

$$w_{h,j}^* = f_{h,j}/f_{c,j} \tag{16}$$

Thus, the relative statistical weights of the three-state model,[12] $w_{h,j}^*$ and $v_{\epsilon,j}^*$, can be calculated for all the amino acids (j = 1 to 20) on the basis of the statistical analyses performed in section I.

B. Relationship between the Statistical Weights of the Two-State and Three-State Models. In the two-state model of the helix–coil transition, the statistical weight u may be obtained by integrating $e^{-E/kT}$ over the entire (ϕ,ψ) space, excluding the small range corresponding to the h state. On the other hand, in the three-state model, the statistical weight u is obtained by integrating $e^{-E/kT}$ over the entire (ϕ,ψ) space, excluding the regions corresponding to both the h and ϵ states. Therefore, we introduce superscripts to distinguish $u^{(2)}$ for the c state of the two-state model from $u^{(3)}$ for that of the three-state model. In a similar manner, we introduce the symbols $w_h^{(2)*}$, $w_h^{(3)*}$, and $v_\epsilon^{(3)*}$.

It then follows that $u^{(2)}$ (or 1, in relative statistical weights) corresponds to $u^{(3)} + v_\epsilon^{(3)}$ [or to $1 + v_\epsilon^{(3)*}$ in relative statistical weights]. Therefore, the relative statistical weights for the two- and three-state models are related as follows:

$$w_h^{(2)*} = w_h^{(3)*}/(1 + v_\epsilon^{(3)*}) \tag{17}$$

$$v_h^{(2)*} = v_h^{(3)*}/(1 + v_\epsilon^{(3)*}) \tag{18}$$

(It should be noted the superscript (3) on $v_\epsilon^{(3)}$ is superfluous since the ϵ state is included in the c state of the two-state model.) All statistical weights of eq 17 and 18 (which are described in more detail in section IV of paper II) may include a subscript j to specify the jth species of amino acid.

It should also be noted that the statistical weights $w_{h,j}^{(2)*}$ and $v_{h,j}^{(2)*}$ are calculable directly from data on protein conformation, in the same manner as described for $w_{h,j}^{(3)*}$ and $v_{h,j}^{(3)*}$ in section IIA and in reference 48.

C. Nucleation of Helical Sequences and Asymmetric Properties. We consider now the nucleation of helical sequences, and the asymmetric character of the nucleation process. Recently, we developed a general treatment of a conformational transition in which we took into account the asymmetric character of the nucleation of two different conformational states,[49] and applied it to the conformational transition between form I and form II of poly(L-proline).[50,51] As noted earlier,[49] the same treatment is applicable to the helix–coil transition in polyamino acids, which we now consider.

We begin with the two-state model, and consider the sequence

$$\begin{array}{l} \ldots\ c,c,c,c,h,h,h,h,h,h,h,h,h,h,c,c,c,c,c,h,h,h,\ldots \\ \ldots\ 1,1,1,3,7,7,7,8,8,8,6,6,6,2,1,1,1,3,7,7,7,\ldots \\ \qquad\qquad\quad \uparrow \qquad\quad \uparrow \qquad \uparrow \\ \qquad\qquad\quad v_N^{(h)}\ v_C^{(h)} \qquad v_N^{(c)} \qquad v_C^{(c)} \end{array} \tag{19}$$

where the symbols h and c in the first line denote the helical and coil conformations of each residue, and the numer-

Table II
Conformational States and Statistical Weights of the Tanaka–Scheraga Model of the Helix–Coil Transition[a]

Conformational state of			Statistical wt[b]	Rel statistical wt for specific-sequence copolymer	Dummy statistical wt[c]
$i-1$	i	$i+1$			
h	h	h	w	s	q_1
h	h	c	$\sigma_C^{(h)}w$	$\sigma_C^{1/4}\beta_C^{1/4}s$	q_2
h	c	h	$\sigma_N^{(c)}\sigma_C^{(c)}u$	$\sigma_N^{1/4}\sigma_C^{1/4}\beta_N^{-1/4}\beta_C^{-1/4}$	q_3
c	h	h	$\sigma_N^{(h)}w$	$\sigma_N^{1/4}\beta_N^{1/4}s$	q_4
h	c	c	$\sigma_N^{(c)}u$	$\sigma_N^{1/4}\beta_N^{-1/4}$	q_5
c	h	c	$\sigma_N^{(h)}\sigma_C^{(h)}w$	$\sigma_N^{1/4}\sigma_C^{1/4}\beta_N^{1/4}\beta_C^{1/4}s$	q_6
c	c	h	$\sigma_C^{(c)}u$	$\sigma_C^{1/4}\beta_C^{-1/4}$	q_7
c	c	c	u	1	q_8

[a] See ref 49 for a more detailed description. [b] See eq 33–36 for relations between the v's and the σ's. [c] It should be noted that the subscripts 1 to 8 on the q's are *not* related to the states (1 to 8) in the last column of Table I, but pertain to those of Table I of ref 49.

als in the second line denote the conformational states described in section IA and summarized in Table I. Then, we define the statistical weights $v_N^{(h)}$ and $v_N^{(c)}$, and $v_C^{(h)}$ and $v_C^{(c)}$ for the left- and right-end residues of the helical and coil sequences, respectively. The correspondence between the statistical weights presented above and those introduced in our earlier theory[49] is summarized in Table II.

We may define the corresponding relative statistical weights as

$$v_{C_j}^{(h)*} = v_{C_j}^{(h)}/u_j \quad (20)$$

$$v_{N_j}^{(h)*} = v_{N_j}^{(h)}/u_j \quad (21)$$

$$v_{C_j}^{(c)*} = v_{C_j}^{(c)}/u_j \quad (22)$$

and

$$v_{N_j}^{(c)*} = v_{N_j}^{(c)}/u_j \quad (23)$$

Using similar reasoning to that used to obtain eq 14, the value of $v_{C_j}^{(h)*}$ of eq 20 can be computed as

$$v_{C_j}^{(h)*} = f_j'^{(6)}/f_j^{(1)} \quad (24)$$

with the aid of

$$v_{C_j}^{(h)*} = [v_{C_j}^{(h)}/z_j]/(u_j/z_j) \quad (25)$$

where

$$f_j'^{(6)} = [N_j^{(6)}/3]/N_j \quad (26)$$

and

$$f_j^{(1)} = N_j^{(1)}/N_j \quad (27)$$

where $N_j^{(6)}$ and $N_j^{(1)}$ are the number of times that the jth amino acid is found in states 6 and 1, respectively, as described in section IA and shown in Table I for $i = 1$ to 8, and N_j is the total number of times that the jth amino acid is found in the proteins listed in Table XII, as defined previously in section IA. As seen in the sequence of expression 19, state 6 is assigned to the three residues at the C terminus of the helical sequence.[52] Therefore, the number of occurrences of an h state at the C terminus of a helical sequence (at the junction with a coil sequence) is equal to one-third of the number of residues in state 6. Hence, in eq 26, the value of $[N_j^{(6)}/3]$ is used; however, it is possible to use an alternative quantity[52] instead of $[N_j^{(6)}/3]$ in eq 26. In a similar manner,[52] $v_{N_j}^{(h)*}$, $v_{C_j}^{(c)*}$, and $v_{N_j}^{(c)*}$, defined by eq 21–23, can be computed by

$$v_{N_j}^{(h)*} = f_j'^{(7)}/f_j^{(1)} \quad (28)$$

in which

$$f_j'^{(7)} = [N_j^{(7)}/3]/N_j \quad (29)$$

$$v_{C_j}^{(c)*} = f_j^{(3)}/f_j^{(1)} \quad (30)$$

$$v_{N_j}^{(c)*} = f_j^{(2)}/f_j^{(1)} \quad (31)$$

in which, as in eq 27,

$$f_j^{(i)} = N_j^{(i)}/N_j \quad (32)$$

We consider next the initiation parameters $\sigma_N^{(h)}$, $\sigma_C^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$ introduced in the Tanaka–Scheraga theory of the helix–coil transition,[49] and summarized in terms of the notation used here in Table II. If the parameters, $v_C^{(h)}$, $v_N^{(h)}$, $v_C^{(c)}$, and $v_N^{(c)}$ introduced above are related to the statistical weights defined in Table II (and in ref 49), we find

$$v_N^{(h)} = \sigma_N^{(h)}w \quad (33)$$

$$v_C^{(h)} = \sigma_C^{(h)}w \quad (34)$$

$$v_N^{(c)} = \sigma_N^{(c)}u \quad (35)$$

and

$$v_C^{(c)} = \sigma_C^{(c)}u \quad (36)$$

Then, $\sigma_N^{(h)}$ can be computed from eq 33 for the jth amino acid as

$$\sigma_{N_j}^{(h)} = v_{N_j}^{(h)}/w_j = [v_{N_j}^{(h)}/z_j]/(w_j/z_j) = f_j'^{(7)}/f_j'^{(5,6,7,8)} \quad (37)$$

in which $f_j'^{(7)}$ is computed by eq 29, and $f_j'^{(5,6,7,8)}$ by

$$f_j'^{(5,6,7,8)} = N_j'^{(5,6,7,8)}/N_j \quad (38)$$

and

$$N_j'^{(5,6,7,8)} = n_{h,j} - \{[N_j^{(6)} + N_j^{(7)}]/3\} \quad (39)$$

Equation 39 is used[53] to obtain the number of *interior* helical residues, and $n_{h,j}$ of eq 39 is obtained from eq 8, because the number of residues located at the N and C terminal ends of the helical sequences, i.e., $[N_j^{(7)}/3]$ and $[N_j^{(6)}/3]$, must be subtracted from the total number of the jth amino acid residue in the helical conformations found in proteins, i.e., $n_{h,j}$. Thus, it should be noted that the relative statistical weight $w_{h_j}^{(2)*}$ for the two-state model has been defined by

$$w_{h_j}^{(2)*} = w_{h_j}^{(2)}/u_j^{(2)} = f_j'^{(5,6,7,8)}/f_j^{(1)} \quad (40)$$

rather than by eq 16 (see ref 53). In a manner similar to that used in obtaining eq 37, we have obtained the following equations related to $\sigma_{C_j}^{(h)}$, $\sigma_{N_j}^{(c)}$, and $\sigma_{C_j}^{(c)}$, by using eq 34–36:

$$\sigma_{C_j}^{(h)} = v_{C_j}^{(h)}/w_j = [v_{C_j}^{(h)}/z_j]/(w_j/z_j) = f_j'^{(6)}/f_j'^{(5,6,7,8)} \quad (41)$$

$$\sigma_{N_j}^{(c)} = v_{N_j}^{(c)}/u_j = [v_{N_j}^{(c)}/z_j]/(u_j/z_j) = f_j^{(2)}/f_j^{(1)} \quad (42)$$

$$\sigma_{C_j}^{(c)} = v_{C_j}^{(c)}/u_j = [v_{C_j}^{(c)}/z_j]/(u_j/z_j) = f_j^{(3)}/f_j^{(1)} \quad (43)$$

Table III
The Number and the Frequencies of the Amino Acids Found in Proteins, and Their Occurrence in Various Helical Conformational States

Amino acid j	N_j[a]	P_j[b]	i = (2–8) $N_j(i)$[c]	i = (2–8) $f_j(i)$[d]	i = (3, 7) $N_j(i)$[c]	i = (3, 7) $f_j(i)$[d]	i = (2, 6) $N_j(i)$[c]	i = (2, 6) $f_j(i)$[d]	i = (5, 8) $N_j(i)$[c]	i = (5, 8) $f_j(i)$[d]
Ala	250	0.0959	135	0.540	30	0.120	40	0.160	65	0.260
Arg	83	0.0318	27	0.325	2	0.024	12	0.145	13	0.157
Asn	134	0.0514	43	0.321	19	0.142	14	0.105	10	0.075
Asp	124	0.0475	44	0.355	23	0.186	8	0.065	13	0.105
Cys	54	0.0207	17	0.315	6	0.111	5	0.093	6	0.111
Gln	98	0.0376	40	0.408	10	0.102	12	0.122	18	0.184
Glu	117	0.0449	67	0.573	35	0.299	14	0.120	18	0.154
Gly	226	0.0867	50	0.221	17	0.075	10	0.044	22	0.097
His	76	0.0291	38	0.500	7	0.092	18	0.237	12	0.158
Ile	114	0.0437	47	0.412	9	0.079	13	0.114	25	0.219
Leu	204	0.0782	98	0.480	21	0.103	32	0.157	45	0.221
Lys	188	0.0721	80	0.426	12	0.064	36	0.192	32	0.170
Met	32	0.0123	14	0.438	1	0.031	6	0.188	7	0.219
Phe	90	0.0345	37	0.411	7	0.078	9	0.100	20	0.222
Pro	89	0.0341	22	0.247	22	0.247	0	0.0	0	0.0
Ser	215	0.0824	66	0.307	23	0.107	24	0.112	18	0.084
Thr	164	0.0629	52	0.317	22	0.134	10	0.061	20	0.122
Trp	48	0.0184	21	0.438	6	0.125	3	0.063	12	0.250
Tyr	102	0.0391	26	0.255	12	0.118	9	0.088	5	0.049
Val	200	0.0767	81	0.405	17	0.085	25	0.125	39	0.195

[a] This analysis is based on the two-state scheme (h or c) in which the c state includes the ϵ conformation. N_j is the number of the jth species of amino acid found in the 16 proteins listed in Table XII. The total number of amino acids in the 16 proteins is given by $\sum_{j=1}^{20} N_j = 2608$. [b] The frequency of occurrence of the jth amino acid in these 16 proteins, P_j, is defined by $P_j = N_j/(\sum_{j=1}^{20} N_j)$. [c] $N_j(i)$ in each column is the sum over $N_j(i)$ for those values of i given in the column heading, where the individual $N_j(i)$'s are given in Appendix I. [d] $f_j(i) = N_j(i)/N_j$.

in which $f_j'^{(6)}$, $f_j'^{(5,6,7,8)}$, and $f_j^{(i)}$ are given by eq 26, 38, and 32, respectively.

The parameters σ, β_N, and β_C, which were introduced in our previous paper[49] and used in Table II, can be computed, as described in ref 49, by using the parameters $\sigma_N^{(h)}$, $\sigma_C^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$ obtained in eq 37 and 41–43. The parameter σ_j for the jth amino acid may be calculated[54] by

$$\sigma_j = \sigma_{N,j}^{(h)}\sigma_{N,j}^{(c)}\sigma_{C,j}^{(h)}\sigma_{C,j}^{(c)} \quad (44)$$

The parameters $\beta_{N,j}$ and $\beta_{C,j}$ can be calculated by using

$$\beta_{N,j} = [\sigma_{N,j}^{(h)}/\sigma_{N,j}^{(c)}]^2 \quad (45)$$

and

$$\beta_{C,j} = [\sigma_{C,j}^{(h)}/\sigma_{C,j}^{(c)}]^2 \quad (46)$$

which were defined in eq 12 and 13 of our previous paper.[49]

By using the values of s_j, σ_j, $\beta_{N,j}$, and $\beta_{C,j}$, the first one of which is identical with the values of $w_{h,j}^{(2)*}$ calculated by eq 40, and the latter ones being obtained from eq 44, 45, and 46, respectively, the relative statistical weights q_i (i = 1 to 8) (for the two-state model) for the asymmetric transition theory proposed by Tanaka and Scheraga[49] can be obtained by means of the definitions of q_i (i = 1 to 8) given in Table II.

Up to now, in this section, we limited our discussion of all of the nucleation parameters to the two-state model; we now describe all of the same parameters in terms of the three-state model. All of the quantities $v_C^{(3)(h)*}$, $v_N^{(3)(h)*}$, $v_C^{(3)(c)*}$, $v_N^{(3)(c)*}$, and $w_{h,j}^{(3)*}$ can be computed by using eq 24, 28, 30, 31, and 40, in which one may use the same quantities that appeared on the right-hand sides of these equations, except for the value of $f^{(1)}$. In these equations, one has to use

$$f_j'^{(1)} = [N_j^{(1)} - n_{\epsilon,j}]/N_j = f_j^{(1)} - f_{\epsilon,j} \quad (47)$$

in which $N_j^{(1)}$ and $f_j^{(1)}$ are the same quantities used in eq 27, and $n_{\epsilon,j}$ and $f_{\epsilon,j}$ are the same values used in eq 6. The parameters $\sigma_{N,j}^{(c)}$ and $\sigma_{C,j}^{(c)}$ for the three-state model, i.e., $\sigma_{N,j}^{(3)(c)}$ and $\sigma_{C,j}^{(3)(c)}$, may be computed by eq 42 and 43, respectively, by using $f'^{(1)}$ of eq 47 instead of $f^{(1)}$ that appeared in eq 42 and 43. Thus, all of the parameters that appeared in the two-state model (i.e., $v_C^{(h)*}$, $v_N^{(h)*}$, $v_C^{(c)*}$, $v_N^{(c)*}$, $\sigma_{N,j}^{(h)}$, $\sigma_{N,j}^{(c)}$, $\sigma_{C,j}^{(h)}$, $\sigma_{C,j}^{(c)}$, σ_j, $\beta_{N,j}$, and $\beta_{C,j}$) should be changed only by use of eq 47. Hence, among these parameters, only $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$ are common to both the two- and three-state models because these parameters do not involve the factor $f^{(1)}$ [hence, $f'^{(1)}$ of eq 47] as seen in eq 37 and 38. Therefore, we will not use the superscripts (2) and (3) in the expressions for $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$, hereafter. The parameters $\sigma_j^{(3)}$, $\beta_{N,j}^{(3)}$, and $\beta_{C,j}^{(3)}$ for the three-state model may be calculated by using $\sigma_{N,j}^{(h)}$, $\sigma_{C,j}^{(h)}$, $\sigma_{N,j}^{(3)(c)}$, and $\sigma_{C,j}^{(3)(c)}$. Finally, we calculate the relative statistical weights q_i (i = 1 to 8) (given in Table II) which can be used in the three-state model when one wants to combine the previous Tanaka–Scheraga theory[49] and the three-state model proposed by Tanaka–Scheraga in the accompanying paper[12] (see section V of ref 12 for more details). By using s_j, σ_j, $\beta_{N,j}$, and $\beta_{C,j}$ [or $w_{h,j}^*$, $\sigma_{N,j}^{(h)}$, $\sigma_{C,j}^{(h)}$, $\sigma_{N,j}^{(c)}$, and $\sigma_{C,j}^{(c)}$], the quantities q_i (i = 1 to 8) given in Table II may then be computed. In addition to the three-state conformations (the α-helical, ϵ, and other states), we can effectively take the asymmetric nucleation of helical sequences into account when these statistical weights are used. We may also use v_ϵ^* for the quantity q_9 introduced in section V of paper II.[12]

III. Results and Discussion

A. Tentative Values of Statistical Weights. Since the data set on protein structures is being continually enlarged and improved, the computed frequencies will likewise undergo revision. Thus, the statistical weights evaluated here should be regarded as tentative ones.

The data of Appendix I provide information about the relative tendencies of all 20 naturally occurring amino acids to adopt the α-helical conformation. The total number, N_j, of the jth amino acid, and its frequency of occurrence, P_j, in the 16 proteins of Table XII are given in the second and third columns of Table III. The values of $N_j(i)$ and $f_j(i)$ (for i = 2 to 8; see footnote c of Table II) are the number and

Table IV
The Number and the Frequencies of Amino Acids Found in the α-Helical, ϵ, and Other States, and the Corresponding Statistical Weights

Amino acid j	$n_{h,j}$[a]	$f_{h,j}$[b]	$n_{\epsilon,j}$[a]	$f_{\epsilon,j}$[b]	$f_{c,j}$[c]	Statistical weights		
						$v_{\epsilon,j}$*[d]	$w_{h,j}^{(3)*}$[e]	$w_{h,j}^{(2)*}$[f]
Ala	127	0.508	41	0.164	0.328	0.500	1.549	1.033
Arg	22	0.265	14	0.169	0.566	0.298	0.468	0.361
Asn	28	0.209	14	0.105	0.687	0.152	0.304	0.264
Asp	37	0.298	10	0.0806	0.621	0.130	0.481	0.425
Cys	12	0.222	15	0.278	0.500	0.556	0.444	0.286
Gln	35	0.357	19	0.194	0.449	0.432	0.795	0.555
Glu	57	0.487	12	0.103	0.410	0.250	1.188	0.950
Gly	36	0.159	31	0.137	0.704	0.195	0.226	0.189
His	23	0.303	10	0.132	0.566	0.233	0.535	0.434
Ile	41	0.360	27	0.237	0.404	0.587	0.891	0.562
Leu	90	0.441	47	0.230	0.328	0.702	1.343	0.790
Lys	69	0.367	24	0.128	0.505	0.253	0.726	0.580
Met	12	0.375	8	0.250	0.375	0.667	1.000	0.600
Phe	32	0.356	14	0.145	0.489	0.318	0.727	0.552
Pro	17	0.191	18	0.202	0.607	0.333	0.315	0.236
Ser	46	0.214	32	0.149	0.637	0.234	0.336	0.272
Thr	41	0.250	39	0.238	0.512	0.464	0.488	0.333
Trp	21	0.438	8	0.167	0.396	0.421	1.105	0.778
Tyr	16	0.157	25	0.245	0.598	0.410	0.262	0.186
Val	73	0.365	56	0.280	0.355	0.789	1.028	0.575

[a] The quantities $n_{h,j}$ and $n_{\epsilon,j}$ are defined in eq 8 and 9, respectively. [b] $f_{h,j}$ and $f_{\epsilon,j}$ are defined in eq 7 and 6, respectively. [c] $f_{c,j}$ is defined in eq 5, in which $n_{c,j}$ is given by eq 10. [d] $v_{\epsilon,j}$* is defined by eq 12, and calculated from eq 14 by using the values of $f_{\epsilon,j}$ and $f_{c,j}$ given in the fifth and sixth columns of this table. [e] $w_{h,j}^{(3)*}$ is defined by eq 11 for the three-state model, and computed from eq 16 by using the values of $f_{h,j}$ and $f_{c,j}$ given in the third and sixth columns of this table. [f] $w_{h,j}^{(2)*}$ was computed from eq 17 (however, it may also be computed from eq 16′ of ref 48).

frequency of the jth amino acid in helical sequences. The corresponding quantities for i = 3, 7 pertain to residues at the C termini of coil sequences (state 3) and at the N termini of helices (state 7), those for i = 2, 6 pertain to residues at the N termini of coil sequences (state 2) and at the C termini of helices (state 6), and those for i = 5, 8 pertain to residues in the interior of helices (although state 5 also pertains to ends of *short* helices, as shown in Figure 1a). The implications of the data in Table III are discussed in section IIIB.

Using the method described in section IIA, the relative statistical weights for h and ϵ states, $w_{h,j}^{(3)*}$ and $v_{\epsilon,j}$*, for the jth amino acid in the three-state model were computed. For this purpose $n_{h,j}$ and $n_{\epsilon,j}$ were computed from eq 8 and 9, and the results are given in the second and fourth columns of Table IV. Then, $f_{c,j}$, $f_{\epsilon,j}$, and $f_{h,j}$ were computed from eq 5–7, using eq 10 to obtain $n_{c,j}$; these results are given in the sixth, fifth, and third columns, respectively, of Table IV. The statistical weights $w_{h,j}^{(3)*}$ and $v_{\epsilon,j}$* for the jth amino acid were obtained from eq 16 and 14, and are listed in columns 7 and 8 of Table IV. Finally, for the sake of comparison, the values of $w_{h,j}^{(3)*}$ for the three-state model were converted to $w_{h,j}^{(2)*}$ for the two-state model using eq 17 (but see also ref 48), and are given in the last column of Table IV. The magnitudes of the *relative* statistical weights are always such that $w^{(3)*} \geq w^{(2)*}$ because, in eq 17, the statistical weight $v_{\epsilon}^{(3)*}$ is always positive or equal to zero. Of course, the actual statistical weights (i.e., not the *relative* ones) obey the relation $w^{(3)} = w^{(2)}$. The quantity $v_{h,j}$* will be discussed in section IB of paper III.[13]

Using the method described in section IIC, the parameters $v_C^{(h)*}$, $v_N^{(h)*}$, $v_C^{(c)*}$, and $v_N^{(c)*}$, for the jth amino acid in the two-state (i.e., h or c) model were computed from eq 20–32, and are given in Table V. The corresponding parameters for the three-state (i.e., h, ϵ, or c) model (but using eq 47 for the values of $f_j'^{(1)}$, instead of eq 27, when eq 20–32 are used) are given in Table VI.

Using eq 37 and 41–43, the values of $\sigma_{N,j}^{(h)}$, $\sigma_{C,j}^{(h)}$, $\sigma_{N,j}^{(c)}$, and $\sigma_{C,j}^{(c)}$ were computed for the jth amino acid residue in the two-state model, and are given in columns 2 to 5 of Table VII. As mentioned in section IIC, the values of $\sigma_{N,j}^{(h)}$ and $\sigma_{C,j}^{(h)}$ are the same in the two-state and three-state models. Corresponding values of $\sigma_{N,j}^{(c)}$ and $\sigma_{C,j}^{(c)}$ were computed for the jth amino acid residue in the three-state model, using eq 32 and 47 with either eq 42 or 43, as described in section IIC, and are also listed in Table VII.

Substituting the values of $\sigma_{N,j}^{(h)}$, $\sigma_{C,j}^{(h)}$, $\sigma_{N,j}^{(c)}$, and $\sigma_{C,j}^{(c)}$ for both the two-state and three-state models, given in Table VII, into eq 44–46, the values of σ, β_N, and β_C for the jth amino acid residue were computed. The values of β_N and β_C are given in Tables V and VI for the two-state and three-state models, and those of σ are given in Table VII.

Finally, the values of the relative statistical weights q_i (i = 1 to 8), introduced in ref 49, and calculated by the procedure of section IIC, were computed for the jth amino acid, and are listed in Table VIII for the two-state model and in Table IX for the three-state model.

B. Stabilities of α-Helical and Extended Conformations. We now consider the conformational preferences of each of the naturally occurring amino acids. The values of $v_{\epsilon,j}$* and $w_{h,j}^{(3)*}$ in Table IV provide information about the preferences of the extended structure and α-helical conformation, relative to that of the c state, for which u_j = 1 for the jth amino acid residue. From a qualitative point of view, we use the statistical weights of Table IV, and (in Table X) list the amino acid residues in decreasing order of preference for the α-helical conformation (in both the two-state and three-state models) and, likewise, for the ϵ conformation.

The orders of the α-helical tendencies for the two-state and three-state models, in Table X, are not the same because the statistical weights are expressed relative to that of the c state, and the ϵ state is included in the c state in the two-state model but not in the three-state model. Taking Val as an example, it ranks fifth in the three-state model, but seventh in the two-state model because Val has a very strong tendency to adopt the ϵ conformation (in fact, the strongest, with v_{ϵ}* = 0.79; see Table IV). Because of the

Table V
Tentative Values of the Parameters Related to Asymmetric Nucleation in the Two-State Model[a]

Amino acid j	$v_C^{(h)*}$	$v_N^{(h)*}$	$v_C^{(c)*}$	$v_N^{(c)*}$	β_N	β_C
Ala	0.0986	0.0812	0.0174	0.0522	2.83	37.6
Arg	0.0476	0.0060	0.0179	0.0714	0.0603	61.8
Asn	0.0147	0.0513	0.0549	0.110	3.73	1.22
Asp	0.0208	0.0792	0.0500	0.0375	33.92	1.32
Cys	0.0270	0.0270	0.0811	0.0541	3.42	1.52
Gln	0.0517	0.0460	0.0345	0.0517	3.09	8.80
Glu	0.0600	0.200	0.100	0.100	5.17	0.465
Gly	0.0057	0.0208	0.0341	0.0398	8.66	0.876
His	0.0702	0.0263	0.105	0.263	0.0386	1.72
Ile	0.0547	0.0249	0.0597	0.0299	2.45	2.97
Leu	0.0881	0.0535	0.0377	0.0377	4.01	10.9
Lys	0.0802	0.0340	0.0093	0.0926	0.488	272.8
Met	0.0741	0.0185	0.0	0.111	0.0843	∞[b]
Phe	0.0440	0.0314	0.0377	0.0377	2.49	4.88
Pro	0.0	0.0846	0.0746	0.0	∞[b]	0.0
Ser	0.0403	0.0224	0.0872	0.0403	5.10	3.52
Thr	0.0298	0.0327	0.0982	0.0	∞[b]	0.996
Trp	0.0370	0.0741	0.0	0.0	∞[b]	∞[b]
Tyr	0.0263	0.0219	0.0921	0.0395	11.7	3.10
Val	0.0588	0.0364	0.0336	0.0336	4.37	11.4

[a] For an explanation of the parameters in the two-state model (h or c), see section IIC. The superscript (2), which indicates that the parameters pertain to the two-state model, and the subscript j, which specifies the species of amino acid, are omitted for simplicity. [b] These values are very large ($>>1$).

Table VI
Tentative Values of the Parameters Related to Asymmetric Nucleation in the Three-State Model[a]

Amino acid j	$v_C^{(h)*}$	$v_N^{(h)*}$	$v_C^{(c)*}$	$v_N^{(c)*}$	β_N	β_C
Ala	0.153	0.126	0.0270	0.0811	1.17	15.6
Arg	0.0635	0.0079	0.0238	0.0952	0.0339	34.7
Asn	0.0173	0.0606	0.0649	0.130	2.67	0.871
Asp	0.0238	0.0905	0.0571	0.0429	26.0	1.01
Cys	0.0455	0.0455	0.136	0.0909	1.21	0.538
Gln	0.0769	0.0684	0.0513	0.0769	1.40	3.98
Glu	0.0790	0.263	0.132	0.132	2.98	0.269
Gly	0.0069	0.0253	0.0414	0.0483	5.88	0.595
His	0.0952	0.0357	0.143	0.357	0.0210	0.932
Ile	0.0917	0.0417	0.100	0.0500	0.874	1.06
Leu	0.158	0.0960	0.0678	0.0678	1.24	3.37
Lys	0.103	0.0437	0.0119	0.119	0.295	165.0
Met	0.133	0.0333	0.0	0.200	0.0260	∞[b]
Phe	0.0598	0.0427	0.0513	0.0513	1.35	2.64
Pro	0.0	0.116	0.102	0.0	∞[b]	0.0
Ser	0.0513	0.0285	0.111	0.0513	3.14	2.17
Thr	0.0457	0.0502	0.151	0.0	∞[b]	0.423
Trp	0.0526	0.105	0.0	0.0	∞[b]	∞[b]
Tyr	0.0392	0.0327	0.137	0.0588	5.28	1.40
Val	0.111	0.0688	0.0635	0.0635	1.22	3.20

[a] For an explanation of the parameters in the three-state model (h, ϵ, or c), see section IIC. The superscript (3) which indicates that the parameters pertain to the three-state model, and the subscript j, which specifies the species of amino acid, are omitted for simplicity. [b] These values are very large ($>>1$).

large value of v_ϵ^* (i.e., the high stability of the ϵ state for Val), the value of $w_h^{(2)*}$ is reduced, according to eq 17.

In Table X, the amino acid residues have been grouped into three categories according to their *helix-stabilizing* power.[55] Those in the first group may be regarded as *helix stabilizers* ($w_h^* \geq 0.5$), those in the middle group as *helix indifferent* ($0.27 \leq w_h^* < 0.5$ and $0.33 < w_h^* < 0.5$ for the two-state and three-state models, respectively), and those in the last group as *helix destabilizers* ($w_h^* \leq 0.27$ and 0.33 for the two-state and three-state models, respectively). It should be noted that the group in which a particular amino acid residue is included is the same in both the two-state and three-state models.

We now compare the α-helix stabilizing assignments of Table X to those made by other authors. Kotelchuck and Scheraga[56] classified the amino acids Ala, Val, Leu, Ile, Met, Trp, Gln, Glu, Phe, Cys, Arg, Pro as h, and Ser, Thr, Asn, Asp, Tyr, Lys, and His as c. Our present assignments of the helix-stabilizing power are found to be in fairly good agreement with those made by Kotelchuck and Scheraga. As seen in the first column of Table X, the *helix stabilizers*, i.e., the amino acid residues from Ala to Lys, are found among the h units of Kotelchuck and Scheraga, except for Lys which is the weakest of the helix stabilizers in the classification of Table X. The amino acids belonging to the *helix-indifferent* group (in Table X) (including amino acids such as Lys and Pro in the intermediate regions between the groups) are found among either the h or c units of Kotelchuck and Scheraga. As for Pro, as mentioned in section IIIC, it is not a helix stabilizer but a strong helix initiator at the N terminus of the α-helical sequence. Therefore, Pro may be included among the h units, accord-

Table VII
Tentative Values of the Parameters Related to the Nucleation of α-Helical Sequences

Amino acid j	$\sigma_N^{(h)}$[a]	$\sigma_C^{(h)}$[a]	For the two-state model[b]			For the three-state model[c]		
			$\sigma_N^{(c)}$	$\sigma_C^{(c)}$	σ	$\sigma_N^{(c)}$	$\sigma_C^{(c)}$	σ
Ala	0.0878	0.107	0.0522	0.0174	0.85×10^{-5}	0.0811	0.0270	0.21×10^{-4}
Arg	0.0175	0.140	0.0714	0.0179	0.31×10^{-5}	0.0952	0.0238	0.56×10^{-5}
Asn	0.212	0.0606	0.110	0.0549	0.78×10^{-4}	0.130	0.0649	0.11×10^{-3}
Asp	0.218	0.0575	0.0375	0.0500	0.24×10^{-4}	0.0429	0.0571	0.31×10^{-4}
Cys	0.100	0.100	0.0541	0.0811	0.44×10^{-4}	0.0909	0.136	0.12×10^{-3}
Gln	0.0909	0.102	0.0517	0.0345	0.17×10^{-4}	0.0769	0.0513	0.37×10^{-4}
Glu	0.227	0.0682	0.100	0.100	0.15×10^{-3}	0.132	0.132	0.27×10^{-3}
Gly	0.117	0.0319	0.0398	0.0341	0.51×10^{-5}	0.0483	0.0414	0.75×10^{-5}
His	0.0517	0.138	0.263	0.105	0.20×10^{-3}	0.357	0.143	0.36×10^{-3}
Ile	0.0467	0.103	0.0299	0.0597	0.86×10^{-5}	0.0500	0.100	0.24×10^{-4}
Leu	0.0756	0.124	0.0377	0.0377	0.13×10^{-4}	0.0678	0.0678	0.43×10^{-4}
Lys	0.0647	0.153	0.0926	0.0093	0.85×10^{-5}	0.119	0.0119	0.14×10^{-4}
Met	0.0323	0.129	0.111	0.0	0.13×10^{-4}[d]	0.200	0.0	0.42×10^{-4}[d]
Phe	0.0595	0.0833	0.0377	0.0377	0.71×10^{-5}	0.0513	0.0513	0.13×10^{-4}
Pro	0.500	0.0	0.0	0.0746	0[e]	0.0	0.102	0[e]
Ser	0.0909	0.164	0.0403	0.0872	0.52×10^{-4}	0.0513	0.111	0.85×10^{-4}
Thr	0.108	0.0980	0.0	0.0982	0.93×10^{-4}[f]	0.0	0.151	0.22×10^{-3}[f]
Trp	0.111	0.0556	0.0	0.0	0[e]	0.0	0.0	0[e]
Tyr	0.135	0.162	0.0395	0.0921	0.80×10^{-4}	0.0588	0.137	0.18×10^{-3}
Val	0.0703	0.114	0.0336	0.0336	0.90×10^{-5}	0.0635	0.0635	0.32×10^{-4}

[a] The parameters $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$ are the same for the two-state and three-state models (see section IIC). The subscripts j are omitted for simplicity. [b] The superscript (2) and subscript j are omitted for simplicity. [c] The superscript (3) and subscript j are omitted for simplicity. [d] Calculated by using eq a of ref 54. If eq 44 were used, the value of σ would be zero, since the paucity of x-ray data would probably make one of the terms on the right-hand side of eq 44 zero. [e] These values are very small. [f] Calculated by using eq b of ref 54. If eq 44 were used, the value of σ would be zero, for the same reason as in footnote *d*.

ing to the assignment of Kotelchuck and Scheraga, because this residue is found frequently at the N termini of α-helical sequences [see $N_j^{(7)}$ and $f_j^{(7)}$ in Appendix I]. Furthermore, the present assignment of Asn and Tyr as helix destabilizers is in accordance with that made by Kotelchuck and Scheraga. Pain and Robson,[57] using information theory, made the following assignments in the order of decreasing helix-forming power: Gln, Leu, Met, Trp, Ala, His, Ile, Gln, Val, Phe, Cys, Arg, Thr, Asp, Tyr, Lys, Asn, Gly, Ser, (Pro); our present assignments correlate fairly well with these, except for Ala, Lys and His. Lewis et al.[58] grouped the amino acid residues as follows: helix formers (Val, Gln, Ile, His, Ala, Trp, Met, Leu, Glu), helix indifferent (Lys, Tyr, Asp, Thr, Arg, Cys, Phe), and helix breaker (Pro, Ser, Gly, Asn).[38] Fairly good agreement between our present assignment and that of Lewis et al. is found except for amino acids, such as Phe, Lys, His, Ser, in the boundary region between the groups. The only significant exception is the assignment of Tyr, as helix indifferent, while Tyr is a strong helix destabilizer according to the present assignment. Comparing our results with the assignments made by Finkelstein and Ptitsyn,[59] in which Ala, Glu, His, and Leu are classified as what they call helical, Arg, Asp, Gln, Ile, Lys, Met, Phe, Thr, Trp and Val as neutral, and Asn, Cys, Gly, Pro, Ser, and Tyr as antihelical, fairly good agreement between both assignments is again found except for His. According to the analysis of Burgess et al.[4] the assignment of helical tendency is in the order of Glu, Ala, Trp, Leu, Gln, Met, Lys, His, Val, Ile, Phe, Asp, Thr, Arg, Pro, Cys, Ser, Asn, Tyr, and Gly, which is in good agreement with the present assignment, especially in that Tyr is a strong helix destabilizer. A recent analysis made by Chou and Fasman[60] has led to similar results to those obtained here except for His.

Thus, we see that there is a difference in the assignments for His made by various authors. This difference can be explained as follows. As can be seen for $f_j^{(i)}$ (i = 2 to 8) for His in Table III [$f_j^{(i)}$ = 0.500], half of the His residues were found to be in conformational states 2 to 8. Furthermore, as observed from the value of $N_j^{(i)}$ for His in Appendix I, 10, 4, 8, and 3 residues (25 out of a total of the 38 residues found in states 2–8) were found in states 2, 3, 6, 7, respectively (i.e., at the ends of helices), while 11 (respectively, 12) residues were found in state 8 (respectively in states 5 and 8), i.e., in the interior of the α-helical sequences. Thus, the discrepancy in the assignment of His arises from the way in which the α-helical conformational state is defined, i.e., whether (i) all conformational states 2 to 8 are included, which is the procedure often used by x-ray crystallographers and some other authors,[56–60] or (ii) states 2, 3, and 4 are omitted, while states 5, 6, 7, and 8 are assigned as α-helices, on the basis of the values of (ϕ,ψ) and the presence of a hydrogen bond, as mentioned in section IA; the latter is the procedure used here in computing the statistical weights $w_{hj}^{(3)*}$ and $w_{hj}^{(2)*}$ (of course, v_ϵ^* also will be affected through the values of u_j, which include states 1, 2, 3, and 4). Thus, among all of the amino acids, His is especially remarkable in that a considerable number of His residues is found in states 2, 3, and 4 (15 residues) while 11 residues occur in the interior of the helix (i.e., in state 8). Since we have not included states 2, 3, and 4 in the α-helical conformation, the resulting helix-stabilizing power has been reduced; this is the reason why other authors ranked His as a strong helix former rather than as a helix-indifferent residue, which is the assignment made in Table X. However, it is important to realize that such discrepancies as the different assignments of the helical tendency of His, and, more generally, the differences in the magnitudes of the statistical weights depending upon the different definitions of the α-helical conformation, must not affect our main objective, i.e., the prediction of protein conformation; i.e., the assignment of conformational states must be made by the same definition as that adopted when the x-ray data on proteins are interpreted. In other words, since we have omitted states 2, 3, and 4 in defining the α-helical state, these states are not included in the α-helical sequence predicted. Therefore, for example, it can be expected that His will be found in states 2, 3, and 4, i.e., as c states near the ends of α-helical sequences rather than in states 5 and 8, i.e., in the interior of the α-helical sequence, as long as we do not in-

Table X
Tentative Assignments of Conformational Tendencies of the Amino Acids in the α-Helical and Extended Conformations

α-Helical tendency based on the statistical wt of		Extended conformational tendency[c]
Three-state model[a]	Two-state model[b]	
Ala	Ala	Val
Leu	Glu	Leu
Glu	Leu	Met
Trp	Trp	Ile
Val	Met	Cys
Met	Lys	Ala
Ile	Val	
Gln	Ile	Thr
Phe	Gln	Gln
Lys	Phe	Trp
		Tyr
His	His	Pro
Thr	Asp	Phe
Asp	Arg	Arg
Arg	Thr	Lys
Cys	Cys	Glu
Ser	Ser	Ser
		His
Pro	Asn	
Asn	Pro	Gly
Tyr	Gly	Asn
Gly	Tyr	Asp

[a] In the order of the statistical weights, $w_{h,j}^{(3)*}$, given in Table IV (from the large one to the small one). [b] In the order of the statistical weights, $w_{h,j}^{(2)*}$, given in Table IV. [c] In the order of the statistical weights, $v_{\epsilon,j}^{*}$, given in Table IV.

clude states 2, 3, and 4 as α-helical states when predictions are made.[13,14]

Turning to a consideration of extended structures, the amino acid residues were ordered according to their ε-conformation forming power (in the last column of Table X), based on the values of v_ϵ^* given in column 7 of Table IV. Again three groups are distinguished as ε formers ($v_\epsilon^* \geq 0.5$), ε indifferent ($0.2 < v_\epsilon^* < 0.5$), and ε breakers ($v_\epsilon^* \leq 0.2$). The results show that the Val, Leu, Met, Ile, and Cys have a strong ε tendency which is in good agreement with the results of the analysis of Burgess et al.[4] and of Chou and Fasman.[60] Thus, it is seen that, as a general characteristic property, the amino acids Leu, Val, and Met have preference for highly structured, or ordered conformations (i.e., α-helical and ε structures), since these residues are in the top groups in every column of Table X, while Asn and Gly residues have preferences for unstructured, or unordered conformations (i.e., c-state tendency) since they appear at the bottom of every column of Table X.

Finally, the present results will be compared to experimental observations of the conformational properties of polyamino acids in solution in section IIID, together with a discussion of solvent effects, because $w_{h,j}^{(2)*}$ and the Zimm–Bragg parameter s can be affected by solvent. No comparable experimental results (involving the effect of solvent) are available for comparison with the statistical weights v_ϵ^*.

C. Asymmetric Properties of the Amino Acids in Nucleation of the α-Helix. The theory of asymmetric nucleation of the conformational transition has been described in general terms for one-dimensional transition phenomena in our previous paper,[49] and applied there to interpret the conformational transition of poly(L-proline). In section IIC, we described asymmetric nucleation in terms of the helix–coil transition, and calculated various sets of statistical weights such as ($v_{C_j}^{(h)*}$, $v_{N_j}^{(h)*}$, $v_{N_j}^{(c)*}$, $v_{C_j}^{(c)*}$), ($\sigma_{C_j}^{(h)}$, $\sigma_{N_j}^{(h)}$, $\sigma_{N_j}^{(c)}$, $\sigma_{C_j}^{(c)}$), (σ_j, β_{N_j}, β_{C_j}), and the statistical weights q_i (i = 1 to 8). In a general sense, all of these parameters (which pertain to the boundaries between helical and coil sequences) are statistical weights (involving contributions from *both* nucleation and propagation processes, and not only from nucleation processes) for the corresponding states relative to the c state, as described in section IIC and in ref 49. The quantities $\sigma_N^{(h)}$, $\sigma_C^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$, and β_N and β_C designate the contributions of nucleation to the statistical weights assigned to states in the boundary between the different conformational phases; the statistical weights for states in the boundary between different conformational sequences are usually expressed as a product of two terms corresponding to nucleation parameters $\sigma_C^{(h)}$, $\sigma_N^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$ and the propagation parameters w and u, as seen, for example, in the second column of Table II. In order to interpret the physical meanings of these parameters, we will consider several amino acids as examples. After that, we will deduce the general implications about the asymmetric properties of amino acids from the results of Tables V, VI, and VII.

The numerical values of $v_{C_j}^{(h)*}$, $v_{N_j}^{(h)*}$, $v_{C_j}^{(c)*}$, and $v_{N_j}^{(c)*}$ provide information about the stabilities of conformational state 6, 7, 3, and 2 (see expression 19) of the jth amino acid relative to the c state, and the values of $w_{h,j}^*$ specify the stabilities of a helical state in the interior of an α-helical sequence (relative to the c state). Therefore, if we examine the values of $v_C^{(h)*}$ and $v_N^{(h)*}$ of Table V, we know the relative stabilities of the α-helical residues at the C and N terminal ends of the α-helical sequence. For example, the values of $v_C^{(h)*}$ (=0.0986) and $v_N^{(h)*}$ (=0.0812) for Ala (see Table V) show that Ala does not have significant asymmetric properties, since the relative stabilities of these two states are similar. Similarly, Cys has identical values (0.0270) for $v_C^{(h)*}$ and $v_N^{(h)*}$. In contrast, for Arg, the α-helical conformation is considerably more stable at the C terminus [$v_C^{(h)*}$ = 0.0476] of the α-helical sequence than that at the N terminus [$v_N^{(h)*}$ = 0.0060], and vice versa for Glu [$v_C^{(h)*}$ = 0.0600 and $v_N^{(h)*}$= 0.200]. The same information, based on the statistical weights $v_N^{(h)*}$ and $v_C^{(h)*}$, may be obtained, alternatively, from the nucleation parameters $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$ (given in the second and third columns of Table VII).

Therefore, one method to assign the asymmetric properties of the amino acids for the formation of the α-helical sequence may be based on either $v_N^{(h)*}$ and $v_C^{(h)*}$, or $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$, which has often been employed, implicitly or explicitly, in describing the asymmetric properties, i.e., the asymmetric properties of the amino acids have often been deduced only from the number of α-helical residues found at (or near) the N or at (or near) the C terminus of the α-helical sequence. However, from a statistical thermodynamical point of view, the asymmetric properties of amino acids should not be based only on the quantities $v_N^{(h)*}$ and $v_C^{(h)*}$, or $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$, or on the number of amino acids found at the ends of the α-helical sequences, as has often been done by previous authors.[60,61] The reason for this is that the occurrence of ends of the α-helical sequences has to be related not only to $v_N^{(h)*}$ and $v_C^{(h)*}$, or $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$, but also to $v_N^{(c)*}$ and $v_C^{(c)*}$, or $\sigma_C^{(c)}$ and $\sigma_N^{(c)}$ which are related to the occurrence of the ends of c-state sequences. In other words, as described in ref 49, the four parameters $\sigma_N^{(h)}$, $\sigma_C^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$ [or $v_N^{(h)*}$, $v_C^{(h)*}$, $v_N^{(c)*}$, and $v_C^{(c)*}$] are not independent of each other; thus, the asymmetric properties of amino acids should be described on the basis of the four parameters $\sigma_N^{(h)}$, $\sigma_C^{(h)}$, $\sigma_N^{(c)}$, and $\sigma_C^{(c)}$ [or $v_N^{(h)*}$, $v_C^{(h)*}$, $v_N^{(c)*}$, and $v_N^{(h)*}$] rather than the two parameters $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$ [or $v_N^{(h)*}$ and

$v_C^{(h)*}$]. In other words, the number of amino acids found at the ends of the α-helix [on the basis of which we computed the values of $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$ (or $v_N^{(h)*}$ and $v_C^{(h)*}$)] cannot be independent of the number of the amino acids found at the ends of the c-state sequences. For this reason, as defined in eq 45 and 46, β_N and β_C have been defined in terms of the magnitudes of $\sigma_N^{(h)}$ and $\sigma_C^{(h)}$, which measure the nucleations at the N and C terminal ends of the α-helical sequence, relative to the magnitudes of $\sigma_N^{(c)}$ and $\sigma_C^{(c)}$, which measure the nucleations at the N and C terminal ends of the c-state sequences.

The appearance of a certain amino acid, j, at the ends of the α-helical or the c-state sequence can be described in terms of the parameters β_{Nj} and β_{Cj}, in the helix–coil transition theory described in ref 49. Let us consider the conformational transitions in two directions, one from the amino to the carboxyl end, and the other from the carboxyl to the amino end. The parameters $\sigma_N^{(h)}$ and $\sigma_N^{(c)}$ [or $v_N^{(h)*}$ and $v_N^{(c)*}$] measure the nucleation difficulties of the α-helical state and the c state for the conformational transition occurring in the direction from the N to the C terminal ends. Therefore, if β_N [$=\sigma_N^{(h)}/\sigma_N^{(c)}$ as given in eq 45] is larger than 1, nucleation of the α-helical state is preferable to that of the c state, which means that the jth amino acid can serve as a helix initiator at the N terminus of the α-helical sequence. In a similar fashion, for the opposite direction of the conformational transition, if β_{Cj} is larger than 1, the jth amino acid can serve as a helix initiator at the C terminus of the α-helical sequence. On the contrary, if β_{Nj} is smaller than 1, the jth amino acid can serve as a helix terminator (or c-state initiator) at the N terminus. If $\beta_{Cj} < 1$, the jth amino acid can serve as a helix terminator (or c-state initiator) at the C terminus.[62,63] If $\beta_{Nj}/\beta_{Cj} \sim 1$, the jth amino acid is considered to have no asymmetric properties (helix impartial) for the nucleation of the α-helical sequence. Thus, the numerical values of β_{Nj} and β_{Cj} given in the last two columns of Table V (or of Table VI for the three-state model) provide us with quantitative information about the asymmetric properties of the amino acids for the nucleation of α-helical sequences.

In order to simplify the illustration, we will restrict our discussion to the two-state model because the general qualitative trends of the asymmetric properties among the amino acids are not altered in going from the two-state to the three-state model, as seen in Tables V and VI. As described above, no significant asymmetric properties could be observed for the Ala residue, when based on the values of $v_C^{(h)*}$ (=0.0986) and $v_N^{(h)*}$ (=0.0812) (or $\sigma_N^{(h)} = 0.088$ and $\sigma_C^{(h)} = 0.107$). However, the values of β_N (=2.83) and β_C (=37.6) for Ala (given in Table V) show that Ala can have substantial asymmetric properties (though not so strong) for the formation of the α-helical conformation, which is caused primarily because of the difference in the nucleation properties of the c state [see $v_N^{(c)*}$ (=0.052) and $v_C^{(c)*}$ (=0.017) given in Table V (or $\sigma_N^{(c)}$ and $\sigma_C^{(c)}$ given in Table VII)]. As general conclusions, among the 20 amino acids listed in Table V, the values of β_C indicate that Met, Lys, Arg, His, and Ala may be classified as *helix initiators at the C terminus* of the α-helical sequence. This assignment of His has been made on the basis of the difference in the asymmetric properties of nucleation of the α-helical state (e.g., $\beta_C/\beta_N = 1.72/0.0386 \simeq 43$ for His) even though the absolute values of β_C and β_N are very small, because most of the His residues in the conformational states 2 to 8 were found in the c states 2, 3, and 4 rather than in the helical states 5, 6, 7, and 8. Nevertheless, the His residue has an asymmetric property with respect to the initiation of the α-helix. On the other hand, Pro, Thr, Asp, and Glu can be assigned as *helix initiators at the N terminus* of the helical sequence (because $\beta_N > \beta_C$), while Asn, Cys, Gln, Gly, Ile, Leu, Phe, Ser, Trp, Tyr, and Val have *helix-impartial* properties for nucleation of the α-helix (because β_N and β_C do not differ significantly), although each amino acid of this group has slight asymmetric properties as seen in the values of β_N and β_C given in Table V. It should be noted that the present assignment of the asymmetric nucleation property of a jth amino acid does not mean that the jth amino acid is found more frequently at the N or C terminus than the other amino acids; instead it means that the jth amino acid is found more frequently at one end of an α-helical sequence than at the other end.

Some authors have taken the asymmetric nucleation or termination properties of amino acids (for forming α-helices) into account in their empirical prediction schemes. Kotelchuck et al.[62] found that polar amino acids can play a helix-breaking role; e.g., Asp and Asn serve as helix breakers at the C termini of helical sequences. This also has a quantitative basis in the values of β_N and β_C for polar residues (Tables V and VI). Residues with positively charged side chains (Arg, His, and Lys), with $\beta_N < \beta_C$, are helix terminators at the N termini and helix initiators at the C termini of helical sequences. On the other hand, residues with negatively charged side chains (Asp and Glu), with $\beta_N > \beta_C$, are helix initiators at the N termini and helix terminators at the C termini of helical sequences. The asymmetric nucleation and termination properties of polar residues in α-helical sequences may be attributed to electrostatic interactions,[64] i.e., between the charged residues at the N and C termini and the dipole moment along the helix axis.[63] This dipole has its negative end at the C terminus and its positive end at the N terminus. Thus, positively charged residues at the C terminus, and negatively charged residues at the N terminus, stabilize the helix. Blagdon and Goodman[63] took this asymmetric nucleation property into consideration in their empirical scheme to predict α-helical sequences.

Finally, we point out that, in order to obtain information about the asymmetric properties of nucleation, the parameters β_N and β_C are determinable by analyzing experimental data on the helix–coil transition in polyamino acids in a manner similar to that used in the determination of the Zimm–Bragg parameters σ and s, which have been obtained for the naturally occurring amino acids.[8] In such a system, the independent parameters describing the helix–coil transition are the degree of polymerization of the polyamino acid, β_N and β_C, in addition to the parameters s and σ.[49–51]

D. Solvent Effects. It is of interest to compare the present statistical weights, obtained from x-ray data, to those determined from experimental studies on polyamino acids in solution. Since experimental results on helix–coil transitions in polyamino acids are usually treated in terms of a two-state model, whereby s and σ of the Zimm–Bragg theory[9] or w and v of the Lifson–Roig theory[47] are evaluated, we will compare these experimental results with $w_{h_j}^{(2)*}$. For this purpose, we have summarized the experimental values[8,65–87] of s and σ (or w and v) in Table XI.

As seen in Table XI, some of the experimental values of s and σ vary considerably, depending upon the experimental conditions (e.g., solvents, salt concentrations, and temperature) and the experimental methods. Since these variations may not be due only to experimental error, it seems to be a questionable procedure to average these quantities to obtain experimental values of s and σ for the jth amino acid, as Chou and Fasman did.[60] Furthermore, considerable error can be committed by comparing the statistical weights obtained from x-ray data with experimental values of the parameters s and σ of the Zimm–Bragg theory[9] (or

Table XI
Observed Values of s (or w) and σ (or v) from Solution Experiments on Polyamino Acids

Amino acid	Solvent	Exptl[a] method (type of transition obsd)	s (or w)[b] at 10° C	20° C	25° C	30° C	σ (or v)[b]	Ref
Ala	Water	O	1.06*	1.05*	1.04*	1.04*	0.012*	65
	0.06 M NaBr	T			1.09			66
	Water	O	1.08	1.07		1.06	8×10^{-4}	67
Glu	0.1 N KCl, pH 2.3 or water, pH 2.3	O[c]	1.42	1.35		1.28	1×10^{-2}	8
	0.1 N KCl, pH 8	[d]	0.97	0.97		0.97	6×10^{-4}	8
	Water	[d]	0.96	0.96		0.96	6×10^{-4}	8
	0.2 M NaCl	T[c]	1.45	1.36	1.32	1.29		68
	0.2 M, 0.013 M NaCl	T[e]					3×10^{-3}	69
	0.1 M KCl	T[e]	1.30	1.22	1.18	1.14		70
	KCl, KSCN[f]	T[c]	1.61	1.49	1.44	1.39		71
	0.01, 0.05, 0.1, 0.4 M NaCl	T[c]	1.48	1.39	1.35	1.32		72, 73[g]
	0.2 M NaCl	T[c]					2.5×10^{-3}	73
	0.1 M NaCl	T[c]	1.36	1.27	1.24	1.20	5×10^{-3}	74
	0.1 M KCl, 8 M urea	T, O[c]					5×10^{-5}	75
	0.1 M KCl and ethanol vol % of ethanol:							
	10	T[c]	1.33	1.30	1.28	1.27		76
	20		1.44	1.42	1.41	1.40		76
	27		1.53	1.51	1.50	1.49		76
	0.1 M KCl and 3 M methanol	T, O[c]			1.30			
	3 M propanol				1.51			
	1.71 M ethanol	T, O, at 9° C					0.70×10^{-3}	77
		at 35° C					0.90×10^{-3}	77
	3.42 M ethanol	at 14° C					0.90×10^{-3}	77
		at 35° C					1.30×10^{-3}	77
	4.62 M ethanol	at 10° C					0.84×10^{-3}	77
	4.95 M methanol	at 25° C					0.98×10^{-3}	77
	2.01 M propanol	at 25° C					1.25×10^{-3}	77
	0.87 M butanol	at 25° C					1.10×10^{-3}	77
	0.2 M NaCl and dioxane		1.78	1.61	1.53	1.46		73
Leu	Water	O	1.19*	1.20*	1.20*	1.21*	0.05*–0.011*	78
	Water	O	1.12	1.14		1.14	3.3×10^{-4}	79
			1.12	1.15		1.16	1.2×10^{-4}	79
			1.11	1.12		1.12	3.0×10^{-3}	79
	0.05 M KF	T			1.92		6.8×10^{-3}	80
							6.3×10^{-2}	80
Lys	0.1 M KCl	T[c]	1.24	1.18	1.15	1.12		70
	0.1 M KCl	T[c]			1.15			71
	0.1 M KSCN	T[c]			1.16			71
	0.1 M KF	T[c]			1.17			71
	0.06 M NaBr	T[c]			1.27		2.9×10^{-3}	66
	0.2 M NaCl	T[c]	1.05	1.00	0.98	0.95	2.3×10^{-4}	81
	0.05 M KF	T[c]			1.15		4.6×10^{-4}	80
Phe	Water	O	1.08	1.08		1.06	1.8×10^{-3}	82
Val	Water	O	0.90	0.93		0.97	1×10^{-4}	83
His	0.02 M KCl	T[c]	1.25	1.17	1.13	1.09		84
	0.1 M KCl	C[c]			1.00[h]			84
					0.86[h]			84
					0.78[h]			84
Ser	Water	O	0.73	0.76		0.77	1×10^{-4}	85
Gly	Water	O	0.55	0.59		0.62	1×10^{-5}	86
Pro							*i*	87

[a] The experimental methods are designated as: O, optical (optical rotatory dispersion or circular dichroism); T, titration; and C, calorimetric. [b] The parameters w and v of Lifson and Roig[47] are designated by an asterisk in this table ($\sigma \sim v^2$). [c] Transition of the uncharged random coil to the uncharged helix. See ref 8 for a discussion of the transitions in Glu. [d] Transition of the charged coil to the charged helix. [e] Transition of the charged coil (or the extended coil conformation) to the uncharged helix. [f] The authors of ref 71 did not specify the concentration of the salts. [g] Without carrying out any computations, the authors of ref 73 concluded that the values of the enthalpy and entropy change for the transition of the uncharged coil to the uncharged helix were the same as those reported in ref 72, because of the agreement of their plot of the free energy difference vs. temperature with those reported in ref 72. [h] These values (0.78, 0.86, and 1.00) were calculated by using $\Delta H = -0.94$, -1.00, and -1.07 kcal/mol from calorimetric measurements (given in ref 84) and by assuming $\Delta S = -3.65$ eu, which was observed in titration measurements (given in ref 84). [i] Ganser et al.[87] obtained experimental values of s, σ, β_N, and β_C (see ref 49 as to the physical meanings of σ, β_N, and β_C) by observing the form I $\rightleftarrows$ form II transition of poly(L-proline). The results are: $\sigma = 1.0 \times 10^{-5}$ (in benzyl alcohol–1-butanol at 70° C), $10^{-6} < \sigma < 10^{-5}$ (in trifluoroethanol–1-butanol at 70° C), and $\sigma = 5 \times 10^{-6}$ (in trifluoroethanol–1-butanol at 25° C). However, these values should *not* be compared to the values of $w_{h,j}^{(2)*}$, σ_j, $\beta_{N,j}$, and $\beta_{C,j}$ (j = Pro) (see the text and ref 60 for more detail).

the parameters w and v of the Lifson–Roig theory[47]) if attention is not paid to the nature of conformational transition observed experimentally in solution. As one example, one should never compare the statistical weights $w_{h,j}^{(2)}$, $\sigma_j^{(2)}$, and the parameters of asymmetric nucleation such as $\beta_{N,j}$ and $\beta_{C,j}$ (for j = proline) with the experimental values of s, σ, β_N, and β_C observed for the order $\rightleftarrows$ order (helix form I $\rightleftarrows$ helix form II) conformational transition of poly(L-proline) in solution (see the theoretical phenomenological and molecular treatments in our previous papers[49–51] for the definitions and the physical meanings of s, σ, β_N, and β_C, and the correspondence of these quantities to those used in ref 87). The order $\rightleftarrows$ order and helix–coil transitions are two entirely different phenomena. All of the statistical weights of the two-state model for j = proline in this paper are defined for the helix–coil transition and not for the helix–helix transition; the latter involves a conformational transformation accompanied by rotation about the peptide bonds (variation of ω[37]), in addition to variation of ϕ and ψ.[37] Thus, the comparison of s and σ for proline made by Chou and Fasman[60] is erroneous. As another example, as we have indicated in the third column of Table XI, one often observes different conformational transitions although the same notation, s and σ, is used to describe the phenomena. For example, the values of s and σ for Glu are defined for different conformational transitions such as the transition of the uncharged random coil to the uncharged helix, of the charged coil to the charged helix, and of the charged coil (or the extended coil conformation) to the uncharged helix, as indicated in the third column of Table XI; this may be responsible for the variations in the values of s and σ which were pointed out above. Therefore, it is meaningless to average these values together. It is necessary to understand the physical meanings of the statistical weights and the possible factors which can affect their magnitudes both in proteins and in synthetic polyamino acids before any comparisons can be made.

We will first discuss the effects of solvent on the parameters $w_{h,j}^{(2)*}$ (or s) and σ (or the other parameters of asymmetric nucleation of helical sequences described in section IIC). As described by Gō et al.[88] and Tanaka and Scheraga,[50,51] the effects of solvent on the helix–coil transition in polyamino acids can be classified into three types: (1) specific interactions of solvent molecules with *free* NH and CO groups of the amide group of amino acid residues (only the CO group for the proline residue[50,51]); (2) effect of the solvent molecules as the dielectric constant of a bulk medium on electrostatic interactions; and (3) additional interaction energies between atoms or groups of atoms of a protein molecule caused by the presence of solvent molecules. We will discuss these three types of solvent effects in both protein folding and the helix–coil transition in polyamino acids.

Several examples of phenomena caused by interactions of type (1) have been observed in helix–coil transitions of polyamino acids in solution. One example is the inverse thermal helix–coil transition of poly(γ-benzyl L-glutamate) in the two-component solvents system, dichloroacetic acid (DCA) and 1,2-dichloroethane (DCE),[89] in which DCA molecules bind to the NH and CO groups of the amino acid residue in the coil state, while DCE plays the role of an inactive diluent.[90–93] Another example is the conformational transition between the form I and II helices of poly(L-proline), triggered by the different affinities between the different species of alcohol molecules and the CO groups in the form I and II conformations.[49–51,87] In aqueous solution, water can bind to the backbone NH and CO groups (type 1 solvent effect). As discussed elsewhere,[50,51,88,94] a type (1) solvent effect can affect the parameter s for polyamino acids in solution considerably. In native proteins, it is much more difficult to assess the solvent effect of type (1) on the values of s (or the values of $w_{h,j}^{(2)*}$). Generally speaking, the existence of the α-helical state of amino acids located on the surfaces of proteins (or exposed to the solvent) will be determined by the competition between the relative stabilities of the helical and coil states, with the latter being influenced by bound solvent molecules. On the other hand, for the α-helical state of amino acids located in the interior of the native protein, this competition may not exist, especially for those involved in nonpolar cores in which the solvent molecules may not be able to approach the peptide bonds to form the hydrogen bonds [it should be noted that this is described as a type (1) solvent effect, and not as a type (3) hydrophobic effect described below]. Even when solvent molecules can exist around the amino acid residues located in the interior of protein molecules, the values of s in a protein can still differ from those observed in a polyamino acid in solution, because of the different activities of the solvent molecules in solution and in the interior of native proteins (see, for example, eq 37, 38, and the equation given in footnote 27 of ref 50 for the effect of the activity of the solvent molecules on the partition function). Thus, it may be considered that the values of $w_{h,j}^{(2)*}$ (or s) for the jth amino acid, evaluated on the basis of x-ray data, are averaged over the effects mentioned above, which depend on the various environments of the jth amino acid in native proteins. In other words, the statistical weights evaluated here for the jth amino acid pertain to the conformational properties of this residue averaged over the various environments encountered in native proteins [also, including the effects of long-range interactions which, however, can be neglected (see section IIIE)]. Therefore, the values of $w_{h,j}^{(2)*}$ (or s) may correspond to those that are obtained by averaging the values of s observed in the *helix–coil* transitions in polyamino acids in *several* solvent systems which simulate the environments present in native proteins; however, the x-ray values of $w_{h,j}^{(2)*}$ need not correspond to the values of s and σ from polyamino acids in a *single* solvent system, as stated above.

In contrast to the effect on s, a type (1) solvent effect cannot influence the parameter σ,[50,51,88,94] or the asymmetric nucleation parameters β_N and β_C.[50,51] Therefore, as far as the type (1) solvent effect is concerned, the values of σ (and β_N and β_C, although experimental values of these quantities are not yet available) given in Tables V–VII are comparable to those observed in solution (given in the eighth column of Table XI).

The type (2) solvent effect can also affect the values of both s and σ of the helix–coil transition of polyamino acids.[88b] The value of s of an amino acid may differ depending upon the environment (e.g., on whether the medium is aqueous or organic) around the amino acid in the native state, mainly because of the different effective local dielectric constant. This effect, as well as the type (1) solvent effect, may also be able to be taken into account by using the values of s observed in a model polymer–solvent system that simulates the conditions present in the native protein. In this respect, it would be of interest to investigate the variation of s in aqueous and organic media. Actually, some recent studies have detected the type (2) solvent effect on the values of s. For example, the values of s observed can change remarkably when organic solvents such as dioxane and alcohol are added to aqueous solution as can be seen for the values of s for Glu in 0.2 M NaCl (s = 1.32 at 25°C),[68] and in 0.2 M NaCl, dioxane (s = 1.53 at 25°C),[73] in the fourth and last rows of Glu in Table XI. Other examples for the values of s observed for Glu are those in 0.1 M KCl, 3 M methanol[77] (shown in Table XI), and in 0.1 M

KCl, ethanol (shown in the original paper[76]). In both cases, the values of s increase as the amount of methanol and of ethanol increases. More recently, Dubin[94] found that the free energy change (ΔG) in going from the uncharged helix of Glu to the uncharged coil obviously depends on the dielectric constants of the bulk media (see the plot of $-\Delta G$ vs. the dielectric constant in Figure 2 of ref 94). These variations in the values of s (or in the free energy change, $-\Delta G$) may be due partly to a type (2) and also a type (3) solvent effect, although the type (1) solvent effect can also affect the values of s or ΔG by affecting the activities of the solvent, as described above. Roughly speaking, the α-helical state in an organic environment is more stable relative to the coil state than in an aqueous medium because of the absence of solvent-peptide hydrogen bonds which stabilize the coil state in the latter environment. From the experimental results cited above, it appears that the α-helical states of amino acids in organic environments are more stable (i.e., the values of s are larger) than in an aqueous environment in native proteins.

In contrast, the variation in the values of σ seems to be less sensitive to a type (2) solvent effect, perhaps because the nucleation of the α-helix occurs in the early stages of protein folding in which the environment of the amino acids may not be different from that in the bulk solution. In other words, we may expect the values of σ observed in the native state of a protein to be similar to those observed in the helix–coil transition in polyamino acids in solution.

The typical solvent effect of type (3) is the hydrophobic interaction energy.[95] Hydrophobic interactions can stabilize the α-helix in polyamino acids which have nonpolar groups in their side chains. This effect can be reflected in the value of s. The hydrophobic interaction free energy can be evaluated for the amino acids by using the Nemethy–Scheraga theory.[95] For example, for poly(L-alanine), the hydrophobic interaction energy between a β_1 carbon and an α_4 carbon amounts to -0.35 kcal/mol[88b,95] at 27°C, which successfully accounts for experimental results[88b] on the helix–coil transition in poly(L-alanine) in water. The important role of hydrophobic bonds has been described in the theory of the helix–coil transition,[96,97] and confirmed by a number of experimental studies.[65,70,72,78] However, as a matter of course, we cannot straightforwardly assign these hydrophobic interaction energies, estimated on the basis of experiments for homopolymers, to the amino acid residues of protein molecules. The reason for this is that the hydrophobic (and also electrostatic[98]) interactions which stabilize the α-helical sequence are obviously dependent on the species of the ith, $(i + 4)$th, and $(i - 4)$th amino acid, which in turn are dependent on the amino acid sequence of the protein. It is beyond the scope of the present model to take into account the (longer-range, sequence-dependent) hydrophobic interactions in the α-helical sequences of proteins, mainly because the x-ray data are not sufficiently extensive to enable us to evaluate the quantities correlated to pairs of amino acids found in proteins. However, in principle, it is possible to take this effect into account within a one-dimensional Ising model, i.e., in the model of specific-sequence copolymers of amino acids, in which the hydrophobic interaction energy may be assigned depending upon the pair of amino acids found at the ith and $(i + 4)$th positions. The hydrophobic interaction free energy can be evaluated, for example, by the Nemethy–Scheraga theory.[95]

The above discussion of solvent effects would seem to make it almost meaningless, at present, to compare, in a quantitative manner, the statistical weights $w_{h,j}^{(2)*}$, evaluated from x-ray data, with the values of s determined from studies of the helix–coil transition in polyamino acids, because the values of s for polyamino acids have not been determined in many different solvents. Nevertheless, it is of interest to make this comparison. For this purpose, we will discuss the relative helix-forming tendency among the amino acids observed in each of two different systematic studies in a qualitative manner. As seen in Table XI, the values of s observed in the helix–coil transition of polyamino acids in solution show that the amino acids Ala, Glu, Leu, Lys, and Phe behave like helix stabilizers (because $s > 1$ at room temperature), while His and Val behave as helix indifferent residues ($s \simeq 1$). The values of s for Ser and Gly, which are smaller than unity, show that these amino acids serve as helix destabilizers. Among the helix stabilizers, the Lys residue shows a tendency toward weak helix-forming power. These relative helix-stabilizing powers of the amino acids observed in solution are in good agreement with those found in native proteins (except for the Phe residue), as seen in the second column of Table X.

Finally, we will compare the values of σ evaluated on the basis of the x-ray data (given in Table VII) with those observed in the helix–coil transition of polyamino acids in solution. As stated previously, the values of σ are relatively insensitive to solvent effects. Therefore, the numerical values of σ observed in the native state (given in Table VII) are in satisfactory agreement with those observed in solution (see the values of σ of Table XI), if one takes into consideration that the determinations of the values of σ are less sensitive than are those of the values of s when analyzing the experimental data. It is of particular interest to point out that the value of σ for Glu, observed in the native state, is larger (less cooperative) than for the other amino acids except His (see Table VII), which is in good agreement with the experimental observations on σ in solution, in particular, with the large value of $\sigma = 1 \times 10^{-2}$ reported in ref 8.

E. Some Considerations of Possible Long-Range Effects on the Statistical Weights. Based on earlier studies of Kotelchuck et al.[56,62,64] and subsequent work summarized in ref 6, it seems that the conformational preferences of polypeptide chains (for helices, extended structures, bends, and nonregular structures) are determined primarily by short-range interactions, with long-range interactions playing a lesser role. The questions thus arise as to whether the neglect of long-range interactions in the procedures described here can (i) invalidate our use of the one-dimensional Ising model for treating protein conformation, or (ii) affect the statistical weights evaluated in this paper. We consider both of these questions in this section.

In the second paragraph of the introductory section, we have already mentioned some arguments attesting to the approximate validity of the one-dimensional Ising model for treating helix and coil states in proteins. Similarly, conformations in the ϵ region appear to be stabilized primarily by short-range interactions, thereby validating a one-dimensional Ising model treatment that includes ϵ states. The ϵ region of a conformational energy map of an amino acid residue is broad and relatively flat, the main energy contribution being an electrostatic interaction between neighboring amide dipoles.[99–101] In addition, because of the flatness of this region, it is also stabilized entropically. Indeed, as stated by Flory[101] in conjunction with the calculation of the characteristic ratio, $\langle R^2\rangle_0/nl^2$ of the polypeptide chain, the partition function of an amino acid residue in the coil state is determined primarily by the statistical weights attributed to the ϵ region; e.g., the contributions of the ϵ region, the right-hand α-helical region, and the left-handed α-helical region to the partition function (in the absence of long-range interactions) are 93, 6, and 1%, respectively, for alanine.[101] Thus, the extended conformation

Table XII
The α-Helical (h) and Extended (ϵ) Regions Observed by X-Ray Experiments[a]

Protein	Source	No. of amino acid residues	α-Helical (h) region[a]	Extended (ϵ) region[a]
Myoglobin[b]	Sperm whale	153	4–17, 21–35, 37–41, 52–57, 59–76, 87–94, 101–117, 125–148	None
Lysozyme[c]	Hen egg white	129	5–14, 25–35, 81–84, 89–95, 110–113, 121–123	43–46
Ribonuclease S[d]	Bovine	124	4–12, 25–33, 51–55[e]	43–47, 61–65, 77–88, 95–111
Oxyhemoglobin[f] α chain	Horse	141[g]	4–17, 21–35, 37–41, 53–70, 81–88, 95–111, 119–137	None
Oxyhemoglobin[h] β chain	Horse	146	5–17, 20–33, 36–40, 51–55, 57–75, 86–93, 100–116, 124–142	None
α-Chymotrypsin[i] B chain	Bovine	131[j]	None	4–9, 15–20, 27–31, 36–40, 49–53, 67–70, 88–93, 95–100, 103–110, 119–125
C chain		97	17–24, 88–96	6–16, 32–35, 41–44, 49–55, 58–61, 76–82
Carboxypeptidase A[k]	Bovine	307	15–25,[l] 73–87, 95–99,[l] 113–121, 174–186, 216–230, 255–259,[l] 286–305	11–14, 33–37, 45–53, 60–66, 103–107, 191–198, 201–205, 236–242, 266–271
Subtilisin BPN'[m]	*Bacillus subtilis*	275	6–9, 15–19, 65–72, 104–116, 133–144, 224–237, 243–251, 270–274	2–5, 26–30, 88–91, 147–152, 174–181, 190–193, 205–210, 213–219
Insulin[n]	Pig			
A chain		21	3–5, 14–18	None
B chain		30	8–18	2–7, 24–29[o]
Elastase[p]	Pig	240	156–159, 234–239	14–22, 25–35, 38–44, 53–58, 69–80, 93–101, 124–135, 139–142, 145–153, 171–180, 185–196, 199–209, 215–226[o]
Staphylococcal[q] nuclease	*Staphylococcus aureas*	149	55–66, 100–105, 123–133	7–14, 22–27, 30–37, 39–43, 71–77, 88–94, 109–112
Papain[r]	Papaya	212[s]	25–40, 51–56, 68–77, 117–125, 139–142	162–167, 169–175, 185–192
Ferricytochrome c[t]	Horse heart	104	10–12, 15–17, 50–53, 63–69, 72–74, 92–100	None
Cytochrome b_5[u]	Calf liver	93	9–14, 34–37, 43–48, 56–61, 65–73, 81–85	4–8, 21–24, 28–33, 75–80
Myogen[v]	Carp muscle	108	8–14, 27–32, 41–50, 68–70, 79–88, 103–106	74–77
Sea lamprey[w] hemoglobin	*Petromyzon marinus*	148	13–28, 31–43, 68–87, 99–105,[x] 115–126,[x] 133–144[x]	None

[a] The h regions are assigned to sequences of three residues in state 5 (Figure 1a) or more than three residues in states 5, 6, 7, and 8 (Figure 1b–e); these conformational states are defined in section IA of the text and in Table I. See also section IB and ref 4 for the definition of the ϵ regions. [b] Reference 15. [c] References 16 and 17. [d] Reference 18. [e] Since the helical sequence at 56–59 is the α_{II}-helix, those residues are not involved in the α-helical region. [f] References 19 and 20. [g] Perutz et al.[19] reported that residues F6, G4, and GH3 are Asp, Asp, and Asp, whereas they are Asn, Asn, and Asn according to Dickerson and Geis.[36] The former assignment is employed in this paper. [h] Reference 19. [i] References 21 and 22. [j] According to Birktoft and Blow,[22] residues 55 and 77 are Glu and Asp, respectively, whereas they are Gly and Asn according to Matthews et al.[21] The former assignment was employed in this paper. [k] Reference 23. [l] Residues 26–29, 100–103, and 260–262 are in the α_{II}-helical conformation, which are not involved in the α-helical region. [m] References 24 and 25. [n] Reference 26. [o] These ϵ regions are those reported by the original authors. [p] Reference 27. [q] Reference 28. [r] References 29 and 20. [s] Mitchel et al.[29] reported that residues 64, 167, and 169 are Asp, Asn, and Gly, respectively, while, according to Drenth et al.[30] they are Asn, Gly, and Asn, respectively. The latter assignment was used in this paper. [t] Reference 31. [u] References 32 and 33. [v] Reference 34. [w] Reference 35. [x] Residues 45–52, 62–66, 92–98, 111–114, and 145–148 are in the α_{II}-helical conformation, which are omitted in this paper.

is stable without the hydrogen bonds associated with parallel and antiparallel β structures although, of course, such hydrogen bonds confer additional stability on the extended conformation. In a similar manner, the stability of chain reversal conformations can be attributed, in first approximation, to the short-range interactions within the residues comprising the bend.[6,102] Thus, we may conclude that it is a good approximation to treat protein conformation in terms of a one-dimensional Ising model consisting of a limited number of discrete states (two and three in this paper, but more than three in a subsequent paper[14]).

We now turn to the second question, as to how long-range interactions might influence the statistical weights computed here. Since short-range interactions dominate, it appears reasonable to assume that, in passing from the unfolded to the folded native state of a protein, most amino acids remain *within* a given conformational region such as h, ϵ, etc. Long-range interactions (and solvent effects) alter the dihedral angles within a given region but, as long as the residue does not move out of the given region, its statistical

Table VIII. Tentative Numerical Values of the Relative Statistical Weights in the Two-State Model

Amino Acid	Relative Statistical Weights[a]							
	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
Ala	0.925	0.107	0.001	0.081	0.032	0.009	0.017	1.0
Arg	0.339	0.140	0.001	0.006	0.071	0.002	0.018	1.0
Asn	0.242	0.061	0.006	0.051	0.110	0.013	0.055	1.0
Asp	0.363	0.057	0.002	0.079	0.038	0.013	0.050	1.0
Cys	0.270	0.100	0.004	0.027	0.054	0.010	0.081	1.0
Gln	0.506	0.102	0.002	0.046	0.052	0.009	0.034	1.0
Glu	0.880	0.068	0.010	0.200	0.100	0.015	0.100	1.0
Gly	0.178	0.032	0.001	0.021	0.040	0.004	0.034	1.0
His	0.509	0.138	0.028	0.026	0.263	0.007	0.105	1.0
Ile	0.532	0.103	0.002	0.025	0.030	0.005	0.060	1.0
Leu	0.708	0.124	0.001	0.053	0.036	0.009	0.038	1.0
Lys	0.525	0.153	0.001	0.034	0.093	0.009	0.009	1.0
Met	0.574	0.129	0.0	0.019	0.111	0.004	0.0	1.0
Phe	0.528	0.083	0.001	0.031	0.038	0.005	0.036	1.0
Pro	0.169	0.0	0.0	0.085	0.0	0.0	0.075	1.0
Ser	0.246	0.164	0.004	0.022	0.040	0.015	0.087	1.0
Thr	0.304	0.098	0.0	0.033	0.0	0.011	0.098	1.0
Trp	0.667	0.056	0.0	0.074	0.0	0.006	0.0	1.0
Tyr	0.162	0.162	0.004	0.022	0.039	0.022	0.052	1.0
Val	0.518	0.114	0.001	0.036	0.034	0.008	0.034	1.0

(a) See Table II for the actual statistical weights corresponding to the dummy statistical weights q_i (i = 1 to 8).

Table IX. Tentative Numerical Values of the Relative Statistical Weights in the Three-State Model

Amino Acid	Relative Statistical Weights[a]							
	q_1	q_2	q_3	q_4	q_5	q_6	q_7	q_8
Ala	1.437	0.107	0.002	0.126	0.081	0.009	0.027	1.0
Arg	0.452	0.140	0.002	0.008	0.095	0.002	0.024	1.0
Asn	0.286	0.061	0.008	0.061	0.130	0.013	0.065	1.0
Asp	0.414	0.057	0.002	0.090	0.043	0.013	0.057	1.0
Cys	0.455	0.100	0.012	0.045	0.091	0.010	0.136	1.0
Gln	0.752	0.102	0.004	0.068	0.077	0.009	0.051	1.0
Glu	1.258	0.068	0.017	0.283	0.132	0.015	0.132	1.0
Gly	0.216	0.032	0.002	0.025	0.048	0.004	0.041	1.0
His	0.691	0.138	0.051	0.036	0.357	0.007	0.143	1.0
Ile	0.892	0.103	0.005	0.042	0.050	0.005	0.100	1.0
Leu	1.272	0.124	0.005	0.096	0.068	0.009	0.068	1.0
Lys	0.675	0.153	0.001	0.044	0.119	0.010	0.012	1.0
Met	1.033	0.129	0.0	0.033	0.200	0.004	0.0	1.0
Phe	0.718	0.083	0.003	0.043	0.051	0.005	0.051	1.0
Pro	0.231	0.0	0.0	0.116	0.0	0.0	0.102	1.0
Ser	0.313	0.164	0.006	0.028	0.051	0.015	0.111	1.0
Thr	0.466	0.098	0.0	0.050	0.0	0.011	0.151	1.0
Trp	0.947	0.056	0.0	0.105	0.0	0.006	0.0	1.0
Tyr	0.242	0.162	0.008	0.033	0.059	0.022	0.137	1.0
Val	0.979	0.114	0.004	0.069	0.063	0.008	0.063	1.0

(a) See Table II for the actual statistical weights corresponding to the dummy statistical weights q_i (i = 1 to 8). The statistical weights were converted from the two-state to the three-state model by the procedure described in sections IIB and IIC.

Footnotes to Table XIV

(a) The definitions of the i conformational states of the j^{th} amino acid residue (i = 1 to 6) and of a sequence of s conformations are described in section IB of the text. The analysis is based on the two-state scheme (h or ϵ, where ϵ includes the helical conformation).

(b) $N_{s,j}^{(i)}$ is the number of the j^{th} amino acid (j = 1 to 20) in the i^{th} (i = 1 to 6) s conformational state (whose definitions are given in section IB of the text) found in the sixteen proteins listed in Appendix I; $f_{s,j}^{(i)}$ is the fraction of the i^{th} conformational state of the j^{th} amino acid, and is given by $f_{s,j}^{(i)} = N_{s,j}^{(i)}/N_j$ where N_j is the total number of the j^{th} amino acid found in the sixteen proteins listed in Appendix I.

Appendix I

Table XIII. The Number of Amino Acids and the Frequencies of Occurrence in the Helical Region[a]

Amino acids	Conformational State[b]															
	1		2		3		4		5		6		7		8	
	$N_{h,j}^{(1)}$	$f_{h,j}^{(1)}$	$N_{h,j}^{(2)}$	$f_{h,j}^{(2)}$	$N_{h,j}^{(3)}$	$f_{h,j}^{(3)}$	$N_{h,j}^{(4)}$	$f_{h,j}^{(4)}$	$N_{h,j}^{(5)}$	$f_{h,j}^{(5)}$	$N_{h,j}^{(6)}$	$f_{h,j}^{(6)}$	$N_{h,j}^{(7)}$	$f_{h,j}^{(7)}$	$N_{h,j}^{(8)}$	$f_{h,j}^{(8)}$
Ala	115	0.460	6	0.024	2	0.008	0	0.0	8	0.032	34	0.136	28	0.112	57	0.228
Arg	56	0.615	4	0.048	1	0.012	0	0.0	3	0.041	8	0.096	1	0.012	12	0.145
Asn	91	0.679	10	0.075	5	0.037	0	0.0	2	0.015	4	0.030	14	0.105	8	0.060
Asp	80	0.665	3	0.024	4	0.031	0	0.0	3	0.024	5	0.040	19	0.153	10	0.081
Cys	37	0.685	2	0.037	3	0.056	0	0.0	2	0.037	3	0.056	3	0.056	4	0.074
Gln	58	0.592	3	0.032	2	0.020	0	0.0	5	0.051	9	0.092	8	0.082	13	0.133
Glu	50	0.427	5	0.043	5	0.043	0	0.0	1	0.009	9	0.077	10	0.256	17	0.145
Gly	176	0.779	7	0.031	6	0.027	1	0.004	1	0.004	3	0.013	11	0.049	21	0.093
His	38	0.500	10	0.132	4	0.053	1	0.013	1	0.013	8	0.105	3	0.040	11	0.145
Ile	67	0.588	2	0.018	4	0.035	0	0.0	1	0.009	11	0.097	5	0.044	24	0.211
Leu	106	0.520	4	0.020	9	0.044	0	0.0	4	0.020	28	0.137	17	0.083	41	0.201
Lys	108	0.575	10	0.053	1	0.005	0	0.0	2	0.011	26	0.138	11	0.059	30	0.160
Met	18	0.563	2	0.063	0	0.0	0	0.0	0	0.0	4	0.125	1	0.031	7	0.219
Phe	53	0.589	2	0.022	2	0.022	1	0.011	0	0.0	4	0.044	7	0.078	5	0.056
Pro	67	0.755	6	0.0	5	0.056	0	0.0	0	0.0	0	0.0	17	0.191	0	0.0
Ser	149	0.693	6	0.028	13	0.060	1	0.005	0	0.0	18	0.084	10	0.047	18	0.084
Thr	112	0.681	0	0.0	11	0.067	0	0.0	4	0.024	10	0.061	11	0.067	16	0.098
Trp	27	0.563	0	0.0	0	0.0	0	0.0	1	0.021	3	0.063	6	0.125	9	0.188
Tyr	76	0.745	3	0.029	7	0.069	0	0.0	1	0.010	6	0.059	5	0.049	4	0.039
Val	119	0.995	4	0.020	4	0.020	0	0.0	4	0.020	21	0.105	13	0.065	35	0.175

Footnotes to Table XIII

(a) The helical conformational states (i = 1 to 8) of the j^{th} amino acid residue, and the assignment of the helical region, are described in section IA of the text and summarized in Table I. The analysis is based on the two-state scheme (h or ϵ, where ϵ includes the extended conformation defined in section IB).

(b) $N_{h,j}^{(i)}$ is the number of the j^{th} amino acid (j = 1 to 20) in the i^{th} helical conformational state (i = 1 to 8) (see footnote a) found in the sixteen proteins given in Appendix I; $f_{h,j}^{(i)}$ is the fraction of the i^{th} helical conformational state of the j^{th} amino acid, and is given by $f_{h,j}^{(i)} = N_{h,j}^{(i)}/N_j$ where N_j is the total number of the j^{th} amino acid (given in the second column of Table III) found in the sixteen proteins listed in Appendix I.

Appendix II

Table XIV. The Number of Amino Acids and the Frequencies of Occurrence in the Extended Region[a]

Amino acid	1		2		3		4		5		6	
	$N_{e,j}^{(1)}$	$f_{e,j}^{(1)}$	$N_{e,j}^{(2)}$	$f_{e,j}^{(2)}$	$N_{e,j}^{(3)}$	$f_{e,j}^{(3)}$	$N_{e,j}^{(4)}$	$f_{e,j}^{(4)}$	$N_{e,j}^{(5)}$	$f_{e,j}^{(5)}$	$N_{e,j}^{(6)}$	$f_{e,j}^{(6)}$
Ala	28	0.112	5	0.020	8	0.032	15	0.060	10	0.040	184	0.736
Arg	11	0.133	2	0.024	1	0.012	3	0.036	6	0.072	60	0.723
Asn	7	0.052	4	0.030	3	0.022	16	0.119	9	0.067	95	0.709
Asp	6	0.048	3	0.024	1	0.008	7	0.057	3	0.024	100	0.807
Cys	11	0.204	2	0.037	2	0.037	4	0.074	2	0.037	33	0.611
Gln	12	0.122	5	0.051	2	0.020	7	0.071	3	0.031	69	0.704
Glu	9	0.077	1	0.009	2	0.018	6	0.051	3	0.043	96	0.821
Gly	20	0.089	4	0.018	7	0.031	21	0.094	22	0.098	452	0.673
His	8	0.105	0	0.0	2	0.026	2	0.026	2	0.026	57	0.750
Ile	21	0.196	3	0.028	3	0.028	1	0.009	9	0.084	73	0.658
Leu	15	0.127	3	0.015	9	0.044	6	0.029	5	0.025	146	0.716
Lys	17	0.090	5	0.027	2	0.011	17	0.064	10	0.053	142	0.755
Met	5	0.156	2	0.063	1	0.031	0	0.0	1	0.031	23	0.719
Phe	17	0.133	2	0.022	0	0.0	5	0.056	1	0.011	70	0.778
Pro	17	0.115	2	0.073	4	0.045	8	0.090	4	0.045	99	0.661
Ser	18	0.086	8	0.037	6	0.028	19	0.088	17	0.079	147	0.686
Thr	21	0.165	8	0.049	4	0.024	10	0.061	12	0.073	103	0.628
Trp	7	0.146	1	0.021	0	0.0	2	0.042	4	0.083	34	0.708
Tyr	22	0.216	0	0.0	3	0.029	10	0.098	7	0.069	60	0.588
Val	30	0.090	9	0.045	7	0.035	7	0.035	8	0.040	141	0.695

weight (evaluated from x-ray data, as we have done here) will not alter significantly.

The protein folds to its globular form, not only by changing the conformations of its amino acid residues within their given conformational regions[5] (because of long-range and solvent effects), but also because of *large* changes in the conformations of some residues (assumed to be *very few* in number), i.e., a transfer from one conformational region to another, say h to ϵ, again because of long-range and solvent effects.[5] However, since these drastic changes in conformation are presumably so few in number, and because the statistical weights are averages over a *large* number of x-ray data, the average values of the statistical weights are thus relatively unaffected by the few residues that pass from one conformational region to another during folding.

We thus conclude that, as a first approximation, we may use a one-dimensional Ising model, with the statistical weights of a discrete number of states determined by x-ray data on proteins, to predict with reasonably high accuracy the preferred average conformational states of segments of a protein (to be followed, of course, by minimization of the energy of the whole protein). The statistical weights developed here, and the theory presented in paper II,[12] will be used in this manner in paper III.[13]

IV. Conclusion

We summarize here the conclusions reached in this paper. (1) The statistical weights of the 20 naturally occur-

ring amino acids, in the two-state model [α-helical (h) and other states (c)] and in the three-state model [α-helical (h), extended (ϵ), and other states (c)] have been evaluated, based on the x-ray data of native proteins; these will be used in the prediction of protein conformation in an accompanying paper.[13] (2) The asymmetric nucleation parameters of the 20 amino acids have been evaluated on the basis of a previous theory of the helix–coil transition.[49] (3) The conformational properties of the amino acids are discussed in terms of the statistical weights obtained here; the amino acids have been classified as helix stabilizer, helix indifferent, and helix destabilizer on the basis of the values of $w_{hJ}^{(2)*}$, and as helix initiator at the N terminus, helix impartial, and helix initiator at the C terminus on the basis of the asymmetric nucleation parameters. (4) A discussion was presented to compare the statistical weights evaluated from x-ray data on native proteins with the experimental results obtained from studies of the helix–coil transition in polyamino acids in solution. (5) The predominant role of short-range interactions and some possible effects of long-range interactions in determining the statistical weights were discussed in conjunction with the mechanism of protein folding. (6) While we have calculated the statistical weights from x-ray data on proteins, in this paper, these quantities can be obtained from any of a number of alternative theoretical or experimental sources, for use in the statistical mechanical treatment of protein conformation formulated in paper II.

Miniprint Material Available. Full-sized photocopies of the miniprint material from this paper only or microfiche (105 × 148 mm, 24× reduction, negatives) containing all of the miniprint and supplementary material for the papers in this issue may be obtained from the Business Office, Books and Journals Division, American Chemical Society, 1155 16th St., N.W., Washington, D.C. 20036. Remit check or money order for $4.50 for photocopy or $2.50 for microfiche, referring to code number MACRO-76-142.

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) 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.
(4) See A. W. Burgess, P. K. Ponnuswamy, and H. A. Scheraga, *Isr. J. Chem.*, **12**, 239 (1974), for a review and evaluation of these empirical prediction schemes.
(5) This is demonstrated very effectively in Figure 1 of A. W. Burgess and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **72**, 1221 (1975).
(6) H. A. Scheraga, *Pure Appl. Chem.*, **36**, 1 (1973).
(7) The role of medium-range interactions (up to four residues away from a given one) has also been considered by P. K. Ponnuswamy, P. K. Warme, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **70**, 830 (1973), and incorporated into an empirical prediction algorithm by Burgess et al.[4]
(8) See F. R. Maxfield, J. E. Alter, G. T. Taylor, and H. A. Scheraga, *Macromolecules*, 8, 479 (1975), and earlier papers of that series.
(9) B. H. Zimm and J. K. Bragg, *J. Chem. Phys.*, **31**, 526 (1959).
(10) P. N. Lewis, M. Go, N. Go, D. Kotelchuck, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **65**, 810 (1970).
(11) While Lewis et al.[10] assumed tentative values of σ and s for all of the naturally occurring amino acids, these parameters are now being determined directly from experiments on random copolymers by the host–guest technique.[8]
(12) S. Tanaka and H. A. Scheraga, *Macromolecules*, Part II, following paper in this issue.
(13) S. Tanaka and H. A. Scheraga, *Macromolecules*, Part III, following in this issue.
(14) S. Tanaka and H. A. Scheraga, manuscripts in preparation (multistate models).
(15) H. C. Watson, *Prog. Stereochem.*, **4**, 299 (1969).
(16) C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Phillips, and V. R. Sarma, *Nature (London)*, **206**, 757 (1965).
(17) D. C. Phillips, *Proc. Natl. Acad. Sci. U.S.*, **57**, 484 (1967).
(18) H. W. Wyckoff, D. Tsernoglou, A. W. Hanson, J. R. Knox, B. Lee, and F. M. Richards, *J. Biol. Chem.*, **245**, 305 (1970).
(19) M. F. Perutz, H. Muirhead, J. M. Cox, and L. C. G. Goaman, *Nature (London)*, **219**, 131 (1968).
(20) M. F. Perutz, *J. Mol. Biol.*, **13**, 646 (1965).
(21) B. W. Matthews, P. B. Sigler, R. Henderson, and D. M. Blow, *Nature (London)*, **214**, 652 (1967).
(22) J. J. Birktoft and D. M. Blow, *J. Mol. Biol.*, **68**, 187 (1972).
(23) F. A. Quiocho and W. N. Lipscomb, *Adv. Protein Chem.*, **25**, 1 (1971).
(24) C. S. Wright, R. A. Alden, and J. Kraut, *Nature (London)*, **221**, 235 (1969).
(25) R. A. Alden, J. J. Birktoft, J. Kraut, J. D. Robertus, and C. S. Wright, *Biochem. Biophys. Res. Commun.*, **45**, 337 (1971).
(26) 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).
(27) D. M. Shotton and H. C. Watson, *Nature (London)*, **225**, 811 (1970).
(28) A. Arnone, C. J. Bier, F. A. Cotton, V. W. Day, E. E. Hazen, Jr., D. C. Richardson, J. S. Richardson, and A. Yonath, *J. Biol. Chem.*, **246**, 2302 (1971).
(29) R. E. J. Mitchel, I. M. Chaiken, and E. L. Smith, *J. Biol. Chem.*, **245**, 3485 (1970).
(30) J. Drenth, J. N. Jansonius, R. Koekoek, and B. G. Wolthers, *Adv. Protein Chem.*, **25**, 79 (1971).
(31) R. E. Dickerson, T. Takano, D. Eisenberg, O. B. Kallai, L. Samson, A. Cooper, and E. Margoliash, *J. Biol. Chem.*, **246**, 1511 (1971).
(32) F. S. Mathews, M. Levine, and P. Argos, *J. Mol. Biol.*, **64**, 449 (1972).
(33) F. S. Mathews, P. Argos, and M. Levine, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 387 (1972).
(34) C. E. Nockolds, R. H. Kretsinger, C. J. Coffee, and R. A. Bradshaw, *Proc. Natl. Acad. Sci. U.S.*, **69**, 581 (1972).
(35) W. A. Hendrickson, W. E. Love, and J. Karle, *J. Mol. Biol.*, **74**, 331 (1973).
(36) R. E. Dickerson and I. Geis, "The Structure and Action of Proteins", Harper and Row, New York, N.Y., 1969, p 52.
(37) (a) Throughout this paper, we use the recommendations proposed by an IUPAC Commission on Nomenclature, *Biochemistry*, **9**, 3471 (1970); (b) IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, **14**, 449 (1975).
(38) N. Gō, P. N. Lewis, M. Gō, and H. A. Scheraga, *Macromolecules*, **4**, 692 (1971).
(39) Recently, the present authors[40] developed a theory for the helix–coil transition in polyamino acids based on a consideration of intermolecular interactions. They also defined the conformational state of a residue in terms of the dihedral angles (ϕ_i,ψ_i) for the h and c states, and in terms of the state of hydrogen bonding, but in a different manner from that of Gō et al.[38] In the Tanaka–Scheraga formulation,[40] the hydrogen-bonded states were described in terms of the presence or the absence of hydrogen bonds (i) between the CO group of the (i − 2)th residue and the NH group of the (i + 2)th residue, (ii) on the CO group of the (i − 1)th residue, (iii) on the NH and CO groups of the ith residue, and (iv) on the NH group of the (i + 1)th residue,[37] in assigning the statistical weights for the ith residue, rather than in terms of the presence or the absence of hydrogen bonds on the NH and CO groups of the ith residue, as proposed by Gō et al.[38] We will follow the formulation of Gō et al.[38] throughout the present analysis in order to *identify* h states in the x-ray structures of proteins, but will use the Tanaka–Scheraga description of hydrogen bonding[40] when computing statistical weights.
(40) S. Tanaka and H. A. Scheraga, manuscript in preparation (based on statistical weights from potential functions).
(41) Maps showing the arrangement of hydrogen bonds in helical and extended structure regions have been reported for myoglobin,[15] ribonuclease S,[18] α-chymotrypsin,[22] and sea lamprey hemoglobin.[35] Hydrogen bond maps for the extended structure regions of lysozyme,[17] carboxypeptidase,[23] elastase,[27] papain,[30] and cytochrome b_5[32] have been reported.
(42) The hydrogen bonds being ignored here are those involved in the formation of parallel and antiparallel pleated sheets, those involved in the formation of helical structures other than α (or α_I)[43] helices (e.g., α_{II}[43] and 3_{10} helices) (as an exception see ref 44), and those between two different side chains.
(43) G. Nemethy, D. C. Phillips, S. J. Leach, and H. A. Scheraga, *Nature (London)*, **214**, 363 (1967).
(44) Hydrogen bond maps for the α- and β-hemoglobins are not yet available. The original authors[19] reported that helix C is a 3_{10} helix (based on a 2.8-Å resolution map). However, we counted helix C in these two proteins as α helical since each consists of five amino acid residues, which *are* α helical in sperm whale myoglobin[15] and in the homologous (horse) α- and β-hemoglobins[19,20] (based on a 1.4-Å resolution map).
(45) The *observed* conformation of an amino acid in a protein can be affected by the conformation of its neighboring amino acids along the chain because of cooperativity. For example, a given amino acid can adopt the helical conformation more easily the greater the tendency of its neighboring amino acids to form a helical conformation. However, in this paper, we neglect this cooperativity in that we regard the observed data for the amino acid residues in proteins as pertaining to isolated residues; this is consistent with the evidence[6] that the confor-

mations of amino acid residues in proteins are determined *predominantly* by short-range interactions (including the hydrogen bond for the helical state). The cooperativity *is* taken into account when these statistical weights are used in the one-dimensional Ising models formulated in paper II and applied to proteins in paper III. If one wanted to take the cooperativity into account when evaluating the statistical weights, one would have to adjust them to reproduce the observed protein conformation exactly. However, this cannot be done at the present time because there are not enough experimental data to adjust 60 statistical weights, i.e., three (w_h^*, v_h^*, and v_t^*) for each of 20 amino acids.

(46) When an h state is defined by its dihedral angles (ϕ,ψ), it is not yet specified whether it is or is not involved in a hydrogen bond. Thus, f_h is given by eq 4, and the ratio f_h/f_c is $(w_h + v_h)/u$, instead of eq 15. On the other hand, when one can discriminate whether a hydrogen bond is or is not present, then the fractions of helical states with and without a hydrogen bond, f_h' and f_h'', respectively, are given by $f_h' = w_h/z$, $f_h'/f_c = (w_h/z)/(u/z) = w_h/u$ (which is the same as eq 15), and $f_h'' = v_h/z$, $f_h''/f_c = (v_h/z)/(u/z)$. In this paper, we evaluate only w_h (by using x-ray data). The value of v_h will be obtained theoretically,[40] and will be used in paper III.[13] However, in a later paper,[14] we will evaluate both w_h and v_h separately from x-ray coordinates of native proteins.

(47) S. Lifson and A. Roig, *J. Chem. Phys.*, **34**, 1963 (1961).

(48) The relative statistical weights, $w_j^{(2)*}$, can also be computed using eq 5–7 and 16 by means of the following relations:

$$w_j^{(2)*} = f_{hj}/f_{cj} \tag{16'}$$

$$f_{cj} = n_{cj}/N_j \tag{5'}$$

and

$$f_{hj} = n_{hj}/N_j \tag{7'}$$

However, n_{cj} should then be computed by

$$n_{cj} = N_j - n_{hj} \tag{10'}$$

instead of eq 10, and n_{hj} is then given by eq 8.

(49) S. Tanaka and H. A. Scheraga, *Macromolecules*, **8**, 494 (1975).

(50) S. Tanaka and H. A. Scheraga, *Macromolecules*, **8**, 504 (1975).

(51) S. Tanaka and H. A. Scheraga, *Macromolecules*, **8**, 516 (1975).

(52) This statement is correct only in the case of a helical sequence of six or more residues, as seen in Figure 1. Nevertheless, this statement is expected to describe the general aspects of the α-helical sequences found in native proteins, because most of the α-helical sequences are six residues long or longer, as seen in Table XII. It is also possible to obtain an alternative expression to eq 26 by using, instead of $[N_j^{(6)}/3]$ in eq 26, the actual number of amino acid residues in the last (C terminal) position of α-helical sequences in native proteins. This alternative method would be preferable if we had a larger number of x-ray data. However, because of the limited set of x-ray data available, it is preferable to use the procedure described in the text (eq 26). Of course, the number of c states at junctions (i.e., in states 2 and 3) are counted directly (see expression 19). The h states at the N termini of helical sequences (i.e., in state 7) are treated similarly (see eq 29). The use of the prime in eq 24 and 26 and subsequent equations calls attention to the procedure used in eq 26; these primes have a different meaning from that in ref 46.

(53) If we used the actual number of the jth amino acid residue located at the N- and C-terminal ends of the helical sequences found in native proteins,[52] i.e., $N_j'^{(7)}$ and $N_j'^{(6)}$, instead of $[N_j^{(6)}/3]$ and $[N_j^{(7)}/3]$ in eq 26 and 29, then eq 39 would have to be

$$N_j'^{(5,6,7,8)} = n_{hj} - [N_j'^{(6)} + N_j'^{(7)}] \tag{39'}$$

(see also ref 52).

(54) From eq 11 of ref 49, we have

$$\sigma_{Nj} = [\sigma_{Nj}^{(h)} \cdot \sigma_{Nj}^{(c)}]^2 \tag{a}$$

or

$$\sigma_{Cj} = [\sigma_{Cj}\mathrm{supc(h)} \cdot \sigma_{Cj}^{(c)}]^2 \tag{b}$$

Taking into account both contributions from the N- [i.e., $\sigma_{Nj}^{(h)}$ and $\sigma_{Nj}^{(c)}$] and C- [i.e., $\sigma_{Cj}^{(h)}$ and $\sigma_{Cj}^{(c)}$] terminal ends, we may define the new parameter σ by means of eq 44 for a specific-sequence copolymer. This corresponds to the similar approximation made in using the Zimm–Bragg model for a specific-sequence copolymer, in which the physical (molecular) meaning of the parameter σ cannot be given explicitly because σ_{Nj} and σ_{Cj} are not always equal for a specific-sequence copolymer; i.e., $\sigma_{Cj} \neq \sigma_{Nj}$, which was demonstrated in our previous paper.[49] If we take $\sigma_{Nj}^{(c)} = \sigma_{Cj}^{(c)} = 1$ in eq 44–46, and if σ is set equal to $\sigma_{Nj} = \sigma_{Cj}$, the present treatment corresponds to the Zimm–Bragg model for a specific-sequence copolymer. The numerical values of σ_{Nj} and σ_{Cj} may be calculated by using eq a and b (given above in this reference) and eq 37 and 41–43. These values of σ_{Nj} and σ_{Cj} may be used to compute the relative statistical weights of column 3 of Table II, although we calculated the statistical weights q_i (i = 1 to 8) by using the values of $\sigma_{Nj}^{(h)}$, $\sigma_{Cj}^{(h)}$, $\sigma_{Nj}^{(c)}$, and $\sigma_{Cj}^{(c)}$ given in Table VII.

(55) In describing the conformational properties of the amino acid residues, it is desirable to discriminate between the stability and the ease of nucleation of the α helix; thus, we use different terms to distinguish between these properties. To denote the stabilizing power of the α-helical conformation, which may be described in terms of the parameter w_h^* (or of the parameter s of the Zimm–Bragg theory[9]), we use the terms *helix stabilizer, helix indifferent,* and *helix destabilizer* to designate the three groups in Table X. In order to specify the nucleating power of the α-helical sequence, which may be described in terms of the parameters discussed in section IIC of this paper (or the parameter σ of the Zimm–Bragg theory[9]), we use the terms *helix initiator, helix impartial,* and *helix terminator.* The terms helix maker, helix indifferent, and helix breaker,[10,38,56] or helical, antihelical, and neutral,[57] etc., that are frequently used, cannot distinguish between these two properties.

(56) D. Kotelchuck and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **62**, 14 (1969).

(57) (a) R. H. Pain and B. Robson, *Nature (London)* **227**, 62 (1970); (b) B. Robson and R. H. Pain, *J. Mol. Biol.*, **58**, 237 (1971).

(58) Originally, Lewis et al. made the assignments described in ref 10. However, in the present paper, we compare our results with their later assignments, as reported in ref 38.

(59) A. V. Finkelstein and O. B. Ptitsyn, *J. Mol. Biol.*, **62**, 613 (1971).

(60) (a) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 211 (1974). (b) There is a numerical error in eq 4 of ref 60a [see erratum, *Biochemistry*, **14**, 196 (1975)]. However, an ambiguity seems to remain because the argument of their text[60a] is based on eq a and c given below, and the numerical values given in Table II of ref 60a are computed by eq b and c below (see their paper for notation).

$$\langle f_k \rangle = \Sigma f_{j,k}/\Sigma j \tag{a}$$

$$\langle f_k \rangle = \Sigma n_{j,k}/\Sigma n_j \tag{b}$$

and

$$P_{j,k} = f_{j,k}/\langle f_k \rangle \tag{c}$$

Therefore, we recomputed the values of their parameters by both methods, i.e., using eq a and c and eq b and c, and confirmed that their order of helical and extended tendencies of amino acids did not change, no matter which method is used. However, the *numerical values* (especially P_β) do depend on the method used. For example, P_β for Ile is 1.54 from eq a and c, but 1.60 from eq b and c. Such changes will affect the predictions based on the relative values of P_α and P_β. Furthermore, their comparison of P_α, and some corresponding values of P_α for residues near the N and C termini, with experimental values of s and σ, respectively, determined from studies of the helix–coil transition in polyamino acids is conceptually incorrect. σ and s are *statistical weights,* but P_α is the ratio of an *a priori probability* (i.e., $f_{j,k}$ in their notation) to the average value ($\langle f_k \rangle$) of all *a priori probabilities.* The statistical weights s and σ involve averaging the Boltzmann factor $e^{-E/RT}$ over the conformational space of a *single* residue. Chou and Fasman's P_α and P_β involve averages over all 20 residues instead of a single residue. The difference between a statistical weight (eq 55 of paper II[12]) and an a priori probability (eq 57 and 59 of paper II[12]) is discussed in section VIC of paper II.[12] It is important to keep this difference in mind even though Chou and Fasman's values of P_α happen to show apparent agreement with the values of s observed experimentally; this is due primarily to the fact that the quantities P_α for amino acids are distributed in a small range around unity, because P_α (like s) is a relative quantity [$P_{j,\alpha}$ being given by the ratio in eq c above (with $k = \alpha$), and s being defined as the stability of an α-helical state relative to that of a coil state in the helix–coil transition]. Furthermore, the comparison of s and σ for proline to experimental values of s and σ obtained from the helix form I $\rightleftarrows$ helix form II transition is also conceptually erroneous, as discussed in section IIID [see also ref 49–51 for the physical meaning of s, σ (and β_N and β_C) for the order $\leftrightarrow$ order conformational transition in poly(L-proline)]. Further discussion of the Chou and Fasman procedure will be presented in paper III.[13]

(61) O. B. Ptitsyn and A. V. Finkelshtein, *Biofizika*, **15**, 757 (1970).

(62) D. Kotelchuck, M. Dygert, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **63**, 615 (1969).

(63) D. E. Blagdon and M. Goodman, *Biopolymers*, **14**, 241 (1975).

(64) D. Kotelchuck and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **61**, 1163 (1968).

(65) R. T. Ingwall, H. A. Scheraga, N. Lotan, A. Berger, and E. Katchalski, *Biopolymers*, **6**, 331 (1968).

(66) H. Sugiyama and H. Noda, *Biopolymers*, **9**, 459 (1970).

(67) K. E. B. Platzer, V. S. Ananthanarayanan, R. H. Andreatta, and H. A. Scheraga, *Macromolecules*, **5**, 177 (1972).

(68) W. G. Miller and R. E. Nylund, *J. Am. Chem. Soc.*, **87**, 3542 (1965).

(69) R. L. Snipp, W. G. Miller, and R. E. Nylund, *J. Am. Chem. Soc.*, **87**, 3547 (1965).

(70) J. Hermans, Jr., *J. Phys. Chem.*, **70**, 510 (1966).

(71) A. Ciferri, D. Puett, L. Rajagh, and J. Hermans, Jr., *Biopolymers*, **6**, 1019 (1968).

(72) D. S. Olander and A. Holtzer, *J. Am. Chem. Soc.*, **90**, 4549 (1968).
(73) V. E. Bychkova, O. B. Ptitsyn, and T. V. Barskaya, *Biopolymers*, **10**, 2161 (1971).
(74) A. Warashina and A. Ikegami, *Biopolymers*, **11**, 529 (1972).
(75) J. Hermans, Jr., *J. Am. Chem. Soc.*, **88**, 2418 (1966).
(76) G. Conio and E. Patrone, *Biopolymers*, **8**, 57 (1969).
(77) G. Conio, E. Patrone, and S. Brighetti, *J. Biol. Chem.*, **245**, 3335 (1970).
(78) S. E. Ostroy, N. Lotan, R. T. Ingwall, and H. A. Scheraga, *Biopolymers*, **9**, 749 (1970).
(79) J. E. Alter, G. T. Taylor, and H. A. Scheraga, *Macromolecules*, **5**, 739 (1972).
(80) C. R. Snell and G. D. Fasman, *Biopolymers*, **11**, 1723 (1972).
(81) T. V. Barskaya and O. B. Ptitsyn, *Biopolymers*, **10**, 2181 (1971).
(82) H. E. Van Wart, G. T. Taylor, and H. A. Scheraga, *Macromolecules*, **6**, 266 (1973).
(83) J. E. Alter, R. H. Andreatta, G. T. Taylor, and H. A. Scheraga, *Macromolecules*, **6**, 564 (1973).
(84) M. Terbojevich, A. Cosani, E. Peggion, F. Quadrifoglio, and V. Crescenzi, *Macromolecules*, **5**, 622 (1972).
(85) L. J. Hughes, R. H. Andreatta, and H. A. Scheraga, *Macromolecules*, **5**, 187 (1972).
(86) V. S. Ananthanarayanan, R. H. Andreatta, D. Poland, and H. A. Scheraga, *Macromolecules*, **4**, 417 (1971).
(87) V. Ganser, J. Engel, D. Winklmair, and G. Krause, *Biopolymers*, **9**, 329 (1970).
(88) (a) M. Gō, N. Gō, and H. A. Scheraga, *J. Chem. Phys.*, **52**, 2060 (1970); (b) *ibid.*, **54**, 4489 (1971).
(89) P. Doty and J. T. Yang, *J. Am. Chem. Soc.*, **78**, 498 (1956).
(90) J. A. Schellman, *J. Phys. Chem.*, **62**, 1485 (1958).
(91) D. C. Poland and H. A. Scheraga, *Biopolymers*, **3**, 275 (1965).
(92) M. Bixon and S. Lifson, *Biopolymers*, **4**, 815 (1966).
(93) F. E. Karasz and J. M. O'Reilly, *Biopolymers*, **5**, 27 (1967).
(94) P. L. Dubin, *Biopolymers*, **12**, 685 (1973).
(95) G. Nemethy and H. A. Scheraga, *J. Phys. Chem.*, **66**, 1773 (1962).
(96) M. Bixon, H. A. Scheraga, and S. Lifson, *Biopolymers*, **1**, 419 (1963).
(97) D. C. Poland and H. A. Scheraga, *Biopolymers*, **3**, 283, 305, 315, 335 (1965).
(98) F. R. Maxfield and H. A. Scheraga, *Macromolecules*, **8**, 491 (1975).
(99) H. A. Scheraga, *Adv. Phys. Org. Chem.*, **6**, 103 (1968).
(100) G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.*, **23**, 283 (1968).
(101) P. J. Flory, "Statistical Mechanics of Chain Molecules", Interscience, New York, N.Y., 1969, Chapter VII.
(102) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **68**, 2293 (1971).

Statistical Mechanical Treatment of Protein Conformation. II. A Three-State Model for Specific-Sequence Copolymers of Amino Acids[1]

Seiji Tanaka[2] and Harold A. Scheraga*

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

ABSTRACT: A one-dimensional three-state Ising model [involving α-helical (α), extended (ϵ), and coil (or other) (c) states] for specific-sequence copolymers of amino acids has been formulated in order to treat the conformational states of proteins. This model involves four parameters ($w_{h,i}$, $v_{h,i}$, $v_{\epsilon,i}$, and $u_{c,i}$), and requires a 4×4 matrix for generating statistical weights. Some problems in applying this model to a specific-sequence copolymer of amino acids are discussed. A nearest-neighbor approximation for treating this three-state model is also formulated; it requires a 3×3 matrix, in which the same four parameters appear, but (as with the 4×4 matrix treatment) only three parameters (w_h^*, v_h^*, and v_ϵ^*) are required if relative statistical weights are used. The relationship between the present three-state model (3×3 matrix treatment) and models of the helix–coil transition is discussed. Then, the three-state model (3×3 matrix treatment) is incorporated into an earlier (Tanaka–Scheraga) model of the helix–coil transition, in which asymmetric nucleation of helical sequences is taken into account. A method for calculating molecular averages and conformational-sequence probabilities, $P(i|n|\{\rho\})$, i.e., the probability of finding a sequence of n residues in a specific conformational state $\{\rho\}$, starting at the ith position of the chain, is described. Two alternative methods for calculating $P(i|n|\{\rho\})$, that can be applied to a model involving any number of states, are proposed and presented; one is the direct matrix-multiplication method, and the other uses a first-order a priori probability and a conditional probability. In this paper, these calculations are performed with the nearest-neighbor model, and without the feature of asymmetric nucleation. Finally, it is indicated how the three-state model and the methods for computing $P(i|n|\{\rho\})$ can be applied to predict protein conformation.

In order to develop a prediction scheme to obtain an initial conformation of a protein, which can be refined by subsequent energy minimization,[3] we have formulated a statistical mechanical treatment of protein conformation. In paper I[4] of this series, we deduced the statistical weights for various conformations of the naturally occurring amino acids from x-ray data on proteins. In the present paper, we formulate a three-state model [involving α-helical (α), extended (ϵ), and coil (δther) (c) states], and incorporate it into our earlier model[5] of the helix–coil transition in which asymmetric nucleation is taken into account; also, we show how to compute the probabilities of occurrence of helical and extended conformations, respectively. In paper III,[6] we compute these probabilities for specific proteins, using the theory of the present paper and the statistical weights of paper I.[4]

The one-dimensional Ising model is used in this treatment. It is applicable as a first approximation[4] because short-range interactions dominate[7] (although not exclusively) in determining the native conformations of proteins. In fact, the Zimm–Bragg theory[8] of the helix–coil transition has already been applied to the prediction of α-helical regions[9,10] in native proteins. The formation of the α helix is treated in this paper as a cooperative process, because of the formation of hydrogen bonds, but long-range effects in the formation of the extended structure are neglected.[11]

The three-state model is presented in section I, and its application to specific-sequence copolymers of amino acids