(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

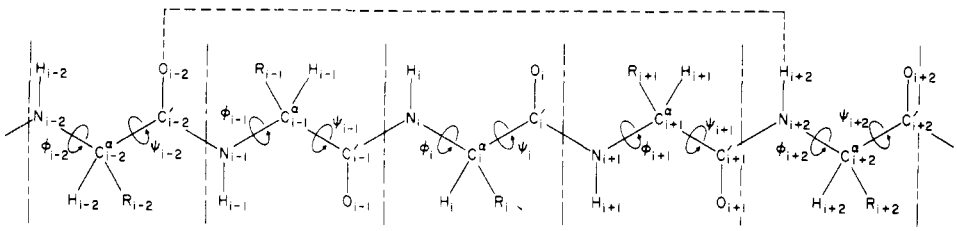

Figure 1. A portion of a polypeptide chain, showing the hydrogen bond that would exist between O_{i-2} and H_{i+2} when residues[12] $i-1$, i, and $i+1$ are in the α-helical conformation.

is discussed in section II. A nearest-neighbor approximation to the thre-state model is presented in section III, and the relation between the two-state and three-state models is discussed in section IV. In section V, the present three-state model is incorporated into a model[5] of the helix–coil transition, in which asymmetric nucleation of helical sequences is taken into account. The methods for computing molecular averages and conformational-sequence probabilities are given in section VI, and a concluding discussion is presented in section VII.

In paper I,[4] we deduced the statistical weights for the various amino acids from x-ray data, using an eight-state and a six-state model for the α-helical and extended conformations, respectively. We also discussed[4] the asymmetric properties of the various amino acids for nucleation of helical sequences. While all of these features can be used in a statistical mechanical treatment of conformational changes in proteins, they would require excessively large matrices (and large amounts of computer time) to evaluate the partition function and compute molecular averages. Therefore, for the present, we confine ourselves to a three-state model, in which the three-state statistical weights were deduced[4] from those computed[4] for the eight-state and six-state models, respectively, by the procedure of section IIA of paper I;[4] also, while we incorporate the feature of asymmetric nucleation of helical sequences in section V, and such a theory can be used (though requiring excessive computation), we compute the conformational-sequence probabilities in section VI (and use them for the prediction of protein conformation in paper III[6]) from the nearest-neighbor approximation to the three-state model, without the feature of asymmetric nucleation, again to save computer time.

I. Three-State Model

Consider a polyamino acid chain of N residues, a portion of which is depicted in Figure 1.[12] In this section, we will discuss a homopolymer of amino acids and, in section II, the present formulation will be applied to a model for a specific-sequence copolymer of amino acids. Because of the partial double bond character of the peptide bonds, C′–N, the amide groups are assumed to be fixed in the planar trans conformation ($\omega = 180°$).[13] Hence, the conformation of the backbone of the polyamino acid chain may be described by a set of dihedral angles for rotation, ϕ_i and ψ_i around the N_i–C_i^α and C_i^α–C_i' bonds,[13] respectively, where $i = 1$ to N. As seen in Figure 1, a hydrogen bond can be formed between the oxygen atom of the CO group of the $(i-2)$th residue[12] and the hydrogen atom of the NH group of the $(i+2)$th residue when residues $i-1$, i, and $i+1$ are in the α-helical conformation. The conformational energy of the chain, $E^{(N)}$, can be expressed as a function of a set of variable dihedral angles of the backbone, $\phi_1, \psi_1, \phi_2, \psi_2, \ldots, \phi_N, \psi_N$, and of the side chains, $\{\chi_i^{j_i}\}$, where i is the residue number ($i = 1$ to N) and j_i is the jth side-chain dihedral angle of the ith amino acid.

To calculate the energy $E^{(N)}$ of the chain, we assume that its conformations are determined by the energy of intra-residue interactions, $E_i^{(1)}[\phi_i, \psi_i, \{\chi_i^{j_i}\}]$, and by the energy of interactions between the residues involved in formation of the α-helical conformation.[5,8,14,15] For the extended (ϵ) and other (c) states, we require[11] only the interaction energy $E_i^{(1)}[\phi_i, \psi_i, \{\chi_i^{j_i}\}]$. For the helical conformation (except for isolated h and hh states), we require $E_i^{(1)}[\phi_i, \psi_i, \{\chi_i^{j_i}\}]$ and the energy of formation of the helical conformation. The latter (inter-residue) interaction energy is expressed as $E_i^{(3)}(\phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \phi_{i+1}, \psi_{i+1})$, and includes the energy of formation of the hydrogen bond between O_{i-2} and H_{i+2} and the nonbonded and electrostatic interactions between all backbone atoms (including the C^β atoms), which depend on these six dihedral angles (but not on $\{\chi_i^{j_i}\}$).[16] $E_i^{(3)}$ is assumed to be zero if the conformations of all three residues, $i-1$, i, and $i+1$, are not in the small range characteristic of the α-helix conformation.[14] The conformational energy of the chain is given by

$$E^{(N)}[\phi_1, \psi_1, \{\chi_1^{j_1}\}, \ldots, \phi_N, \psi_N, \{\chi_N^{j_N}\}] = \sum_{i=1}^{N} E_i^{(1)}[\phi_i, \psi_i, \{\chi_i^{j_i}\}] + \sum_{i=2}^{N-1} E_i^{(3)}[\phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \phi_{i+1}, \psi_{i+1}] \quad (1)$$

In the second term on the right-hand side of eq 1, the reason that the sum of $E_i^{(3)}$ is taken from 2 to $N-1$ (not including 1 and N) is that, while the first and last residues of the chain can be in helical states, they do not contribute a hydrogen-bonding interaction. The conformational partition function, Z, of the chain may then be computed from

$$Z = \int_{\phi_1=0}^{2\pi} \int_{\psi_1=0}^{2\pi} \int_{\chi_1^{j_1}=0}^{2\pi} \cdots \int_{\chi_N^{j_N}=0}^{2\pi} \exp(-E^{(N)}/RT) \times d\phi_1\, d\psi_1\, \{d\chi_1^{j_1}\} \ldots \{d\chi_N^{j_N}\} \quad (2)$$

where $E^{(N)}$ is given by eq 1.

Following Lifson and Roig,[14] the ϕ, ψ conformational space is divided into two regions: one is the small α-helical region (h region), and the other is the rest of the space, i.e., the coil region. In a three-state model (α, ϵ, and c), we divide the original coil region into the ϵ and c regions, as shown in Figure 2. The ϵ region corresponds to the broad one of low energy (shaded in Figure 2), found in computed conformational energy maps of single residues.[17–20] For the present, we do not specify the boundaries of the α, ϵ, and c regions because they may vary from one type of amino acid to another.

A. α-Helical Region. We perform the integration of eq 2 over the helical region of residue i, i.e., when ϕ_{i-1}, ψ_{i-1}, ϕ_{i+1}, and ψ_{i+1}, as well as ϕ_i and ψ_i, are in the helical region and obtain

$$w_{h,i} = \int \cdots \int_\alpha \exp[-\{E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\}) + E_i^{(3)}(\phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \phi_{i+1}, \psi_{i+1})\}/RT]\, d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (3)$$

for $i = 2$ to $N-1$ (but, see ref 21). However, when all dihe-

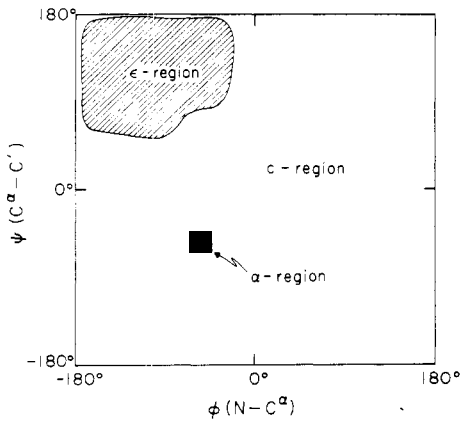

Figure 2. Schematic illustration of the regions of the right-handed α-helical (α or h), extended (ϵ), and other (c) states in the ϕ,ψ conformational space of an amino acid residue.

dral angles ϕ_{i-1}, ψ_{i-1}, ϕ_{i+1}, and ψ_{i+1} are not in the helical region, but ϕ_i, ψ_i *is* helical, the term $E_i^{(3)}$ vanishes as a first approximation, giving

$$v_{h,i} = \int \cdots \int_{\alpha} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (4)$$

for $i = 2$ to $N - 1$. At the ends of the chain, i.e., $i = 1$ or N, the integration over the α-helical region gives

$$v_{h,i} = \int \cdots \int_{\alpha} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (5)$$

B. ϵ Region. Performing the integration over the ϵ region (see Figure 2) we then have

$$v_{\epsilon,i} = \int \cdots \int_{\epsilon} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (6)$$

for $i = 2$ to $N - 1$, and

$$v_{\epsilon,i} = \int \cdots \int_{\epsilon} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (7)$$

for $i = 1$ or N, since the term $E_i^{(3)}$ vanishes when residues $i - 1$, i, and $i + 1$, are not in the α-helical region.

C. c Region. Performance of the integration over the c region, excluding the α and ϵ regions (see Figure 2), gives

$$u_{c,i} = \int \cdots \int_{c} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (8)$$

for $i = 2$ to $N - 1$, and

$$u_{c,i} = \int \cdots \int_{c} \exp[-E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\})/RT] \times d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (9)$$

for $i = 1$ and N. Thus, we have obtained all the statistical weights for the α, ϵ, and c states for the ith residue.

D. Matrix Treatment of Partition Function. By using eq 1 and 2, the partition function of the polyamino acid chain can be rewritten as

$$Z = \sum_{\eta_1=\alpha,\epsilon,c} \cdots \sum_{\eta_i=\alpha,\epsilon,c} \cdots \sum_{\eta_N=\alpha,\epsilon,c} \int_{\eta_1} \cdots \int_{\eta_N} \exp[-E_1^{(1)}/RT] \left\{ \prod_{i=2}^{N-1} \exp[-(E_i^{(1)} + E_i^{(3)})/RT] \right\} \times \exp[-E_N^{(1)}/RT]\, d\phi_1\, d\psi_1\, d\{\chi_1^{j_1}\} \cdots d\phi_N\, d\psi_N \{d\chi_N^{j_N}\} \quad (10)$$

Let us write the statistical weights of the conformational states η_i for residue i ($i = 1$ to N), in a general way, as $m_{i;\eta_{i-1}\eta_i\eta_{i+1}}$. The actual values of these statistical weights were given in eq 3, 4, 6, and 8 for $1 < i < N$, and in eq 5, 7, and 9 for $i = 1$ and N. Then, eq 10 may be rewritten as

$$Z = \sum_{\eta_1=\alpha,\epsilon,c} \cdots \sum_{\eta_i=\alpha,\epsilon,c} \cdots \sum_{\eta_N=\alpha,\epsilon,c} \times \left[m_{1;\eta_1} \left\{ \prod_{i=2}^{N-1} m_{i;\eta_{i-1}\eta_i\eta_{i+1}} \right\} m_{N;\eta_N} \right] \quad (11)$$

where $\Sigma_{\{\eta\}}$ designates the summation over all possible conformational states $\{\eta\}$ of the *chain*. In order to evaluate Z, we will employ the method of matrix multiplication. For this purpose, we will construct the matrix operator, $\mathbf{W}_i$, for the ith residue ($2 \le i \le N - 1$) by correlating the conformational states of the three residues, $i - 1$, i, and $i + 1$. By using the statistical weights for the three-state model given in eq 3, 4, 6, and 8, we then have eq 12,

$\mathbf{W}_i =$

$i-1$	$i+1$	c	α	c	ϵ	c	ϵ	α	ϵ	α
	i	c	c	α	c	ϵ	ϵ	ϵ	α	α
c	c	u_c	u_c	0	u_c	0	0	0	0	0
c	α	0	0	v_h	0	0	0	0	v_h	v_h
α	c	u_c	u_c	0	u_c	0	0	0	0	0
c	ϵ	0	0	0	0	v_ϵ	v_ϵ	v_ϵ	0	0
ϵ	c	u_c	u_c	0	u_c	0	0	0	0	0
ϵ	ϵ	0	0	0	0	v_ϵ	v_ϵ	v_ϵ	0	0
ϵ	α	0	0	v_h	0	0	0	0	v_h	v_h
α	ϵ	0	0	0	0	v_ϵ	v_ϵ	v_ϵ	0	0
α	α	0	0	v_h	0	0	0	0	v_h	w_h

$_i$ (12)

where the subscript i written on the bottom right-hand side of the matrix indicates that the statistical weights, i.e., the elements of the matrix, pertain to the ith residue; this notation serves instead of the symbols $u_{c,i}$, $v_{h,i}$, $w_{h,i}$, and $v_{\epsilon,i}$. This 9 × 9 matrix may be contracted as follows:

$\mathbf{W}_i =$

$i-1$	$i+1$	c$\cup\epsilon\cup\alpha$	c$\cup\epsilon\cup\alpha$	c$\cup\epsilon$	α
	i	c	ϵ	α	α
c	c$\cup\epsilon\cup\alpha$	u_c	v_ϵ	v_h	v_h
ϵ	c$\cup\epsilon\cup\alpha$	u_c	v_ϵ	v_h	v_h
α	c$\cup\epsilon$	u_c	v_ϵ	0	0
α	α	0	0	v_h	w_h

$_i$ (13)

where the symbol $\cup$ means that, for example, c $\cup$ ϵ $\cup$ α should be read as c *or* ϵ *or* α.

For the first residue at the N terminus of the chain, we define the row vector **t**, which consists of the statistical weights for the allowed conformational states of the N terminus given in eq 5, 7, and 9 (keeping in mind that the first residue cannot contribute to the hydrogen-bond energy), as

$$\mathbf{t} = [u_c\ v_\epsilon\ v_h\ v_h] \quad (14)$$

For the last residue of the chain (i.e., the C terminus), we define the column vector, **t***, as

$$\mathbf{t}^* = \begin{bmatrix} u_c + v_\epsilon + v_h \\ u_c + v_\epsilon + v_h \\ u_c + v_\epsilon \\ v_h \end{bmatrix} \quad (15)$$

The elements of eq 15 correspond to the states c $\cup$ ϵ $\cup$ α, c $\cup$ ϵ $\cup$ α, and c $\cup$ ϵ for residue $i + 1$ when $i = N$.

It is now possible to evaluate the partition function of eq 11, by using the set of eq 13–15 for $\mathbf{W}_i$, **t**, and **t***, i.e.,

$$Z = \mathbf{t} \left[\prod_{i=2}^{N-1} \mathbf{W}_i \right] \mathbf{t}^* \quad (16)$$

Since the parameters appearing in eq 14 and 15 do not differ from those in eq 13, the elements of eq 14 can be found in the first row of eq 13; similarly, the elements of eq 15 can be obtained from those of eq 13. Thus, eq 16 may be written as

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N^* \tag{17}$$

where

$$\mathbf{e}_1 = [1\ 0\ 0\ 0] \tag{18a}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \tag{18b}$$

because $\mathbf{t}$ of eq 14 and $\mathbf{t}^*$ of eq 15 are given by

$$\mathbf{t} = \mathbf{e}_1 \mathbf{W}_1 \tag{19a}$$

and

$$\mathbf{t}^* = \mathbf{W}_N \mathbf{e}_N^* \tag{19b}$$

respectively.

II. Application to a Specific-Sequence Copolymer of Amino Acids

In order to treat protein molecules, we shall first discuss some problems that arise when applying the present three-state model to a specific-sequence copolymer of amino acids.

First, in theories of the helix–coil transition in homopolymers, it is always assumed that the α-helical conformation is *regular*, in the sense that all residues in the helix have identical values of ϕ and ψ. We will make the same assumption about regularity for a specific-sequence copolymer as is made in treating homopolymers. This assumption can be removed in an appropriate procedure when the three-dimensional structure of a protein is computed.[22-24]

Second, we assume that the conformational energy of the chain can be divided into two contributions, $E_i^{(1)}$ and $E_i^{(3)}$ (see eq 1). Since $E_i^{(1)}$ is a function of only ϕ_i, ψ_i, and $\{\chi_i^{ji}\}$, which pertain to the ith residue, this term depends only on the species of amino acid in the ith position of the chain. However, $E_i^{(3)}$ is a function of the dihedral angles of *three* residues, at positions $i-1$, i, and $i+1$, and has a nonzero value only when the dihedral angles of all three residues are in the regular α-helical region (see Figure 2). Therefore, the statistical weights $u_{c,i}$, $v_{h,i}$, $v_{\epsilon,i}$ of the three-state model, as well as $v_{h,i}$ and $u_{c,i}$ of the two-state model (or v and u of the original Lifson–Roig model), for which $E_i^{(3)} = 0$, depend only on the chemical species of the amino acid in the ith position of the chain.

Third, we consider the statistical weight $w_{h,i}$, that depends on $E_i^{(3)}$. In the approximation treated in the text,[16] the stability of the α helix depends only on the interactions (nonbonded, electrostatic, hydrogen bonding, and torsional) between the atoms of the backbone (including the C^β atoms), neglecting the contribution from the side chains. Thus, since side chain–side chain and side chain–backbone interactions (except those involving the ith side chain) are neglected, and the helix is assumed to be regular, $w_{h,i}$ depends only on the species in the ith position through the term $E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{ji}\})$. While the term $E_i^{(3)}$ contributes to the stability of the helix, $E_i^{(3)}$ is independent of the type of amino acid residues in positions $i-1$ and $i+1$ because of the assumption of regularity and the neglect of interactions involving the side chains.

III. Nearest-Neighbor Three-State Model

Just as a nearest-neighbor Ising model (2 × 2 matrix treatment) of the helix–coil transition[8,25] is a good approximation to the original Lifson–Roig treatment (3 × 3 matrix), it is also convenient, for the actual computations, to have a smaller size matrix expression of a nearest-neighbor interaction model which can serve as a good approximation of eq 13. As a matter of fact, it was demonstrated earlier[10] that the nearest-neighbor model (2 × 2 matrix treatment) effectively reproduces results similar to those obtained by a treatment with larger matrices, although the correspondence of the statistical weights used in the nearest-neighbor model to the molecular structure is lost. In order to save computational time in applying this theory to proteins,[6] we describe here a nearest-neighbor interaction model which allows us to contract the matrix size and reduce the number of parameters of the three-state model introduced in section I.

In the nearest-neighbor Ising model treatment of the helix–coil transition,[25] the first residue of a helical sequence is assigned a statistical weight v^2/w. With this assignment, eq 13 of the three-state model may be written as

$$\mathbf{W}_i = \left[\begin{array}{c|ccc} & \multicolumn{3}{c}{i} \\ i-1 & \mathrm{c} & \epsilon & \alpha \\ \hline \mathrm{c} & u_c & v_\epsilon & v_h^2/w_h \\ \epsilon & u_c & v_\epsilon & v_h^2/w_h \\ \alpha & u_c & v_\epsilon & w_h \end{array}\right]_i \tag{20}$$

The partition function in the nearest-neighbor interaction scheme, corresponding to eq 17, is

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N^* \tag{21}$$

where

$$\mathbf{e}_1 = [1\ 0\ 0] \tag{22a}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \tag{22b}$$

Equation 22a indicates that the first residue of the chain can be preceded by an ϵ or c state but not by an α state with a hydrogen bond, as in the usual treatment of the helix–coil transition; the first residue of the chain has to be represented by the first row of eq 20 in the form of $\mathbf{e}_1\mathbf{W}_1$. Equation 22b indicates that the last residue of the chain can be preceded by the c, ϵ, or α states; hence, it is represented by the column vector $\mathbf{W}_N\mathbf{e}_N^*$ which contains the sum of the elements of the three row vectors in eq 20.

Since our main interest is to obtain the relative conformational properties of a particular conformational state, the statistical weights may be expressed relative to that of a standard state. Taking the c state as a standard state, we define the relative statistical weights, $w_{h,i}^*$, $v_{h,i}^*$, and $v_{\epsilon,i}^*$, of the α and ϵ states in terms of

$$w_{h,i}^* = w_{h,i}/u_{c,i} \tag{23}$$

$$v_{h,i}^* = v_{h,i}/u_{c,i} \tag{24}$$

and

$$v_{\epsilon,i}^* = v_{\epsilon,i}/u_{c,i} \tag{25}$$

respectively. Using eq 23–25, we may rewrite[26] eq 20 as

$$\mathbf{W}_i = \begin{bmatrix} 1 & v_\epsilon^* & (v_h^*)^2/w_h^* \\ 1 & v_\epsilon^* & (v_h^*)^2/w_h^* \\ 1 & v_\epsilon^* & w_h^* \end{bmatrix}_i \tag{26}$$

The partition function can be calculated by substituting eq 26 for eq 20 when eq 21 and 22 are used.

IV. Relationship between the Three-State and Two-State Models

As stated in section I, the conformational ϕ,ψ space is divided into the three regions, α, ϵ, and c. The coil region, de-

fined for the helix–coil transition, was divided into the ϵ and c regions. The statistical weights $u_{c,i}$ and $v_{\epsilon,i}$ have been obtained by integrating the Boltzmann factors over the conformational regions c and ϵ as seen in eq 6–9. Thus, the statistical weight $u^{(2)}$, or $u_c{}^{(2)}$, of the two-state model of the helix–coil transition corresponds to the sum of the statistical weights $u_c{}^{(3)}$ and $v_\epsilon{}^{(3)}$ of the three-state model; these relationships can be represented as follows:

	Two-state model	Three-state model	
or	$u_c{}^{(2)}$	$u_c{}^{(3)} + v_\epsilon{}^{(3)}$	(27)
	1	$1 + v_\epsilon{}^{(3)*}$	

where we have used the superscripts (2), (3), and * to designate the two-state model, the three-state model, and the relative statistical weight, respectively.

Then, we have the following relationships between the statistical weights of the two-state and three-state models:

$$w_h{}^{(2)*} = w_h{}^{(3)*}/(1 + v_\epsilon{}^{(3)*}) \tag{28}$$

and

$$v_h{}^{(2)*} = v_h{}^{(3)*}/(1 + v_\epsilon{}^{(3)*}) \tag{29}$$

V. Nearest-Neighbor Three-State Model with Asymmetric Nucleation of Helical Sequences

We recently[5] formulated a model of the helix–coil transition in polypeptides, in which account was taken of the different helix nucleation properties at each end of a regular helical sequence (asymmetric nucleation). Then, in paper I,[4] we evaluated the asymmetric nucleation parameters of the amino acids in proteins. Therefore, we now incorporate the asymmetric nucleation property[5] into the nearest-neighbor three-state model.

For this purpose, it is necessary to construct a matrix of statistical weights for the ith residue, in which the states of three consecutive residues, $i-1$, i, and $i+1$, are correlated in a manner similar to that of eq 33 of ref 5. However, it should be noted that the next-nearest-neighbor interaction between residues $i-1$ and $i+1$ is not included in the statistical weights, but only the nearest-neighbor interactions between residues $i-1$ and i and between residues i and $i+1$ (see section IA of ref 5); thus, the statistical weights used here are different from those that appear in section I of this paper (e.g., those in eq 12 and 13).

For the sake of simplicity, we make the following two assumptions: (i) the inter-residue interactions that involve an ϵ state are equivalent to those involving a c state; e.g., the interactions between c and ϵ, ϵ and ϵ, and α and ϵ are equivalent to those between c and c, c and c, and α and c, respectively; (ii) the free energy of an ϵ state is different from that of a c state because of the different intra-residue interactions. With these assumptions, we may use the statistical weights from the earlier theory of asymmetric nucleation of the helix–coil transition (see the eight statistical weights in the third column of Table I of ref 5), and augment them by an additional one, v_ϵ, which pertains to the ϵ state. Since the ϵ state is treated as equivalent to the c state, as far as the inter-residue interactions are concerned, nucleation parameters will appear for the boundaries between α and c states (and vice versa) and between α and ϵ states (and vice versa), but not between c and ϵ states (and vice versa). Furthermore, this same assumption allows us to use the same nucleation parameters for the boundaries between α and c states as between α and ϵ states; otherwise, we would have to introduce additional parameters for the boundaries between α and ϵ states.

The additional parameter, v_ϵ, introduced here for the three-state model with asymmetric nucleation of helical sequences, may be expressed as the relative statistical weight v_ϵ/u_c in the notation of the present paper. This relative statistical weight is used, together with those in columns 4 and 5 of Table I of ref 5, for homopolymers and specific-sequence copolymers, respectively, or in Table II of paper I,[4] and may be designated by the dummy statistical weight q_9, as for q_1 to q_8 of the last column of Table I of ref 5. Using these nine dummy statistical weights, we can construct the statistical weight matrix for the ith residue (including asymmetric nucleation) by correlating the states of three consecutive residues, $i-1$, i, and $i+1$, but not including the next-nearest-neighbor interaction, as mentioned above. The result is eq 30 may be contracted to eq 31. From the

$\mathbf{W}_i =$

$i-1$	i \ $i+1$	c	ϵ	c	α	c	ϵ	α	ϵ	α
c	c	q_8	q_8	0	q_7	0	0	0	0	0
c	ϵ	0	0	q_9	0	0	q_9	q_9	0	0
ϵ	c	q_8	q_8	0	q_7	0	0	0	0	0
c	α	0	0	0	0	q_6	0	0	q_6	q_4
α	c	q_5	q_5	0	q_3	0	0	0	0	0
ϵ	ϵ	0	0	q_9	0	0	q_9	q_9	0	0
ϵ	α	0	0	0	0	q_6	0	0	q_6	q_4
α	ϵ	0	0	q_9	0	0	q_9	q_9	0	0
α	α	0	0	0	0	q_2	0	0	q_2	q_1

$_i$ (30)

$\mathbf{W}_i =$

$i-1$	i \ $i+1$	c∪ϵ	c∪ϵ∪α	α	c∪ϵ	α
		c	ϵ	c	α	α
c	c∪ϵ	q_8	q_9	q_7	0	0
ϵ	c∪ϵ∪α	q_8	q_9	q_7	q_6	q_4
c	α	0	0	0	q_6	q_4
α	c∪ϵ	q_5	q_9	q_3	0	0
α	α	0	0	0	q_2	q_1

$_i$ (31)

statistical weights for the allowed conformational states of the first (i.e., amino terminal) residue, we obtain the vector $\mathbf{t}_1$

$$\mathbf{t}_1 = [q_8\ q_9\ q_7\ q_6\ q_4]_1 \tag{32a}$$

where each element corresponds to the amino-terminal residue in a state where residue $i-1$ is in state c or ϵ (viz., the combined first and third rows in eq 31). For the last (i.e., carboxyl terminal) residue, we have

$$t_N{}^* = \begin{bmatrix} q_8 + q_9 \\ q_8 + q_9 + q_6 \\ q_6 \\ q_5 + q_9 \\ q_2 \end{bmatrix}_N \tag{32b}$$

where each element corresponds to the carboxyl-terminal residue in a state where residue $i+1$ is in state c $\cup$ ϵ, c $\cup$ ϵ $\cup$ α, and c $\cup$ ϵ.

Using eq 31 and 32, the partition function can be written as

$$Z = \mathbf{t}_1 \left[\prod_{i=2}^{N-1} \mathbf{W}_i \right] \mathbf{t}_N{}^* \tag{33}$$

or

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N{}^* \tag{34}$$

where

$$\mathbf{e}_1 = [1\ 0\ 1\ 0\ 0] \tag{35a}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} \tag{35b}$$

together with

$$\mathbf{t}_1 = \mathbf{e}_1\mathbf{W}_1 \tag{36}$$

and

$$\mathbf{t}_N = \mathbf{W}_N\mathbf{e}_N^* \tag{37}$$

Finally, it should be noted that the numerical values of the set of nucleation parameters in the two-state model (α and c) are different from those in the three-state model (α, ϵ, and c), as described in sections IIB and IIC of paper I,[4] and as seen in Tables V–IX of paper I.[4] Thus, in eq 31–34, one has to use the statistical weights q_1 to q_8 for the three-state model, given in Table IX of paper I,[4] and the statistical weight v_ϵ^* (i.e., q_9), given in Table IV of paper I.[4]

VI. Molecular Averages and Conformational-Sequence Probability

It is now possible to compute molecular averages by using the partition function given in eq 16, 17, or 21. For example, the average fractions of helical residues, θ_h, and of extended conformation residues, θ_ϵ, can be computed by using the partition function of the chain. It is of particular interest to calculate the conformational-sequence probability in conjunction with our later studies[6] in which we will apply the present model to predict the conformations of proteins.

A. Conformational-Sequence Probability. Let us define the conformational-sequence probability as $P(i|n|\{\rho\})$ where $\{\rho\}$ is a general expression for a conformational sequence consisting of n residues; for the three-state model, $\{\rho\}$ is a set of combinations of α, ϵ, and c states, n is the number of residues in a sequence, i is the location of the first residue of a sequence in the conformational state $\{\rho\}$. For example, consider a protein molecule for which we want to compute the probability of finding, say, a sequence of six residues in the α-helical conformation, beginning at the ith residue in the chain; i.e., $n = 6$ and $\{\rho\} = \alpha\alpha\alpha\alpha\alpha\alpha$, starting at the ith residue, and the probability is $P(i|6|\alpha\alpha\alpha\alpha\alpha\alpha)$. In this subsection, we will consider the method for computing $P(i|n|\{\rho\})$, and, in subsection VIB, we will consider the computation of θ_h and θ_ϵ.

We may express $\{\rho\}$ as

$$\{\rho\} = \eta_i\eta_{i+1}\cdots\eta_{i+n-1} \tag{38}$$

where η_i (for $1 \leq i \leq N$) is h or c in the two-state model of the helix–coil transition, and is α, ϵ, or c in the three-state model. The probability $P(i|n|\{\rho\})$ of finding residues i to $i + n - 1$ in a specific conformational state $\{\rho\}$, given by expression 38, can be computed as the quotient of the sum of the products of the statistical weights for residues i to $i + n - 1$ in the conformational state $\{\rho\}$ (and for all other residues in all possible states) divided by the partition function Z. It is important to note here that the value of $P(i|n|\{\rho\})$, for n residues, depends on Z, i.e., on the possible states of *all* residues in the chain.

We use the nearest-neighbor interaction model of eq 20–26, and rewrite eq 11 as

$$Z = \sum_{\eta_1}\cdots\sum_{\eta_j}\cdots\sum_{\eta_N}\left[m_{1;\eta_1}\left\{\prod_{j=2}^{N-1} m_{j;\eta_{j-1}\eta_j}\right\}m_{N;\eta_N}\right] \tag{39}$$

where η_j = (h or c) or (α, ϵ, or c) for the two-state or three-state models, respectively. The probability can then be computed as

$$P(i|n|\{\rho\}) = Z^{-1}\sum_{\eta_1}\cdots\sum_{\eta_h\,(h\neq i\text{ to }i+n-1)}\cdots\sum_{\eta_N}\left[m_{1;\eta_1}\left\{\prod_{j=2}^{i-1} m_{j;\eta_{j-1}\eta_j}\right\}\times \left\{\prod_{k=i}^{i+n-1} m_{k;\eta_{k-1}\eta_k}\right\}_{\{\rho\}}\left\{\prod_{l=i+n}^{N-1} m_{l;\eta_{l-1}\eta_l}\right\}m_{N;\eta_N}\right] \tag{40}$$

where the index h denotes j and l, but not k, and $\Sigma_{\eta_1}\ldots\Sigma_{\eta_h(h\neq i\text{ to }i+n-1)}\ldots\Sigma_{\eta_N}$ denotes the sum of the statistical weights of the possible conformational states of the chain in which the sequence of n residues (starting at the ith) is fixed in the specific conformation $\{\rho\}$, given in expression 38, and the remaining residues of the chain are allowed to be in all possible states. The subscript $\{\rho\}$ in the middle product of matrices on the right-hand side of eq 40 indicates that the product within the brackets (from $k = i$ to $i + n - 1$) is computed for the specific conformational sequence $\{\rho\}$.

It should be kept in mind that, in performing the summation in eq 40, one must include all possible conformational states η_h for $h = 1$ to $i - 1$ [which includes η_{i-1} in the first term $m_{i;\eta_{i-1}\eta_i}$ of $\{\Pi_{k=i}^{i+n-1} m_{k;\eta_{k-1}\eta_k}\}_{\{\rho\}}$] and for $h = i + n$ to N [which includes η_{i+n-1} in the first term $m_{i+n;\eta_{i+n-1}\eta_{i+n}}$ of $\{\Pi_{l=i+n}^{N-1} m_{l;\eta_{l-1}\eta_l}\}$], while the summation over the states η_h (for $h = i$ to $i + n - 1$) is not carried out since this term is specified by the subscript $\{\rho\}$ on the product over k. The sum of the products of the statistical weights appearing in eq 40 may then be replaced by the products of the statistical weight matrices, such as eq 20, 26, or 30 (or more generally, by eq 55 below), as follows:

$$P(i|n|\{\rho\}) = Z^{-1}\mathbf{e}_1\left[\prod_{j=1}^{i-1}\mathbf{W}_j\right]\times \left[\frac{\partial\mathbf{W}_i}{\partial\ln(\mathbf{m}_{i;\eta_i})}\prod_{k=i+1}^{i+n-1}\frac{\partial\mathbf{W}_k}{\partial\ln(m_{k;\eta_{k-1}\eta_k})}\right]_{\{\rho\}}\times \left[\prod_{l=i+n}^{N}\mathbf{W}_l\right]\mathbf{e}_N^* \tag{41}$$

where the subscript $\{\rho\}$ has the same meaning as in eq 40, and $\mathbf{m}_{i;\eta_i}$ is the η_ith column vector of the statistical weight matrix (eq 20, 26, 30, or 55); hence, $\partial\mathbf{W}_i/\partial\ln(\mathbf{m}_{i;\eta_i})$ indicates that all possible conformational states, η_{i-1}, of the $(i - 1)$th residue are taken into account, i.e., are summed, as pointed out in the explanation of eq 40. Similarly, $m_{k;\eta_{k-1}\eta_k}$ is the $(\eta_{k-1}\eta_k)$ element of the statistical weight matrix (eq 20, 26, 30, or 55) corresponding to the conformational states η_{k-1} and η_k of the $(k - 1)$th and kth residues; hence, $\Pi\partial\mathbf{W}_k/\partial\ln(m_{k;\eta_{k-1}\eta_k})$ indicates a product of statistical weights for the specific conformational states $\{\rho\}$ of residue $i + 1$ to $i + n - 1$. A simpler form of eq 41 will be given later (in eq 48). Thus, with eq 41, it is possible to calculate the conformational-sequence probability $P(i|n|\{\rho\})$ for any specific conformational sequence $\{\rho\}$ of any length n, starting at any position i of the chain, for any model for the conformational states (i.e., α, ϵ, c, etc.).

The probability $P(i|n|\{\rho\})$ may be regarded as an nth-order a priori probability, i.e., the probability of finding a sequence of n residues in a certain conformational state $\{\rho\}$. We may also consider first- and second-order a priori probabilities, $F_{i;\eta_i}$ and $F_{i;\eta_{i-1}\eta_i}$, defined as follows:

$$F_{i;\eta_i} = P(i|1|\eta_i) \tag{42}$$

and

$$F_{i;\eta_{i-1}\eta_i} = P(i-1|2|\eta_{i-1}\eta_i) \tag{43}$$

In computing $F_{i;\eta_i}$ using eq 41, it should be noted that, inasmuch as $n = 1$, only the term $\partial \mathbf{W}_i/\partial \ln (\mathbf{m}_{i;\eta_i})$ remains in eq 41, and the product $\Pi_{k=i+1}^{i+n-1}$ vanishes. On the other hand, for probabilities of second and higher order, i.e., $n \geq 2$, including $F_{i;\eta_{i-1}\eta_i}$, the term $\partial \mathbf{W}_i/\partial \ln (\mathbf{m}_{i;\eta_i})$ is not always needed in order to compute numerical values (see eq 47 and 48 below). Nevertheless, eq 41 is of sufficiently general form to compute values of $P(i|n|\{\rho\})$ for any order of a priori probability.

Letting $n = 1$ in eq 41, eq 42 may be written as

$$F_{i;\eta_i} = Z^{-1}\mathbf{e}_1 \left[\prod_{j=1}^{i-1} \mathbf{W}_j\right] \left[\frac{\partial \mathbf{W}_i}{\partial \ln (\mathbf{m}_{i;\eta_i})}\right]_{|\rho|} \times \left[\prod_{l=i+1}^{N} \mathbf{W}_l\right] \mathbf{e}_N^* \quad (44)$$

where the definition and meaning of $\mathbf{m}_i$ is the same as in eq 41. Equation 44 is equivalent to eq 41, if we let $n = 1$.

In a similar fashion, letting $n = 2$ in eq 41, eq 43 can be written as

$$F_{i;\eta_{i-1}\eta_i} = Z^{-1}\mathbf{e}_1 \left[\prod_{j=1}^{i-2} \mathbf{W}_j\right] \left\{\left[\frac{\partial \mathbf{W}_{i-1}}{\partial \ln (\mathbf{m}_{i-1;\eta_{i-1}})}\right] \times \left[\frac{\partial \mathbf{W}_i}{\partial \ln (\mathbf{m}_{i;\eta_{i-1}\eta_i})}\right] \left[\frac{\partial \mathbf{W}_{i+1}}{\partial \ln (\mathbf{m}_{i+1;\eta_i})}\right]\right\}_{|\rho|} \times \left[\prod_{l=i+2}^{N} \mathbf{W}_l\right] \mathbf{e}_N^* \quad (45)$$

where $\mathbf{m}_{i-1;\eta_{i-1}}$ and $\mathbf{m}_{i+1;\eta_i}$ represent the vector of the η_{i-1}th column of the statistical weight matrix and the vector of the η_ith row of the statistical weight matrix, respectively (as defined in eq 20, 26, or 30 or, more generally, eq 55). However, in eq 45, $\partial \mathbf{W}_i/\partial \ln (m_{i;\eta_{i-1}\eta_i})$ gives a matrix in which all the elements except the (η_{i-1}, η_i)th are zero.

The relation

$$\frac{\partial \mathbf{W}_{i-1}}{\partial \ln (\mathbf{m}_{i-1;\eta_{i-1}})} \frac{\partial \mathbf{W}_i}{\partial \ln (m_{i;\eta_{i-1}\eta_i})} \frac{\partial \mathbf{W}_{i+1}}{\partial \ln (\mathbf{m}_{i+1;\eta_i})} = \mathbf{W}_{i-1} \frac{\partial \mathbf{W}_i}{\partial \ln (m_{i;\eta_{i-1}\eta_i})} \mathbf{W}_{i+1} \quad (46)$$

makes it possible to rewrite eq 45 as

$$F_{i;\eta_{i-1}\eta_i} = Z^{-1}\mathbf{e}_1 \left[\prod_{j=1}^{i-1} \mathbf{W}_j\right] \times \left[\frac{\partial \mathbf{W}_i}{\partial \ln (m_{i;\eta_{i-1}\eta_i})}\right]_{|\rho|} \left[\prod_{l=i+1}^{N} \mathbf{W}_l\right] \mathbf{e}_N^* \quad (47)$$

From a similar treatment as that used for eq 44 to 47, the equality of eq 46 converts eq 40 to the following:

$$P(i|n|\{\rho\}) = Z^{-1}\mathbf{e}_1 \left[\prod_{j=1}^{i} \mathbf{W}_j\right] \times \left[\prod_{k=i+1}^{i+n-1} \frac{\partial \mathbf{W}_k}{\partial \ln (m_{k;\eta_{k-1}\eta_k})}\right]_{|\rho|} \left[\prod_{l=i+n}^{N} \mathbf{W}_l\right] \mathbf{e}_N^* \quad (48)$$

Thus, for the computation of a priori probabilities of second order or higher (i.e., $n \geq 2$), either $\partial \mathbf{W}_i/\partial \ln (\mathbf{m}_{i;\eta_i})$ of eq 41 or $\mathbf{W}_i$ of eq 47 or 48 may be used. However, the term $\partial \mathbf{W}_i/\partial \ln (\mathbf{m}_{i;\eta_i})$, which cannot be replaced by $\mathbf{W}_i$, is necessary for the computation of $F_{i;\eta_i}$ or $P(i|1|\eta_i)$, as seen in eq 44.

B. Average Fractions of Helical and Extended Conformations. If we substitute α for η, e.g., in eq 42, we have the probability of finding the α-helical state at the ith residue, viz., $F_{i,h}$. For the two-state model of the helix–coil transition, the value of θ_h is then given by

$$\theta_h = \frac{1}{N} \sum_{i=1}^{N} F_{i;\alpha} \quad (49)$$

In a similar manner, we obtain, for the three-state model

$$\theta_\alpha = \frac{1}{N} \sum_{i=1}^{N} F_{i;\alpha} \quad (50)$$

and

$$\theta_\epsilon = \frac{1}{N} \sum_{i=1}^{N} F_{i;\epsilon} \quad (51)$$

In general, the average fraction θ_η of the state η is given by

$$\theta_\eta = \frac{1}{N} \sum_{i=1}^{N} F_{i;\eta_i} \quad (52)$$

C. A priori and Conditional Probabilities. We now introduce the conditional probability $P_{i;\eta_{i-1}\eta_i}$, i.e., the probability that, given that residue $i - 1$ is in the conformational state η_{i-1}, then residue i will be in conformational state η_i. We may then express the second-order probability as

$$F_{i;\eta_{i-1}\eta_i} = F_{i-1;\eta_{i-1}} P_{i;\eta_{i-1}\eta_i} \quad (53)$$

Using a first-order a priori probability $F_{i;\eta_i}$, and the conditional probability $P_{\eta_i\eta_{i+1}}$, the nth-order probability $P(i|n|\{\rho\})$ may be expressed as

$$P(i|n|\{\rho\}) = F_{i;\eta_i} \left[\prod_{k=i}^{i+n-1} P_{k+1;\eta_k\eta_{k+1}}\right]_{|\rho|} \quad (54)$$

Equation 54 illustrates the advantageous point that, if one once obtains the first-order a priori probability $F_{i;\eta_i}$ and the conditional probability $P_{\eta_i\eta_{i+1}}$, then the probability $P(i|n|\{\rho\})$ for any specific conformational sequence of n residues, starting at the ith position of the chain, can be calculated easily, without requiring the time-consuming repetitive matrix multiplication of eq 48. This will be especially advantageous when one wants to compare the probabilities of occurrence of various specific conformational sequences of various lengths, starting at various positions i in the chain; this problem will be encountered in the next paper[6] on the prediction of protein conformations. Thus, the problem of computing $P(i|n|\{\rho\})$ has been reduced to the computation of a first-order a priori probability and a conditional probability, as seen in eq 54.

A method for calculating first-order and second-order a priori probabilities and conditional probabilities for the helix–coil transition has been described by Tanaka and Nakajima for both infinite and finite chain lengths.[27] In the present anticipated application, we are interested mainly in the treatment of finite chain length because most protein molecules have low enough molecular weights that fall within the range of finite chain lengths. However, we will not repeat the Tanaka–Nakajima derivation here but, instead, will describe their procedure in general terms and apply it to the three-state model. For this purpose, we first rewrite the statistical weight matrix (eq 20 and 26 for the three-state model, or eq 3 and 4 of ref 27 for the two-state helix–coil transition model) in the general form

$$\mathbf{W}_i = [m_{i;\eta_{i-1}\eta_i}] \quad (55)$$

Then the conditional probability matrix $\mathbf{P}_i$ and the second-order a priori probability matrix $\mathbf{F}_i$ for the ith residue are given be eq 56 and 57, respectively; the elements in the matrices in eq 56 and 57 are those of the corresponding conformational states in $\mathbf{W}_i$ of eq 55.

$$\mathbf{P}_i = [P_{i;\eta_{i-1}\eta_i}] \quad (56)$$

and

$$\mathbf{F}_i = [F_{i;\eta_{i-1}\eta_i}] \quad (57)$$

The elements of eq 56, $P_{i;\eta_{i-1}\eta_i}$ are the conditional probabilities that the $(i - 1)$th and ith residues will be found in the

η_{i-1} and η_i conformational states. The elements of eq 57 are the second-order a priori probabilities for finding residues $i-1$ and i in the η_{i-1} and η_i conformational states. For example, explicit expressions for eq 56 and 57 for the helix–coil transition are given in eq 6 and 7 of ref 27, and those for the three-state model can be written as

$$\mathbf{P}_i = \begin{bmatrix} P_{i;cc} & P_{i;c\epsilon} & P_{i;c\alpha} \\ P_{i;\epsilon c} & P_{i;\epsilon\epsilon} & P_{i;\epsilon\alpha} \\ P_{i;\alpha c} & P_{i;\alpha\epsilon} & P_{i;\alpha\alpha} \end{bmatrix} \tag{58}$$

and

$$\mathbf{F}_i = \begin{bmatrix} F_{i;cc} & F_{i;c\epsilon} & F_{i;c\alpha} \\ F_{i;\epsilon c} & F_{i;\epsilon\epsilon} & F_{i;\epsilon\alpha} \\ F_{i;\alpha c} & F_{i;\alpha\epsilon} & F_{i;\alpha\alpha} \end{bmatrix} \tag{59}$$

The elements of the matrix of eq 58 are computed by the procedure described below. The elements of the matrix of eq 59 are computed from eq 47. The stochastic character of the probabilities leads to the following relations.

(i) The sum of the first-order a priori probabilities is unity, i.e.,

$$\sum_{\eta_{i-1}} F_{i-1;\eta_{i-1}} = 1 \tag{60}$$

(ii) The sum of the second-order a priori probabilities is unity, i.e.,

$$\sum_{\eta_{i-1}} \sum_{\eta_i} F_{i;\eta_{i-1}\eta_i} = 1 \tag{61}$$

(iii) Each of the sums of conditional probabilities that a state will be followed by all possible states is unity, i.e.,

$$\sum_{\eta_i} P_{i;\eta_{i-1}\eta_i} = 1 \tag{62}$$

which yields η_{i-1} linear equations related to the conditional probabilities. Equation 62 indicates that the sums of the elements of each row vector of the matrix of eq 56 (as a general expression), or of eq 58 for the three-state model, are unity; hence, one obtains as many linear equations of the form of eq 62 as the number of row vectors in the matrix of eq 56 or 58. Explicit examples of eq 60 to 62 can be seen in eq 9 of ref 27 for the case of the helix–coil transition, in which the number of equations of the type of eq 62 is 2.

We may rewrite eq 53 as

$$P_{i;\eta_{i-1}\eta_i} = \frac{F_{i;\eta_{i-1}\eta_i}}{F_{i-1;\eta_{i-1}}} \tag{63}$$

which implies that the conditional probabilities $P_{i;\eta_{i-1}\eta_i}$ can be calculated by using the first- and second-order a priori probabilities $F_{i-1;\eta_{i-1}}$ and $F_{i;\eta_{i-1}\eta_i}$, respectively. On the other hand, the first-order a priori probability $F_{i-1;\eta_{i-1}}$ can be calculated easily by using the value of $P(i-1|1|\eta_{i-1})$ of eq 42, which requires eq 41. In a similar fashion, the second-order a priori probability $F_{i;\eta_{i-1}\eta_i}$ can be computed using eq 47. However, in order to compute all of the elements of the conditional probability matrix of eq 56, it is unnecessary to compute all of the values of $F_{i;\eta_{i-1}}$ *and* $F_{i;\eta_{i-1}\eta_i}$ because of the relations of eq 60–62. In order to obtain all of the conditional probabilities $P_{i;\eta_{i-1}\eta_i}$ (of which there are $\eta_{i-1}\eta_i$), it is necessary to calculate the values of $\eta_{i-1}\eta_i - \eta_{i-1}$ conditional probabilities out of a total of $\eta_{i-1}\eta_i$ values, because of the η_{i-1} independent linear equations of eq 62. To obtain the $\eta_{i-1}(\eta_i - 1)$ values of $P_{i;\eta_{i-1}\eta_i}$, one has to calculate only the $\eta_{i-1}(\eta_i - 1) - 2$ values out of the $\eta_{i-1} + \eta_{i-1}\eta_i$ values of $F_{i-1;\eta_{i-1}}$ and $F_{i;\eta_{i-1}\eta_i}$ because of the two equations, eq 60 and 61, related to $F_{i-1;\eta_{i-1}}$ and $F_{i;\eta_{i-1}\eta_i}$, respectively. Thus, it is necessary to compute the $\eta_{i-1}(\eta_i - 1) - 2$ values of $F_{i-1;\eta_{i-1}}$ and $F_{i;\eta_{i-1}\eta_i}$ based on eq 48, 42, and 43.

In a similar manner to that used to obtain eq 49–51, the average fraction, $\theta_{\{\rho\}}^{(n)}$, of the specific conformation $\{\rho\}$ for a sequence of n residues can be calculated by

$$\theta_{\{\rho\}}^{(n)} = \frac{1}{N-n+1} \sum_{i=1}^{N-n+1} P(i|n|\{\rho\}) \tag{64}$$

where the values of $P(i|n|\{\rho\})$ are given by eq 54, and $N-n+1$ is the maximum number of sequences of n residues in a chain of N residues.

Finally, it is of particular importance to point out that the conformational-sequence probability $P(i|n|\{\rho\})$ of a sequence of n residues, or the nth order a priori probability [in other words, the probability of finding a sequence of n residues from the ith to the $(i+n-1)$th residues in a certain conformational state $\{\rho\}$ $(=\eta_i\eta_{i+1}\cdots\eta_{i+n-1})$], is never equal to the product of the first-order a priori probabilities $F_{i;\eta_i}$ for i to $(i+n-1)$, each of which is the probability of finding the ith residue in a certain conformational state η_i. The above statement is expressed as

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

or

$$\neq \left[\prod_{k=i}^{i+n-1} P(i|1|\eta_i)\right]_{\{\rho\}} \tag{65}$$

This inequality arises because of the cooperativity of helix formation in the chain molecule; if, instead, there were no cooperativity (model of independent units), then the inequality would become an equality. The inequality of eq 65 can be understood easily because the left-hand side of eq 65 should be computed by eq 40 or 54, whereas the right-hand side of eq 65 is given, by using eq 40 and 42,

$$\prod_{k=i}^{i+n-1} F_{i;\eta_i} = \prod_{k=i}^{i+n-1} \times \left[Z^{-1} \sum_{\eta_1} \cdots \sum_{\eta_h (h \neq i)} \cdots \sum_{\eta_N} \left[m_{i;\eta_1} \prod_{j=2}^{i-1} m_{j;\eta_{j-1}\eta_j}\right] \times m_{i;\eta_{i-1}\eta_i} \left[\prod_{l=i+1}^{N-1} m_{l;\eta_{l-1}\eta_l}\right] m_{N;\eta_N}\right] \tag{66}$$

Furthermore, it also should be pointed out that, although the values of $P(i|n|\{\rho\})$ are calculated by using such quantities as $F_{i;\eta_i}$ and $P_{k+1;\eta_k\eta_{k+1}}$ which are assigned to the individual amino acid residues when the present formulation is applied to proteins, the quantity $P(i|n|\{\rho\})$ is not to be thought of as reflecting the properties of only those amino acids located between the ith and $(i+n-1)$th residues; instead, it has to be thought of as reflecting the conformational properties of the overall protein molecule, because the values of $F_{i;\eta_i}$ and $P_{k+1;\eta_k\eta_{k+1}}$ are obtained as molecular averages (from Z), and are not obtained as a product of the probabilities of independent conformational occurrences assigned to the i to $(i+n-1)$th amino acid residues of the protein.

It is expected that the present three-state model will give us useful information about the α-helical (α), extended (ϵ) and other (c) conformational states when it is applied to the prediction of protein conformations, which will be discussed in our next paper.[6] In other words, within the framework of the nearest-neighbor Ising model, the partition function can be used to describe the equilibrium state of the protein, as an approximation. The treatment presented here can be extended to include more states (e.g., chain-reversal, bridge-region, left-handed helical, etc., conformations). The description can then be improved[24] and the conformation of the protein altered (computationally) by introducing the long-range interactions that are not included in the nearest-neighbor Ising model.

VII. Concluding Discussion

In this paper, we have presented a three-state model for conformational transitions in homopolymers and specific-sequence copolymers of amino acids in order to compute the conformational states of proteins. The three-state model was discussed in comparison with the two-state model of the helix–coil transition of polyamino acids. In the Tanaka–Scheraga model[5] of the conformational transition in a polymer chain, the asymmetric nucleation properties of a certain conformational state were taken into account; indeed the asymmetric properties of the helical conformation of the amino acids in proteins have been demonstrated in paper I.[4] The present three-state model of helical (α), extended (ϵ), and other (c) states has been incorporated into the Tanaka–Scheraga model[5] for asymmetric nucleation of the helical state from the ϵ and c states.

The present three-state model and the computational method for obtaining molecular averages, i.e., conformational-sequence probabilities, will be applied (without the feature of asymmetric nucleation) to the empirical prediction of protein conformations in the accompanying paper.[6] However, it is also true that the present three-state model is too simple to describe protein conformations completely. Therefore, we are now developing multi-state models (with more than three states) to provide a more detailed treatment of protein molecules, and will apply them to the prediction of the conformation of proteins whose primary structures are known.

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) For recent reviews, see (a) H. A. Scheraga, *Chem. Rev.*, **71**, 195 (1971), and (b) "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) S. Tanaka and H. A. Scheraga, *Macromolecules*, preceding paper in this issue.
(5) S. Tanaka and H. A. Scheraga, *Macromolecules*, **8**, 494 (1975).
(6) S. Tanaka and H. A. Scheraga, *Macromolecules*, following paper in this issue.
(7) H. A. Scheraga, *Pure Appl. Chem.*, **36**, 1 (1973).
(8) B. H. Zimm and J. K. Bragg, *J. Chem. Phys.*, **31**, 526 (1959).
(9) P. N. Lewis, N. Gō, M. Gō, D. Kotelchuck, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **65**, 810 (1970).
(10) N. Gō, P. N. Lewis, M. Gō, and H. A. Scheraga, *Macromolecules*, **4**, 692 (1971).
(11) As described in section IIIE of paper I,[4] the extended conformation (ϵ) is stabilized primarily by short-range interactions, and the hydrogen-bonding interactions between parallel or antiparallel extended chains provide only additional stabilization of the ϵ conformation.
(12) The IUPAC-IUB definition of a residue[13] is used here, i.e., a residue extends from the NH to the CO group (see Figure 1).
(13) IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, **9**, 3471 (1970).
(14) S. Lifson and A. Roig, *J. Chem. Phys.*, **34**, 1963 (1961).
(15) See D. Poland and H. A. Scheraga, "Theory of Helix-Coil Transitions in Biopolymers", Academic Press, New York, N.Y., 1970, for a summary of theories of the helix–coil transition.
(16) One may consider three levels of approximation for calculating the inter-residue interaction energy: (i) $E_i^{(3)}(\phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \phi_{i+1}, \psi_{i+1})$, which is described in the text; (ii) $E_i^{(3)}[\phi_{i-1}, \psi_{i-1}, \{\chi_{i-1}^{j_{i-1}}\}, \phi_i, \psi_i, \{\chi_i^{j_i}\}, \phi_{i+1}, \psi_{i+1}, \{\chi_{i+1}^{j_{i+1}}\}]$ includes possible side chain–side chain and side chain–backbone interactions, in addition to those included in case (i); (iii) $E_i^{(3)}[\phi_{i-1}, \psi_{i-1}, \phi_i, \psi_i, \{\chi_i^{j_i}\}, \phi_{i+1}, \psi_{i+1}]$ includes all interactions of case (ii) except side chain–side chain interactions. The general treatment of section I is valid for all three cases (i)–(iii). However, for case (ii), the statistical weight, $w_{h,i}$, for the ith amino acid residue would depend on the species of amino acids of its neighbors.
(17) H. A. Scheraga, *Adv. Phys. Org. Chem.*, **6**, 103 (1968).
(18) G. N. Ramachandran and V. Sasisekharan, *Adv. Protein Chem.*, **23**, 283 (1968).
(19) D. A. Brant and P. J. Flory, *J. Am. Chem. Soc.*, **87**, 2791 (1965).
(20) D. A. Brant, W. G. Miller, and P. J. Flory, *J. Mol. Biol.*, **23**, 47 (1967).
(21) From a mathematical point of view, eq 3 (and the corresponding equations defining w and v in the Lifson–Roig theory[14]) has no meaning because the integration is carried out only over ϕ_i, ψ_i, and $\chi_i^{j_i}$, while the term $E_i^{(3)}$ involves ϕ_{i-1}, ψ_{i-1}, ϕ_{i+1}, and ψ_{i+1}. Since the entropy of a helical sequence cannot be separated into contributions from each residue, the statistical weight of a helical sequence (in principle) cannot be expressed as a product of the statistical weights w_h and v_h. Nevertheless, as in the Lifson–Roig theory,[14] we assume that the entropy of a helical sequence *can* be separated into contributions from each residue, in essence by neglecting the entropy of hydrogen-bond formation. In such a case, eq 3 should be written as

$$w_{h,i} = \int \cdots \int_{\alpha} \exp[-\{E_i^{(1)}(\phi_i, \psi_i, \{\chi_i^{j_i}\}) + E_i^{(3)}(\phi_{i-1}^0, \psi_{i-1}^0, \phi_i, \psi_i, \phi_{i+1}^0, \psi_{i+1}^0)\}/RT]\, d\phi_i\, d\psi_i\, \{d\chi_i^{j_i}\} \quad (3')$$

where ϕ_{i-1}^0, ψ_{i-1}^0, ϕ_{i+1}^0, and ψ_{i+1}^0 indicate that these values are fixed at those of the helical state. The entropy terms that are responsible for the variation of (ϕ_{i-1}, ψ_{i-1}) or (ϕ_{i+1}, ψ_{i+1}), when we separate the entropy into contributions from each residue, are included in $w_{h,i-1}$ or $w_{h,i+1}$ (or $v_{h,i-1}$ or $v_{h,i+1}$ when residues $i-1$ or $i+1$ are at the end of a helical sequence). These points are discussed in more detail in ref 40 of paper I.
(22) Alternatively, one might have treated the helix by minimizing its energy with respect to *all* dihedral angles, without imposing the restriction of regularity. However, this alternative procedure is not used because (1) we do not have precise enough information to judge how irregular the helices in proteins are, and (2) once such "regular" helices are located in a protein by an empirical predictive algorithm, the restriction of regularity is subsequently removed anyway when the energy of the whole protein is minimized.[23]
(23) A. W. Burgess and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **72**, 1221 (1975).
(24) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.*, **72**, 3802 (1975).
(25) D. Poland and H. A. Scheraga, *J. Chem. Phys.*, **43**, 2071 (1965).
(26) The (1,3) and (2,3) elements of eq 26 can be obtained as follows: these elements should be $v_{h,i}^2/w_{h,i}u_{c,i}$ which is $(v_{h,i}/u_{c,i})^2/(w_{h,i}/u_{c,i})$, or $(v_{h,i}^*)^2/w_{h,i}^*$.
(27) S. Tanaka and A. Nakajima, *Macromolecules*, **5**, 714 (1972).