(26) See paragraph at end of paper regarding supplementary material.
(27) A. W. Burgess, F. A. Momany, and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 1456 (1973).
(28) K. Wüthrich, Ch. Grathwohl, and R. Schwyzer, "Peptides, Polypeptides, and Proteins", E. R. Blout, F. A. Bovey, M. Goodman, and N. Lotan, Ed., Wiley, New York, N.Y., 1974, p 300.
(29) S. S. Zimmerman and H. A. Scheraga, "Peptides: Chemistry, Structure, and Biology", R. Walter and J. Meienhofer, Ed., Ann Arbor Science Publishers, Ann Arbor, Mich., 1975, p 263.
(30) Y. Grenie, M. Avignon, and C. Garrigou-Lagrange, *J. Mol. Struct.*, **24**, 293 (1975).
(31) G. M. Crippen and J. T. Yang, *J. Phys. Chem.*, **78**, 1127 (1974).
(32) A. W. Burgess and H. A. Scheraga, *Biopolymers*, **12**, 2177 (1973).
(33) See, for example, G. N. Ramachandran, ref 28, p 14.
(34) See, for example, B. Pullman and B. Maigret, *Jerusalem Symp. Quantum Chem. Biochem.*, **5**, 13 (1973).
(35) B. K. Vijayalakshmi and R. Srinivasan, *Acta Crystallogr., Sect. B*, **31**, 999 (1975).

Statistical Mechanical Treatment of Protein Conformation. 5. A Multistate 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 March 19, 1976

ABSTRACT: One-dimensional short-range interaction models for specific-sequence copolymers of amino acids have been developed in this series of papers. In the present paper, a multistate model [involving right-handed helical (h_R), extended (ϵ), chain-reversal (R and S), left-handed helical (h_L), right-handed bridge-region (ζ_R), left-handed bridge-region (ζ_L), and coil (or other) (c) states] is developed for the prediction of protein backbone conformation. This model involves ten parameters (w_{hR}, v_{hR}, v_ϵ, v_R, v_S, w_{hL}, v_{hL}, $u_{\zeta R}$, $u_{\zeta L}$, and u_c) and requires a 10 × 10 statistical weight matrix. Assuming that the left-handed helical *sequence* cannot occur in proteins, this 10 × 10 matrix can be reduced to a 9 × 9 matrix with nine parameters (w_{hR}, v_{hR}, v_ϵ, v_R, v_S, v_{hL}, $u_{\zeta R}$, $u_{\zeta L}$, and u_c). A nearest neighbor approximation of this multistate model is also formulated; with the omission of left-handed helical *sequences*, and the inclusion of the *left*-handed bridge region in the c state, this approximate model requires a 7 × 7 matrix with statistical weights w_{hR}^*, v_{hR}^*, v_R^*, v_S^*, v_{hL}^*, $u_{\zeta R}^*$, and u_c^*, expressed as values relative to the statistical weight of the ϵ state. The statistical weights for the multistate model are evaluated from the atomic coordinates of the x-ray structures of 26 native proteins. These statistical weights and the multistate model are applied in the prediction of the backbone conformations of proteins. The conformational probabilities of finding a residue in h_R, ϵ, R, S, h_L, ζ_R, or c states, defined as relative values with respect to their average values over the whole molecule, are calculated for bovine pancreatic trypsin inhibitor and clostridial flavodoxin, in order to select the most probable conformation for each residue of these proteins. The predicted results are compared to experimental observations and are discussed together with the reliability of the statistical weights. In the Appendix, the property of asymmetric nucleation of helical sequences is introduced into the (nearest neighbor) multistate model.

In this series of papers,[3-6] we have developed a statistical mechanical treatment of protein conformation within the context of one-dimensional short-range interaction models. These will be referred to here as papers I,[3] II,[4] III,[5] and IV,[6] with equations designated as I-1, II-1, III-1, etc.

In paper I, we presented a method for evaluating empirical statistical weights for various conformational states of amino acid residues on the basis of the reported (x-ray) conformations of native proteins.[3] In paper II,[4] we divided the conformational space of a residue into helical (h), extended (ϵ), and coil (other) (c) regions and formulated a three-state model for specific-sequence polypeptides, to predict protein conformation in paper III.[5] In paper IV,[6] we extended this treatment to a four-state model that included chain-reversal (R and S) states, as well as h, ϵ, and c states; also, in paper IV (as well as in the present paper), we used the X-ray coordinates (instead of the crystallographers' description of the conformational states) to evaluate the statistical weights. Since many (conformationally undefined) amino acid residues remain in the c state, even in the four-state model (see the last column of Table II of paper IV[6]), we can specify these conformations more precisely by further dividing the c region, in the present paper, and thereby develop a multistate model for treating protein conformation.

Even though the multistate model developed here specifies the conformational state of a residue more precisely, this does not render the four-state model obsolete for two reasons: (a) the four-state model is more convenient to use (because of its smaller size statistical weight matrix) and gives valid information about helical, extended, and chain-reversal conformations, and (b) the multistate model will be much improved when more x-ray data become available to provide more accurate statistical weights for the many states involved.

In section I of this paper, we formulate a multistate model for treating polypeptide conformation. In section II, a nearest neighbor treatment of the multistate model is presented. X-ray data on native proteins are analyzed in section III and are used in section IV to evaluate the statistical weights. The method of calculating conformational probabilities (for detecting backbone conformations in proteins) is presented in section V and applied and discussed in section VI. Finally, in section VII, we summarize the results obtained in papers I–V in relation to their applicability to the computation of the three-dimensional structures of proteins. In the Appendix, the (nearest neighbor) multistate model, developed in section II, is incorporated into our earlier model[7] of the helix–coil transition, in which asymmetric nucleation of helical sequences is taken into account.

(I) Formulation of Multistate Model

In our four-state model,[6] we considered right-handed α-helical[8] (h_R), extended (ϵ), chain-reversal (R and S), and other (c) states and assigned statistical weights w_{hR}, v_{hR}, v_ϵ, v_R, v_S, and u_c to a right-handed helical state with a hydrogen bond, a right-handed helical state without a hydrogen bond, an extended state, an R [for an $(i-1)$th residue] and S (for an ith residue) state of a chain reversal, and other (c) states, respectively. We retain the earlier[6] definitions[8] of the h_R, ϵ, R, and S states and the statistical weights w_{hR}, v_{hR}, v_ϵ, v_R, and v_S, in the present multistate model.

We now divide the conformational space of the c state of the four-state model into left-handed helical (h_L) and right-handed (ζ_R) and left-handed (ζ_L) bridge regions and a remaining (new) c region; the boundaries of these new regions are defined in section III. The statistical weights for the left-handed helical states are w_{hL} and v_{hL}; these correspond to w_{hR} and v_{hR} for the right-handed helical states (see eq II-3 to II-5). By substituting ζ_R, ζ_L, and (new) c for the (old) c appearing in eq II-8 and II-9, integration over the ζ_R, ζ_L, and c regions leads to the statistical weights $u_{\zeta R}$, $u_{\zeta L}$, and u_c. We thus have the complete set of statistical weights w_{hR}, v_{hR}, v_ϵ, w_{hL}, v_{hL}, v_R, v_S, $u_{\zeta R}$, $u_{\zeta L}$, and u_c required for the present multistate model.

Incorporating the three additional states, h_L, ζ_R, and ζ_L, into eq IV-1, we obtain eq 1 for the statistical weight matrix of the multistate model, where the states of three consecutive residues, $i-1$, i, and $i+1$, are correlated; in the matrix of eq 1, all blank elements are zero.

In a manner similar to that in which eq IV-2 was obtained from eq IV-1, eq 1 can be contracted as in eq 2, where the symbol ∪ means that, for example, c ∪ h_R ∪ ϵ ∪ R should be read as c or h_R or ϵ or R.

Actually, in the x-ray structures of proteins determined thus far, isolated (single and double) h_L states (without a hydrogen bond), but no left-handed helical *sequences*, are found. Within the context of the short-range interaction model, the h_L conformation occurs in response to intra-residue side chain–backbone interactions. Thus, we assign only the statistical weight v_{hL} to the ith residue in an h_L state [independent of the states of residues $(i-1)$ and $(i+1)$, an approximation made for all statistical weights other than w_{hR}] and omit w_{hL} from eq 2, to obtain eq 3.

For the first residue at the N terminus of the chain,[9] we define the row vector **t**, whose elements are the statistical weights for the allowed conformational states at the N terminus (see eq IV-3) as eq 4.

$$W_i = \left(\begin{array}{cc|cccccccccccccccccccccccccccccccccccccccccccccccccc}
 & i+1 & c & h_R & \epsilon & R & h_L & \zeta_R & \zeta_L & c & c & c & c & c & c & h_R & \epsilon & R & h_L & \zeta_R & \zeta_L & h_R & h_R & h_R & h_R & h_R & \epsilon & R & h_L & \zeta_R & \zeta_L & \epsilon & \epsilon & \epsilon & \epsilon & S & R & R & R & R & h_L & \zeta_R & \zeta_L & h_L & \zeta_R & \zeta_L & h_L & h_L & \zeta_R & \zeta_L & \zeta_R & \zeta_L \\
i-1 & i & c & c & c & c & c & c & c & h_R & \epsilon & S & h_L & \zeta_R & \zeta_L & h_R & h_R & h_R & h_R & h_R & h_R & \epsilon & S & h_L & \zeta_R & \zeta_L & \epsilon & \epsilon & \epsilon & \epsilon & \epsilon & S & h_L & \zeta_R & \zeta_L & R & S & h_L & \zeta_R & \zeta_L & S & S & S & h_L & h_L & h_L & \zeta_R & \zeta_L & \zeta_L & \zeta_R & \zeta_R & \zeta_L \\
\hline
c & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
c & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
c & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
c & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
c & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
c & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
c & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
h_R & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
\epsilon & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
S & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
h_L & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
\zeta_R & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
\zeta_L & c &u_c&u_c&u_c&u_c&u_c&u_c&u_c \\
h_R & h_R &&&&&&&&v_{hR}&&&&&&w_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
h_R & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
h_R & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
h_R & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
h_R & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
h_R & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
\epsilon & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
S & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
h_L & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
\zeta_R & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
\zeta_L & h_R &&&&&&&&v_{hR}&&&&&&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR}&v_{hR} \\
\epsilon & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
\epsilon & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
\epsilon & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
\epsilon & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
\epsilon & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
S & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
h_L & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
\zeta_R & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
\zeta_L & \epsilon &&&&&&&&&v_\epsilon&&&&&&&&&&&v_\epsilon&&&&&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon&v_\epsilon \\
R & S &&&&&&&&&&v_S&&&&&&&&&&&v_S&&&&&&&&&v_S&&&&&v_S&&&&v_S&v_S&v_S \\
S & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
h_L & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
\zeta_R & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
\zeta_L & R &&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&&v_R \\
S & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
S & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
S & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
h_L & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&w_{hL}&v_{hL}&v_{hL} \\
h_L & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
h_L & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
\zeta_R & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
\zeta_L & h_L &&&&&&&&&&&v_{hL}&&&&&&&&&&&v_{hL}&&&&&&&&&v_{hL}&&&&&v_{hL}&&&&&&v_{hL}&v_{hL}&v_{hL} \\
\zeta_L & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
\zeta_R & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L} \\
\zeta_R & \zeta_R &&&&&&&&&&&&u_{\zeta R}&&&&&&&&&&&u_{\zeta R}&&&&&&&&&u_{\zeta R}&&&&&u_{\zeta R}&&&&&&&&u_{\zeta R}&&&u_{\zeta R}&u_{\zeta R} \\
\zeta_L & \zeta_L &&&&&&&&&&&&&u_{\zeta L}&&&&&&&&&&&u_{\zeta L}&&&&&&&&&u_{\zeta L}&&&&&u_{\zeta L}&&&&&&&&u_{\zeta L}&u_{\zeta L}&&&u_{\zeta L}
\end{array} \right)_i \tag{1}$$

$\mathbf{W}_i =$

$i-1$	$i+1$ i	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ c	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ϵ	$c\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ h_R	h_R h_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ S	S R	$c\cup h_R\cup\epsilon\cup R$ $\cup\zeta_R\cup\zeta_L$ h_L	h_L h_L	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ζ_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ζ_L
c	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
ϵ	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
h_R	$c\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	0	0	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
h_R	h_R	0	0	v_{hR}	w_{hR}	0	0	0	0	0	0
S	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
R	S	0	0	0	0	v_S	0	0	0	0	0
h_L	$c\cup h_R\cup\epsilon\cup R$ $\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	0	0	$u_{\zeta R}$	$u_{\zeta L}$
h_L	h_L	0	0	0	0	0	0	v_{hL}	w_{hL}	0	0
ζ_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
ζ_L	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$

(2)

$\mathbf{W}_i =$

$i-1$	$i+1$ i	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ c	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ϵ	$c\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ h_R	h_R h_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ S	S R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ h_L	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ζ_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$ ζ_L
c	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
ϵ	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
h_R	$c\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	0	0	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
h_R	h_R	0	0	v_{hR}	w_{hR}	0	0	0	0	0
S	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
R	S	0	0	0	0	v_S	0	0	0	0
h_L	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
ζ_R	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$
ζ_L	$c\cup h_R\cup\epsilon\cup R$ $\cup h_L\cup\zeta_R\cup\zeta_L$	u_c	v_ϵ	v_{hR}	v_{hR}	0	v_R	v_{hL}	$u_{\zeta R}$	$u_{\zeta L}$

(3)

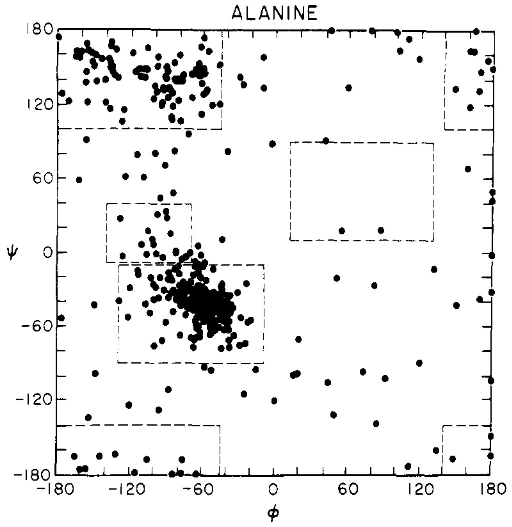

Figure 1. Distribution of the dihedral angles, ϕ and ψ, of the alanine residues observed in 26 proteins. The ϕ,ψ regions for the right-handed helical (h_R), left-handed helical (h_L), extended (ϵ), and right-handed bridge (ζ_R) conformations are enclosed by broken lines; the boundaries of these regions are defined in the text.

$$\mathbf{t}_1 = [u_c \quad v_\epsilon \quad v_{hR} \quad v_{hR} \quad 0 \quad v_R \quad v_{hL} \quad u_{\zeta R} \quad u_{\zeta L}]_1 \qquad (4)$$

For the last residue of the chain (i.e., the C terminus),[9] we define the column vector $\mathbf{t}_N{}^*$ as

$$\mathbf{t}_N{}^* = \begin{bmatrix} u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ u_c + v_\epsilon + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ v_{hR} \\ u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ v_S \\ u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \\ u_c + v_\epsilon + v_{hR} + v_R + v_{hL} + u_{\zeta R} + u_{\zeta L} \end{bmatrix}_N \qquad (5)$$

The elements of eq 5 correspond to the state $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, S, $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup h_R \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$ for residue $i + 1$ when $i = N$.

It is now possible to evaluate the partition function Z of a polypeptide chain (of N residues) by using eq 3–5 for $\mathbf{W}_i$, $\mathbf{t}_1$, and $\mathbf{t}_N{}^*$, i.e.,

$$Z = \mathbf{t}_1 \left[\prod_{i=2}^{N-1} \mathbf{W}_i \right] \mathbf{t}_N{}^* \qquad (6)$$

Since the parameters appearing in eq 4 and 5 do not differ from those in eq 3, the elements of eq 4 can be found in the first row of eq 3; similarly, the elements of eq 5 can be obtained from those of eq 3. Thus, eq 6 may be written as

$$Z = \mathbf{e}_i \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N{}^* \qquad (7)$$

in which

$$\mathbf{e}_1 = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \qquad (8a)$$

$$\mathbf{e}_N{}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \qquad (8b)$$

because $\mathbf{t}_1$ of eq 4 and $\mathbf{t}_N{}^*$ of eq 5 can be obtained by

$$\mathbf{t}_1 = \mathbf{e}_1 \mathbf{W}_1 \qquad (9a)$$

and

$$\mathbf{t}_N{}^* = \mathbf{W}_N \mathbf{e}_N{}^* \qquad (9b)$$

respectively.

Table I
The Number of Amino Acids Occurring in Various Conformational States in 26 Proteins[a]

Amino acid		Right-handed helical	Isolated right-handed helical	Extended	Chain reversal[c]			Left-handed helical	Bridge region		Other
j	N_j[b]	$N_{hR,j}$	$N_{hR',j}$	$N_{\epsilon,j}$	$N_{R,j}$	$N_{S,j}$	$N_{D,j}$	$N_{hL,j}$	$N_{\zeta R,j}$	$N_{\zeta L,j}$	$N_{c,j}$
Ala	407	175	8	104	43	21	19	2	2	0	33
Arg	128	40	3	45	9	9	11	3	4	0	4
Asn	223	48	2	60	13	48	12	12	5	0	23
Asp	257	75	7	44	24	40	21	4	13	0	29
Cys	100	22	1	41	6	14	5	3	0	0	8
Gln	149	43	5	50	17	11	8	6	0	0	9
Glu	207	104	3	43	17	14	7	1	0	0	18
Gly	401	60	3	94	36	77	18	29	3	12	69
His	111	37	5	31	6	17	6	2	2	0	5
Ile	225	74	6	99	12	11	2	1	7	0	13
Leu	323	124	8	116	17	22	13	0	2	0	21
Lys	320	118	8	76	30	31	20	7	4	0	26
Met	68	26	0	24	5	3	3	0	1	0	6
Phe	150	52	3	51	5	21	6	1	3	0	8
Pro	161	26	7	65	45	3	6	0	1	0	8
Ser	368	79	4	130	47	44	21	1	6	2	34
Thr	270	56	11	106	28	21	20	1	7	0	20
Trp	75	22	3	27	3	10	3	0	2	0	5
Tyr	180	35	3	78	13	24	11	2	0	0	14
Val	353	108	3	163	20	24	11	1	5	1	17

[a] Based on the calculated values of ϕ and ψ, using the x-ray data of the 26 proteins[10] given in Table I of paper IV.[6] The original papers, in which the x-ray data were reported, are cited in the footnotes of Table I of paper IV.[6] [b] Total number of amino acid j found in 26 proteins.[10] [c] See sections IIIA and IIIB of paper IV[6] for the definition of the R, S, and D states.

(II) Nearest Neighbor Multistate Model

In papers II[4] and IV,[6] we formulated nearest neighbor interaction approximations of the three-[4] and four-state[6] models, in order to reduce the size of the matrices, for rapid computation. We can also do this here, and thus obtain a good approximation of eq 3, for use in eq 7. However, before doing so, we can make a further simplification. From an analysis of the x-ray data of native proteins, all residues except Gly, Ser, and Val rarely exhibit the ζ_L conformation [the points in the ζ_L region in Figure 1 (for Ala) and most of those in the ζ_L region in Figure 3 (for Val) belong to the chain-reversal conformation and are not assigned to the ζ_L state; see Table I for the number of occurrences in the ζ_L region]. Therefore, in order to reduce the size of the matrix further, we will incorporate the ζ_L conformation (which was introduced in section I for generality) into the c state (for *all* residues) and drop the subscript R from ζ_R (the incorporation of the ζ_L conformation into the c state is not a feature of the nearest neighbor approximation but of the full multistate model). Hence, in essence, in this multistate model, we are introducing only two new states (h_L and ζ) and redefining the c state, compared to the four-state model. Thus, the statistical weight matrix for a nearest neighbor treatment of the multistate model may be written as

$$W_i = \left[\begin{array}{c|ccccccc} & & & & i & & & \\ i-1 & c & h_R & \epsilon & R & S & h_L & \zeta \\ \hline c & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \\ h_R & u_c & w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \\ \epsilon & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \\ R & 0 & 0 & 0 & 0 & v_S & 0 & 0 \\ S & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \\ h_L & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \\ \zeta & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL} & u_\zeta \end{array}\right]_i \tag{10}$$

In eq 10, we have again neglected hydrogen-bonded h_L states (i.e., we have omitted w_{hL}), as we did in eq 3. However, if it is desired to include w_{hL} in some applications (e.g., in treating copolymers with D residues, which can form h_L sequences), then eq 10 should be rewritten as eq 11.

$$W_i = \left[\begin{array}{c|ccccccc} & & & & i & & & \\ i-1 & c & h_R & \epsilon & R & S & h_L & \zeta \\ \hline c & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL}^2/w_{hL} & u_\zeta \\ h_R & u_c & w_{hR} & v_\epsilon & v_R & 0 & v_{hL}^2/w_{hL} & u_\zeta \\ \epsilon & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL}^2/w_{hL} & u_\zeta \\ R & 0 & 0 & 0 & 0 & v_S & 0 & 0 \\ S & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL}^2/w_{hL} & u_\zeta \\ h_L & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & w_{hL} & u_\zeta \\ \zeta & u_c & v_{hR}^2/w_{hR} & v_\epsilon & v_R & 0 & v_{hL}^2/w_{hL} & u_\zeta \end{array}\right]_i \tag{11}$$

The conformational state ζ_L that is not involved in eq 10 or 11 can be incorporated easily into these equations simply by adding a column and a row corresponding to ζ_L, with a statistical weight given by $u_{\zeta L}$ (it should be noted that ζ means ζ_R in eq 10 and 11, as mentioned above).

The partition function in the nearest neighbor Ising treatment of the multistate model, corresponding to eq 7, is

$$Z = \mathbf{e}_1 \left[\prod_{i=1}^{N} \mathbf{W}_i \right] \mathbf{e}_N{}^* \tag{12}$$

where

$$\mathbf{e}_1 = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0] \tag{13a}$$

and

$$\mathbf{e}_N{}^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{13b}$$

and $\mathbf{W}_i$ is given by either eq 10 or 11. Equation 13a 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 10 or 11 in the form of $\mathbf{e}_1\mathbf{W}_1$. Equation 13b indicates that the last residue (the C terminus) of the chain can be preceded by c, h_R, ϵ, R, S, h_L, or ζ states; hence, it is represented by the column vector $\mathbf{W}_N\mathbf{e}_N{}^*$ which contains the sum of the elements of the seven row vectors in eq 10 or 11.

Since our main interest is to obtain the relative conformational properties of a particular conformational state, the statistical weights appearing in eq 10 (or eq 11) may be expressed relative to that of a reference state. Choosing the extended (ϵ) state, as we did in our four-state[6] model (but not in our three-state model[3–5]), we now define the relative statistical weights pertaining to h_R, R, S, h_L, ζ (i.e., ζ_R), and c states by the following relations:

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

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

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

$$v_S{}^* = v_S/v_\epsilon \tag{17}$$

$$v_{hL}{}^* = v_{hL}/v_\epsilon \tag{18}$$

$$u_\zeta{}^* = u_\zeta/v_\epsilon \tag{19}$$

and

$$u_c{}^* = u_c/v_\epsilon \tag{20}$$

respectively, where a subscript j should be understood as appearing on all of the statistical weights to specify the species of the jth amino acid. Using the relative statistical weights of eq 14–20, eq 10 becomes eq 21. The corresponding form of eq 11 may be obtained by placing an asterisk superscript on the statistical weights of eq 11 and substituting 1 for v_ϵ in the third column of eq 11.

The partition function can be evaluated by substituting eq 21 for eq 11 when eq 12 and 13 are used.

$$W_i = \left[\begin{array}{c|ccccccc} & & & & i & & & \\ i-1 & c & h_R & \epsilon & R & S & h_L & \zeta \\ \hline c & u_c^* & v_{hR}^{*2}/w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \\ h_R & u_c^* & w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \\ \epsilon & u_c^* & v_{hR}^{*2}/w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \\ R & 0 & 0 & 0 & 0 & v_S^* & 0 & 0 \\ S & u_c^* & v_{hR}^{*2}/w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \\ h_L & u_c^* & v_{hR}^{*2}/w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \\ \zeta & u_c^* & v_{hR}^{*2}/w_{hR}^* & 1 & v_R^* & 0 & v_{hL}^* & u_\zeta^* \end{array}\right]_i \qquad (21)$$

(III) Analysis of X-Ray Data of Native Proteins

In this section, we analyze the x-ray data of native proteins, so that they may be used (in section IV) to evaluate the statistical weights required for the multistate model. We consider the x-ray data for the same 26 proteins,[10] used to evaluate the statistical weights in paper IV.[6] First, the dihedral angles[9] ϕ and ψ were computed from the Cartesian coordinates of these 26 proteins. The results for several amino acids are plotted in Figures 1–4, as examples. We used the same definitions for residues involved in h_R sequences, isolated h_R residues (without a hydrogen bond), for ϵ states, and for R and S states of a chain reversal as given in paper IV.[6] For the h_L, ζ_R, and ζ_L regions, we use the definitions of Burgess et al.,[11] viz., the h_L region is that in which the dihedral angles are in the range of $10° \le \phi \le 130°$ and $10° \le \psi \le 90°$, the ζ_R region is that in which the dihedral angles are in the range of $-140° \le \phi \le -70°$ and $-9° \le \psi \le 40°$, and the ζ_L region is that in which the dihedral angles are in the range of $70° \le \phi \le 140°$ and $-40° \le \psi \le 9°$. The h_R, ϵ, h_L, and ζ_R regions are shown on the ϕ, ψ maps of Figures 1–4; the ζ_L region is not shown for the reason mentioned in the first paragraph of section II.

We then counted the number of each type of amino acid j (where j = 1 to 20 for the naturally occurring amino acids) found in right-handed α-helical sequences ($N_{hR,j}$), in right-handed α-helical states without formation of a hydrogen bond ($N_{hR',j}$), in extended (ϵ) states ($N_{\epsilon,j}$), in the R and S states of a chain reversal [$N_{R,j}$ and $N_{S,j}$, as well as $N_{D,j}$ (the latter referring to duplicately assigned R and S states, as described in section IIIB of paper IV[6])], in the right-handed (ζ_R) bridge region ($N_{\zeta R,j}$), in the left-handed (ζ_L) bridge region ($N_{\zeta L,j}$) and in other (c) states ($N_{c,j}$). The results for the 20 amino acids are summarized, together with the total number N_j of each amino acid in the 26 proteins surveyed,[10] in Table I. It should be noted that, in Table I, the numbers of amino acids occurring in the right-handed helical ($N_{hR,j}$ and $N_{hR',j}$), extended ($N_{\epsilon,j}$), and chain-reversal (N_R, N_S, and N_D) regions are the same as in Table II of paper IV.[6]

(IV) Calculation of Statistical Weights for Multistate Model

The statistical weights $w_{hR,j}^*$, $v_{hR,j}^*$, $v_{R,j}^*$, and $v_{S,j}^*$ for the jth type of amino acid, relative to the statistical weight for the ϵ state, can be calculated using eq IV-18 to IV-20 and IV-22

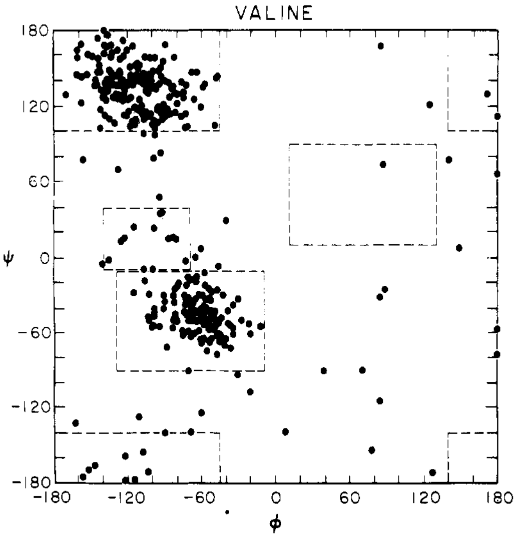

Figure 3. Same as Figure 1, but for valine residues.

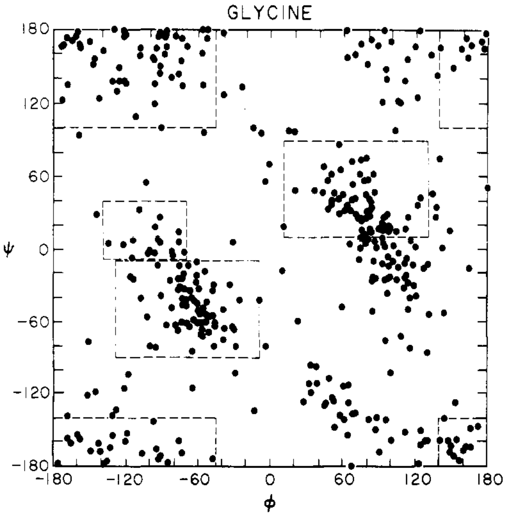

Figure 2. Same as Figure 1, but for glycine residues.

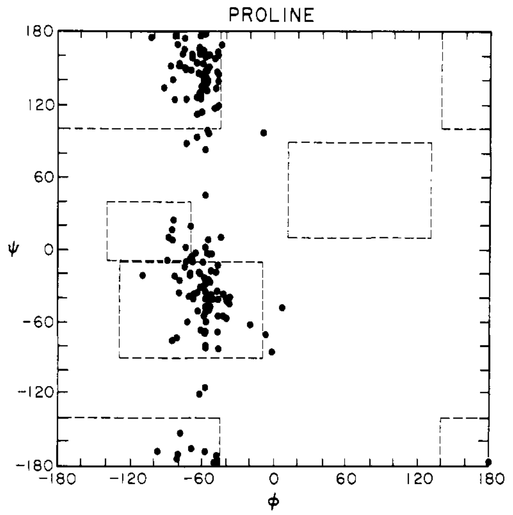

Figure 4. Same as Figure 1, but for proline residues.

Table II
Statistical Weights for the Multistate Model

Amino acid	Statistical weight[a]						
j	$w_{\mathrm{hR},j}{}^*$	$v_{\mathrm{hR},j}{}^*$	$v_{\mathrm{R},j}{}^*$	$v_{\mathrm{S},j}{}^*$	$v_{\mathrm{hL},j}{}^*$	$u_{\zeta\mathrm{R},j}{}^*$	$u_{\mathrm{c}',j}{}^{*b}$
Ala	1.683	0.077	0.505	0.293	0.019	0.019	0.317
Arg	0.889	0.067	0.322	0.322	0.067	0.089	0.089
Asn	0.800	0.033	0.317	0.900	0.200	0.083	0.383
Asp	1.705	0.159	0.784	1.148	0.091	0.295	0.659
Cys	0.537	0.024	0.207	0.402	0.073	0.0	0.195
Gln	0.860	0.100	0.420	0.300	0.120	0.0	0.180
Glu	2.419	0.070	0.477	0.407	0.023	0.0	0.419
Gly	0.638	0.032	0.479	0.915	0.309	0.032	0.862
His	1.194	0.161	0.290	0.645	0.065	0.065	0.161
Ile	0.747	0.061	0.131	0.121	0.010	0.071	0.131
Leu	1.069	0.069	0.203	0.246	0.0	0.017	0.181
Lys	1.553	0.105	0.526	0.539	0.092	0.053	0.342
Met	1.083	0.0	0.271	0.187	0.0	0.042	0.250
Phe	1.020	0.059	0.157	0.471	0.020	0.059	0.157
Pro	0.400	0.108	0.738	0.092	0.000	0.015	0.123
Ser	0.608	0.031	0.442	0.419	0.008	0.046	0.277
Thr	0.528	0.104	0.358	0.292	0.009	0.066	0.189
Trp	0.815	0.111	0.167	0.426	0.0	0.074	0.185
Tyr	0.449	0.038	0.237	0.378	0.026	0.0	0.179
Val	0.663	0.018	0.156	0.181	0.006	0.031	0.110

[a] The statistical weights are relative to the extended (ϵ) state; i.e., $v_{\epsilon,j}{}^* = 1.0$ for all 20 amino acid residues. [b] As described in section IV of the text, the left-handed bridge conformation (ζ_{L}) is incorporated into the other (c) state when the values of $u_{\mathrm{c}',j}{}^*$ of this table are evaluated (see text for more detail).

to IV-25, together with the values of N_j, $N_{\mathrm{hR},j}$, $N_{\mathrm{hR}',j}$, $N_{\epsilon,j}$, $N_{\mathrm{R},j}$, $N_{\mathrm{S},j}$ (and $N_{\mathrm{D},j}$) given in Table I. In a similar manner, the statistical weights $v_{\mathrm{hL},j}{}^*$, $u_{\zeta\mathrm{R},j}{}^*$, $u_{\zeta\mathrm{L},j}{}^*$, and $u_{\mathrm{c},j}{}^*$ (relative to the ϵ state) can be calculated from the corresponding data of Table I.

As seen in column 11 of Table I, the ζ_{L} conformation is a rare occurrence, $N_{\zeta\mathrm{L},j}$ being nonzero only for Gly, Ser, and Val (with the values for Gly being the only significantly frequent ones). That is why we incorporated the ζ_{L} state into the c state, a procedure which requires the introduction of a new quantity $N_{\mathrm{c}',j}$ defined as

$$N_{\mathrm{c}',j} = N_{\zeta\mathrm{L},j} + N_{\mathrm{c},j} \tag{22}$$

where $N_{\zeta\mathrm{L}j}$ and $N_{\mathrm{c},j}$ are given in Table I. Hereafter, we will use the symbol c′ to represent the "other" state when the ζ_{L} conformation is incorporated into the c state.

In a manner similar to the use of eq IV-20, we can calculate $v_{\mathrm{hL},j}{}^*$ (relative to the ϵ state) as

$$v_{\mathrm{hL},j}{}^* = f_{\mathrm{hL},j}/f_{\epsilon,j} \tag{23}$$

Likewise,

$$u_{\zeta\mathrm{R},j}{}^* = f_{\zeta\mathrm{R},j}/f_{\epsilon,j} \tag{24}$$

In eq 23 and 24, the values $f_{\mathrm{hL},j}$, $f_{\epsilon,j}$, and $f_{\zeta\mathrm{R},j}$ are computed by means of eq IV-19, i.e.,

$$f_{\eta,j} = N_{\eta,j}/N_j \tag{25}$$

where η stands for $\mathrm{h_L}$, ϵ, or ζ_{R}, and the values of N_j and $N_{\eta,j}$ (for $\eta = \epsilon$, $\mathrm{h_L}$, and ζ_{R}) are given in columns 2, 5, 9, and 10 of Table I. The value of $u_{\mathrm{c},j}{}^*$ is computed from

$$u_{\mathrm{c},j}{}^* = f_{\mathrm{c},j}/f_{\epsilon,j} \tag{26}$$

where $f_{\mathrm{c},j}$ and $f_{\epsilon,j}$ are obtained from eq 25 with η = c and ϵ, respectively. With the ζ_{L} conformation incorporated into the c state, the use of eq 25 (to obtain $u_{\mathrm{c}',j}{}^*$) requires the introduction of $N_{\mathrm{c}',j}$ of eq 22, instead of $N_{\mathrm{c},j}$ which is given in the last column of Table I.

The resulting statistical weights (relative to the ϵ state) for the multistate model are given in Table II. These statistical weights will be used in the prediction of protein conformation in sections V and VI.

(V) Calculation of Conformational Probabilities

The method for calculating the conformational probability for finding a residue in a certain state or a sequence in a certain set of conformational states was formulated for the helix–coil transition model in polypeptides[12] and extended to a more general treatment that can be applied to a model with any number of conformational states in paper II.[4] These methods were used to evaluate conformational-sequence probabilities in the prediction of protein conformation in papers III[5] and IV.[6]

The first-order a priori probabilities that a residue i in a protein will be found in states $\mathrm{h_R}$, ϵ, R, S, $\mathrm{h_L}$, ζ_{R}, or c′ are $F_{i;\mathrm{hR}}$, $F_{i;\epsilon}$, $F_{i;\mathrm{R}}$, $F_{i;\mathrm{S}}$, $F_{i;\mathrm{hL}}$, $F_{i;\zeta\mathrm{R}}$, or $F_{i;\mathrm{c}'}$, respectively; these can be computed with the aid of eq II-44, viz.,

$$F_{i;\eta_i} = Z^{-1}\mathbf{e}_1 \left[\prod_{j=1}^{i-1} \mathbf{W}_j\right]\left[\frac{\partial \mathbf{W}_i}{\partial \ln{(\mathbf{m}_{i;\eta_i})}}\right]_{\{\rho\}} \times \left[\prod_{l=i+1}^{N} \mathbf{W}_l\right]\mathbf{e}_N{}^* \tag{27}$$

together with eq 12, 13, and 21 and the statistical weights of Table II, where η designates the specific conformational states and $\{\rho\}$ is a specific sequence of conformational states; since only a first-order a priori probability is used in this paper, $\{\rho\}$ pertains to the conformation of one residue here, i.e., η_i. The lower order matrix of the nearest neighbor model is used instead of the larger matrices of eq 2 or 3 to save computer time.

The average probabilities θ_η over a whole protein chain (where $\eta = \mathrm{h_R}$, ϵ, R, S, $\mathrm{h_L}$, ζ_{R}, and c′) are computed from eq II-52, i.e.,

$$\theta_\eta = \frac{1}{N}\sum_{i=1}^{N} F_{i;\eta} \tag{28}$$

where N is the number of amino acid residues in the protein.

In a manner similar to eq III-13, III-14, and IV-29, we define the probabilities relative to the average probabilities θ_η as

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

(VI) Results and Discussion

In order to obtain quantitative information about the tendencies for a *residue* of a protein to be in one of the conformational states $\mathrm{h_R}$, ϵ, R, S, $\mathrm{h_L}$, or c′, we use the conformational probability for a residue, $F_{i;\eta}$, as described in section V. To detect a helical or extended conformational *sequence* in a protein, we use the conformational-sequence probability defined in paper III[5] (this point will be discussed below). However, extensive computational effort[13] is necessary to obtain conformational probabilities of order higher than first order when large-order statistical weight matrices are used (compared to the smaller matrices in models with fewer states[3-5,7,12]). In addition to the extensive computational effort, ad hoc empirical rules are required to select between different lengths of helical and extended sequences and to determine the conformations of such sequences without duplication or ambiguity, as discussed in section III of paper III.[4] To avoid extensive computations and empirical rules, such as those introduced in section III of paper III,[4] in this paper we use the first-order conformational probabilities relative to θ_η, i.e., $P_{i;\eta}{}^*$, the computation of which was described in section V. Thus, the most probable conformational state of any residue i in a protein is taken as the one having the largest value of $P_{i;\eta}{}^*$, where η is $\mathrm{h_R}$, ϵ, R, S, $\mathrm{h_L}$, ζ_{R}, or c′.

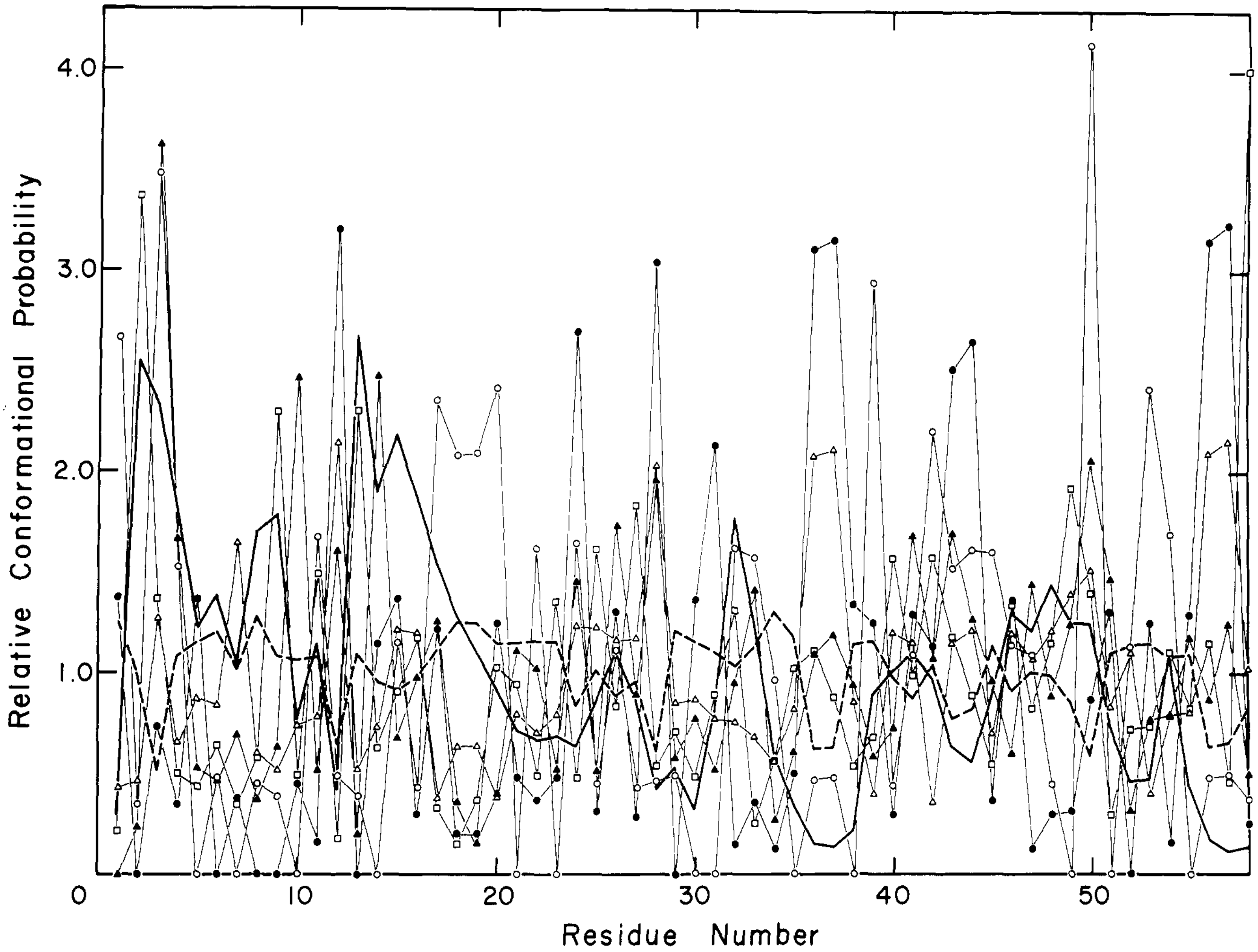

Figure 5. The relative probability of finding each amino acid residue of bovine pancreatic trypsin inhibitor in the conformational states h_R (—), ϵ (- - -), the R (□) and S (▲) states of a chain-reversal, h_L (●), ζ_R (○), and c′ (△).

Using eq 12, 13, 21, and 27–29, and the statistical weights of Table II, we computed the relative conformational probabilities for a residue i to be in h_R, ϵ, R, S, h_L, ζ_R, and c′ states for bovine pancreatic trypsin inhibitor (BPTI) and clostridial flavodoxin.[14] The values of $P_{i;\eta}{}^*$ computed for BPTI are plotted as a function of residue number i in Figure 5. The most probable conformation of BPTI (shown in column 2 of Table III) was taken as the one with the highest value of $P_{i;\eta}{}^*$ for each residue. Similar data are shown in column 2 of Table IV for clostridial flavodoxin. In Tables III and IV, we also summarize the predicted results on helical and extended sequences of paper III[5] (quoted from Table IV of paper III[5]) and on chain-reversal conformations of paper IV[6] (quoted from Table VI of paper IV[6]).

The above predictions were based on the values of $P_{i;\eta}{}^*$, computed from eq 29, using the first-order a priori probabilities $F_{i;\eta}$ (and the average values $\bar{\theta}_\eta$ given by eq 28). However, it is more accurate to use conformational probabilities of higher order than first order to locate helical and extended *sequences*, as pointed out in section IIB of paper III,[5] and to locate chain-reversal conformations [R and S states at the (i − 1)th and ith residues, respectively], as pointed out in section VA of paper IV.[6] Only the first-order probabilities were used in the above calculations on BPTI and clostridial flavodoxin simply to reduce the computer time required to obtain the higher order probabilities. The advantages of using probabilities of higher order can be understood from the following two examples.

As a first example, consider residues 4–6 of BPTI. As seen in column 2 of Table III, residues 4–6 are predicted to be h_R, h_L, and h_R, respectively, on the basis of single-residue probabilities. On the other hand, residues 4–7 were predicted to form an h_R sequence on the basis of the conformational-sequence probability of a triad, as seen in the third column of Table III; i.e., the weak tendency of residue 5 to adopt the h_R state (shown by the fact that it prefers the h_L state, as seen in column 2 of Table III) does not prevent the triad 4–6 from forming an h_R sequence (based on the conformational-sequence probability of a triad). This difference in predicted conformations arises because the probability of occurrence of a given conformational *sequence* is *not* given by the product of first-order conformational probabilities, as illustrated in section VIC of paper II[4] (see eq II-65).

As a second example, we compare the results of this paper and paper IV[6] for chain-reversal conformations. A chain reversal consists of two consecutive residues, the (i − 1)th and ith, in R and S states, respectively. Therefore, to locate a chain-reversal conformation in a protein, one must compute the second-order a priori probability $P_{i;RS}{}^*$ that residue (i − 1) is in the R state and, at the same time, residue i is in the S state (see paper IV[6]). If, on the other hand, one uses only the first-order a priori probabilities (that are converted into those that are relative to the average probabilities), $P_{i;R}{}^*$ and $P_{i;S}{}^*$, to locate R and S states, one obtains isolated R and S states, i.e., R states not followed by S states, and S states not preceded by R states, as seen in column 2 of Tables III and IV. In such cases, we have indicated in parentheses in column 2 of Tables III and IV the conformations having the next highest single-residue probability.

These examples show that it is necessary to compute higher

Table III
Predicted and Experimentally observed conformations of Pancreatic Trypsin Inhibitor

<table>
<tr><th></th><th colspan="3">Predicted results</th><th colspan="2"></th></tr>
<tr><th>Residue No.</th><th>Multi-state model[a]</th><th>Three-state model[b,c]</th><th>Four-state model[c,d]</th><th colspan="2">Obsd[c,e] conformation</th></tr>
<tr><td>1</td><td>ζ</td><td></td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>R</td><td></td><td rowspan="2">R
S</td><td></td><td>ϵ</td></tr>
<tr><td></td><td>S</td><td></td><td></td><td rowspan="3">h_R</td></tr>
<tr><td></td><td>h_R</td><td rowspan="4">h_R</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td rowspan="2">R
S</td></tr>
<tr><td></td><td>h_R</td><td></td><td></td></tr>
<tr><td></td><td>c'</td><td></td><td></td><td rowspan="4">ϵ</td></tr>
<tr><td></td><td>h_R</td><td></td><td></td><td></td></tr>
<tr><td></td><td>R</td><td></td><td rowspan="2">R
S</td><td></td></tr>
<tr><td>10</td><td>S</td><td></td><td></td></tr>
<tr><td></td><td>ζ</td><td></td><td></td><td></td><td>h_R</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td></td><td>c'</td></tr>
<tr><td></td><td>h_R</td><td rowspan="4">h_R</td><td></td><td></td><td>ζ</td></tr>
<tr><td></td><td>S (h_R)</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>h_R</td><td></td><td></td><td>ζ</td></tr>
<tr><td></td><td>h_R</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>ζ</td><td rowspan="5">ϵ</td><td></td><td></td><td>c'</td></tr>
<tr><td></td><td>ζ</td><td></td><td></td><td rowspan="7">ϵ</td></tr>
<tr><td></td><td>ζ</td><td></td><td></td></tr>
<tr><td>20</td><td>ζ</td><td></td><td></td></tr>
<tr><td></td><td>ϵ</td><td></td><td></td></tr>
<tr><td></td><td>ζ</td><td></td><td></td><td></td></tr>
<tr><td></td><td>R (ϵ)</td><td rowspan="6">h_R</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td></tr>
<tr><td></td><td>R</td><td></td><td></td><td rowspan="3">h_R</td></tr>
<tr><td></td><td>S</td><td></td><td></td></tr>
<tr><td></td><td>R (c')</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td>h_L</td></tr>
<tr><td>29</td><td>ϵ</td><td rowspan="8">ϵ</td><td></td><td></td><td rowspan="7">ϵ</td></tr>
<tr><td>30</td><td>h_L</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td></tr>
<tr><td></td><td>h_R</td><td></td><td></td></tr>
<tr><td></td><td>ζ</td><td></td><td></td></tr>
<tr><td></td><td>ϵ</td><td></td><td></td></tr>
<tr><td></td><td>ϵ</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td>h_R</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td></td><td>c'</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>ζ</td><td></td><td></td><td></td><td>h_L</td></tr>
<tr><td>40</td><td>R</td><td></td><td rowspan="2">R
S</td><td></td><td>ϵ</td></tr>
<tr><td></td><td>S</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>ζ</td><td></td><td rowspan="2">R
S</td><td rowspan="2">R
S</td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td rowspan="7">h_R</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>ζ</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td>h_R</td></tr>
<tr><td></td><td>S (h_R)</td><td></td><td></td><td>ϵ</td></tr>
<tr><td></td><td>h_R</td><td></td><td></td><td rowspan="7">h_R</td></tr>
<tr><td></td><td>R (c')</td><td></td><td></td></tr>
<tr><td>50</td><td>ζ</td><td></td><td></td></tr>
<tr><td></td><td>S (h_L)</td><td rowspan="4">ϵ</td><td></td><td></td></tr>
<tr><td></td><td>ϵ</td><td></td><td></td></tr>
<tr><td></td><td>ζ</td><td></td><td></td></tr>
<tr><td></td><td>ζ</td><td rowspan="2">R
S</td><td rowspan="2">R
S</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td></td><td>h_R</td></tr>
<tr><td></td><td>h_L</td><td></td><td></td><td></td><td>c'</td></tr>
<tr><td></td><td>R (c')</td><td></td><td></td><td></td><td>ϵ</td></tr>
</table>

[a] Results of the present paper. The conformations given in parentheses are the ones with the next highest probability for residues for which isolated R and S states have the highest probability. [b] Results of paper III,[5] quoted from Table IV of paper III.[5] [c] The rectangles designate h_R or ϵ sequences or RS chain reversals. [d] Results of rule II of paper IV,[6] quoted from Table VI of paper IV.[6] [e] Experimental results from x-ray data (see Table I of paper IV[6]). The regions designated here differ from those given in Table XII of paper I since we used different definitions of the conformational states, as described in papers I and IV. In paper IV, these ordered conformations were determined from the x-ray coordinates (see Table I of paper IV[6])

order conformational-sequence probabilities to locate helical and extended sequences and chain-reversal conformations. On the other hand, to locate isolated h_L, ζ, and c' states (and, of course, isolated h_R and ϵ states), it is sufficient to consider the first-order conformational probabilities. Since the procedure of paper III[5] (the three-state model) already locates h_R and ϵ *sequences*, and that of paper IV[6] (the four-state model) already locates chain reversals, we will focus attention here on the prediction of h_L, ζ, and c' states. For this purpose, as illustrated above, the first-order probabilities suffice; i.e., h_R and ϵ sequences are located by the higher order probabilities of paper III and chain reversals by the higher order probabilities of paper IV, and first-order probabilities suffice for h_L, ζ, and c' states. As seen in Table III, the predictive results for h_L, ζ, and c' states for BPTI are not very good; however, as seen in Table IV, they are fairly good for clostridial flavodoxin.

There are three important points to consider in evaluating this and previous papers[3–6] of this series. First, we have developed a framework for treating protein conformation with a short-range interaction model. Given the assumptions usually used in such a model, we believe that the matrix treatment has been formulated properly. Second, in order to treat proteins properly, medium- and long-range interactions must be introduced (see the last paragraph of this section). Third, the statistical weights have been obtained from x-ray data. These data are not as extensive as would be required (especially for the less frequently occurring h_L and ζ states) but can be improved as more x-ray data become available. In addition, the x-ray data on proteins, used to obtain statistical weights for a short-range interaction model, reflect also the medium- and long-range interactions that do not appear in the short-range-interaction model. However, this is a point that can be tested when more extensive x-ray data become available, viz., the applicability of a short-range interaction model (with statistical weights deduced in this or in any other manner) to proteins. For the three- and four-state models, where the x-ray data are more extensive, the dominance of short-range interactions in determining protein structure, in first approximation, is well established. It remains to be seen whether this is also true of the multistate model. For the above reasons, we see no point, at present, in dividing the c' region among additional states, although this can be done, of course, by extending the present formalism, when enough x-ray data become available to make this extension worthwhile.

In section III of paper III,[5] we introduced ad hoc empirical rules to interpret the conformational-sequence probabilities of h_R triads and ϵ tetrads, i.e., to assign the backbone conformations of proteins without ambiguity. Without these rules, the predictions of paper III would sometimes be inconclusive. On the other hand, however, it is desirable to eliminate such ad hoc empirical rules from a predictive scheme. This can be done by using the one-dimensional short-range interaction models to compute conformational-sequence probabilities for *long* sequences. While such computations would require much computer time, they would provide unambiguous assignments of backbone conformations. Efforts are now in progress[15] to compute conformational-sequence probabilities for *long* sequences, without resort to empirical rules.

Recently,[16,17] we proposed a hypothesis for protein folding

Table IV
Predicted and Experimentally Observed Conformations of Clostridial Flavodoxin

	Predicted results					Predicted results			
Residue No.	Multistate model[a]	Three-state model[b,c]	Four-state model[c,d]	Obsd[c,e] conformation	Residue No.	Multistate model[a]	Three-state model[b,c]	Four-state model[c,d]	Obsd[c,e] conformation
1	ε	ε		ε	70	h_R	h_R		h_R
	h_L					h_R			
	ζ					h_R			
	ε					ζ			
	ε					ε			R
	ζ					ζ		R	D
	R (ε)		R			h_L		S	S
	h_L		S	R		ζ			ζ
	R (ζ)			S		R (ε)		R	R
10	h_L			c′		h_L		S	S
	h_L	h_R		h_R	80	R	ε		ε
	h_R					S			
	h_R					ε			
	h_R					c′			
	h_R					ε			
	h_R					ζ			
	h_R					h_L			
	h_R					R		R	
	h_R					S		S	
20	R (h_L)					h_L			R
	R (h_L)				90	ζ			S
	h_L	h_R				h_L			ε
	ζ					ζ		R	R
	ζ					h_L	h_R	S	S
	c′					h_L			h_R
	R (ε)					ζ			
	h_L			h_L		ε			
	R (h_L)		R	ε		ζ			
	ζ		S			ζ	h_R		
30	ε					h_R			
	h_L				100	h_R			
	ζ					h_R			
	ζ					ζ			
	h_L					R (ε)		R	
	ε			R		h_L		S	
	R (ε)		R	S		h_L			
	ζ		S	ζ		ε	ε		
	ε			ε		h_L			h_L
	h_L	h_R		c′		h_L			ε
40	ζ			R		ε			ε
	ζ			D	110	ε			ε
	S (h_R)			S		ε			c′
	h_R			R		c′			ε
	R (h_R)			D		ζ			
	h_L			S		R			
	R (c′)	ε		ε		S			
	ζ			R		ζ			
	ζ			S		ε			
	ε			ε		h_L			
50	ζ					h_L			c′
	ε				120	c′			ε
	h_L					h_R			h_R
	h_L					h_R			c′
	ε					h_R			R
	c′					h_R			S
	R (ε)					h_R			h_R
	h_L			R		h_L			
	ζ			D		ζ			
	h_R			S		h_L			
60	h_R			ε		ζ			
	h_R			ε	130	R		R	
	h_R			c′		S		S	
	h_R			h_R		h_L	h_R		
	h_R					R			
	R		R			S			
	S		S			ζ			
	c′					R (c′)		R	R
	R	h_R				h_L		S	S
69	S				138	ζ			ε

[a–e] Footnotes are the same as in Table III.

involving a three-step mechanism. In step A, a one-dimensional short-range interaction model (such as those developed in this series of papers) is used to assign initial conformational states to each residue. However, in order to predict the three-dimensional structure of a protein, step A must be followed by other steps (e.g., steps B and C of our recently developed procedure[16,17]) that incorporate the medium- and long-range interactions that are not present in the short-range interaction models. Efforts along these lines are also in progress.

(VII) Summary of This Series

In this series of five papers (I–IV of ref 3–6 and the present paper V), the following results have been obtained:

(1) Theoretical formulations have been provided for one-dimensional short-range interaction models [three states (h, ϵ, c) in paper II,[4] four states (h, ϵ, R–S, c) in paper IV,[6] and multistates (h_R, ϵ, R–S, h_L, ζ_R, c′) in the present paper].

(2) Statistical weights, based on x-ray data on native proteins, have been evaluated (in papers I,[3] IV,[6] and V) for use in the matrices required for computations with each of these models. However, as pointed out in earlier papers of this series,[3-5] the statistical weights can be obtained from other sources than the x-ray data on proteins, e.g., from experimental studies of model polypeptides in solution or from theoretical calculations using empirical conformational energy functions.

(3) Using the statistical weights mentioned in (2) above, the one-dimensional short-range interaction models were applied to predict the backbone conformations of proteins in papers III,[5] IV,[6] and V.

(4) The one-dimensional short-range interaction models mentioned in (1) above can provide the backbone conformations in step A of the three-step mechanism[16,17] of protein folding. The models developed in papers I–V are now being incorporated into this three-step mechanism to try to predict the three-dimensional structures of native proteins.

Appendix

Nearest Neighbor Multistate Model with Asymmetric Nucleation of Helical Sequence. We recently[7] 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. This asymmetric nucleation of helical sequences was incorporated into the three-[4] and four-state[6] models. In a similar manner, the asymmetric nucleation properties of helical sequences can be incorporated into the present multistate model.

In this Appendix, and in this Appendix only, we use the c state as a reference, instead of the ϵ state. We then define statistical weights (relative to the c state) in a manner similar to that used in obtaining q_9, q_{10}, and q_{11} in eq A-1 to A-3 of paper IV,[6] viz.,

$$q_{12} = v_{\mathrm{hL}}/u_{\mathrm{c}} \tag{A-1}$$

$$q_{13} = u_{\zeta\mathrm{R}}/u_{\mathrm{c}} \tag{A-2}$$

and

$$q_{14} = u_{\zeta\mathrm{L}}/u_{\mathrm{c}} \tag{A-3}$$

for the left-handed helical and the right- and left-handed bridge-region conformations. Assuming the same nucleation parameters (although it is possible to assign different ones) at the boundaries between a right-handed helical sequence and h_L, ζ_R, and ζ_L states, as was done in the case of ϵ, R, and S states in papers II[4] and IV,[6] we can now construct the statistical weight matrix of the nearest neighbor multistate model with asymmetric properties, corresponding to eq 3, as eq A-4, where q_1 to q_8 were defined in ref 7 (see the summary of the

$$\mathbf{W}_i = \left[\begin{array}{ll|cccccccccccccc}
 & i+1 & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_R} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_R} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_R} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_R} & \mathrm{S} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_R} & \begin{array}{c}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & \mathrm{h_L} \\
i-1 & i & \mathrm{c} & \mathrm{c} & \epsilon & \epsilon & \mathrm{h_R} & \mathrm{h_R} & \mathrm{S} & \mathrm{S} & \mathrm{R} & \mathrm{h_L} & \zeta_\mathrm{R} & \zeta_\mathrm{R} & \zeta_\mathrm{L} & \zeta_\mathrm{L} \\
\hline
\mathrm{c} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\mathrm{c} & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\epsilon & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\epsilon & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{h_R} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_5 & q_3 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\mathrm{h_R} & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_2 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{S} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\mathrm{S} & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{R} & \mathrm{S} & 0 & 0 & 0 & 0 & 0 & 0 & q_{11} & q_{11} & 0 & 0 & 0 & 0 & 0 & 0 \\
\mathrm{h_L} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\zeta_\mathrm{R} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\zeta_\mathrm{R} & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\zeta_\mathrm{L} & \begin{array}{l}\mathrm{cU\epsilon URUh_L,}\\ \mathrm{U\zeta_RU\zeta_L}\end{array} & q_8 & q_7 & q_9 & q_9 & 0 & 0 & 0 & 0 & q_{10} & q_{12} & q_{13} & q_{13} & q_{14} & q_{14} \\
\zeta_\mathrm{L} & \mathrm{h_R} & 0 & 0 & 0 & 0 & q_6 & q_4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]_i \tag{A-4}$$

statistical weights in Table I of ref 7), q_9 was defined in paper II,[4] and q_{10} and q_{11} were defined in eq A-2 and A-3 of paper IV.

The statistical weight vector $\mathbf{t}_1$ for the first (N terminal) residue is constructed from the conformations allowed for this residue as

$$\mathbf{t}_1 = (q_8 \quad q_7 \quad q_9 \quad q_9 \quad q_6 \quad q_4 \quad 0 \quad 0 \quad q_{10} \quad q_{12} \quad q_{13} \quad q_{13} \quad q_{14} \quad q_{14}) \tag{A-5}$$

For the last (C terminal) residue, we have

$$\mathbf{t}_N{}^* = \begin{bmatrix} q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_6 \\ q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_6 \\ q_5 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_2 \\ q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_6 \\ q_{11} \\ q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_6 \\ q_8 + q_9 + q_{10} + q_{12} + q_{13} + q_{14} \\ q_6 \end{bmatrix}_N \tag{A-6}$$

where each element corresponds to the carboxyl terminal residue in a state where residue $i + 1$ is in state $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, S, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$, and $c \cup \epsilon \cup R \cup h_L \cup \zeta_R \cup \zeta_L$.

Using eq A-4 to A-6, the partition function may be written as

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

or as

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

where

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

and

$$\mathbf{e}_N{}^* = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \tag{A-10}$$

since $\mathbf{t}_1$ and $\mathbf{t}_N{}^*$ of eq A-5 and A-6 are given by $\mathbf{e}_1\mathbf{W}_1$ and $\mathbf{W}_N\mathbf{e}_N{}^*$, respectively.

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) to whom requests for reprints should be addressed.

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

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

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

(6) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9,** 812 (1976).

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

(8) In the present multistate model, we will take the left-handed helical (h_L) conformation into account explicitly. Therefore, we will consider the helical (h) state, that has been used in past treatments of the helix–coil transition (i.e., two-state) model, and in three-[3–5] and four-state[6] models, as a right-handed helical state (designated by h_R), to distinguish it from the left-handed helical state (designated by h_L in this paper). Thus, the statistical weights w_h and v_h used in the three-[4] and four-state[6] models will be expressed as w_{hR} and v_{hR} in the multistate model. The range of dihedral angles in the ϕ,ψ space,[9] used to define the various conformational states, is quite arbitrary as long as most of the observed conformations lie in the regions chosen (see Figures 1–4). For example, a broad range of values of ϕ and ψ was used to define the α-helical state 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 or free energy of the whole protein is calculated (see point ii in section VI of paper IV[6]). The same comments apply to the definitions of the extended region and to the chain-reversal conformation.

(9) Throughout this paper, we use the recommendations proposed by an IUPAC-IUB Commission on Biochemical Nomenclature [*Biochemistry*, **9,** 3471 (1970)].

(10) The proteins used for the present analysis are tabulated in column 1 of Table I of paper IV.[6] References to the original papers, in which the x-ray data are reported, are also given in the footnotes of Table I of paper 4.[6]

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

(12) S. Tanaka and A. Nakajima, *Macromolecules*, **5,** 714 (1972).

(13) In general, the execution time required for matrix multiplication on a computer is proportional to the cube of the order of the matrix.

(14) Bovine pancreatic trypsin inhibitor and clostridial flavodoxin were included in the protein data set used to evaluate the statistical weights in this paper and in paper IV[6] but were not included in the evaluation of the statistical weights in the prediction scheme of papers I–III.

(15) S. Tanaka and H. A. Scheraga, work in progress.

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

(17) S. Tanaka and H. A. Scheraga, *Macromolecules*, in press. (paper on hypothesis about the mechanism of protein folding).