rate of 10 K/min, Cowie[13] found values of T_g close to 266 K for atactic head-to-tail polypropylene fractions of molecular weights in the range 2×10^4–6×10^4. At the same heating rate we measured a T_g close to 249 K for hydrogenated 1,4-PDMB indicating a slightly higher chain mobility for this polymer compared to head-to-tail polypropylene.

Acknowledgments. The authors thank the Canadian 220-MHz NMR Centre for making the proton NMR measurements and Mr. R. Mayer for making the carbon NMR spectra. This work was supported by the National Research Council of Canada and the Quebec Ministry of Education.

References and Notes

(1) A. Laramée, P. Goursot, and J. Prud'homme, *Makromol. Chem.*, **176**, 3079 (1975).
(2) J. Prud'homme, J. E. L. Roovers, and S. Bywater, *Eur. Polym. J.*, **8**, 901 (1972).
(3) D. Blondin, J. Regis, and J. Prud'homme, *Macromolecules*, **7**, 187 (1974).
(4) G. Gianotti, G. Dall'Asta, A. Valvassori, and V. Zamboni, *Makromol. Chem.*, **149**, 117 (1971).
(5) L. A. Mango and R. W. Lenz, *Makromol. Chem.*, **163**, 13 (1972).
(6) G. Natta, G. Dall'Asta, G. Mazzanti, I. Pasquon, A. Valvassori, and A. Zambelli, *J. Am. Chem. Soc.*, **83**, 3343 (1961).
(7) G. Natta, G. Allegra, I. W. Bassi, P. Corradini, and P. Ganes, *Makromol. Chem.*, **58**, 242 (1962).
(8) J. C. Falk, *J. Polym. Sci., Part A-1*, **9**, 2617 (1971).
(9) S. J. Lapporte, *Ann. N.Y. Acad. Sci.*, **158**, 510 (1969).
(10) C. J. Carman, A. R. Tarpley, and J. H. Goldstein, *Macromolecules*, **6**, 719 (1973).
(11) Y. Tanaka and K. Hatada, *J. Polym. Sci., Part A-1*, **11**, 2057 (1973).
(12) A. Zambelli, G. Gatti, S. Sacchi, W. O. Crain, and J. D. Roberts, *Macromolecules*, **4**, 475 (1971).
(13) J. M. G. Cowie, *Eur. Polym. J.*, **9**, 1041 (1973).

Statistical Mechanical Treatment of Protein Conformation. 4. A Four-State Model for Specific-Sequence Copolymers of Amino Acids[1]

Seiji Tanaka[2a] and Harold A. Scheraga*[2b]

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received February 3, 1976

ABSTRACT: One-dimensional short-range interaction models for specific-sequence copolymers of amino acids are being developed in this series of papers. In this paper, our earlier three-state model [involving helical (h), extended (ϵ), and coil (or other) (c) states] is extended to a four-state model by preserving the h and ϵ states, introducing the chain-reversal state (R and S), and redefining the c state. This model involves six parameters (w_h, v_h, v_ϵ, v_R, v_S, and u_c) and requires a 6×6 statistical weight matrix. A nearest-neighbor approximation of the four-state model is also formulated; it requires a 5×5 matrix, involving the same six parameters. By expressing the statistical weights relative to that of the ϵ state, only five parameters (w_h^*, v_h^*, v_R^*, v_S^*, and u_c^*) are required in both the 6×6 and 5×5 matrices. The statistical weights for the four-state model are evaluated from the atomic coordinates of the x-ray structures of 26 native proteins. These statistical weights, and the four-state model, are used to develop a procedure to predict the backbone conformations of proteins. Since the prediction of helical and extended conformations is carried out by the procedure described in papers 1–3 of this series, we focus particular attention on chain-reversal conformations in this paper. The conformational-sequence probabilities of finding a residue in h, ϵ, R, S, or c states, and of finding two consecutive residues in a chain-reversal conformation, defined as relative values with respect to their average values over the whole molecule, are calculated for 23 proteins. By comparing these conformational-sequence probabilities to experimental X-ray observations, it was found that, in addition to the prediction of helical and extended conformations (reported in paper 3), 219 chain-reversal regions out of 372 observed by x-ray diffraction studies of 23 proteins were predicted correctly. These results suggest that the assumption of the dominance of short-range interactions in determining chain-reversal (as well as helical or extended) conformations in proteins, on which the predictive scheme is based, is a reasonable one. Finally, in the Appendix, the property of asymmetric nucleation of helical sequences is introduced into the (nearest-neighbor) four-state model.

In this series of papers,[3–5] we have developed a scheme to predict protein conformations in terms of a one-dimensional model based on short-range interactions. More specifically, we described a method[3] for evaluating the statistical weights of conformational states, based on conformational information from x-ray crystal structures of native proteins, and formulated a three-state model[4] for polypeptide chains, which included α-helical (h), extended (ϵ), and coil (or other) (c) states. This model was used to predict the occurrence of h, ϵ, and c states in proteins.[5] These will be referred to here as papers 1,[3] 2,[4] and 3,[5] with equations designated as 1-1, 2-1, etc. In the present paper, we extend the method to a four-state model, consisting of α-helical (h), extended (ϵ), chain-reversal[6] (R and S), and coil (or other) (c) states. In this extension, we preserve the h and ϵ states of the three-state model,[3–5] and subdivide the old[3–5] c state into chain reversals (R and S states, which previously appeared in the c state) and a new (more restricted) c state.

In paper 3,[5] it was found that a close correlation exists between the regions of high probability of occurrence of h and ϵ states (calculated with the three-state model) and the helical and extended regions observed experimentally. This attests to the approximate validity of the one-dimensional short-range interaction model for helical and extended conformations in proteins, despite its omission of long-range interactions. Since it was argued previously[6b,7] that short-range interactions also play a dominant role in determining chain-reversal[6] conformations, it is of interest to include chain-reversal states in the short-range interaction model. Such chain-reversal conformations are included here in an extension of our earlier three-state model.[4]

In section I of this paper, a theoretical formulation of the four-state model is presented. A nearest-neighbor approximation of the four-state model of section I is discussed in section II. [The property of asymmetric nucleation of helical sequences is incorporated into the four-state model in the Appendix.] In section III, the x-ray data (atomic coordinates) of native proteins are analyzed and, in section IV, these are used to compute the statistical weights of the four-state model. In section V, we consider two procedures for prediction of

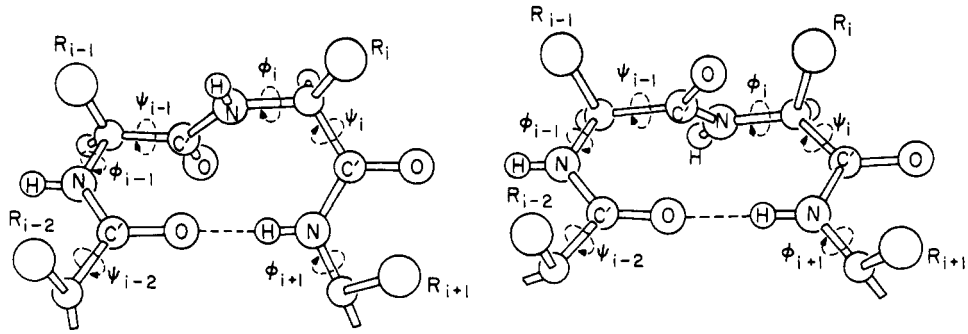

Figure 1. The two typical chain-reversal conformations found in native proteins, as proposed by Venkatachalam.[6a] Conformation (a) is called a type I and (b) a type II chain reversal.

chain reversals in the backbones of proteins. The first (using rule I) predicts only chain-reversal conformations; the second (using rule II) preserves the predictions of helical and extended conformations of paper **3**, and augments these with predictions of chain-reversal and (newly defined) c conformations. The results are presented and discussed in section VI.

I. Formulation of Four-State Model

We consider first the formulation of the four-state model, consisting of helical, extended, chain-reversal, and other conformations. The statistical weights w_h, v_h, and v_ϵ (and not the *relative* statistical weights w_h^*, v_h^*, and v_ϵ^*, because we will choose the reference state differently here), defined in section I of paper 2,[4] can be used without alteration in the four-state model. However, the c state in the four-state model differs from that in the three-state model since we now remove the chain-reversal conformation (*only*) from the c state of the earlier three-state model. Thus, the statistical weight u_c for the c state in the present model has a correspondingly different meaning from that in the three-state model.

Two typical chain-reversal conformations are depicted in Figure 1. Although the precise definition of such a conformation differs from author to author,[6] all authors describe the chain reversal in terms of the conformations of two consecutive amino acid residues in the protein. For example, as seen in Figure 1, a chain-reversal conformation may be defined by the dihedral angles[8] ϕ_{i-1}, ψ_{i-1}, ϕ_i, and ψ_i of the two consecutive residues, $i-1$ and i, assuming that the peptide bonds are fixed in the planar trans conformation. We will use the definition of a chain reversal given by Lewis et al.,[6b] viz., a conformation in which the distance[8] $R_{i-2,i+1}$ between C^α_{i-2} and C^α_{i+1} is less than 7 Å; as seen in Figure 1, this distance is a function of ϕ_{i-1}, ψ_{i-1}, ϕ_i, and ψ_i. If residues $i-1$ and i satisfy this condition, *and also* both $i-1$ and i are *not* involved in helical or extended sequences, they are assigned as a chain reversal. Thus, the chain reversal is defined by the conformational states of two consecutive residues. We will designate the conformational states of residues $i-1$ and i of a chain reversal as R and S, respectively. The conformational state S is allowed for an ith residue only if residue $i-1$ is in an R conformational state; likewise, an R state can be followed only by an S state. We then assign statistical weights v_R and v_S to the R and S states of residues $i-1$ and i, respectively.

Parenthetically, it should be noted that we could have chosen the statistical weight of a chain reversal to be dependent on *both* of the amino acid species of residues $i-1$ and i. However, because of the lack of enough x-ray data on native proteins, it is impossible at the present time to obtain a set of 400 statistical weights for all possible pairs of 20 amino acids. Therefore, in this paper, we divide the statistical weight of a chain reversal into two parts, one for residue $i-1$ and one for residue i. However, this is not an independent-residue treatment because account is taken of a correlation, to the extent that an S state is allowed to follow only an R state, and nothing but an S state is allowed to follow an R state, if the conformation is to be regarded as a chain reversal.

Using the statistical weights (and the correlation of R and S states) introduced above, and the correlation of the states of three residues in order to assign w_h, we can construct the statistical weight matrix as an extension of eq 2-12, viz.

$$U_i = \begin{array}{c|cccccccccccccccc|} & i+1 & c & h & \epsilon & R & c & c & c & h & \epsilon & R & h & h & \epsilon & R & \epsilon & R & S \\ i-1 & i & c & c & c & c & h & \epsilon & S & h & h & h & \epsilon & S & \epsilon & \epsilon & S & S & R \\ \hline c & c & u_c & u_c & u_c & u_c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ c & h & 0 & 0 & 0 & 0 & v_h & 0 & 0 & v_h & v_h & v_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ c & \epsilon & 0 & 0 & 0 & 0 & 0 & v_\epsilon & 0 & 0 & 0 & 0 & v_\epsilon & 0 & v_\epsilon & v_\epsilon & 0 & 0 & 0 \\ c & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_R \\ h & c & u_c & u_c & u_c & u_c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \epsilon & c & u_c & u_c & u_c & u_c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S & c & u_c & u_c & u_c & u_c & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ h & h & 0 & 0 & 0 & 0 & v_h & 0 & 0 & w_h & v_h & v_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ h & \epsilon & 0 & 0 & 0 & 0 & 0 & v_\epsilon & 0 & 0 & 0 & 0 & v_\epsilon & 0 & v_\epsilon & v_\epsilon & 0 & 0 & 0 \\ h & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_R \\ \epsilon & h & 0 & 0 & 0 & 0 & v_h & 0 & 0 & v_h & v_h & v_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ S & h & 0 & 0 & 0 & 0 & v_h & 0 & 0 & v_h & v_h & v_h & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \epsilon & \epsilon & 0 & 0 & 0 & 0 & 0 & v_\epsilon & 0 & 0 & 0 & 0 & v_\epsilon & 0 & v_\epsilon & v_\epsilon & 0 & 0 & 0 \\ \epsilon & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_R \\ S & \epsilon & 0 & 0 & 0 & 0 & 0 & v_\epsilon & 0 & 0 & 0 & 0 & v_\epsilon & 0 & v_\epsilon & v_\epsilon & 0 & 0 & 0 \\ S & R & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_R \\ R & S & 0 & 0 & 0 & 0 & 0 & 0 & v_S & 0 & 0 & 0 & 0 & v_S & 0 & 0 & v_S & v_S & 0 \end{array}_i \quad (1)$$

The statistical weight matrix of eq 1 may be contracted as in eq 2 where the symbol U means that, for example, c U ϵ U R

$$W_i = \begin{array}{c|cc|ccccc|} i-1 & i+1 \\ & i & \text{cUhU}\epsilon\text{UR} & \text{cUhU}\epsilon\text{UR} & \text{cU}\epsilon\text{UR} & \text{h} & \text{cUhU}\epsilon\text{UR} & \text{S} \\ & & \text{c} & \epsilon & \text{h} & \text{h} & \text{S} & \text{R} \\ \hline \text{c} & \text{cUhU}\epsilon\text{UR} & u_c & v_\epsilon & v_h & v_h & 0 & v_R \\ \epsilon & \text{cUhU}\epsilon\text{UR} & u_c & v_\epsilon & v_h & v_h & 0 & v_R \\ \text{h} & \text{cU}\epsilon\text{UR} & u_c & v_\epsilon & 0 & 0 & 0 & v_R \\ \text{h} & \text{h} & 0 & 0 & v_h & w_h & 0 & 0 \\ \text{S} & \text{cUhU}\epsilon\text{UR} & u_c & v_\epsilon & v_h & v_h & 0 & v_R \\ \text{R} & \text{S} & 0 & 0 & 0 & 0 & v_S & 0 \end{array}_i \quad (2)$$

should be read as c or ϵ or R. For the first residue at the N terminus of the chain,[9] we define the row vector $\mathbf{t}_1$, which consists of the statistical weights for the allowed conformational states of the N terminus in a similar manner to eq 2-14, as

$$\mathbf{t}_1 = [u_c \quad v_\epsilon \quad v_h \quad v_h \quad 0 \quad v_R]_1 \tag{3}$$

(by keeping in mind that the first residue cannot contribute to the hydrogen bond energy or to the S state of a chain reversal conformation). For the last residue of the chain (i.e., the C terminus[9]), we define the column vector, $\mathbf{t}^*$, as

$$\mathbf{t}_N^* = \begin{bmatrix} u_c + v_\epsilon + v_h + v_R \\ u_c + v_\epsilon + v_h + v_R \\ u_c + v_\epsilon + v_R \\ v_h \\ u_c + v_\epsilon + v_h + v_R \\ v_S \end{bmatrix}_N \tag{4}$$

The elements of eq 4 correspond to the states c U ϵ U h U R, c U ϵ U h U R, c U ϵ U R, c U h U ϵ U R, and S for residue $i + 1$ when $i = N$.

It is now possible to calculate the partition function Z by using the set of eq 2–4 for $\mathbf{W}_i$, $\mathbf{t}_1$, and $\mathbf{t}_N^*$, i.e.,

$$Z = \mathbf{t}_1 \left[\prod_{i=2}^{N-1} \mathbf{W}_i \right] \mathbf{t}_N^* \tag{5}$$

Since the parameters appearing in eq 3 and 4 do not differ from those in eq 2, the elements of eq 3 can be found in the first row of eq 2; similarly, the elements of eq 4 can be obtained from those of eq 2. Thus, eq 5 may be written as

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N^* \tag{6}$$

where

$$\mathbf{e}_1 = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \tag{7a}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \tag{7b}$$

because $\mathbf{t}_1$ of eq 3 and $\mathbf{t}_N^*$ of eq 4 are given by

$$\mathbf{t}_1 = \mathbf{e}_1 \mathbf{W}_1 \tag{8a}$$

$$\mathbf{t}_N^* = \mathbf{W}_N \mathbf{e}_N^* \tag{8b}$$

respectively.

II. Nearest-Neighbor Four-State Model

In the previous paper 2,[4] we formulated a nearest-neighbor interaction model to reduce the size of the matrix and thus make the computations easier. We can also do this here and thus obtain a good approximation of eq 2.

The statistical weight matrix for a nearest-neighbor Ising model treatment of the four-state model, corresponding to eq 2-20 of the three-state model, may be written as

$\mathbf{W}_i =$

$i-1$ \ i	c	h	ϵ	R	S
c	u_c	v_h^2/w_h	v_ϵ	v_R	0
h	u_c	w_h	v_ϵ	v_R	0
ϵ	u_c	v_h^2/w_h	v_ϵ	v_R	0
R	0	0	0	0	v_S
S	u_c	v_h^2/w_h	v_ϵ	v_R	0

$_i$ (9)

The partition function in the nearest-neighbor interaction model, corresponding to eq 6, is

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N^* \tag{10}$$

in which

$$\mathbf{e}_1 = [1 \quad 0 \quad 0 \quad 0 \quad 0] \tag{11a}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{11b}$$

Equation 11a indicates that the first residue (the N terminus) of the chain can be preceded by a c state, and has to be represented by the first row of eq 9 in the form of $\mathbf{e}_1\mathbf{W}_1$. Equation 11b indicates that the last residue (the C terminus) of the chain can be preceded by c, h, ϵ, R, or S states; hence, it is represented by the column vector $\mathbf{W}_N\mathbf{e}_N^*$ which contains the sum of the elements of the five row vectors in eq 9.

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 reference state. In contrast to section III of paper 2,[4] where the c state was taken as the reference state, we take the ϵ state as the reference here, because the c state loses its earlier significance in the redefinition introduced here; also, by taking the ϵ state as the reference, the numerical values of the relative statistical weights will lie in a convenient range. With the ϵ state as the reference, we define the relative statistical weights, w_h^*, v_h^*, v_R^*, v_S^*, and u_c^* pertaining to the h, R, S, and c states, in terms of

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

$$v_h^* = v_h/v_\epsilon \tag{13}$$

$$v_R^* = v_R/v_\epsilon \tag{14}$$

$$v_S^* = v_S/v_\epsilon \tag{15}$$

and

$$u_c^* = u_c/v_\epsilon \tag{16}$$

respectively. Using eq 12–16, we may rewrite eq 9 as

$\mathbf{W}_i =$

$i-1$ \ i	c	h	ϵ	R	S
c	u_c^*	v_h^{*2}/w_h^*	1	v_R^*	0
h	u_c^*	w_h^*	1	v_R^*	0
ϵ	u_c^*	v_h^{*2}/w_h^*	1	v_R^*	0
R	0	0	0	0	v_S^*
S	u_c^*	v_h^{*2}/w_h^*	1	v_R^*	0

$_i$ (17)

The partition function can be calculated by substituting eq 17 for eq 9 when eq 10 and 11 are used.

We have thus formulated a four-state model involving helical, extended, chain-reversal, and other states in section I, and a nearest-neighbor approximation of it (involving a smaller-size matrix[10]) in this section. The method for calculating molecular averages and conformational-sequence probabilities was already described in general terms in section VI of paper 2.[4] Therefore, we will not repeat these details here. In order to compute these quantities, we then evaluate the statistical weights by using x-ray coordinates of native proteins in sections III and IV.

The phenomenon of asymmetric nucleation[11–13] of helical

sequences is introduced, for completeness, in the Appendix.

III. Analysis of X-Ray Data of Native Proteins

As in paper 1,[3] we obtain the statistical weights from the x-ray structures of native proteins. However, whereas in paper 1 we relied on the crystallographers' statements as to which residues were in h and ϵ states, and common criteria may not have been used in assigning such states, here we compute the dihedral angles directly from the x-ray coordinates in order to assign the conformational states. The x-ray coordinates of 26 proteins[14-39] were used (those listed in column I of Table I), and the conformational states of each residue were assigned from the values of the backbone dihedral angles (ϕ, ψ), using the criteria summarized below. In Table I, the α and β chains of hemoglobin are regarded as separate proteins; however, the B and C chains of chymotrypsin are counted as one protein.

(A) Conformational States. Helical State. We use the criterion of Burgess et al.[40] to define a right-handed α-helical state (h), viz., as one whose backbone dihedral angles lie within the range of $-130° \leq \phi \leq -10°$ and $-90° \leq \psi \leq -10°$. Furthermore, as in paper 1,[3] if at least three consecutive residues have dihedral angles in this range, they are considered to be in a helical sequence; less than three consecutive helical residues constitute the isolated helical state(s). We consider a residue in a helical sequence to be in an h state, with a statistical weight w_h (or w_h^*), and an isolated residue(s) to be in an h′ state, with a statistical weight v_h (or v_h^*).[41]

Extended State. We use the range of dihedral angles proposed by Burgess et al.[40] to define an extended state (ϵ) of an amino acid *residue*, viz., as one whose backbone dihedral angles lie within the range of $-180° \leq \phi \leq -45°$ 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 $-180° \leq \psi \leq -140°$. All *residues* in ϵ states contribute to the statistical weight v_ϵ. In the present paper, we do not distinguish between isolated extended (ϵ) residues or those that are in extended sequences (by an "extended *sequence*", we mean one that consists of four or more consecutive residues in ϵ states,[42] as in paper 1[3]).

Chain-Reversal Conformation. The definition used here for a chain-reversal conformation was given[45] in section I. The two residues in such a conformation are considered to be in R and S states, with statistical weights v_R and v_S, respectively.

(B) Statistical Analysis. Using the x-ray coordinates of 26 proteins,[14-39] the values of (ϕ, ψ) and the distance $R_{i-2,i+1}$ between C^α_{i-2} and C^α_{i+1} were computed. With these values and the criteria described in section I, the conformational states h, h′, ϵ, R, S, and c were assigned to every residue in the 26 proteins. The results of this analysis are presented in Table I. To save space, h, hh, ϵ, $\epsilon\epsilon$, and $\epsilon\epsilon\epsilon$ sequences are not shown; however, the number of h′ residues ($N_{h',j}$) are given in column 4 of Table II, and the ϵ, $\epsilon\epsilon$, and $\epsilon\epsilon\epsilon$ sequences are included (together with extended ϵ sequences) in the values of $N_{\epsilon,j}$ in column 5 of Table II. Also, the R and S states are combined as "chain reversal" in Table I, but $N_{R,j}$ and $N_{S,j}$ are listed separately in Table II.

Some chain reversals consist of more than two pairs of residues with $R_{i-2,i+1}$, $R_{i-1,i+2}$, etc., all <7 Å. These are multiple chain-reversal regions where a duplicate assignment of both R and S would occur (see Table I); e.g., residues $i-1$ and i would be assigned as R and S, respectively, but, at the same time, residues i and $i+1$ would be assigned as R and S, with residue i being duplicately assigned as S and R. Such duplicately assigned residues will be considered to be in a D state. Thus, a sequence of more than two residues in a multiple chain-reversal region will be designated as RDD···DDS, to indicate that the multiple chain reversal can extend over many residues in D states.

A duplicate assignment can occur at the ends of helical and extended sequences, where the duplication is between chain reversal and helical or between chain reversal and extended. These residues (enclosed in parentheses in the last column of Table I) are considered as helical or extended, and not as chain reversal, in counting the number of residues in each type of conformational state. For example, residues 4–19 and 21–35 of myoglobin are helical and residues 19–20 constitute a chain reversal; i.e., residue 19 can act both as a helical one and as a chain reversal (see Table I). Residue 19 is counted as helical (instead of as an R state of a chain reversal) in order to preserve the definition of the statistical weight of a helical state from the three-state model; i.e., we are now preserving the h and ϵ states of the three-state model, and dividing the old c state into chain-reversal and new c states. Residue 20 is counted as an S state of a chain reversal. If the R or S states of a chain reversal precede or follow an h or ϵ sequence (such as Rhh···hh, or hh···hhS, or R$\epsilon\epsilon$···$\epsilon\epsilon$, or $\epsilon\epsilon$···$\epsilon\epsilon$S), the combination Rh, hs, Rϵ, or ϵS may be regarded as a chain reversal (if the distance $R_{i-2,i+1} \leq 7$ Å). Thus, we list the chain reversal as (19)–20 in Table I. In reporting the results of predictions, such duplicate assignments are treated in the same manner, as shown by the parentheses in column 6 of Tables IV–VI.

The above situation arises because of an ambiguity in the definition of a chain reversal, when it occurs at either end of a helical or extended sequence. We have chosen to resolve this ambiguity by an approximation that counts such chain reversals when computing the statistical weights v_R and v_S but omits Rh, ϵS, etc., states from the statistical weight matrix of eq 17 (in order to keep the order of the matrix small); while it would have been more consistent to use a higher-order matrix, to include such Rh, ϵS, etc., states, the paucity of x-ray data would have made such a sophistication unwarranted at this time.

The results of the present analysis are summarized in Table II, where N_j is the total number of the jth type of amino acid found in the 26 proteins,[46] and $N_{h,j}$, $N_{h',j}$, $N_{\epsilon,j}$, $N_{R,j}$, $N_{S,j}$, $N_{D,j}$, and $N_{c,j}$ designate the total number of the jth type of amino acid found in h, h′, ϵ,[42] R, S, D, and other (c) states.

IV. Computation of Statistical Weights

We evaluate the statistical weights from the data of Table II, on the basis of the concept that short-range interactions dominate in determining protein conformation (see second paragraph of the introductory section; also see ref 45 and section IIIE of paper 1,[3] and ref 7).

Following an argument similar to that made in section IIA of paper 1,[3] in deriving eq 1–4 of paper 1, and keeping in mind that the ϵ state (rather than the c state, as in paper 1[3]) is taken as the standard state, we obtain the following analogue of eq 1-16

$$w_{h,j}^* = f_{h,j}/f_{\epsilon,j} \tag{18}$$

where $f_{h,j}$ and $f_{\epsilon,j}$ can be obtained by substituting h and ϵ for η in

$$f_{\eta,j} = N_{\eta,j}/N_j \tag{19}$$

since $w_{h,j}^* = w_{h,j}/v_{\epsilon,j}$, as given in eq 12 (see also eq 1-15). The numerical values of N_j, $N_{h,j}$, and $N_{\epsilon,j}$ are given in Table II. In a similar manner, the statistical weights $v_{h,j}^*$ and $u_{c,j}^*$ can be evaluated by

$$v_{h,j}^* = f_{h',j}/f_{\epsilon,j} \tag{20}$$

and

$$u_{c,j}^* = f_{c,j}/f_{\epsilon,j} \tag{21}$$

where $f_{h',j}$, $f_{\epsilon,j}$, and $f_{c,j}$ can be obtained by substituting h′, ϵ, and c, respectively, for η in eq 19. The numerical values of $N_{h',j}$, $N_{\epsilon,j}$, and $N_{c,j}$ are given in Table II.

Table I
Experimentally Observed Backbone Structures[a] of Proteins

Protein	Source	No. of amino acid residues	Backbone structures[a]		
			Helical	Extended	Chain reversal[b]
Myoglobin[c]	Sperm whale	153	4–19, 21–35, 37–42, 44–48, 54–57, 59–77, 83–95, 101–117, 120–122, 125–149	none	(19)–20, (35)–36, 52–53, (77)–78, 79–81, (95)–97, (117)–119, (122)–123
Lysozyme[d]	Hen egg white	129	5–14, 25–35, 80–84, 89–99, 109–113, 120–123	1–4, 43–46	(14)–15, 18–19, 20–22, (35)–36, 37–38, 40–41, 55–56, 60–62, 67–68, 70–71, 75–76, 86–87, (99)–101, 104–108, (113)–115, 116–117, (123)–124, 125–126
Ribonuclease-S[e]	Bovine	124	4–12, 25–33, 51–55, 57–60	43–47, 61–65, 78–87, 95–111	16–18, 37–38, 55–56, 66–68, 76–77, 88–89, 92–94, 113–114
Deoxyhaemoglobin[f] α-chain	Horse	141	4–13, 23–26, 28–35, 54–58, 62–67, 69–71, 85–91, 96–108, 110–112, 114–116, 119–138	none	(13)–17, 18–20, 21–22, (26)–27, (35)–36, 37–43, 44–45, 50–51, 53–(54), (58)–61, 67–68, 71–72, 73–75, 76–80, 81–84, 95–(96), (108)–109, (112)–113, 116–117, (138)–139
Deoxyhaemoglobin[f] β-chain	Horse	146	8–16, 22–31, 51–56, 58–69, 71–75, 81–91, 101–110, 112–118, 125–141	none	5–7, (16)–17, 21–(22), (31)–35, 36–37, 38–42, 43–45, 70–(71), (75)–77, 78–80, (91)–93, 100–(101), (110)–111, 119–122, 124–(125), (141)–142
α-Chymotrypsin[g] B-chain	Bovine	131	41–43	4–8, 15–19, 27–31, 36–40, 49–52, 65–70, 88–93, 103–110, 119–125	2–3, 9–10, 11–12, 13–14, 21–22, 34–35, 47–48, 53–54, 58–59, 61–62, 77–78, 81–83, 85–86, 94–95, 100–102, 111–112, 117–118
C-chain		97	17–20, 89–95	7–16, 32–35, 40–43, 49–54, 58–61, 76–82	5–6, (20)–24, 25–27, 30–31, 38–39, 44–46, 47–48, 55–57, 70–71, 74–75, 83–86, 87–88, (95)–96
Carboxypeptidase-A[h]	Bovine	307	15–26, 73–80, 84–88, 94–100, 113–122, 174–186, 216–229, 254–261, 286–305	11–14, 33–37, 45–53, 60–66, 103–107, 191–198, 201–205, 236–242, 266–272	4–6, 9–10, (26)–29, 30–32, 42–43, 57–58, 68–69, 70–72, (80)–83, (88)–89, 90–92, (100)–102, 109–110, 111–112, 124–125, 143–146, 149–150, 151–152, 154–155, 160–161, 163–164, 170–171, 207–208, 214–215, (229)–234, 243–247, 251–252, (261)–262, 264–265, 274–275, 276–277, 278–280, 283–285, (305)–306
Subtilisin BPN'[i]	*Bacillus subtilis*	275	2–5, 13–17, 64–68, 104–116, 133–145	6–9, 26–30, 88–91, 147–152, 174–181, 190–193, 205–210, 213–219, 222–237, 243–251, 260–263, 270–274	(9)–11, (17)–19, 24–25, 37–38, 40–41, 52–53, 57–58, 61–62, (68)–73, 84–85, 86–87, 98–100, (116)–117, 120–121, 159–161, 167–170, 172–173, 182–183, 188–189, 194–195, 203–204, 211–212, 220–221, (237)–238, 239–240, (251)–253, 264–265
Elastase[j]	Pig	240	155–159	4–8, 15–21, 25–28, 30–34, 39–43, 51–57, 69–74, 76–80, 94–99, 102–106, 110–116, 119–122, 125–131, 139–142, 146–154, 172–175, 190–195, 199–202, 220–225, 230–238	2–3, 9–10, 11–12, 13–14, 23–24, 37–38, 46–47, 48–49, 60–61, 62–63, 81–82, 85–87, 88–90, 100–101, 107–109, 117–118, 123–124, 135–136, 137–138, 143–144, 161–163, 165–166, 170–171, 179–181, 185–187, 188–189, 196–198, 211–212, 214–215, 217–218, 226–229
Staphylococcal nuclease[k]	*Staphylococcus aureas*	142[l]	55–61, 64–67, 99–106, 122–134	7–14, 22–26, 30–36, 39–43, 72–77, 90–93, 109–112	2–3, 4–5, 20–21, 27–29, 37–38, 47–49, 50–51, 53–54, (61)–63, (67)–68, 70–71, 84–85, 94–96, 116–119, 120–121, (134)–135, 138–141

Table I (Continued)

Protein	Source	No. of amino acid residues	Backbone structures[a]		
			Helical	Extended	Chain reversal[b]
Papain[m]	Papaya	212	25–42, 50–56, 68–78, 97–100, 118–127, 140–143	44–49, 79–82, 91–95, 109–114, 128–135, 159–163, 170–175, 186–191, 205–212	3–4, 7–10, 20–21, 58–60, 62–64, 83–86, 115–116, 136–137, 139–(140), 168–169, 179–180, 182–185, 196–197, 199–201, 202–203
Ferricytochrome c[n]	Horse heart	104	2–9, 52–54, 71–74, 88–100	none	(9)–13, 14–17, 22–23, 27–29, 33–34, 35–37, 43–45, 50–51, 62–63, 64–70, (74)–75, 76–77, 79–80, (100)–102
Lactate dehydrogenase[o]	Dogfish	329	3–7, 32–42, 55–69, 84–86, 106–126, 140–151, 164–178, 211–215, 225–242, 248–260, 307–323	15–19, 22–26, 48–52, 76–80, 91–96, 131–135, 186–190, 269–272, 283–287, 289–292, 299–303	(7)–8, 30–31, 46–47, (69)–71, 81–82, 83–(84), 87–88, (126)–128, 138–139, 155–158, 162–163, 182–185, 194–196, 201–203, 206–208, 219–220, 246–247, (260)–261, 274–275, 276–278, 293–294, 297–298
Cytochrome b_5[p]	Calf liver	93[q]	9–14, 35–38, 43–48, 55–60, 65–73, 82–86	4–8, 21–25, 77–80	(14)–15, 18–20, 26–27, 33–34, 40–41, (48)–49, 50–51, (60)–61, (73)–74, 81–(82)
Thermolysin[r]	*Bacillus thermoproteolyticus*	316	70–87, 139–150, 160–173, 175–179, 234–244, 260–274, 281–296, 302–312	1–12, 16–23, 39–43, 53–57, 98–103, 112–115, 119–123, 254–257	13–14, 25–27, 33–35, 36–37, 45–46, 50–51, 58–59, 65–67, 68–69, (87)–88, 104–105, 108–109, 116–118, 127–129, 133–136, 137–138, (150)–153, 159–(160), (173)–174, (179)–180, 182–183, 188–189, 190–192, 195–196, 198–199, 205–207, 208–211, 217–219, 225–229, 230–232, 233–(234), (244)–246, 250–252, 277–278, 298–299, 301–(302), (312)–313
Concanavalin[s]	*Canavalia ensiformis*	237	56–58, 81–84	3–10, 22–28, 37–40, 46–55, 60–67, 70–79, 93–97, 100–103, 108–117, 124–130, 140–143, 157–160, 169–175, 189–201, 209–216, 219–222, 233–237	11–12, 15–17, 29–30, 32–33, 35–36, 44–45, 68–69, 87–88, 98–99, 118–119, 135–136, 138–139, 144–145, 148–149, 150–152, 161–162, 167–168, 184–185, 187–188, 202–203, 204–205, 217–218, 223–224, 227–229, 230–232
Myogen[t]	Carp muscle	108	8–17, 26–32, 40–50, 60–64, 68–70, 79–88, 100–107	74–78	3–5, 6–7, (17)–18, 21–22, (32)–33, 35–37, (50)–51, 52–53, 55–56, (64)–65, 66–67, (70)–71, 72–73, (88)–89, 91–92, 99–(100)
Sea lamprey hemoglobin[u]	*Petromyzon marinus*	148	15–21, 23–28, 31–44, 46–50, 63–65, 68–87, 98–105, 115–124, 132–144	7–12	13–14, (21)–22, (28)–29, 30–(31), (44)–45, (50)–51, 53–54, 56–57, 61–62, (65)–66, 88–90, 92–97, 107–108, 110–111, 112–113, 114–(115), (124)–126, 146–147
Rubredoxin[v]	*Clostridium pasteurianum*	54	15–17, 30–32	2–6, 49–52	7–9, 20–22, 26–27, 35–36, 40–42, 46–48
Cytochrome C_2[w]	*Rhodospirillum rubrum*	112	3–14, 54–56, 65–70, 97–105	none	2–(3), 15–17, 22–23, 27–29, 33–34, 36–37, 40–41, 44–45, 50–53, (56)–58, 64–(65), (70)–73, 74–81, 85–86, (105)–111
Ferredoxin[x]	*Peptococcus aerogenes*	54	15–17	2–5	6–7, 19–20, 26–27, 33–34, 40–44, 46–47
Trypsin[y]	Bovine	223	78–80, 146–151, 214–221	15–18, 25–28, 34–37, 84–89, 101–106, 116–120, 135–142	9–10, 11–12, 13–14, 20–21, 30–31, 32–33, 39–41, 52–54, 55–56, 74–75, 91–92, 97–99, 107–(108), 112–113, 124–127, 128–129, 133–134, 144–145, (151)–152, 158–159, 167–168, 174–175, 177–178, 183–185, 195–197, 198–201, 209–213, (221)–222
Pancreatic trypsin inhibitor[z]	Bovine	58	3–5, 25–27, 48–54	7–10, 18–24, 29–35	(5)–6, 42–43, (54)–55

Table I (Continued)

Protein	Source	No. of amino acid residues	Backbone structures[a] Helical	Extended	Chain reversal[b]
Glyceraldehyde phosphate dehydrogenase[a']	Lobster	333	11–23, 36–44, 47–49, 78–80, 101–108, 129–132, 149–163, 210–215, 218–220, 251–263, 312–327, 329–332	2–5, 28–31, 55–58, 70–73, 114–118, 125–128, 143–147, 169–173, 202–207, 224–231, 236–245, 297–300, 303–310	10–(11), 32–33, (44)–45, 53–54, 59–61, 65–67, 76–77, 83–86, 89–90, 96–98, (108)–111, 122–123, 133–134, 138–140, 141–142, 148–(149), (163)–164, 179–180, 183–184, 190–191, 198–200, 209–(210), (215)–217, 221–223, 264–266, 267–269, 276–277, 281–282, 283–284, 288–289, 293–294, 295–296, 301–302, (327)–328
Clostridial flavodoxin[b']	*Clostridium*	138	11–26, 63–74, 94–106, 125–136	1–7, 28–34, 49–56, 80–88, 112–118	8–9, 35–36, 40–42, 43–45, 47–48, 57–59, (74)–76, 78–79, 89–90, 92–93, 123–124, (136)–137
High potential iron protein[c']	*Chromatium vinosum*	85	12–16, 28–31	60–63, 69–72, 80–83	4–5, 9–10, (16)–17, 21–22, 23–25, 38–40, 43–45, 47–48, 51–52, 54–55, 58–59, 64–66, 67–68, 73–74, 78–79

[a] See section IIIA of the text and ref 42 for the present definitions of the backbone structures of the helical, extended, and chain-reversal conformations. [b] The residue numbers in parentheses are those that could be duplicately assigned as chain reversals and helical or extended conformations. These residues are regarded as helical or extended ones rather than as chain reversals in evaluating statistical weights (see section IIIB). The sequences in this column that consist of more than two residues pertain to a multiple chain-reversal region in which a duplicate assignment of both R and S would occur. Such a duplicately assigned residue is designated as a D conformation in the text. Thus, a sequence consisting of more than two residues in a multiple chain-reversal region would be designated as RDD · · · DDS. [c] Reference 15. [d] Reference 16. [e] Reference 17. [f] Reference 18. [g] Reference 19. [h] Reference 20. [i] Reference 21. [j] Reference 22. [k] Reference 23. [l] In ref 23, the authors reported 149 (which was used in our papers **1** and **3**). However, in the recent x-ray data from the Brookhaven Data Bank,[14] this number was reported as 142. [m] Reference 24. [n] Reference 25. [o] Reference 26. [p] Reference 27. [q] The x-ray coordinates were reported only for 85 residues [from residues 3 (Ala) to 87 (Ile)]. [r] Reference 28. [s] Reference 29. [t] Reference 30. [u] Reference 31. [v] Reference 32. [w] Reference 33. [x] Reference 34. [y] Reference 35. [z] Reference 36. [a'] Reference 37. [b'] Reference 38. [c'] Reference 39.

Since each D residue acts as both an R and an S, we consider that half of the residues found in the duplicate conformation D of a chain reversal contribute to the R state, and the other half of the D residues contribute to the S state; hence, we evaluate the total number of residues in R and S states, $N'_{R,j}$ and $N'_{S,j}$, by using

$$N'_{\eta,j} = N_{\eta,j} + [N_{D,j}/2] \tag{22}$$

where R and S are substituted for η to obtain $N'_{R,j}$ and $N'_{S,j}$, respectively. The statistical weights $v_{R,j}^*$ and $v_{S,j}^*$ can be computed from

$$v_{R,j}^* = f'_{R,j}/f_{\epsilon,j} \tag{23}$$

and

$$v_{S,j}^* = f'_{S,j}/f_{\epsilon,j} \tag{24}$$

where $f'_{R,j}$ and $f'_{S,j}$ are given by

$$f'_{\eta,j} = N'_{\eta,j}/N_j \tag{25}$$

where R and S are substituted for η, and $N'_{\eta,j}$ is given in eq 22. The values of $f_{\epsilon,j}$ are given by substituting ϵ for η in eq 19 (and not in eq 25). The numerical values of $N_{R,j}$, $N_{S,j}$, and $N_{D,j}$ (used in eq 22), as well as $N_{\epsilon,j}$ (used in eq 19, 23, and 24), are given in Table II.

The results obtained for the statistical weights relative to the ϵ state are given for 20 amino acids in Table III. These statistical weights for the four-state mode will be used in the prediction of protein conformation in section VI.

V. Prediction of Chain-Reversal Conformation

In this section, we describe the methods to predict the chain-reversal conformation, using the four-state model formulated in sections I and II and the statistical weights evaluated in section IV. For this purpose, we describe the method to compute the conformational-sequence probability in section VA, and propose two rules for the prediction of the chain-reversal conformation in section VB.

(A) Calculation of Conformational-Sequence Probability. A method for calculating a conformational probability for a residue or a sequence to be found in a certain conformational state or conformational sequence was formulated for a two-state model (h or c) by Tanaka and Nakajima,[47] and generalized to a model with any number of conformational states in our paper **2**.[4]

The first-order a priori probability that a residue i will be found in a helical state ($F_{h;i}$), an extended state ($F_{\epsilon;i}$), an R state ($F_{R;i}$), an S state ($F_{S;i}$), and other state ($F_{c;i}$) can be calculated by using eq **2**-44, viz.

$$F_{i;\eta_i} = Z^{-1}\mathbf{e}_1\left[\prod_{j=1}^{i-1}\mathbf{W}_j\right]\left[\frac{\partial \mathbf{W}_i}{\partial \ln (\mathbf{m}_{i;\eta_i})}\right]_{\{\rho\}}\left[\prod_{l=i+1}^{N}\mathbf{W}_l\right]\mathbf{e}_N^* \tag{26}$$

where the partition function Z is obtained from eq 10 [since we will use the nearest-neighbor interaction model, with a low-order matrix,[10] for the predictions, we will use eq 9 (or 17)–11 rather than eq 2–5 or 6–8]. The statistical weight matrices $\mathbf{e}_1$, $\mathbf{W}_i$, and $\mathbf{e}_N^*$ of eq 11a, 17, and 11b, respectively, are used in eq 26. Then, the conformational states η_i and $\{\rho\}$ of eq 26 are replaced by h, ϵ, R, S, or c to obtain the values of $F_{h;i}$, $F_{\epsilon;i}$, $F_{R;i}$, $F_{S;i}$, or $F_{c;i}$.

In order to predict the chain reversal, it is necessary to compute the second-order a priori probability $F_{i;RS}$ that the $(i-1)$th residue is found in an R state and the ith residue in an S state. The value of $F_{i;RS}$ can be calculated by substituting R and S for η_{i-1} and η_i, respectively, in eq **2**-47 (i.e., $\{\rho\}$ = RS).

Table II
The Number of Amino Acids Occurring in Helical, Extended, Chain-Reversal, and Other States in 26 Proteins

Amino acid j	N_j	No. of residues in conformational states: Helical $N_{h,j}$	Isolated helical $N_{h',j}$	Extended[a] $N_{\epsilon,j}$	Chain Reversal $N_{R,j}$	Chain Reversal $N_{S,j}$	Chain Reversal $N_{D,j}$	Other $N_{c,j}$
Ala	407	175	8	104	43	21	19	37
Arg	128	40	3	45	9	9	11	11
Asn	223	48	2	60	13	48	12	40
Asp	257	75	7	44	24	40	21	46
Cys	100	22	1	41	6	14	5	11
Gln	149	43	5	50	17	11	8	15
Glu	207	104	3	43	17	14	7	19
Gly	401	60	3	94	36	77	18	113
His	111	37	5	31	6	17	6	9
Ile	225	74	6	99	12	11	2	21
Leu	323	124	8	116	17	22	13	23
Lys	320	118	8	76	30	31	20	37
Met	68	26	0	24	5	3	3	7
Phe	150	52	3	51	5	21	6	12
Pro	161	26	7	65	45	3	6	9
Ser	368	79	4	130	47	44	21	43
Thr	270	56	11	106	28	21	20	28
Trp	75	22	3	27	3	10	3	7
Tyr	180	35	3	78	13	24	11	16
Val	353	108	3	163	20	24	11	24

[a] See ref 42.

Table III
Statistical Weights for the Four-State Model[a]

Amino acid j	Relative statistical weight[b]: $w_{h,j}^*$	$v_{h,j}^*$	$v_{R,j}^*$	$v_{S,j}^*$	$u_{c,j}^*$
Ala	1.683	0.077	0.505	0.293	0.356
Arg	0.889	0.067	0.322	0.322	0.244
Asn	0.800	0.033	0.317	0.900	0.667
Asp	1.705	0.159	0.784	1.148	1.045
Cys	0.537	0.024	0.207	0.402	0.268
Gln	0.860	0.100	0.420	0.300	0.300
Glu	2.419	0.070	0.477	0.407	0.442
Gly	0.638	0.032	0.479	0.915	1.202
His	1.194	0.161	0.290	0.645	0.290
Ile	0.747	0.061	0.131	0.121	0.212
Leu	1.069	0.069	0.203	0.246	0.198
Lys	1.553	0.105	0.526	0.539	0.487
Met	1.083	0.000	0.271	0.187	0.292
Phe	1.020	0.059	0.157	0.471	0.235
Pro	0.400	0.108	0.738	0.092	0.138
Ser	0.608	0.031	0.442	0.419	0.331
Thr	0.528	0.104	0.358	0.292	0.264
Trp	0.815	0.111	0.167	0.426	0.259
Tyr	0.449	0.038	0.237	0.378	0.205
Val	0.663	0.018	0.156	0.181	0.147

[a] This four-state model consists of helical (h), extended (ϵ), chain-reversal (R and S states), and other (c) states. [b] The statistical weights relative to the extended (ϵ) state; i.e., $v_{\epsilon,j}^* = 1.0$ for all 20 amino acid residues.

In eq 2-47, Z can be obtained from eq 9, 11, and 17. The average probabilities θ_h, θ_ϵ, θ_R, θ_S, or θ_c over a whole protein chain can be obtained by substituting h, ϵ, R, S, and c for η in eq 2-52, i.e., from

$$\theta_\eta = \frac{1}{N}\sum_{i=1}^{N} F_{i;\eta i} \tag{27}$$

where N is the number of residues in the protein. In a similar manner to eq 27, the average probability for residues $i-1$ and i to be found in a chain-reversal conformation RS can be calculated from

$$\theta_{RS} = \frac{1}{(N-1)}\sum_{i=2}^{N} F_{i;RS} \tag{28}$$

In a similar manner to eq 3-13 and 3-14, we then define the relative probabilities

$$P_{i;\eta}^* = F_{i;\eta}/\theta_\eta \tag{29}$$

for η = h, ϵ, R, S, or c. For RS states at residues $i-1$ and i of a chain reversal, we define the relative probability

$$P_{i;RS}^* = F_{i;RS}/\theta_{RS} \tag{30}$$

Using the quantities $P_{i;\eta}^*$ (η = h, ϵ, R, S, or c) and $P_{i;RS}^*$, we will develop the methods to predict chain reversals in section VB.

(B) Two Predictive Schemes for the Chain-Reversal Conformation. In this subsection, we propose two rules, one to predict only chain-reversal conformations by means of a four-state model (rule I), and a second to predict helical sequences, extended sequences, and chain reversals by means of the four-state model (rule II).[48]

Rule I. A chain-reversal conformation at residues $i-1$ and i (R state at residue $i-1$ and S state at residue i) will be predicted if the following two conditions are satisfied: (i) The relative probability $P_{i;RS}^*$ is greater than unity, i.e., $P_{i;RS}^* \geq 1$. (ii) The relative probability of an R state at residue $i-1$, $P_{i-1;R}^*$, is greater than any of the probabilities $P_{i-1;h}^*$, $P_{i-1;\epsilon}^*$, $P_{i-1;S}^*$, and $P_{i-1;c}^*$ at residue $i-1$; i.e.,

$$P_{i-1;R}^* > P_{i-1;h}^*, P_{i-1;\epsilon}^*, P_{i-1;S}^*, P_{i-1;c}^*$$

and, at the same time, the relative probability of an S state at residue i, $P_{i;S}^*$, is greater than any of the probabilities $P_{i;h}^*$, $P_{i;\epsilon}^*$, $P_{i;R}^*$, and $P_{i;c}^*$ at residue i, i.e.,

$$P_{i;S}^* > P_{i;h}^*, P_{i;\epsilon}^*, P_{i;R}^*, P_{i;c}^*$$

The application of rule I can predict the possible position of a chain-reversal conformation without any ambiguity.

In order to predict the location of helical and extended sequences and chain-reversal conformations (using the four-state model), we will formulate rule II below by combining the empirical rules of paper 3 (see section III of paper 3), with which one can choose possible helical and extended sequences,

Table IV
Comparison of Predicted and Observed Results for the Chain-Reversal Conformation[a]

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be		Predicted by rule II[d]	Obsd[b,e] chain reversal
			Helical	Extended		
Myoglobin	153	6–7		6–12		
		15–16				
		19–20	13–20			19–20
		22–23			22–23	
		26–27		27–32	26–(27)	
		35–36			35–36	(35)–36
		37–38	37–41, 50–57			52–53
		59–60				
		63–64		65–72		
		77–78	73–77, 80–87,		77–78	(77)–78
		92–93	88–91			79–81
		96–97			92–93	
		102–103	103–107		96–97	95–97
		105–106			102–(103)	
		108–109	108–111			
		118–119		112–117	118–119	117–119
		120–121			120–121	(122)–123
		125–126	124–126	127–130		
		131–132	132–137		131–(132)	
		140–141	141–145		140–(141)	
		144–145				
		147–148	146–149			
Lysozyme	129		4–13			(14)–15
		18–19			18–19	18–19
			26–34			20–22
						(35)–36
		36–37			36–37	37–38
		43–44		40–43	(43)–44	40–41
		45–46			45–46	
		47–48			47–48	
		51–52			51–52	
				53–59		55–56
		60–61			60–61	60–62
		66–67			66–67	67–68
		70–71			70–71	70–71
		73–74			73–74	75–76
		79–80	80–84		79–(80)	
		86–87		88–92	86–87	86–87
		95–96			95–96	
		100–101			100–101	(99)–101
		103–104			103–(104)	104–108
		105–106	104–111			
		110–111				
		112–113			112–113	(113)–115
				118–125		116–117
						(123)–124
						125–126
Ribonuclease S	124		1–10			
		20–21	17–21			16–18
		23–24			23–24	
		31–32	27–30			
		33–34			33–34	
		37–38			37–38	37–38
		52–53	49–57	45–48		55–56
		61–62		62–66		
		64–65				
		66–67			(66)–67	66–68
		70–71			70–71	
		75–76			75–76	76–77
		82–83		78–82	(82)–83	
		90–91			90–91	88–89
		93–94			93–94	92–94
		96–97		95–98		
		102–103	99–103			
		104–105		104–111		
		109–110				
		114–115		114–119		113–114
		122–123			122–123	
Deoxyhemoglobin[f]	141			1–4		
α chain		5–6			5–6	
		8–9	9–14		8–(9)	
		13–14				

Table IV (Continued)

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be: Helical	Predicted[c] to be: Extended	Predicted by rule II[d]	Obsd[b,e] chain reversal
		15–16			15–16	(13)–17
						18–20
		23–24	24–32		23–(24)	21–22
						(26)–27
		37–38			37–38	(35)–36
						37–43
		44–45			44–45	44–45
		49–50			49–50	50–51
		53–54	51–57			53–54
		57–58			(57)–58	(58)–61
		60–61	59–63			
		63–64		65–72	(63)–64	
						67–68
						71–72
		74–75			74–75	73–75
		77–78			77–(78)	76–80
			78–80			
		81–82			81–82	81–84
		84–85	84–89			
			97–101			95–(96)
		102–103		103–114	102–(103)	
		114–115			(114)–115	(112)–113
			119–123			116–117
		130–131			130–131	
				132–138		
		138–139			(138)–139	(138)–139
Deoxyhemoglobin[f] β chain	146		5–17			5–7
						(16)–17
			19–24			21–(22)
			25–28	29–40		(31)–35
		36–37				36–37
						38–42
		44–45			44–45	43–45
		49–50	51–53			
		58–59	58–62			
		62–63			62–63	
		65–66			65–66	
		70–71			70–71	70–71
						(75)–77
		79–80			79–80	78–80
			84–89			(91)–93
				103–107		100–(101)
				108–117		(110)–111
		116–117			116–117	119–122
		120–121			120–121	
			124–128			124–125
		131–132		130–135		
		135–136	136–142			
		138–139				
		142–143			(142)–143	(141)–142
		144–145			144–145	
α-Chymotrypsin[g] B chain	131	5–6	3–8			2–3
		9–10			9–10	9–10
		11–12		12–19	11–(12)	11–12
		13–14				13–14
		19–20			(19)–20	21–22
		34–35	29–32	35–43	34–(35)	34–35
		41–42			41–42	
		48–49		43–47	48–(49)	47–48
		53–54		49–54		53–54
		55–56			55–56	
						58–59
		60–61		67–71	60–61	61–62
		75–76			75–76	
		77–78			77–78	77–78
				88–94		81–83
			94–98			85–86
						94–95
		98–99		99–111	(98)–(99)	100–102
		109–110	112–117			111–112
		113–114				
						117–118
				119–125		

Table IV (Continued)

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be		Predicted by rule II[d]	Obsd[b,e] chain reversal
			Helical	Extended		
Carboxypeptidase A	307	1–2			1–2	
		4–5			4–5	4–6
		6–7			6–7	
		10–11		9–15		9–10
		20–21	19–22	23–28		
		28–29	29–32		(28)–(29)	(26)–29
		34–35		33–36		30–32
		41–42		45–51	41–42	42–43
		51–52			(51)–52	
		57–58		59–61	57–58	57–58
						68–69
		70–71	70–74			70–72
		72–73				
			79–85	75–78		80–83
		88–89			88–89	(88)–89
		92–93			92–93	90–92
		94–95			94–(95)	
			95–101	104–112		(100)–102
						109–110
		111–112				
		113–114			113–114	111–112
		117–118			117–118	
		122–123			122–123	124–125
		127–128		129–133	127–128	
		130–131				
		135–136		135–141		
		143–144			143–144	143–146
		145–146			145–146	149–150
			152–155			151–152
		156–157			156–157	154–155
		158–159			158–159	
		160–161			160–161	160–161
		162–163			162–163	163–164
		165–166			165–166	
		168–169			168–169	
		170–171	171–177		170–(171)	170–171
		172–173				
		177–178			(177)–178	
		181–182			181–182	
		184–185	188–193		184–185	
		191–192				
		197–198		198–206	197–(198)	
		199–200			199–200	
		205–206		208–212	205–206	207–208
		211–212				
		214–215	214–220			214–215
		231–232	220–231		(231)–232	(229)–234
		237–238			237–238	
		239–240			239–240	
				242–251		243–247
						251–252
						(261)–262
		264–265			264–265	264–265
		266–267		266–271		
		272–273			272–273	274–275
		276–277			276–(277)	276–277
			288–295	277–287		278–280
				296–302		283–285
		302–303			(302)–303	(305)–306
Subtilisin BPN′	275			1–6		
		2–3				
		5–6				
			7–17			(9)–11
						(17)–19
		24–25		26–32	24–25	24–25
		38–39			38–39	37–38
		40–41	41–45		40–(41)	40–41
		48–49			48–49	
		52–53			52–53	52–53
		56–57			56–57	57–58

Table IV (Continued)

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be		Predicted by rule II[d]	Obsd[b,e] chain reversal
			Helical	Extended		
		61–62			61–62	61–62
		63–64			63–64	
		69–70		68–76		68–73
				81–87		84–85
		86–87				86–87
				88–97		98–100
		103–104			103–104	
		105–106	110–116		105–106	
		116–117			(116)–117	(116)–117
				121–124		120–121
		129–130	131–144	147–152	129–130	
		151–152				
		161–162			161–162	159–161
		168–169			168–169	167–170
		170–171			170–171	
		172–173		174–181	172–173	172–173
		183–184			183–184	182–183
		188–189			188–189	188–189
		190–191		190–193		194–195
		206–207	193–200	205–210		203–204
		210–211			(210)–211	211–212
		213–214		213–219		
		216–217				
		223–224	225–235		223–224	220–221
		237–238			237–238	(237)–238
		239–240			239–240	239–240
		248–249			248–249	
		251–252			251–252	(251)–253
		260–261			260–261	
			269–275			264–265
Staphylococcal nuclease	142	1–2			1–2	2–3
				10–16		4–5
				22–27		20–21
		31–32			31–(32)	27–29
		33–34		32–41		
						37–38
		42–43			42–43	
		45–46			45–46	
		47–48			47–48	47–49
		52–53			52–53	50–51
						53–54
			55–61			
		60–61	62–67			(61)–63
		67–68			(67)–68	(67)–68
		69–70	71–75		69–70	70–71
		82–83			82–83	84–85
		90–91		87–94		
		94–95	96–99		(94)–95	94–96
		109–110	100–104	108–115		
		112–113				
		117–118			117–118	116–119
				121–125		
			126–137			
Papain	212					3–4
						7–10
		17–18			17–18	
		21–22			21–22	20–21
		24–25	24–26	27–35		
		27–28	36–40			
		47–48	48–52			
		60–61			60–61	58–60
			67–77			62–64
		77–78			(77)–78	
		83–84			83–84	83–86
		85–86			85–86	
		87–88		91–95	87–88	
		92–93				
		97–98	102–107	110–114	97–98	
		115–116	118–122		115–116	115–116
		126–127	123–126		(126)–127	

Table IV (Continued)

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be		Predicted by rule II[d]	Obsd[b,e] chain reversal
			Helical	Extended		
		137–138		128–137	(137)–138	136–137
		139–140	140–143		139–(140)	139–(140)
		152–153		155–166	152–153	
		168–169			168–169	168–169
		174–175			174–175	
		176–177			176–177	179–180
		183–184			183–184	182–185
		196–197		186–189	196–197	196–197
				199–203		199–201
						202–203
		206–207			206–207	
Ferricytochrome C	104		1–4			
		7–8		8–22	7–(8)	(9)–13
		12–13				
		16–17				14–17
						22–23
		25–26			25–26	27–29
		30–31			30–31	33–34
						35–37
		44–45		44–49		43–45
		49–50			(49)–50	50–51
		51–52			51–52	
		53–54			53–54	
			56–66			62–63
		69–70	67–69		(69)–70	64–70
		71–72			71–72	
		73–74			73–74	(74)–75
		76–77	77–79		76–(77)	76–77
				80–85		79–80
		83–84				
		86–87		93–96	86–87	
		96–97	97–101			
		99–100				
		102–103			102–103	(100)–102
Cytochrome b_5	93[h]	5–6	7–14	1–6		
		14–15			14–15	(14)–15
		18–19			18–19	18–20
			31–38	20–30		26–27
						33–34
		38–39			38–39	
		40–41	41–50		40–(41)	40–41
		43–44				
						(48)–49
						50–51
		59–60	65–70		59–60	(60)–61
		71–72		72–77		
		73–74	78–80			(73)–74
		81–82			81–82	81–(82)
		85–86			85–86	
Myogen	108	1–2		1–6		
		3–4				3–5
			7–21			6–7
		9–10				
						(17)–18
		21–22			21–22	21–22
		23–24			23–24	
		31–32		28–36		32–33
		37–38			37–38	35–37
		40–41			40–41	
		44–45	43–45			
		46–47		46–50		
						(50)–51
		52–53			52–53	52–53
			56–58			56–57
						64–65
		68–69	69–73		68–(69)	66–67
						70–71
		72–73		74–78	72–73	72–73
		78–79			(78)–79	
		82–83		83–87	82–(83)	
						(88)–89
		91–92			91–92	91–92
		101–102		101–108		99–(100)

Table IV (Continued)

[a] In this table, the results are given for proteins whose helical and extended conformations had already been predicted in paper **3**.[5] [b] The references to the x-ray data for these proteins are given in Table I of this paper. [c] These predicted results are cited from Table III of paper **3**.[5] [d] The residues in parentheses are those that are duplicately assigned to be in chain-reversal conformations and at the ends of either helical or extended sequences. These are assigned to be helical or extended sequences. [e] See footnote *b* of Table I for the meaning of the parentheses in this column. [f] The predictions of helical and extended regions were made for oxyhemoglobin α and β chains in paper **3**.[5] However, the numbers in columns 4 and 5 are the same as those for oxyhemoglobin because the same amino acid sequence occurs in both oxy- and deoxyhemoglobins. [g] The prediction for α-chymotrypsin C chain was not carried out because a tosyl group is bound covalently to the side chain of Ser 195 (see also ref 49). The individual amino acid residues provide the statistical weights (from the numbers N_h, N_ϵ, etc.), and a tosylated residue (which is omitted from the data set) does not interfere with this computation. However, the prediction method (using the statistical weight matrix) requires operation on the *complete* sequence; for this purpose, the tosylated residue cannot be omitted. Since we have not introduced the tosylated residue into the statistical weight matrix, we cannot make any predictions for proteins containing this residue. This problem does not arise for noncovalently bound ligands, such as the heme group. [h] See footnote *q* of Table I.

with rule I above (by which a possible chain reversal can be selected[48]).

The purpose of rule I is solely to identify *possible* chain-reversal conformations, some of which may have been assigned (duplicately) to helical or extended sequences in paper **3**; however, such duplication is resolved by rule II (see below). Thus, the question of "testing" rule I should not be raised, since rule I is intended to be used only in combination with rule II. The duplication, referred to above, arises because (in this paper) we have used only first-order a priori probabilities for h, ϵ, R, S, and c states (and a second-order a priori probability for an RS state) to predict chain reversals by rule I. Such a duplication would not arise if we were to use higher-order probabilities, not only for RS, but also for h, ϵ, and c.

Rule II. Chain-reversal conformations assigned by rule I are discarded if the residues of a chain reversal are involved in helical or extended sequences predicted by the empirical rules of paper **3** (see section III of paper **3**). Inother words, the predictions of paper **3** for helical and extended sequences take preference over the prediction rule I introduced here for chain reversals. Thus, the predictions of helical and extended sequences in paper **3** are not altered by now taking chain-reversal conformations into consideration.[48]

In using the present rule II, a chain-reversal conformation predicted by rule I should be regarded only as a first possibility of a chain-reversal conformation rather than as a finally determined assignment. However, if one is interested *only* in the prediction of *possible* chain reversals, rule I is explicit enough to yield a final assignment of a chain reversal. Therefore, in order to assess the predictability of the present model, we will predict chain-reversal coformations using rule I, and also using rule II.

VI. Results and Discussion

The numerical values of the statistical weights relative to the ϵ state are tabulated in Table III. The statistical weights v_R^* and v_S^* provide an indication of the tendency of amino acids to form the chain-reversal conformation in proteins. By comparing v_R^* and v_S^*, it can be seen that amino acids such as Ala, Gln, Glu, Met, and Pro can contribute more to a chain-reversal conformation at position $i-1$ (R conformation) than at position i (S conformation) since $v_R^* > v_S^*$. The Pro residue shows the largest such tendency, viz., $v_R^* = 0.738$ and $v_S^* = 0.092$. On the other hand, amino acids such as Asn, Asp, Cys, Gly, His, Phe, Trp, and Tyr have the reverse preference, since $v_R^* < v_S^*$. The other amino acids such as Arg, Ile, Leu, Lys, Ser, Thr, and Val are impartial in this respect, since $v_R^* \simeq v_S^*$. Another interesting aspect of the conformational tendencies of amino acids is seen in the results for amino acids such as Asn ($v_S^* = 0.900$ and $w_h^* = 0.800$), Asp ($v_R^* = 0.784$, $v_S^* = 1.148$, and $v_\epsilon = 1.000$), Gly ($v_S^* = 0.915$ and $w_h^* = 0.638$), and Pro ($v_R^* = 0.738$ and $w_h^* = 0.400$), which have a relatively stronger tendency for a chain-reversal conformation compared to helical, extended, or other states than that found for other amino acids.

By using the four-state model formulated in sections I and II, the set of statistical weights evaluated in section IV, and rules I and II proposed in section V, predictions of chain-reversal conformations were made for 23 proteins.[49] To assess the predictability of the present prediction scheme, these proteins may be classified into four groups. The only purpose in introducing this classification is to distinguish whether the protein is involved in the original data set or not, and whether predicted results for h and ϵ from the three-state model are available from paper **3** (if not, then the experimentally observed h and ϵ structures are used; see Table V). As a first group (group 1), 13 (those in column 1 of Table VII) out of 23 proteins[49] were included in the original data set of both this paper and paper **3**[5] (see also Table XII of paper **1**[3]) to evaluate the statistical weights, and the predictions of helical and extended conformational sequences were reported in paper **3** (see also Table XII of paper **1**). Seven out of the 23 proteins, as a second group (group 2), were not included in the original data set in paper **3** (see also Table XII of paper **1**), but were included in the original data set in this paper (see Table I) to evaluate the statistical weights. No predictions were made in paper **3** for these proteins of group 2 (see column 5 of Table VII). In the third group (group 3) or proteins (see column 9 of Table VII), there are two proteins that were not included in the original data set of paper **3** (see also Table XII of paper **1**) and were included in the present original data set of proteins; however, the prediction of helical and extended sequences was carried out for these two proteins in paper **3**. Lastly, the new protein, in the sense that we did not include it in the original data set of proteins to evaluate the statistical weights in both paper **3** and the present paper, is classified into the fourth group (group 4), which contains only adenylate kinase.

The results for the prediction of chain-reversal conformations in the proteins of group 1 are summarized in Table IV, together with the observed chain-reversal regions. The predicted results, using rule I, are given in the third column of Table IV. To predict the chain-reversal conformation by rule II, the predictions of helical and extended sequences made in paper **3** (quoted from Table III of paper **3**) are given in columns 4 and 5 of Table IV.[50] Deleting the predicted chain reversals in the duplicately assigned regions (between helical and chain reversal or extended and chain reversal), by employing rule II, the results in column 6 of Table IV are obtained.

The results for the prediction of chain-reversal conformations in the proteins of group 2 are summarized in Table V. The predictions of chain-reversal conformations, using only rule I, are given in column 3 of Table V. To assess the pre-

Table V
Assessment of Predictability of Chain-Reversal Conformations

Protein[a]	No. of amino acid residues	Predicted by rule I	Obsd[b]		Predicted by rule II[c]	Obsd[d] chain reversal
			Helical	Extended		
Thermolysin	316	4–5		1–12		
						13–14
		18–19		16–23		
						25–27
		32–35			32–35	33–35
		37–38			37–38	36–37
		44–45		39–43	44–45	45–46
		51–52			51–52	50–51
		59–60		53–57	59–60	58–59
		64–65			64–65	65–67
			70–87			68–69
		73–74				
		77–78				
		85–86				
		88–89			88–89	87–88
		92–93			92–93	
		94–95			94–95	
		98–99		98–103		
		102–103				104–105
				112–115		108–109
		113–114				
		118–119		119–123	118–(119)	116–118
		124–125			124–125	
		126–127			126–127	127–129
			139–150			133–136
						137–138
		149–150				
		153–154			153–154	(150)–153
		158–159			158–159	159–(160)
			160–173			
		169–170				
			175–179			173–174
		177–178				
		180–181			180–181	(179)–180
		182–183			182–183	182–183
		184–185			184–185	
						188–189
		190–191			190–191	190–192
		195–196			195–196	195–196
						198–199
		206–207			206–207	205–207
						208–211
		214–215			214–215	
		218–219			218–219	217–219
		226–227			226–227	225–229
			234–244			230–232
						233–(234)
		241–242			241–242	(244)–246
		249–250		254–257	249–250	250–252
		260–261	260–274			
		277–278			277–278	277–278
		279–280	281–296		279–280	
		293–294				
			302–312			298–299
		304–305				301–(302)
						312–(313)
Rubredoxin	54	2–3		2–6		
			15–17			7–9
		20–21			20–21	20–22
		26–27			26–27	26–27
		28–29			28–29	
		31–32	30–32			
		34–35			34–35	35–36
				49–52		40–42
						46–48
Cytochrome C_2	112			None		2–(3)
		8–9	3–14			
		11–12				
		16–17			16–17	15–17
		19–20			19–20	22–23
		25–26			25–26	27–29

Table V (Continued)

Protein[a]	No. of amino acid residues	Predicted by rule I	Obsd[b] Helical	Obsd[b] Extended	Predicted by rule II[c]	Obsd[d] chain reversal
		30–31			30–31	33–34
		37–38			37–38	36–37
						40–41
		44–45			44–45	44–45
		47–48			47–48	
		51–52			51–52	50–53
			54–56			(56)–58
						64–(65)
		69–70	65–70			
		72–73			72–73	(70)–73
		80–81			80–81	74–81
						85–86
		89–90			89–90	
		102–103	97–105			
		106–107			106–107	(105)–111
Ferredoxin	54	1–2		2–5		
						6–7
		10–11			10–11	
		13–14			13–14	
			15–17			19–20
						26–27
		32–33			32–33	33–34
		40–41			40–41	40–44
		42–43			42–43	
						46–47
		50–51			50–51	
Trypsin	223	9–10			9–10	9–10
						11–12
		13–14			13–14	13–14
				15–18		20–21
				25–28		30–31
				34–37		32–33
		39–40			39–40	39–41
		53–54			53–54	52–54
						55–56
		67–68			67–68	
		74–75			74–75	74–75
			78–80	84–89		91–92
		98–99			98–99	97–99
		101–102		101–106		
		108–109			108–109	107–108
		110–111			110–111	
		112–113			112–113	112–113
				116–120		124–127
		130–131			130–131	128–129
		132–133		135–142	132–133	133–134
		144–145			144–145	144–145
		146–147	146–151			
		149–150				
		151–152			(151)–152	(151)–152
		153–154			153–154	
		158–159			158–159	158–159
		160–161			160–161	
		170–171			170–171	167–168
						174–175
						177–178
						183–185
		192–193			192–193	195–197
		200–201			200–201	198–201
		203–204			203–204	
			214–221			209–213
						(221)–222
Glyceraldehyde	333	1–2		2–5	1–2	
phosphate		6–7			6–7	
dehydrogenase		21–22	11–23			10–11
		24–25			24–25	
		32–33	36–44	28–31	32–33	32–33
		42–43				
		44–45	47–49		44–45	44–45
		48–49				

Table V (Continued)

Protein[b]	No. of amino acid residues	Predicted by rule I	Predicted[c] to be: Helical	Predicted[c] to be: Extended	Predicted by rule II[d]	Obsd[b,e] chain reversal
				55–58		53–54
		59–60			59–60	59–61
		66–67			66–67	65–67
		68–69			68–69	
			78–80	70–73		76–77
		82–83			82–83	
		84–85			84–85	83–86
		86–87			86–87	
		88–89			88–89	89–90
		93–94	101–108		93–94	96–98
		112–113			112–113	108–111
		120–121		114–118	120–121	
		122–123	129–132	125–128	122–123	122–123
		134–135			134–135	133–134
		138–139			138–139	138–140
				143–147		141–142
		144–145				
		146–147	149–163			148–149
		150–151				(163)–164
				169–173		179–180
		185–186			185–186	183–184
		187–188			187–188	
		190–191			190–191	190–191
		198–199			198–199	198–200
		200–201		202–207	200–201	
		204–205				
		210–211	210–220			209–210
						(215)–217
		221–222			221–222	221–223
		228–229		224–231		
		232–233		236–245	232–233	
		234–235			234–235	
		248–249				
		250–251			250–(251)	
		252–253	251–263			
						264–266
						267–269
		275–276			275–276	276–277
		280–281			280–281	281–282
		285–286			285–286	283–284
		288–289			288–289	288–289
		293–294			293–294	293–294
		295–296		297–300	295–296	295–296
		300–301		303–310	300–301	301–302
		308–309				
		311–312	312–327		311–(312)	
		313–314				
		318–319				
			329–332			(327)–328
High potential iron protein	85	4–5			4–5	4–5
		10–11	12–16		10–11	9–10
		18–19			18–19	16–17
		21–22			21–22	21–22
			28–31			23–25
		32–33			32–33	
		34–35			34–35	
						38–40
		44–45			44–45	43–45
		47–48			47–48	47–48
		51–52			51–52	51–52
						54–55
		57–58		60–63	57–58	58–59
						64–66
		67–68			67–68	67–68
				69–72		73–74
		78–79			78–79	78–79
				80–83		

[a] The references to the x-ray data for these proteins are given in Table I. [b] The locations of observed helical and extended regions are cited from columns 4 and 5 of Table I. [c] The results of prediction under the assumption that the predictions for helical and extended conformations were made with 100% accuracy. [d] From column 6 of Table I.

dictability of the present model, the helical and extended regions observed by x-ray experiments were used instead of those obtained by a prediction method; in other words, we used the observed x-ray data and thus, effectively, assumed that the prediction of helical and extended conformations was made with 100% accuracy, when rule II was applied. Hence,

Table VI
The Results of Predicted and Experimentally Observed Chain-Reversal Conformation Regions of Proteins

Protein[a]	No. of amino acid residues	Predicted by rule I	Predicted to be[b]		Predicted by rule II	Obsd[c] chain reversal
			Helical	Extended		
Bovine pancreatic[d] trypsin inhibitor	58	2–3	4–7		2–3	(5)–6
		9–10			9–10	
		25–26	13–16	17–21		
		40–41	23–28	29–36	40–41	
		42–43			42–43	42–43
		46–47	44–50			
		49–50				
		54–55		51–54	(54)–55	(54)–55
Clostridial[d] flavodoxin	138	7–8	11–21	1–7	(7)–8	8–9
		28–29	22–25		28–29	
		36–37			36–37	35–36
			39–44			40–42
						43–45
		46–47		46–53		47–48
						57–59
		65–66	68–73		65–66	
		68–69				
		75–76			75–76	(74)–76
		78–79			78–79	78–79
		80–81		80–87		
		87–88			(87)–88	89–90
		92–93	93–96		92–93	92–93
		94–95	98–101			
		103–104		106–119	103–104	
		114–115				
		118–119				
						123–124
		130–131			130–131	
		133–134	132–136			
		136–137			136–137	(136)–137
Adenylate kinase[e,f]	194	2–3	1–6			g
		6–7			(6)–7	
		8–9			8–9	
		11–12		10–15		
		20–21			20–21	
		23–24	24–29		23–(24)	
		25–26	40–47	64–72		
		83–84	54–63			
		89–90	73–76	89–92		
		92–93	77–84		(92)–93	
		103–104	96–101			
		105–106	102–107			
		108–109			108–109	
		112–113		111–119		
		114–115				
		122–123	122–125			
		125–126		126–130		
		135–136			135–136	
		140–141			140–141	
		142–143	145–149		142–143	
		146–147				
		149–150		150–153		
		151–152	154–156	157–165		
		168–169		167–173		
		180–181	174–176	182–190	180–181	
		182–183				
		186–187				
		188–189				

[a] The references to the x-ray data for these proteins are given in Table I. [b] These predicted results are cited from Table IV of paper **3**.[5] [c] These regions are cited from column 6 of Table I. [d] These proteins were not included in the original data set to evaluate the statistical weights in papers **1**[3] and **3**,[5] but were included in the present data set, as seen in Table I (group 3). [e] This protein was not included in the original data sets in both paper **1**[3] (or **3**[5]) and the present paper (group 4). [f] From G. E. Schulz, M. Elzinga, F. Marx, and R. H. Schirmer, *Nature (London)*, **250**, 120 (1974). [g] The values of ϕ and ψ were not determined since the x-ray coordinates were not available to us.

the chain-reversal conformations that were duplicately assigned between helical regions (given in column 4 of Table V) or extended regions (given in column 5 of Table V) and chain-reversal conformations predicted by rule I (given in column 3) were omitted to obtain the results given in column 6 (from rule II). In the last column of Table V, the chain-reversal regions observed by x-ray experiments are given for comparison.

The results for the prediction of chain-reversal conformations in the proteins belonging to group 3, together with those for adenylate kinase (group 4), are given in Table VI. The predictions, using only rule I, are summarized in column 3. Those obtained by using rule II are given in column 6 [by omitting the chain-reversal regions assigned duplicately with the helical and extended sequences[50] predicted in paper **3** (cited from columns 4 and 7 of Table IV of paper **3**)]. The

Table VII
Summary of Predicted and Experimentally Observed Chain-Reversal Regions

Results of Table IV				Results of Table V				Results of Table VI			
Protein[a]	$n_{(obsd)}$	$n_{(cor)}$	$n_{(over)}$	Protein[a]	$n_{(obsd)}$	$n_{(cor)}$	$n_{(over)}$	Protein[b]	$n_{(obsd)}$	$n_{(cor)}$	$n_{(over)}$
Myoglobin	8	6	5	Thermolysin	37	21	6	Bovine pancreatic trypsin inhibitor	3	3	2
Lysozyme	18	11	5	Rubredoxin	6	3	1				
Ribonuclease S	8	5	5	Cytochrome C_2	15	10	2	Clostridial flavodoxin	12	7	4
Deoxyhemoglobin				Ferrodoxin	6	2	3				
α chain	19	12	5	Trypsin	28	16	5				
β chain	16	5	4	Glyceraldehyde phosphate dehydrogenase	34	22	10				
α-Chymotrypsin B chain	17	10	2								
Carboxypeptidase A	34	19	15	High potential iron protein	15	10	2				
Subtilisin BPN'	27	17	7								
Staphylococcal nuclease	14	10	2								
Papain	15	10	8								
Ferricytochrome C	14	8	3								
Cytochrome b_5	10	5	2								
Myogen	16	7	4								
Subtotals	216	125	67		141	84	29		15	10	6

$\Sigma n_{(obsd)} = 372$, $\Sigma n_{(cor)} = 219$, $\Sigma n_{(over)} = 102$

[a] The references to the x-ray data for these proteins are given in Table I. [b] See also footnotes *e* to *g* of Table VI.

observed chain-reversal regions are given in the last column for comparison (except for adenylate kinase whose x-ray coordinates were not available to us).

A summary of predicted and experimentally observed chain-reversal regions, given in Tables IV (for the proteins of group 1), V (for those of group 2), and VI (only for group 3), is provided in Table VII. The symbols $n_{(obsd)}$, $n_{(cor)}$, and $n_{(over)}$ refer to the number of chain-reversal regions that are observed, predicted correctly, and over predicted, respectively. As seen in Table VII, for the proteins of group 1, 125 chain-reversal regions (out of 216) were assigned correctly, with an over prediction (i.e., predicted but not observed experimentally) of 67 regions; for group 2, 84 chain-reversal regions out of 141 were predicted correctly, with 29 over predictions; for group 3, 10 out of 15 were predicted correctly, with 6 over predictions. In total, 219 chain-reversal regions out of 372 were predicted correctly, with 102 over predictions.

There are several other methods for predicting chain-reversal conformations in proteins. As far as we know, these are the procedures of Lewis et al.,[6b] Burgess et al.,[40] and Chou and Fasman.[51] All of these authors used a probability for a *single* residue to be found in a chain-reversal conformation to calculate the probability of a tetrad[6b,51] or nonamer[40] to be found in a chain reversal, by multiplying the probabilities for single residues. On the other hand, we have developed a statistical mechanical treatment of polypeptide chains to treat the conformations of protein molecules (focusing particularly on the chain-reversal conformation in this paper, and on helical and extended conformations in papers **1**, **2**, and **3**). In the models presented in these papers, we use the *statistical weights*, and not the probabilities (see ref 60 of paper 1 for a discussion of the difference between a statistical weight and a probability of occurrence of a certain conformational state). Using the statistical weights, we calculate the *probability* of finding a chain-reversal conformation by using a statistical mechanical procedure for averaging over a *whole molecule*, as has been described in section VI of paper **2**,[4] and briefly in section VA of this paper. The difference between a statistical mechanical treatment, such as that presented in this paper, and a prediction method using a probability as presented in ref 6b, 40, and 51, was discussed in section VIC of paper **2**.[4]

As described in this section, there appears to be a close relationship between the regions of high probability of finding a chain-reversal conformation and those observed by x-ray experiments. These results suggest that the assumption of the dominance of short-range interactions in determining a chain-reversal conformation in proteins,[6b,7] on which the predictive scheme presented in this paper is based, is a reasonable one. When more x-ray structures become available, it will be necessary to test this four-state model further, by applying it to proteins (other than adenylate kinase) that were not included in the original data set to evaluate the statistical weights.

However, it is also true that the present four-state model, as well as the three-state model,[4] is not sufficient to describe protein conformation completely in the following three respects. (i) First, even within the framework of a short-range one-dimensional model, a region of conformational space, viz., the other (i.e., c) state, does not provide a precise enough definition of the chain conformation. In other words, even if the c states were predicted with a high accuracy, the structure of a protein could not be determined, because the conformations of the residues predicted to be in c states can vary within the wide range of the conformational space of the c state. In order to predict the structure of a native protein, it will be necessary to divide the conformational space [in the present context, the space of the other (c) state] as finely as possible. In other words, it is necessary to remove other conformational states that occur in proteins from the other (c) state. This

point (i) will be discussed in a subsequent paper,[44] in which a multi-state model has been developed to provide a more detailed treatment of protein molecules. (ii) Second, it should be noted that a broad range of ϕ and ψ was used to define the α-helical state. A wide range of values was selected because the α helices observed in x-ray structures are rarely regular; if the observed data for α-helical structures fell in a narrower range, we would have reduced our defined α-helical range correspondingly. In any event, the imposition of a restriction of regularity is not necessary in a theoretical computation of protein structure, since such a restriction is removed subsequently when the energy (see ref 22 of paper 2[4]) or the free energy (see footnote ¶ of ref 52) of the whole protein is calculated. The same comments apply to the definitions of the extended region and to the chain-reversal conformation. (iii) Third, no attention is paid to long-range interactions in the one-dimensional nearest-neighbor model. The description of protein conformation can be improved, and the conformation of the protein altered, by introducing the long-range interactions that are not included in the nearest-neighbor interaction model.[52] (See also Addendum of paper 3.[5])

Acknowledgment. We acknowledge the Brookhaven Data Bank for the x-ray coordinates of some of the proteins used in this work. We are indebted to Shirley Rumsey for filing the x-ray coordinates, and also to Marcia Pottle for her general management of the computer facilities used in this laboratory.

Appendix

Nearest-Neighbor Four-State Model with Asymmetric Nucleation of Helical Sequences. We recently[11] 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). This model was applied to the form I $\rightleftarrows$ form II transition of poly(L-proline), in which asymmetric nucleation played a role.[12,13] Furthermore, in paper 1,[3] asymmetric nucleation was observed for the amino acids in proteins. Therefore, we now incorporate the asymmetric nucleation property[11] of amino acids into the nearest-neighbor four-state model [it should be noted that the term "nearest neighbor" is used in such a way that the interactions beyond nearest neighbor are not included, even though the matrix is constructed for three consecutive residues $i-1$, i, and $i+1$ (see ref 11 and section V of paper 2[4])].

In this Appendix, and in this Appendix only, we use the c state as a reference, instead of the ϵ state. Of course, if desired, one could convert a set of relative statistical weights with a c state as a reference to a set with an ϵ state as a reference.

Following the assumptions made about the ϵ state in the third paragraph of section V of paper 2,[4] the statistical weight for the extended state was introduced as

$$q_9 = v_\epsilon/u_c \tag{A-1}$$

in addition to q_1 to q_8 defined in our earlier model (see the summary of Table I of ref 11 for the definitions of q_1 to q_8). For R and S states, the statistical weights q_{10} and q_{11} relative to the c state are given by

$$q_{10} = v_R/u_c \tag{A-2}$$

and

$$q_{11} = v_S/u_c \tag{A-3}$$

Using the statistical weights q_1 to q_{11}, the statistical weight matrix for the four-state model with asymmetric nucleation of helical sequences can be constructed as

$$\mathbf{W}_i = \left[\begin{array}{cc|cccccccccc} i-1 & i+1 \\ & i & \begin{array}{c}\text{cU}\epsilon\text{UR}\\ \text{c}\end{array} & \begin{array}{c}\text{h}\\ \text{c}\end{array} & \begin{array}{c}\text{cU}\epsilon\text{UR}\\ \epsilon\end{array} & \begin{array}{c}\text{h}\\ \epsilon\end{array} & \begin{array}{c}\text{cU}\epsilon\text{UR}\\ \text{h}\end{array} & \begin{array}{c}\text{h}\\ \text{h}\end{array} & \begin{array}{c}\text{cU}\epsilon\text{UR}\\ \text{S}\end{array} & \begin{array}{c}\text{h}\\ \text{S}\end{array} & \begin{array}{c}\text{S}\\ \text{R}\end{array} \\ \hline \text{c} & \text{cU}\epsilon\text{UR} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} \\ \text{c} & \text{h} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 \\ \epsilon & \text{cU}\epsilon\text{UR} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} \\ \epsilon & \text{h} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 \\ \text{h} & \text{cU}\epsilon\text{UR} & q_5 & q_3 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} \\ \text{h} & \text{h} & 0 & 0 & 0 & 0 & q_2 & q_1 & 0 & 0 & 0 \\ \text{S} & \text{cU}\epsilon\text{UR} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} \\ \text{S} & \text{h} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 \\ \text{R} & \text{S} & 0 & 0 & 0 & 0 & 0 & 0 & q_{11} & q_{11} & 0 \end{array}\right]_i \tag{A-4}$$

Equation A-4 can be obtained by constructing a matrix like that given in eq 1, where the statistical weights q_1 to q_{11} are substituted for u_c, v_h, w_h, etc., in eq 1, and then contracting. From the statistical weights for the allowed conformational states of the first (i.e., amino terminal[9]) residue, we obtain the vector $\mathbf{t}_1$ as

$$\mathbf{t}_1 = [q_8 \quad q_7 \quad q_9 \quad q_9 \quad q_6 \quad q_4 \quad 0 \quad 0 \quad q_{10}]_1 \tag{A-5}$$

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 four rows in eq A-4). For the last (i.e., carboxyl terminal[9]) residue, we obtain

$$\mathbf{t}_N{}^* = \begin{bmatrix} q_8 + q_9 + q_{10} \\ q_6 \\ q_8 + q_9 + q_{10} \\ q_6 \\ q_5 + q_9 + q_{10} \\ q_2 \\ q_8 + q_9 + q_{10} \\ q_6 \\ q_{11} \end{bmatrix}_N \tag{A-6}$$

where each element corresponds to the carboxyl-terminal residue in a state where residue $i+1$ is in state c U ϵ U R, c U ϵ U R, c U ϵ U R, c U ϵ U R, and S.

Using eq A-4, A-5, and A-6, the partition function can be written as

$$Z = \mathbf{t}_1 \left[\prod_{i=2}^{N-1} \mathbf{W}_i\right] \mathbf{t}_N{}^* \tag{A-7}$$

or

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i\right] \mathbf{e}_N{}^* \tag{A-8}$$

where

$$\mathbf{e}_1 = [1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \tag{A-9}$$

and

$$\mathbf{e}_N^* = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \qquad \text{(A-10)}$$

since

$$\mathbf{t}_1 = \mathbf{e}_1\mathbf{W}_1 \qquad \text{(A-11)}$$

and

$$\mathbf{t}_N = \mathbf{W}_N\mathbf{e}_N^* \qquad \text{(A-12)}$$

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 (BMS75-08691).

(2) (a) From Kyoto University, 1972–1975. (b) Direct requests for reprints to this author.

(3) Paper 1: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 142 (1976).

(4) Paper 2: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 159 (1976).

(5) Paper 3: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 168 (1976).

(6) In this paper we use the term chain reversal to mean the same as the terms β turn or β bend. Such conformations have been described in the following papers. (a) C. B. Venkatachalam, *Biopolymers*, **6**, 1425 (1968); (b) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **68**, 2293 (1971); (c) I. D. Kuntz, *J. Am. Chem. Soc.*, **94**, 4009 (1972); (d) P. N. Lewis, F. A. Momany, and H. A. Scheraga, *Biochim. Biophys. Acta*, **303**, 211 (1973); (e) J. L. Crawford, W. N. Lipscomb, and C. G. Shellman, *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 538 (1973); (f) K. Nishikawa, F. A. Momany, and H. A. Scheraga, *Macromolecules*, **7**, 797 (1974); (g) S. S. Zimmerman and H. A. Scheraga, *Macromolecules*, **9**, 408 (1976).

(7) H. A. Scheraga, *Pure Appl. Chem.*, **36**, 1 (1973).

(8) IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, **9**, 3471 (1970).

(9) The IUPAC-IUB definition of a residue[8] is used throughout this paper, i.e., a residue extends from the NH to the CO group.

(10) See ref 9 of paper 3.[5]

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

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

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

(14) Some of the x-ray data were obtained from the Brookhaven Data Bank. The original papers describing the data for these proteins are cited in ref 15–39.

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

(16) (a) C. C. F. Blake, D. F. Koenig, G. A. Mair, A. C. T. North, D. C. Philips, and V. R. Sarma, *Nature (London)*, **206**, 757 (1965); (b) D. C. Philips, *Proc. Natl. Acad. Sci. U.S.A.*, **57**, 484 (1967).

(17) H. W. Wyckoff, D. Tsernoglou, A. W. Hanson, J. R. Knox, B. Lee, and F. M. Richards, *J. Biol. Chem.*, **245**, 305 (1970).

(18) As far as we are aware, the coordinates of deoxyhemoglobin have not been published. They were obtained from the Brookhaven Data Bank.

(19) (a) B. W. Matthews, P. B. Sigler, R. Henderson, and D. M. Blow, *Nature (London)*, **214**, 652 (1967); (b) J. J. Birktoft and D. M. Blow, *J. Mol. Biol.*, **68**, 187 (1972).

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

(21) (a) C. S. Wright, R. A. Alden, and J. Kraut, *Nature (London)*, **221**, 235 (1969); (b) R. A. Alden, J. J. Birktoft, J. Kraut, J. D. Robertus, and C. S. Wright, *Biochem. Biophys. Res. Commun.*, **45**, 337 (1971).

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

(23) 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).

(24) (a) R. E. J. Mitchel, I. M. Chaiken, and E. L. Smith, *J. Biol. Chem.*, **245**, 3485 (1970); (b) J. Drenth, J. N. Jansonius, R. Koekoek, and B. G. Wolthers, *Adv. Protein Chem.*, **25**, 79 (1971).

(25) R. E. Dickerson, T. Takano, D. Eisenberg, O. B. Kallai, L. Samson, A. Cooper, and E. Margaliash, *J. Biol. Chem.*, **246**, 1511 (1971).

(26) M. G. Rossman, M. J. Adams, M. Buehner, G. C. Ford, M. L. Hackert, P. J. Lentz, Jr., A. McPherson, Jr., R. W. Shevitz, and I. E. Smiley, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 179 (1971).

(27) (a) F. S. Matthews, M. Levine, and P. Argos, *J. Mol. Biol.*, **64**, 449 (1972); (b) F. S. Matthews, P. Argos, and M. Levine, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 387 (1972).

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

(29) (a) G. M. Edelman, B. A. Cunningham, G. N. Reeke, Jr., J. W. Becker, M. J. Waxdal, and J. L. Wang, *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 2580 (1972); (b) G. N. Reeke, Jr., J. W. Becker, and G. M. Edelman, *J. Biol. Chem.*, **250**, 1525 (1975).

(30) C. E. Nockolds, R. H. Kretsinger, C. J. Coffee, and R. A. Bradshaw, *Proc. Natl. Acad. Sci. U.S.A.*, **69**, 581 (1972).

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

(32) K. D. Watenpaugh, L. C. Sieker, J. R. Herriott, and L. H. Jensen, *Acta Crystallogr., Sect. B*, **29**, 943 (1973).

(33) F. R. Salemme, S. T. Freer, Ng. H. Xuoung, R. A. Alden, and J. Kraut, *J. Biol. Chem.*, **248**, 3910 (1973).

(34) E. T. Adman, L. C. Sieker, and L. H. Jensen, *J. Biol. Chem.*, **248**, 3987 (1973).

(35) R. M. Stroud, L. M. Kay, and R. E. Dickerson, *J. Mol. Biol.*, **83**, 185 (1974).

(36) (a) R. Huber, D. Kukla, A. Ruehlman, and W. Steigemann, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 141 (1972); (b) R. Huber, private communication.

(37) M. Buehner, G. C. Ford, D. Moras, K. W. Olsen, and M. G. Rossman, *J. Mol. Biol.*, **82**, 563 (1975).

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

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

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

(41) In the present evaluation of w_h and v_h, *all* residues of a helical sequence contribute to w_h, and only *isolated* helical residues contribute to v_h; thus, the contributions of residues in a helical sequence to helix nucleation (i.e., to v_h) are neglected. This procedure was adopted for the following reasons: (i) Even with the x-ray coordinates available, it is difficult to establish precisely which is the first and last residue of a helical sequence; thus, every residue of a helical sequence is treated as an interior one, in evaluating statistical weights (w_h), because of the lack of information about the locations of the hydrogen bonds. (ii) One cannot determine (by observation of a helical sequence) whether it was nucleated at the N or C terminus or in the interior of the sequence; despite the omission of the contribution of helical sequences to v_h, and the aforementioned uncertainty as to the site of helix nucleation, we nevertheless assign the parameter v_h to the N and C termini of a helical sequence in the theoretical formulation of the model (i.e., in the statistical-weight matrix, as in eq 1 or 2, for example). In a homopolymer the statistical weight of a helical sequence of i residues is v^2w^{i-2}, and it does not matter which two residues of a helical sequence are assigned the statistical weight v. However, in a specific-sequence copolymer where v_h as well as w_h depends on the type j of the amino acid residue, one would have to know where the helix was nucleated in order to know which two residues should be assigned the statistical weight v_h. Since we do not have this information, we arbitrarily assign v_h to the residues at the N and C termini, in the statistical-weight matrix (see the discussion in section IB of paper 3). (iii) By omitting the contribution of helix nucleation of helical sequences from the values of v_h, we have underestimated the values of v_h. However, from our experience in trial calculations, we have found that this underestimate does not significantly affect the values of $P_{i;h}^*$, $P_{i;\epsilon}^*$, $P_{i;R}^*$, $P_{i;S}^*$, and $P_{i;c}^*$ (defined in section V); i.e., the final predicted results are insensitive to the nucleation parameter v_h. This point will be discussed in more detail in a future publication.

(42) The statistical weight $v_{\epsilon,j}$ introduced in sections I and II does not involve cooperativity, in the sense that no interactions beyond the single residue under consideration are taken into account. Therefore, we do not distinguish between the ϵ states of an extended sequence and those in isolated extended states (or in extended sequences shorter than four residues).[43] To save space in Table I, we did not list ϵ sequences shorter than four residues; however, they *are included* in the values of $N_{\epsilon,j}$ listed in Table II. (For similar reasons, only helical sequences of three or more residues are listed in Table I, but the short, h and hh, sequences appear in $N_{h',j}$ of Table II.) When applying rule II of section VB, ϵ *sequences* (and h *sequences*) are predicted by the three-state model of papers 1–3, and chain reversals by the four-state model of this paper. However, this does not mean that ϵ sequences shorter than four residues (or h sequences shorter than three residues) are assigned to the c state. If one were interested in computing the appropriate probabilities, one could predict whether those residues (not assigned to ϵ *sequences*, h *sequences*, or chain reversals) are in ϵ, $\epsilon\epsilon$, $\epsilon\epsilon\epsilon$, h, hh, or c states; this will be done in our forthcoming multistate model.[44] Therefore, one should *not* regard those parts of a protein chain (not predicted to be in ϵ or h sequences, or chain reversals) as being in c states.

(43) In paper 1,[3] we used only the number of residues in ϵ *sequences*, rather than $N_{\epsilon,j}$, since it was impossible to detect isolated ϵ states because the x-ray *coordinates* were not available to us. Nevertheless, it is expected that the prediction results obtained in paper 3[5] are valid, since the same criteria that were used in evaluating the statistical weights in paper 1 were used in paper 3[5] when predictions were made about the ϵ state in proteins. In other words, the statistical weights deduced from information about the ϵ sequences were used to predict ϵ sequences.

(44) S. Tanaka and H. A. Scheraga, *Macromolecules*, submitted (multistate model), paper 5.

(45) It should be noted that, for the present analysis, any of the definitions of a chain-reversal conformation proposed by other authors[6] could be used.

(46) As in papers 1–3, we use j to designate the species of amino acid (j = 1 to 20, as seen in column 1 of Table II) and i to designate the position of an amino acid in the protein chain.

(47) S. Tanaka and A. Nakajima, *Macromolecules*, **5**, 714 (1972).
(48) One can use the present four-state model (h, ϵ, chain reversal, and c state) to predict these conformations, without recourse to the three-state model of papers **1–3**, because all of these states are included explicitly in the matrix of eq 17, for example. However, it is more convenient (and requires a smaller-size matrix) to first use the three-state model (h, ϵ, and c) to identify h and ϵ, and then (with the redefinition of c in the four-state model) use the four-state model to identify the chain-reversal and (new) c states.
(49) Predictions are made in this paper for 23 proteins [the sum of 13 (group 1), 7 (group 2), and 2 (group 3) out of the 26 proteins listed in Table I, and the 1 protein (group 4) not listed in TableI]. As to other proteins, no prediction is made since, for two proteins (lactate dehydrogenase and concanavalin), the amino acid sequences have not been determined completely, and only the known parts of the sequences (and their x-ray structures) were used to compute statistical weights. No prediction is made for sea lamprey hemoglobin simply because there are many other globin homologues as seen in column 1 of Table I, and this protein was omitted to save computer time. As for the α-chymotrypsin C chain, see footnote *g* of Table IV. Elastase also includes a tosyl residue, and no prediction was made for it. However, a prediction *was* made for the B chain of α-chymotrypsin.
(50) These regions of helical and extended sequences predicted in paper **3**[5] should be regarded as tentative assignments, since the statistical weights used in paper **3** are tentative as stated in paper **3**.[5]
(51) P. Y. Chou and G. D. Fasman, *Biochemistry*, **13**, 222 (1974).
(52) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 3802 (1975).

Structural Variations and Multiple Charge Transfer Transitions between Chloranil and Carbazole Derivatives

Uzi Landman, A. Ledwith, D. G. Marsh,* and D. J. Williams

Xerox Webster Research Laboratories, Rochester, New York 14580.
Received May 17, 1976

ABSTRACT: Asymmetric charge transfer (CT) spectra from combinations of carbazole derivatives and chloranil are shown to consist of two bands originating from the highest (HOMO) and second highest (HOMO 2) energy occupied molecular orbitals of carbazole. Symmetry arguments are used to indicate that of two possible (parallel plane) alignments of donor and acceptor, that which is totally symmetric gives rise to only the lower energy CT transition whereas the unsymmetrical alignment which is energetically preferred permits the two observable CT transitions. Low molecular weight model carbazoles all show the asymmetric CT bands which have been resolved into two Gaussian components by means of a computer assisted analysis. Significantly, poly(*N*-vinylcarbazole) (PVCA) gives rise to CT bands much less asymmetric than corresponding model systems and it is concluded that steric interactions in PVCA greatly reduce the possibility for interaction of the carbazole units with chloranil in the 1:1 asymmetric arrangement preferred by the model compounds. As expected from the theoretical argument, poly(*N*-ethyl-3-vinylcarbazole) and poly(*N*-ethyl-2-vinylcarbazole), both of which are derived from unsymmetrically substituted carbazoles, give with chloranil highly asymmetric CT spectra which are very similar to those of the appropriate model compounds.

Carbazole and its ring and *N*-alkylated derivatives like other aromatic amines exhibit long wavelength charge transfer (CT) transitions with acceptors because of relatively high energies for the highest occupied molecular orbitals.[1] This property is manifest in low ionization potentials,[1] low oxidation potentials (~1.2 V (SCE)),[2] and a very high propensity to oxidative coupling.[3]

Observation of CT transitions when organic electron donors and acceptors are allowed to interact is now a commonplace phenomenon,[4] it being frequently concluded that these transitions arise from so-called CT complexes without proper regard for the magnitude and nature of the binding forces in such complexes. Carbazoles, no less than other good organic electron donor molecules, readily participate in this type of intermolecular association with electron acceptors, the required close approach of donor and acceptor being favored by the planarity of the carbazole ring system. Particular interest in the formation and properties of CT complexes of carbazole derivatives arise from (mainly two) quite different types of study. Hoegl[5] was the first to show that poly(*N*-vinylcarbazole) (PVCA) was a useful organic photoconductor and, more importantly, light absorption and photoconductivity were improved when PVCA was mixed with a variety of organic acceptor molecules. Several years later Ellinger[6] showed that many organic electron acceptors, including chloranil, were useful initiators for the polymerization of monomeric *N*-vinylcarbazole (NVCA). Related observations were published independently at about the same time by Scott, Miller, and Labes.[7]

These early disclosures stimulated research studies of the formation and photoelectrical properties of CT complexes of PVCA, leading ultimately[8,9] to a commercial process for electrophotography based on compositions formed from PVCA and 2,4,7-trinitrofluorenone (TNF). In complete contrast, and despite extensive studies by several groups of workers,[10] the reactions of NVCA with organic electron acceptors have not yet resulted in developments significant in other than a purely mechanistic sense; a critical survey of the scope and value of such studies has been given recently by Hyde and Ledwith.[11]

Reactions of NVCA and chloranil have been extensively investigated. Originally it was claimed by Ellinger[6] that chloranil was a useful initiator for cationic polymerization of NVCA in toluene, but subsequently other workers[12] found that purified chloranil was inactive in this particular system. Rather, it was shown that the strong protonic acid, 1,4-dihydroxy-2,3,5,6-tetrachlorobenzene (the dihydro reduction product of chloranil), thought to be an impurity in chloranil, was the true initiator. In more polar solvents the situation is even more confused.[13] Another original observation by Ellinger[6] was that mixtures of NVCA and chloranil in acetone gave rise to formation of the cyclodimer of NVCA when exposed to uv light or strong sunlight. This result has been amply confirmed by subsequent studies[14,15] and it is now known that cyclodimerization of NVCA oc-