Statistical Mechanical Treatment of Protein Conformation. 6. Elimination of Empirical Rules for Prediction by Use of a High-Order Probability. Correlation between the Amino Acid Sequences and Conformations for Homologous Neurotoxin Proteins[1]

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

Department of Chemistry, Cornell University, Ithaca, New York 14853.
Received April 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 this paper, a general method for predicting protein conformation (that is based on a one-dimensional short-range interaction model, and eliminates the need for the empirical rules introduced in papers III and IV) is described. The present method involves the use of conformational (or conformational-sequence) probabilities of higher order than the first- or second-order probabilities used in papers IV and V, i.e., it treats a sequence of any number of residues; it thus alters the predictive methods that involved empirical rules in papers III and IV, and low-order (first- or second-order) probabilities in papers IV and V. The general method is applied here to the prediction of the backbone conformations of proteins, using the three-state model [helical (h), extended (ϵ), and other or coil (c) states] proposed in the theoretical formulation of paper II. The statistical weights in the three-state model are evaluated from the atomic coordinates of the x-ray structures of 26 proteins. The conformational-sequence probabilities (taken for three consecutive residues for numerical computation in this paper) are calculated for all possible triads (i.e., for all possible combinations of the three states, h, ϵ, and c for each residue) for bovine pancreatic trypsin inhibitor and clostridial flavodoxin, in order to select the most probable conformations of these proteins. The predicted results for these proteins are compared to those predicted in paper III and to those observed experimentally. The method is applied further to the prediction of the backbone structures of homologous neurotoxin proteins whose amino acid sequences are known but whose x-ray structures are not. The effects of variation in the amino acid sequence on the conformations of the backbones are discussed from the point of view of the homologies in the amino acid sequences of 19 neurotoxins. Application of the present general predictive method to a four- and a multistate model is also described.

In this series of papers, we have developed a statistical mechanical treatment of protein conformations within the context of one-dimensional short-range interaction models. These will be referred to here as papers I,[3] II,[4] III,[5] IV,[6] and V,[7] with equations designated as I-1, II-1, III-1, etc. These models are intended to be used in a first step of a protein folding procedure,[8] in which longer range interactions are introduced in the subsequent steps.

To be more specific, in paper I, we presented a method for evaluating statistical weights of various conformational states of amino acid residues on the basis of conformational information (but not atomic coordinates) from x-ray crystal structures of native proteins.[3] In paper II, a three-state model for specific-sequence polypeptides that included helical (h), extended (ϵ), and other or coil (c) states was formulated,[4] and its application to the prediction of protein conformation in terms of h, ϵ, and c states was described in paper III.[5] In paper IV, we extended this treatment to a four-state model that included helical (h), extended (ϵ), chain-reversal (R and S), and other or coil (c) states and evaluated the statistical weights for these states on the basis of the x-ray coordinates of native proteins.[6] Subsequently, in paper V, a multistate model that, in principle, can include any number of conformational states, but actually dealt only with the right-handed helical (h_R), extended (ϵ), left-handed helical (h_L), chain-reversal (R and S), right-handed bridge region (ζ_R), and other or coil (c) states, was formulated and applied to the prediction of protein conformation.[7] When these short-range interaction models are applied to the prediction of protein conformation, several problems arise; these are summarized below.

(1) Even though the theory is a statistical mechanical one, it was necessary to introduce empirical rules to predict h, ϵ, and c states in proteins (see section III of paper III[5]). Empirical rules were necessary in paper III, since we had conformational information available only about helical and extended *sequences* (and not about individual isolated residues) because the x-ray coordinates were not available and resort was had to crystallographers' statements about the location of helical and extended sequences. Therefore, in this paper, we develop a procedure which eliminates the need for such empirical rules.

(2) When the three-state model[4] was extended to predict chain-reversal conformations by rule I (see section V of paper IV[6] and section V of paper V[7]), there was no ambiguity in the predictive scheme, even though a first- and second-order probability was used [see point (4) below]. However, in rule II of section V of paper IV,[6] we combined the empirical rules of section III of paper III[5] with rule I of paper IV[6] to eliminate duplicate assignments between helical sequences and chain reversals (R and S states) or between extended sequences and chain reversals; the combination of the rules of papers III and IV in this manner is itself an empirical procedure. The procedure developed here enables one to predict h, ϵ, R, S, and c states simultaneously, without resort to these empirical rules, since the statistical weight matrix of the four-state model[6] includes all of these states.

(3) In papers II[4] and III[5] (see eq II-54 and III-9), we used a high-order probability, $P(i|n|\{\rho\})$, to predict *regular* helical or extended sequences of n residues (starting at the ith residue) where

$$\{\rho\} = h_i h_{i+1} \ldots h_{i+n-1} \quad (1a)$$

$$\{\rho\} = \epsilon_i \epsilon_{i+1} \ldots \epsilon_{i+n-1} \quad (1b)$$

and

$$\{\rho\} = c_i c_{i+1} \ldots c_{i+n-1} \quad (1c)$$

Helical sequences of less than three residues, or extended sequences of less than four residues, could not be assigned by the predictive scheme of paper III[5] and were assigned to the c state (when computing the statistical weights in paper I[3] and when predicting helical and extended sequences with the

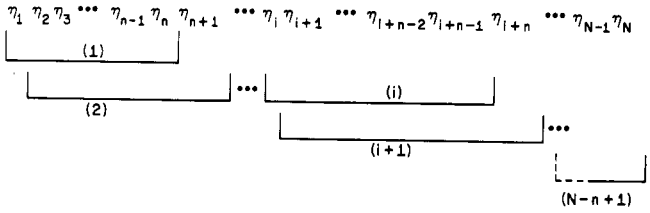

Figure 1. Illustration of the conformational states of a chain of N residues. The residues are considered in groups of n each, beginning at the 1st, 2nd, ith, $(i + 1)$th, and $(N - n + 1)$th, respectively.

three-state model[4] in paper III[5]). The reason for treating h, hh, ϵ, $\epsilon\epsilon$, and $\epsilon\epsilon\epsilon$ sequences in this manner is that they could not be identified because the x-ray coordinates were not available at the time that the computations were carried out. With the subsequent availability of x-ray coordinates on many proteins, the statistical weights were recomputed,[6,7] and short helical and extended sequences could then be identified. In fact, it then became unnecessary to distinguish among extended sequences of different lengths.[6,7] However, it is still necessary[6,7] to distinguish h and hh sequences from helical sequences of three or more residues because the h and hh sequences do not involve the hydrogen bond that is characteristic of the α helix. Therefore, using the three-state model for illustrative purpose, we are concerned here only with isolated helical states (h and hh), helical sequences of three residues or more, extended sequences of *any* length, and c states. However, we no longer wish to be confined to the *regular* sequences of eq 1. In order to predict the conformational states of, say, n residues, it is necessary to use an nth order a priori probability rather than the first-order a priori probability for the state of a *single* residue; i.e., we seek the value of $P(i|n|\{\rho\})$, where

$$\{\rho\} = \eta_i\eta_{i+1}\eta_{i+2}\cdots\eta_{i+n-1} \tag{2}$$

and η_i, η_{i+1}, etc., can be *any* one of three states,[9] h, ϵ, or c (in the three-state model to be considered here as an example). Thus, $\{\rho\}$ no longer will be restricted to *regular* sequences, as in eq 1, but will be any combination of three states, such as

$$\{\rho\} = \text{chhc}\ldots\epsilon\epsilon\text{hhhh}\epsilon\ldots\text{c} \tag{3}$$

and so on.

(4) As indicated in point (2) above, there was no ambiguity in using a first- and second-order probability (described as rule I in paper IV[6]) to predict chain reversals, but, in using rule II,[6] an empirical procedure was used. There is also no ambiguity in using a first-order a priori probability in the predictive scheme of the multistate model of paper V.[7] However, in papers IV and V, we used *low-order* probabilities [first and second order in the predictive scheme of the four-state model (see section VB of paper IV[6]) and only first order in the predictive scheme of the multistate model (see section V of paper V[7])]. It thus becomes desirable to use higher order probabilities to detect helical sequences; however, first-order probabilities suffice for *isolated* h, ϵ, and c states, and second-order probabilities (see eq IV-30) for chain reversals. Similarly, in using the multistate model,[7] first-order a priori probabilities suffice for the prediction of *isolated* right (h_R)- and left (h_L)-handed helical, *isolated* extended (ϵ), right (ζ_R)- and left (ζ_L)-handed bridge region, and other (c) states; however, it is desirable to use higher than first order probabilities to predict helical and extended sequences, even in the multistate model.

The purpose of this paper is to eliminate the empirical rules described in (1) and (2) above and to use higher order probabilities to improve the situation described in (3) and (4). We will use probabilities of high order to predict helical and extended sequences in all models and to predict isolated h, isolated ϵ, and isolated c states in the three-state model, chain reversals (in addition to the states of the three-state model) in the four-state model, and isolated h_L and isolated ζ_R (in addition to the states of the four-state model) in the multistate model. For this purpose, all possible conformational states of a sequence of n residues will be taken into account, as pointed out in (3) above.

In section IA of this paper, a general predictive scheme, that can be applied to one-dimensional short-range interaction models with any number of allowed states, is presented; the method for calculating conformational-sequence probabilities is described briefly in section IB. In section II, the general predictive scheme described in section I is applied to the three-state (h, ϵ, and c) model, and the statistical weights for this model are evaluated using x-ray coordinates (to replace the tentative values given in paper I[3]). In section III, the backbone conformations of bovine pancreatic trypsin inhibitor (BPTI) and clostridial flavodoxin are predicted (in terms of h, ϵ, and c), using the method described in sections I and II and the statistical weights of section IIB; these are compared to the predictive results of the previous paper[5] and to the experimentally observed conformations.[10,11] In section IV, predictions are made for a series of homologous neurotoxins, whose structures have not yet been determined experimentally, in order to investigate the effect of amino acid sequence homologies on the backbone conformations of proteins. Finally, in section V, we indicate how to apply the present general predictive method to a four-[6] and a multistate[7] model and present a concluding discussion of the role that is intended for these short-range interaction models in determining protein conformation.[8]

I. General Predictive Method

(A) A General Predictive Scheme. In this section, we describe a general predictive method that is applicable to a model consisting of any number of states and will restrict it to a three-state model in section II.

We assume that the conformation of the polymer chain can be described by a Markov process; i.e., the conformational state of the ith residue depends on the conformational states of the preceding ones ($i - 1$, $i - 2$, etc.) but not on the conformational states of the succeeding residues ($i + 1$, $i + 2$, etc.). The residues are numbered sequentially, from 1 to N, from the N to the C terminus of the chain[12,13] and the conformational states may be represented as in Figure 1.

The above statement, however, does not mean that cooperativity is neglected. It is important to realize that, in the treatment of the conformation of a chain, even in a short-range interaction model, we do *not* determine the conformation of a sequence, say, from η_i to η_{i+n-1} without taking account of the effect of the other portions of the molecule (viz., η_1 to η_{i-1} and η_{i+n} to η_N; see Figure 1). In other words, one cannot determine the conformational states of residues i to $i + n - 1$ by taking the product of statistical-weight matrices or by computing conformational energies for the segment from residue i to $i + n - 1$ (even if its energy of interaction over the whole molecule is taken into account) (this is true, in principle; however, in practice it may be a valid first approximation; see footnote 45 of paper I). In order to determine the conformational states for residues i to $i + n - 1$, one has to take into consideration all possible states of the rest of the molecule (viz., η_1 to η_{i-1} and η_{i+n} to η_N). For example, in order to determine the most probable conformation for residues i to $i + n - 1$, the conformational energy has to be calculated by allowing for all possible states for residues 1 to $i - 1$ and $i + n$ to N, while maintaining the sequence of residues i to $i + n - 1$ in the given conformational state $\{\rho\}$. Even using statistical weights, one cannot determine the most probable conformation of residues i to $i + n - 1$ simply by multiplying the sta-

tistical-weight matrices for this segment only. However, the Markov treatment of polymer conformation (to be used in the present predictive scheme) *can* provide the most probable conformation of the segment form i to $i + n - 1$, by considering only that segment, for the following reason. In the Markov process, we describe the conformation in terms of an a priori probability and conditional probabilities, both of which are evaluated by averaging over all possible conformational states of the whole molecule. In other words, the conformational behavior of the whole molecule is taken into account in computing the most probable conformational state for a segment of the molecule in the present predictive scheme using a Markov process. Thus, when calculating the a priori and conditional probabilities [hence, $P(i|n|\{\rho\})$ from eq 5 of section IB], the cooperativity is taken into account. While statistical weights (such as those of Table I) are required to compute the a priori and conditional probabilities, the predictions are not based on the statistical weights but on $P(i|n|\{\rho\})$ which is calculated from the a priori and conditional probabilities.

On the basis of the assumption stated above (viz., that the conformation of the polymer chain can be described by a Markov process), the conformation η_{i+n-1} of residue $i + n - 1$ depends on that of residue $i + n - 2$, and hence on the preceding i to $i + n - 2$ residues (i.e., on $\eta_i, \eta_{i+1}, \ldots, \eta_{i+n-2}$), when $P(i|n|\{\rho\})$ is calculated using the nearest-neighbor approximation. Therefore, in order to determine the most probable conformation of residue $(i + n - 1)$, it is necessary to first know the conformations $\eta_i, \eta_{i+1}, \ldots, \eta_{i+n-2}$ and then the effect of the conformational state $\eta_i\eta_{i+1}\ldots\eta_{i+n-2}$ on the state η_{i+n-1}.

With this formulation, we may consider a method to predict the most probable conformation of a protein (with a short-range interaction model) in more detail. Two general steps are required. (1) First, since we are correlating the state of any residue with that of its preceding one (using the short-range interaction model[4,6,7]), and because the residue at the N terminus[13] of the chain has no preceding one, we will begin by determining the most probable conformation for a sequence of residues 1 to n [see (1) of Figure 1] out of all possible conformations of this sequence; i.e., we will allow $\eta_1, \eta_2, \ldots, \eta_n$ each to vary over all possible states allowed for in the model (h, ϵ, and c in the three-state model,[3–5] h, ϵ, R, S, and c in the four-state model,[6] and h_R, h_L, ϵ, R, S, ζ_R, and c in the multistate model).[7] We will use the symbol η^* to designate the number of states in the model ($\eta^* = 3$, 5, and 7 in the three-, four-, and multistate models, respectively). The number of possible conformational states of n residues is thus η^{*n}. The most probable conformation of the sequence of residues 1 to n can be determined by considering the probabilities of occurrence of all possible η^{*n} conformational states among the first n residues, as will be discussed more explicitly in section IB.

(2) Second, having determined the most probable conformation of residues 1 to n in step (1), we also know the most probable conformation of residues 2 to n, which will affect the conformational state η_{n+1} of residue $n + 1$ (see Figure 1). Thus, we fix $\eta_2, \eta_3, \ldots, \eta_n$, as determined in step (1), and calculate the probabilities of occurrence of the sequence $\eta_2\eta_3\ldots\eta_{n+1}$, by allowing only η_{n+1} to take on all allowed conformations, as will be described in section IB. When evaluating η_{n+1}, with η_2 to η_n fixed the (earlier) fixed value of η_1 (which influenced η_2 to η_n) has no influence on η_{n+1}, i.e., the statistical mechanical averaging takes into account all possible values of η_1 when computing η_{n+1} (even though a prediction has already been made for η_1). Thus, only η^* conformational sequences (i.e., the number of states allowed for residue $n + 1$) have to be examined to determine the most probable one for residues 2 to $n + 1$. This procedure is then repeated over and over again, each time fixing the conformations of residues i to $i + n - 2$ in order to determine the most probable conformation η_{i+n-1}, until the most probable conformational states of all N residues of the protein chain are determined; i.e., the value of i is varied from $2 \leq i \leq N - n + 1$ to determine η_{n+1}, $\eta_{n+2}, \ldots, \eta_N$ (since the conformations $\eta_1, \eta_2, \ldots$, and η_n were determined in step 1, described above). The model described above assumes that, while a change in conformation can originate in *any* part of the chain, the direction of the conformational change is $i \rightarrow j$, where $1 \leq i < j \leq N$.

Table I
Relative Statistical Weights of the jth Type of Amino Acid in the Three-State Model[a]

Amino Acid j	Relative Statistical Weight[b]		
	$w_{h,j}^*$	$v_{h,j}^*$	$v_{\epsilon,j}^*$
Ala	1.46	0.067	0.867
Arg	1.00	0.075	1.13
Asn	0.425	0.018	0.531
Asp	0.573	0.053	0.336
Cys	0.611	0.028	1.14
Gln	0.843	0.098	0.980
Glu	1.83	0.053	0.754
Gly	0.246	0.012	0.385
His	0.974	0.132	0.816
Ile	1.61	0.130	2.15
Leu	1.65	0.107	1.55
Lys	1.00	0.068	0.644
Met	1.44	0.0	1.33
Phe	1.18	0.068	1.16
Pro	0.413	0.111	1.03
Ser	0.510	0.026	0.839
Thr	0.577	0.113	1.09
Trp	0.957	0.130	1.17
Tyr	0.547	0.047	1.22
Val	1.37	0.038	2.06

[a] The statistical weights given in this table[14] were evaluated from the dihedral angles computed from the x-ray coordinates of the 26 proteins listed (together with references to the original papers) in Table I of paper IV. [b] The statistical weights in this table are expressed relative to those of the c state; i.e., $u_{c,j}^* = 1$ for all amino acids.[14]

(B) Calculation of Conformational-Sequence Probability. In this section, we describe the method for computing the conformational-sequence probability of a sequence of n residues, to be used to detect the most probable conformation of a protein by the predictive method described in section IA. Since a general method for computing the conformational-sequence probability was described in section VIA of paper II,[4] we will discuss it here only briefly insofar as it will be applied in the predictive method presented in section IA. The comments in the third paragraph of section IA, about statistical mechanical averaging over the whole molecule, are pertinent here.

Consider a specific conformational sequence $\{\rho\}$ given in eq 2, and designate the probability of finding residues i to $i + n - 1$ in this conformational state as $P(i|n|\{\rho\})$. The value of $P(i|n|\{\rho\})$ can be calculated from eq II-48, viz.,

$$P(i|n|\{\rho\}) = Z^{-1}\mathbf{e}_1\left[\prod_{j=1}^{i}\mathbf{W}_j\right] \times \left[\prod_{k=i+1}^{i+n-1}\frac{\partial \mathbf{W}_k}{\partial \ln (m_{k;\eta_{k-1}\eta_k})}\right]_{|\rho|}\left[\prod_{l=i+n}^{N}\mathbf{W}_l\right]\mathbf{e}_N^* \quad (4)$$

where Z is the partition function of the protein chain, and can be computed form eq II-39 for a general model, from eq II-21 for the three-state model, from eq IV-10 for the four-state model, and from eq V-12 for the multistate model. The quantity $m_{k;\eta_{k-1}\eta_k}$ is the statistical weight of the conformational state η_k. Alternatively, to avoid the time-consuming *repetitive* matrix multiplication of eq 4, the values of

$P(i|n|\{\rho\})$ can be computed from eq II-54, viz.:

$$P(i|n|\{\rho\}) = F_{i;\eta_i}\left[\prod_{k=i}^{i+n-1} P_{k+1;\eta_k\eta_{k+1}}\right]_{\{\rho\}} \quad (5)$$

where $F_{i;\eta_i}$ is the first-order a priori probability (given by eq II-44), and $P_{k+1;\eta_k\eta_{k+1}}$ is the conditional probability (given by eq II-63). As discussed in section VIC of paper II,[4] eq 5 is more convenient to use than eq 4, from a computational point of view, but both lead to the same results. Also, as pointed out in paper II,[4] one cannot compute $P(i|n|\{\rho\})$ as a product of the first-order a priori probabilities for the n residues.

The mean value (or average fraction) of $P(i|n|\{\rho\})$, for the specific conformation $\{\rho\}$, over the whole chain is given by eq II-64 as

$$\theta_{\{\rho\}}(n) = \frac{1}{N-n+1}\sum_{i=1}^{N-n+1} P(i|n|\{\rho\}) \quad (6)$$

As pointed out in footnote 45 of paper I,[3] it is difficult, at present, to evaluate the cooperativity that is represented by the statistical weight $v_{h,j}$* in the present models. If the cooperativity were taken into account properly by $v_{h,j}$*, it would suffice to compare the values of $P(i|n|\{\rho\})$ in order to determine the most probable conformation of a given residue. Fortunately, however, although the absolute values of $P(i|n|\{\rho\})$ are sensitive to the values of $v_{h,j}$*, the *relative* values [$P^*(i|n|\{\rho\})$, defined in eq 7] are fairly insensitive to the values of $v_{h,j}$*, because the shape of the curve of $P(i|n|\{\rho\})$, independent of its absolute value, is determined primarily by the values of $w_{h,j}$*. In other words, variation in the values of $v_{h,j}$* serves mainly to alter the mean value, $\theta_{\{\rho\}}{}^{(n)}$, without altering the overall shape of the conformational-probability profile to a great extent. Therefore, by introducing the relative quantity $P^*(i|n|\{\rho\})$, defined by eq III-15 as

$$P^*(i|n|\{\rho\}) = P(i|n|\{\rho\})/\theta_{\{\rho\}}{}^{(n)} \quad (7)$$

the most probable conformational state of a given residue or sequence can be chosen, even though the absolute values of $P(i|n|\{\rho\})$ are not accurate. Work is now in progress to try to improve the accuracy of the values of $v_{h,j}$*; this will enable $P(i|n|\{\rho\})$ to be used directly for predicting conformation. Thus, the use of $P^*(i|n|\{\rho\})$ instead of $P(i|n|\{\rho\})$ should be regarded as a tentative procedure.

In step (1) of section IA, the values of $P^*(1|n|\{\rho\})$ are computed for all η^{*n} possible conformational sequences of the first n residues of the chain. The most probable conformational state of the first n residues is then the one with the largest value of $P^*(1|n|\{\rho\})$. In step (2), and in all succeeding steps, the values of $P^*(i|n|\{\rho\})$ are computed for all of the η^* conformational states allowed for n residues, since there are η^* states for residue $i+n-1$ and $\eta_i \ldots \eta_{i+n-2}$ is fixed. Again, the most probable conformational state for these n residues is the one with the largest value of $P^*(i|n|\{\rho\})$; thus, the most probable conformational state η_{i+n-1} of residue $i+n-1$ is determined.

The method described in sections IA and IB can be applied to models with any number η^* of states allowed for each residue. This predictive scheme will allow us to eliminate the empirical rules that were used to predict *sequences* of h, ϵ, and c states in section III of paper III[5] (see section II, below), as indicated in point (1) of the introductory section. It will also enable us to improve the situation described in points (2) to (4) of the introductory section.

II. Application to the Three-State Model

(A) Model and Predictive Method. In this section, we apply the general predictive scheme of section I to the three-state model (h, ϵ, and c states) that was treated in papers I–III, but now we can dispense with the empirical rules of paper III. Since this three-state model was described in detail in paper II, we will not repeat its formulation here.

In this application here, $\eta^* = 3$ (three-state model) and we will take the sequence length n (for which to compute the conformational-sequence probability) as 3. This value of n corresponds to the minimum length sequence that can form an α-helical conformation (see, for example, paper II). (Of course, larger values of n can be considered, but this increases the required computer time considerably.) Therefore, in step (1) of the predictive method described in section IA, we have to compute the values of $P^*(1|n|\{\rho\})$, where $n = 3$, for the various possible triads of the conformational states h, ϵ, and c, i.e., for $\eta^{*n} = 27$ different conformational sequences such as $\{\rho\}$ = hhh, hεh, hcε, εεε, etc.; for this purpose, we use eq 7, together with eq 5 and 6 and eq II-21, II-22, and II-26. Thus, the most probable conformation for the three residues (residues 1, 2, and 3) at the N terminus[13] of the protein chain can be determined. Then, in the repetitive use of step (2) of the predictive method of section IA, the values of $P^*(i|n|\{\rho\})$, where $n = 3$, are calculated for the triad of residues $i, i+1$, and $i+2$. In this calculation, the conformations of residues i and $i+1$ are fixed at those determined in the previous step (in which the triads of residues $i-1$, i, and $i+1$ were investigated), and η_{i+2} is varied over all three states h, ϵ, and c to obtain all the possible conformational sequences $\{\rho\}$ for $n = 3$. Thus, only three values of $P^*(i|n|\{\rho\})$ are calculated, using the same equations employed above in the computation of $P^*(1|n|\{\rho\})$. When η_{i+2} is determined, the process is repeated until the conformation of the whole chain is determined. It should be remembered that, even though η_i and η_{i+1} are fixed when computing η_{i+2}, the values of $P^*(i|n|\{\rho\})$ used to obtain η_{i+2} involve averaging over *all* possible states of the rest of the molecule.

(B) Statistical Weights for the Three-State Model. In order to compute numerical values of $P^*(i|n|\{\rho\})$, we need the statistical weights for the amino acid residues in the three-state model (in eq II-26). These are $w_{h,j}$*, $v_{h,j}$*, and $v_{\epsilon,j}$* for the jth type of amino acid, defined in paper II.[4] As in papers I–III, these statistical weights are expressed relative to a value of unity for the statistical weight of the c state.

As in papers IV and V, but in contrast to papers I–III, the statistical weights are computed from the x-ray coordinates of 26 proteins (see Table I of paper IV[6] for the references to these 26 proteins). The definitions used here for the h and ϵ states are those given in section IIIA of paper IV.[6] Hence, the numbers of helical ($N_{h,j}$), isolated helical ($N_{h',j}$), and extended ($N_{\epsilon,j}$) states are the same as those given in Table II of paper IV. The number of c states for the jth type of amino acid ($N_{c,j}$) is given by

$$N_{c,j} = N_j - (N_{h,j} + N_{h',j} + N_{\epsilon,j}) \quad (8)$$

where N_j is the total number of the jth type of amino acid found in the 26 proteins and is given in the second column of Table II of paper IV. Using these numbers, the statistical weights (relative to that of the c state) were computed by the following relations:

$$w_{h,j}{}^* = f_{h,j}/f_{c,j} \quad (9)$$

$$v_{h,j}{}^* = f_{h',j}/f_{c,j} \quad (10)$$

and

$$v_{\epsilon,j} = f_{\epsilon,j}/f_{c,j} \quad (11)$$

where

$$f_{\eta,j} = N_{\eta,j}/N_j \quad (12)$$

and η = h, h′, ϵ, or c (see footnote 41 of paper IV for a discussion of the procedure to evaluate $f_{h,j}$ and $f_{h',j}$, and hence $w_{h,j}$*

Table II
Conformational-Sequence Probabilities for all Possible Conformational Triads of the First Three Residues of BPTI

$\eta_1\eta_2\eta_3$	$P^*(1\|n\|\{\rho\})$[a]
ccc	1.13
cεc	1.40
chc	5.67
cce	0.44
cεε	0.50
chε	2.28
cch	1.39
cεh	1.61
chh	5.78
εcc	1.43
εεc	1.65
εhc	6.85
εce	0.57
εεε	0.61
εhε	2.84
εch	1.57
εεh	1.81
εhh	7.14[b]
hcc	0.65
hεc	0.82
hhc	0.38
hce	0.24
hεε	0.28
hhε	0.14
hch	0.68
hεh	0.83
hhh	0.33

[a] For $n = 3$. [b] The εhh triad is the one with the highest conformational-sequence probability.

and $v_{h,j}$*). The numerical values of the statistical weights for the three-state model,[14] evaluated in this manner for the 20 naturally occurring amino acids, are given in Table I.

III. Results for BPTI and Clostridial Flavodoxin

Using the three-state model and the statistical weights of section II, we carry out here the prediction of the h, ε, and c states for two proteins and compare the results to those of paper III[5] and to those observed experimentally.[10,11] The predictive procedure will be described in detail for BPTI in section IIIA, and the results for BPTI and clostridial flavodoxin will be summarized in section IIIB.

(A) Details of the Predictive Scheme. In order to illustrate the details of the predictive scheme, we use BPTI as an example because of its short chain length ($N = 58$). According to step (1) of section IA, applied to the three-state model (h, ε, and c states), we compute $P^*(1|n|\{\rho\})$, with $n = 3$, for the 27 possible conformational triads of the first three residues of the chain. These values of $P^*(1|n|\{\rho\})$ are really relative probabilities of occurrence of the given conformational states. The values are listed in Table II, where it can be seen that εhh is the most probable conformation for residues 1, 2, and 3. This conformation for the first three residues is shown in Figure 2, where $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$ is plotted against residue number, i.

Proceeding to step (2) of section IA (applied to the three-state model), we keep $\eta_2\eta_3$ fixed as hh and compute the values of $P^*(2|n|\{\rho\})$, for $n = 3$, for the conformational triads of residues 2, 3, and 4, viz., $\eta_2\eta_3\eta_4$ = hhh, hhε, and hhc. The resulting values of $P^*(2|3|\{\rho\})$ are indicated by the symbols ○, △, and ●, respectively (for $i = 2$), in Figure 2. It can be seen that the largest value of $P^*_{\eta_2\eta_3\eta_4}$ is that for the triad indicated by the symbol ● at $i = 2$, viz., hhc. Thus, the most probable conformation of residue 4 is c, and that for the first four residues is εhhc. Repetitive application of this procedure gives all values of $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$, for $2 \le i \le N - 2$; these are plotted as a function of i in Figure 2. The most probable conformation for η_{i+2} is that with the largest value of $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$; these are connected by the solid line (with a dashed line for residues 1, 2, and 3) in Figure 2, and represent the most probable conformation of BPTI, in *this three-state model.*

(B) Comparison of Predicted and Experimentally Observed Conformations. The results for BPTI, obtained as indicated in section IIIA, are summarized in Table III, together with the experimentally observed ones and those obtained by the predictive scheme of paper III.[5] The results for clostridial flavodoxin, obtained in a similar manner, are shown in the same way in Table IV.

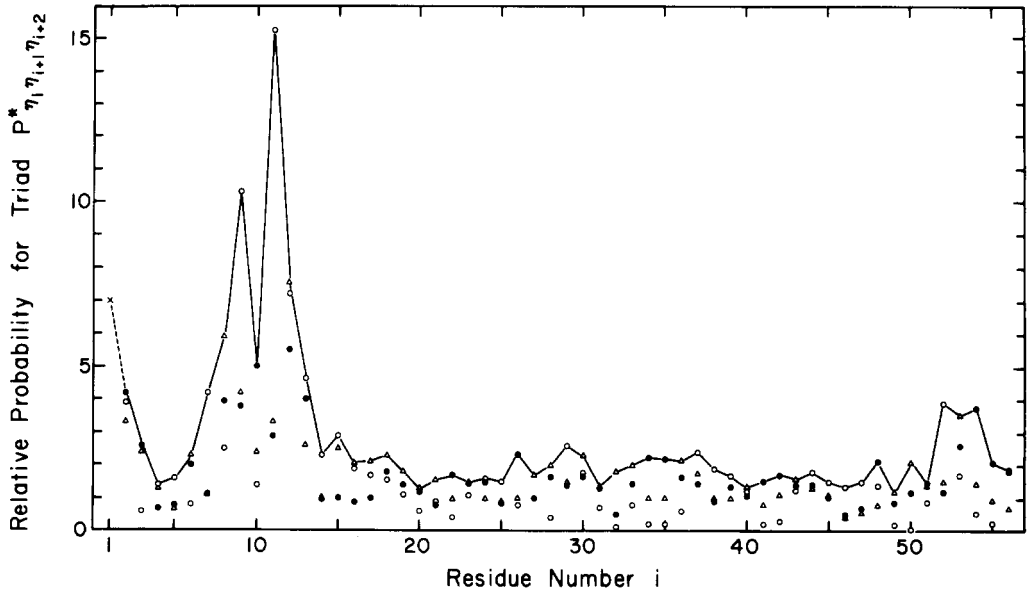

Figure 2. Conformational-sequence probability, $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$, of a triad as a function of residue number, i, for BPTI. The symbols ○, △ and ● designate h, ε, and c states, respectively. The conformation predicted by $P_{\eta_i\eta_{i+1}\eta_{i+2}}$ is actually that of residue $i + 2$; thus, e.g., η_N is obtained when $i = N - 2$. The symbol × at $i = 1$ designates the value of $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$ for the first triad (residues 1, 2, and 3), which is intended to imply that the most probable state of this triad is εhh. In calculating $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$, η_i and η_{i+1} are fixed at the conformations determined in the previous step, and only η_{i+2} is varied (see section IA). The solid line (and dashed line for residues 1, 2, and 3) denotes the most probable conformational state of BPTI.

Table III
Experimentally Observed and Predicted Conformations (Three-State Model) for Bovine Pancreatic Trypsin Inhibitor

Residue Number	Observed[a,b] Conformation	Predicted Results by Three-State Model Using: Predictive[b] Scheme of This Paper	Predictive[b,c] Scheme of Paper III
1	ε	ε	
	ε	h	
	h	h	
	h	c	h
	h	c	h
	c	h	h
	ε	h	h
	ε	ε	
	ε	h	
10	ε	ε	
	h	h	
	c	c	
	c	h	h
	ε	ε	h
	c	h	h
	ε	h	h
	c	h	ε
	ε	ε	ε
	ε	ε	ε
20	ε	ε	ε
	ε	ε	ε
	ε	ε	
	ε	ε	h
	ε	c	h
	h	ε	h
	h	h	h
	h	h	h
28	c	c	h
29	ε	ε	ε
30	ε	ε	ε
	ε	h	ε
	ε	ε	ε
	ε	ε	ε
	ε	ε	ε
	ε	ε	ε
	h	c	ε
	c	c	
	ε	ε	
	c	h	
40	ε	h	
	ε	h	
	c	ε	
	c	c	
	ε	c	h
	ε	ε	h
	h	h	h
	ε	h	h
	h	h	h
	h	h	h
50	h	c	h
	h	ε	ε
	h	ε	ε
	h	c	ε
	h	h	ε
	c	ε	
	h	c	
	c	c	
58	ε	c	

[a] The observed conformation of BPTI cited in this table is that of Table III of paper V,[7] where helical and extended sequences of *all* lengths are given. However, in Table I of paper IV,[6] the only helical and extended sequences that are listed are those that are ≥3h and 4ε residues, respectively; the procedure for counting h, hh, ε, εε, and εεε sequences properly is described in section IIIB of paper 4.[6] [b] The symbols h, ε, and c designate helical, extended, and other (coil) states, respectively. [c] These results were taken from Table IV of paper III.[5] The prediction scheme of paper III can detect only h and ε *sequences* and not isolated h and ε states. Hence, in this column, the blank spaces include not only the c states but also the isolated h and ε states.

These results should be considered in light of the question: "how applicable is the short-range prediction model, and the statistical weights used therein?" The statistical weights will become more accurate as more x-ray data become available (see section IIIE of paper I[3] for a discussion of possible long-range effects on the statistical weights). The inaccuracy in the statistical weight $v_{h,j}$* (which led to the introduction of $P^*(i|n|\{\rho\})$; see eq. 7) makes it difficult to locate *long* helical sequences, as indicated by the inability of the model to predict the long helical sequences at residues 93–106 and 125–136 of clostridial flavodoxin (see Table IV). While the framework of the theory is a reasonable one, its reliability can be improved by computing even higher order probabilities (i.e., $n > 3$) than those used here and by using more accurate statistical weights in the future. The incorporation of long-range interactions in a protein predictive scheme is discussed in section V.

Finally, in evaluating the results presented here and in comparing them with those of the empirical prediction schemes discussed in papers I[3] and III,[5] it should be realized that most empirical predictive schemes pertain to helical, extended, or coil *sequences*. Thus, *short* sequences of h's or ε's would be relegated to the c state in such schemes. However, in the procedure described here, we consider any combination $\{\rho\}$ of three states h, ε, and c in predicting the conformational state of every residue.

IV. Correlation between Amino Acid Sequences and Predicted Conformations of Homologous Neurotoxins

In this section, we apply the predictive method of sections I and II to proteins whose amino acid sequences are known, but whose structures are not. In order to investigate the correlation between amino acid sequences and predicted backbone conformations of proteins, a series of homologous neurotoxins was chosen.

The amino acid sequences of the 19 homologous neurotoxins investigated here are given in Table V in terms of the one-letter code recommended by an IUPAC–IUB commission.[15] (See Table VIII for the homologous amino acids in these neurotoxins.) Using the same procedures as applied in section III, the backbone conformations of all 19 homologues[16–30] were predicted and are shown in the last column of Table V.

In order to see how the amino acid sequence affects the predicted conformation in these homologous neurotoxins, we summarize in Table VI the degree of sequence homology (expressed as number of identical residues between pairs of proteins) and in Table VII the degree of (predicted) conformational identity between pairs of proteins. In Table V, the homologies are indicated by lining up the amino acid sequences in optimal fashion, with dashes indicating deletions.

Table IV

Experimentally Observed and Predicted Conformations (Three-State Model) for Clostridial Flavodoxin

Residue Number	Observed[a,b] Conformation	Predicted Results by Three-State Model Using: Predictive[b] Scheme of This Paper	Predicted Results by Three-State Model Using: Predictive[b,c] Scheme of Paper III
1	ε	ε	ε
	ε	c	ε
	ε	ε	ε
	ε	ε	ε
	ε	ε	ε
	ε	b	ε
	ε	c	ε
	c	c	
	c	h	
10	c	c	
	h	c	h
	h	h	h
	h	h	h
	h	h	h
	h	h	h
	h	h	h
	h	ε	h
	h	h	h
20	h	h	h
	h	c	h
	h	c	h
	h	ε	h
	h	h	h
	h	h	h
	h	c	
	c	c	
	ε	c	
	ε	c	
30	ε	ε	
	ε	c	
	ε	h	
	ε	ε	
	ε	c	
	c	ε	
	c	c	
	c	c	
	ε	ε	
	c	c	h
40	c	h	h
	c	c	h
	c	c	h
	c	ε	h
	c	ε	h
45	c	c	

Residue Number	Observed[a,b] Conformation	Predicted Results by Three-State Model Using: Predictive[b] Scheme of This Paper	Predicted Results by Three-State Model Using: Predictive[b,c] Scheme of Paper III
46	ε	c	ε
	c	c	ε
	c	ε	ε
	ε	ε	ε
50	ε	ε	ε
	ε	ε	ε
	ε	c	ε
	ε	ε	ε
	ε	c	
	ε	c	
	ε	ε	
	c	c	
	c	h	
	c	h	
60	ε	ε	
	ε	h	
	c	h	
	h	h	
	h	c	
	h	c	
	h	ε	
	h	h	h
	h	c	h
70	h	h	h
	h	h	h
	h	h	h
	h	ε	h
	h	c	
	c	h	
	c	c	
	c	ε	
	c	c	
	c	c	
80	ε	h	ε
	ε	h	ε
	ε	ε	ε
	ε	c	ε
	ε	ε	ε
	ε	ε	ε
	ε	c	ε
	ε	c	ε
	ε	ε	
	c	c	
90	c	h	
	ε	c	
92	c	c	

Residue Number	Observed[a,b] Conformation	Predicted Results by Three-State Model Using: Predictive[b] Scheme of This Paper	Predicted Results by Three-State Model Using: Predictive[b,c] Scheme of Paper III
93	c	c	h
	h	c	h
	h	h	h
	h	ε	h
	h	h	
	h	c	h
	h	h	h
100	h	c	h
	h	c	h
	h	ε	
	h	ε	
	h	c	
	h	c	
	h	ε	ε
	c	c	ε
	ε	ε	ε
	ε	ε	ε
110	ε	ε	ε
	c	ε	ε
	ε	c	ε
	ε	h	ε
	ε	ε	ε
	ε	ε	ε
	ε	ε	ε
	ε	ε	ε
	ε	h	ε
	c	c	ε
120	ε	c	
	h	h	
	c	h	
	c	h	
	c	h	
	h	h	
	h	ε	
	h	c	
	h	ε	
	h	h	
130	h	h	
	h	c	
	h	c	h
	h	c	h
	h	c	h
	h	ε	h
	h	c	h
	c	c	
138	ε	ε	

[a] The observed conformation of clostridial flavodoxin was given in Table IV of paper V.[7] See also footnote *a* of Table III. [b] See footnote *b* of Table III. [c] See footnote *c* of Table III.

Table V
Comparison of the Amino Acid Sequences and Predicted Conformations of Homologous Neurotoxins

	Protein	Number of Residues	Amino Acid Sequence[a]	Predicted Conformation[b]
			20 40 60	20 40 60
1	Naja naja atra; cobrotoxin[b]	62	LECHNQQSSQTPTTTGCSGGETNCYKKRWRD-H---RGYRTERGC--GCPSVK-NGIEINCCTT-DRCNN	ccehchcccehehehccccccchcccccchhh-h---cccchhccc--cchccc-cchhccchc-cccccc
2	Laticauda semifasciata; erabutoxin a[d]	62	RICFNQHSSQPQTTKTCPSGSESCYNKQWSD-F---RGTIIERGC--GCPTVK-PGIKLSCCES-EVCNN	ccccchhcchehehhhchccccccccchccc-c---cchhhhccc--cchccc-hccccccccc-ccccc
3	Laticauda semifasciata; erabutoxin b[e]	62	RICFNQHSSQPQTTKTCPSGSESCYHKQWSD-F---RGTIIERCG--GCPTVK-PGIKLSCCES-EVCNN	ccccchhcchehehhhchccccccchhhccc-c---cchhhhccc--cchccc-hccccccccc-ccccc
4	Laticauda semifasciata; erabutoxin c[f]	62	RICFNQHSSQPQTTKTCPSGSESCYHKQWSD-F---RGTIIERGC--GCPTVK-PGINLSCCES-EVCNN	ccccchhcchehehhhchccccccchhhccc-c---cchhhhccc--cchhhc-hccccccccc-ccccc
5	Laticauda laticaudata; laticotoxin a[g]	62	RRCFNHPSSQPQTNKSCPPGENSCYNKQWRD-H---RGTITERGC--GCPTVK-PGIKLTCCQS-EDCNN	ccccchccchehcccchccccccccchhhc-h---cchhhhccc--cchccc-hcccchccchh-hcccc
6	Laticauda laticaudata; laticotoxin a'[h]	62	RRCFNHPSSQPQTNKSCPPGENSCYNKQWRD-H---RGTITERGC--GCPQVK-SGIKLTCCQS-DDCNN	ccccchccchehcccchccccccccchhhc-h---cchhhhccc--cchhcc-cccccchccchc-cccccc
7	Naja nigricollis; toxin α^{i}	61	LECHNQQSSQPPTTKTCP-GETNCYKKVWRD-H---RGTIIERGC--GCPTVK-PGIKLNCCTT-DKCNN	ccehchcccchehechch-cchccccccchhc-h---cchhhhccc--cchccc-hcccccccchc-cccccc
8	Naja haje haje; toxin α & Naja nivea; toxin δ^{j}	61	LECHNQQSSQPPTTKTCP-GETNCYKKRWRD-H---RGSITERGC--GCPSVK-KGIEINCCTT-DKCNN	ccehchccchehehhhcch-cchccccccchhc-h---ccccchhccc--cchccc-cchhccccchc-cccccc
9	Hemachatus haemachatus; toxin II[k]	61	LECHNQQSSQPPTTKSCP-GDTNCYNKRWRD-H---RGTIIERGC--GCPTVK-PGINLKCCTT-DRCNN	ccehchccchehehhhcch-cchccccccchhc-h---cchhhhccc--cchccc-hcccccccchc-cccccc
10	Hemachatus haemachatus; toxin IV[l]	61	LECHNQQSSQTPTTQTCP-GETNCYKKQWSD-H---RGSRTERGC--GCPTVK-PGIKLKCCTT-DRCNK	ccehchccchehehehch-cchccccchccc-h---ccccchhccc--cchccc-hcccccccchc-cccccc
11	Naja nivea; toxin β^{m}	61	MICHNQQSSQRPTIKTCP-GETNCYKKRWRD-H---RGTIIERGC--GCPSVK-KGVGIYCCKT-DKCNR	cccchchccchehcccch-cchccccccchhc-h---cchhhhccc--cchccc-ccccccccch-ccccc
12	Naja melanoleuca; toxin d^{n}	61	MECHNQQSSQPPTTKTCP-GETNCYKKQWSD-H---RGTIIERGC--GCPSVK-KGVKINCCTT-DRCNN	ccchchccchehehhhch-cchccccchccc-h---cchhhhccc--cchccc-ccccccccchc-ccccc
13	Naja naja oxiana; toxin II & toxin α^{o}	61	LECHNQQSSQPPTTKTCS-GETNCYKKWWSD-H---RGTIIERGC--GCPKVK-PGVNLNCCRT-DRCNN	ccchchccchehehhhcc-cchccccchccc-h---cchhhhccc--cchhcc-hccccccccch-ccccc
14	Naja haje annulifera; toxin CM-14 (VIN2)[p]	61	MICHNQQSSQPPTIKTCP-GETNCYKKRWRD-H---RGTIIERGC--GCPSVK-KGVGIYCCKT-NKCNR	cccchchccchehcccch-cchccccccchhc-h---cchhhhccc--cchccc-ccccccccch-ccccc
15	Enhydrina schistosa; toxin 4^{q}	60	MTCCNQQSSQPKTTTNCA--ESSCYKKTWSD-H---RGTRIERGC--GCPQVK-PGIKLECCHT-NECNN	chccchccchcchehccc--ccccccchcccc-h---cchhhhccc--cchhcc-hchhhhhcchc-ccccc
16	Enhydrina schistosa; toxin 5^{r}	60	MTCCNQQSSQPKTTTNCA--ESSCYKKTWSD-H---RGTRIERGC--GCPQVK-SGIKLECCHT-NECNN	chccchccchcchehccc--ccccccchccc-h---cchhhhccc--cchhcc-ccchhhccchc-ccccc
17	Dendroaspis polylepis polylepis; toxin α^{s}	60	RICYNHQSTTRATTKSCE--ENSCYKKYWRD-H---RGTIIERGC--GCPKVK-PGVGIICCQS-DKCNY	ccccchcchcccchcccc--ccccccccchhc-h---cchhhhccc--cchhcc-hccccccchc-ccccc
18	Dendroaspis viridis; toxin $4.11.3^{t}$	60	RICYNHQSTTPATTKSC--GENSCYKKTWSD-H---RGTIIERGC--GCPKVK-RGVHLHCCQS-DKCNN	ccccchcchchhhchccc--ccccccccchccc-h---cchhhhccc--cchhhh-cccchhhcchc-ccccc
19	Dendroaspis jamesonii; toxin I^{u}	60	RICYNHQSTTPATTKSC--GENSCYKKTWSD-H---RGTIIERGC--GCPKVK-QGIKLHCCQS-DKCNN	ccccchcchchhhchccc--ccccccccchccc-h---cchhhhccc--cchhcc-hcccchcchc-ccccc

[a] The sequence is given in terms of the one-letter code recommended by an IUPAC–IUB commission.[15] All 19 amino acid sequences are numbered from 1 to 70, with the inclusion of dashes wherever there are deletions. Some dashes appear in all 19 sequences, because these proteins were taken from a larger list that contains some sequences in which these deletions did not occur. [b] The symbols h, ϵ, and c designate helical, extended, and other (coil) states, respectively. [c] The amino acid sequence is quoted from ref 16. [d] From ref 17. [e] From ref 17. [f] From ref 18. [g] From ref 19. [h] From ref 19. [i] From ref 20. [j] From ref 21. [k] From ref 22. [l] From ref 22. [m] From ref 23. [n] From ref 24. [o] From ref 25. [p] From ref 26. [q] From ref 27. [r] From ref 27. [s] From ref 28. [t] From ref 29. [u] From ref 30.

Table VI
Homologous Amino Acid Residues in Homologous Neurotoxins[a]

Protein	Number of Residues	Naja naja atra; cobrotoxin	Laticauda laticaudata; laticotoxin a'	Laticauda semifasciata; erabutoxin a	Laticauda semifasciata; erabutoxin b	Laticauda semifasciata; erabutoxin c	Laticauda laticaudata; laticotoxin a	Naja nigricollis; toxin α	Naja haje haje; toxin α & Naja nivea; toxin δ	Hemachatus haemachatus; toxin II	Hemachatus haemachatus; toxin IV	Naja nivea; toxin β	Naja melanoleuca; toxin d	Naja naja oxiana; toxin II & toxin α	Naja haje annulifera; toxin CM-14	Enhydrina schistosa; toxin 4	Enhydrina schistosa; toxin 5	Dendroaspis polylepis polylepis; toxin α	Dendroaspis viridis; toxin 4.11.3	Dendroaspis jamesonii; toxin I
Naja naja atra; cobrotoxin	62		36	35	33	33	35	48	53	47	49	44	48	47	43	39	39	34	34	35
Laticauda laticaudata; laticotoxin a'	62	0.53		45	48	61	59	43	38	43	40	37	42	41	38	40	39	36	40	42
Laticauda semifasciata; erabutoxin a	62	0.53	0.98		44	47	44	44	38	42	40	37	42	41	38	40	39	36	40	42
Laticauda semifasciata; erabutoxin b	62	0.53	0.97	0.98		61	46	43	38	43	39	37	41	42	38	39	38	36	40	41
Laticauda semifasciata; erabutoxin c	62	0.56	0.77	0.76	0.74		59	43	39	43	39	36	39	38	37	38	37	40	42	44
Laticauda laticaudata; laticotoxin a	62	0.58	0.73	0.71	0.69	0.95		42	40	42	38	37	40	38	37	38	39	40	43	45
Laticauda laticaudata; laticotoxin a'	62	0.77	0.71	0.71	0.69	0.69	0.68		54	54	51	48	53	53	48	43	42	39	41	43
Naja nigricollis; toxin α	61	0.85	0.61	0.61	0.63	0.65	0.65	0.89		50	49	50	53	49	50	38	38	37	38	39
Naja haje haje; toxin α & Naja nivea; toxin δ	61	0.76	0.69	0.68	0.69	0.69	0.68	0.89	0.82		49	45	49	51	45	40	39	37	39	40
Hemachatus haemachatus; toxin II	61	0.79	0.65	0.65	0.63	0.63	0.63	0.84	0.80	0.80		42	49	48	41	41	40	33	35	37
Hemachatus haemachatus; toxin IV	61	0.71	0.60	0.60	0.60	0.58	0.60	0.79	0.82	0.74	0.69		51	46	59	37	37	42	39	38
Naja nivea; toxin β	61	0.77	0.68	0.66	0.66	0.63	0.65	0.87	0.87	0.80	0.80	0.84		53	51	42	42	38	41	41
Naja melanoleuca; toxin d	61	0.76	0.66	0.66	0.68	0.61	0.61	0.87	0.80	0.84	0.79	0.75	0.87		46	42	41	39	43	42
Naja naja oxiana; toxin II & toxin α	61	0.69	0.61	0.61	0.61	0.60	0.60	0.79	0.82	0.74	0.67	0.97	0.84	0.75		39	39	40	39	38
Naja haje annulifera; toxin CM-14	61	0.63	0.65	0.63	0.61	0.63	0.61	0.84	0.80	0.80	0.67	0.67	0.61	0.69	0.69	0.64		59	35	39	41
Enhydrina schistosa; toxin 4	60	0.63	0.65	0.63	0.61	0.61	0.61	0.70	0.62	0.66	0.67	0.61	0.69	0.67	0.64	0.98			34	39	41
Enhydrina schistosa; toxin 5	60	0.55	0.58	0.58	0.65	0.65	0.65	0.64	0.61	0.61	0.54	0.69	0.62	0.64	0.66	0.58	0.57		51	50
Dendroaspis polylepis polylepis; toxin α	60	0.55	0.65	0.65	0.65	0.68	0.69	0.67	0.62	0.64	0.57	0.64	0.67	0.70	0.64	0.65	0.65	0.85		57
Dendroaspis viridis; toxin 4.11.3	60	0.56	0.68	0.68	0.66	0.71	0.73	0.70	0.64	0.66	0.61	0.62	0.67	0.69	0.62	0.68	0.68	0.83	0.95	
Dendroaspis jamesonii; toxin I	60																			

[a] The upper triangular half of the table represents the number of amino acids that are identical in the two proteins listed in the given row and column. The lower triangular half represents the ratio of these numbers to the number of residues in the protein (given by the "number of residues" in the protein in the row index in the last column).

Table VII
Effect of Homologous Amino Acids of Homologous Neurotoxins on the Predicted Conformations[a]

Protein	Number of Residues
Naja naja atra; cobrotoxin	62
Laticauda semifasciata; erabutoxin a	62
Laticauda semifasciata; erabutoxin b	62
Laticauda semifasciata; erabutoxin c	62
Laticauda laticaudata; laticotoxin a	62
Laticauda laticaudata; laticotoxin a'	62
Naja nigricollis; toxin α	61
Naja haje haje; toxin α & Naja nivea; toxin δ	61
Hemachatus haemachatus; toxin II	61
Hemachatus haemachatus; toxin IV	61
Naja nivea; toxin β	61
Naja melanoleuca; toxin d	61
Naja naja oxiana; toxin II & toxin α	61
Naja haje annulifera; toxin CM-14	61
Enhydrina schistosa; toxin 4	60
Enhydrina schistosa; toxin 5	60
Dendroaspis polylepis polylepis; toxin α	60
Dendroaspis viridis; toxin 4.11.3	60
Dendroaspis jamesonii; toxin I	60

	Naja naja atra; cobrotoxin	Laticauda semifasciata; erabutoxin a	Laticauda semifasciata; erabutoxin b	Laticauda semifasciata; erabutoxin c	Laticauda laticaudata; laticotoxin a	Laticauda laticaudata; laticotoxin a'	Naja nigricollis; toxin α	Naha haje haje; toxin α & Naja nivea; toxin δ	Hemachatus haemachatus; toxin II	Hemachatus haemachatus; toxin IV	Naja nivea; toxin β	Naja melanoleuca; toxin d	Naja naja oxiana; toxin II & toxin α	Naja haje annulifera; toxin CM-14	Enhydrina schistosa; toxin 4	Enhydrina schistosa; toxin 5	Dendroaspis polylepis polylepis; toxin α	Dendroaspis viridis; toxin 4.11.3	Dendroaspis jamesonii; toxin I
41	39	37	43	44	44	54	51	49	43	48	44	43	40	40	40	37	37	39	
0.66	60	58	48	48	43	44	50	50	45	53	52	45	44	44	44	40	44	50	
0.63	0.97	60	46	46	41	42	48	48	43	51	50	43	42	42	42	38	42	48	
0.60	0.94	0.97	44	46	39	40	46	46	42	49	50	42	42	42	42	38	44	48	
0.69	0.77	0.74	0.71	58	47	46	50	46	49	47	46	49	45	45	45	45	45	49	
0.71	0.77	0.74	0.74	0.94	46	47	49	45	50	48	46	50	46	48	48	47	48	51	
0.71	0.69	0.66	0.63	0.76	0.74	48	53	49	50	49	46	50	43	43	43	46	41	45	
0.87	0.71	0.68	0.65	0.74	0.76	0.79	53	51	47	52	47	47	42	42	42	38	38	41	
0.82	0.81	0.77	0.74	0.81	0.79	0.87	0.87	54	50	56	53	50	44	44	43	43	41	47	
0.79	0.81	0.77	0.74	0.74	0.73	0.80	0.84	0.89	46	56	53	46	43	43	43	37	41	46	
0.69	0.73	0.69	0.68	0.79	0.81	0.82	0.77	0.82	0.75	48	47	61	40	42	42	46	40	43	
0.77	0.85	0.82	0.79	0.76	0.77	0.80	0.85	0.92	0.92	0.79	56	48	45	47	47	38	43	48	
0.71	0.84	0.81	0.81	0.74	0.74	0.75	0.77	0.87	0.87	0.77	0.92	47	45	45	45	40	43	49	
0.69	0.73	0.69	0.68	0.79	0.81	0.82	0.77	0.82	0.75	1.00	0.79	0.77	40	42	58	46	46	43	
0.65	0.71	0.68	0.68	0.73	0.74	0.70	0.69	0.72	0.70	0.66	0.74	0.74	0.66	58	46	46	46	48	
0.65	0.71	0.68	0.68	0.73	0.77	0.70	0.69	0.72	0.70	0.69	0.77	0.74	0.69	0.97	46	46	47	48	
0.60	0.65	0.61	0.61	0.73	0.76	0.75	0.62	0.70	0.61	0.75	0.62	0.66	0.75	0.77	0.77	45	45	47	
0.60	0.71	0.68	0.71	0.73	0.77	0.67	0.62	0.67	0.67	0.66	0.70	0.70	0.66	0.77	0.78	0.75	54		
0.63	0.81	0.77	0.77	0.79	0.82	0.74	0.67	0.77	0.75	0.70	0.79	0.80	0.70	0.80	0.80	0.78	0.90		

[a] The upper triangular half of the table represents the number of residues of any type (but in homologous positions) that have identical (predicted) conformations in the two proteins listed in the given row and column. The lower triangular half represents the ratio of these numbers to the number of residues in the protein (given by the "number of residues" in the protein in the row index in the last column).

Table VIII
Identical Amino Acids in Homologous Neurotoxins

Position	Residue	Position	Residue
3	C (Cys)	43	R (Arg)
5	N (Asn)	44	G (Gly)
8	S (Ser)	45	C (Cys)
13	T (Thr)	48	G (Gly)
17	C (Cys)	49	C (Cys)
24	C (Cys)	50	P (Pro)
25	Y (Tyr)	52	V (Val)
27	K (Lys)	53	K (Lys)
29	W (Trp)	56	G (Gly)
31	D (Asp)	60	C (Cys)
37	R (Arg)	61	C (Cys)
38	G (Gly)	68	C (Cys)
42	E (Glu)	69	N (Asn)

Table IX
Identical (Predicted) Conformations in Homologous Positions in the Amino Acid Sequences of Homologous Neurotoxins

Position	Predicted conf[a]	Position	Predicted conf[a]
1	ϵ	42	h
3	ϵ	43	ϵ
5	c	44	c
6	h	45	ϵ
17	ϵ	48	c
21	c	49	ϵ
22	c	50	h
24	c	53	c
25	ϵ	56	c
26	ϵ	61	ϵ
37	ϵ	62	ϵ
38	c	68	ϵ
41	h	69	c

[a] These are the predicted conformations at the given positions, even though the type of residue *may* differ from one protein to another (cf. the amino acid sequences and predicted conformations given in Table V).

In Table VII, the conformations at identical positions in the sequence are compared, no matter whether the residues are the same or not; the position of each conformation (including dashes for deletions) corresponds to the numbering system (including dashes for deletions) in Tables V and VI. Since the fractions in the lower half of Table VII are, in general, greater than the fractions in the lower half of Table VI, it appears that the neurotoxins are more homologous in (predicted) conformation than they are in their amino acid sequences, except in a few cases where the amino acid sequence homology is very high [e.g., dendroaspis viridis; toxin 4.11.3 and dendroaspis jamesonii; toxin I have 57 homologous residues (95%), but only 54 residues (90%) are predicted to have the same conformation]. In other words, the amino acid sequence can change in homologous proteins without drastic alterations in their conformations, and such proteins can be more similar in conformation than in amino acid sequence. This conclusion is examined in more detail in the remainder of this section.

The amino acid residues that are identical in all 19 homologous neurotoxins are shown in Table VIII; there are 26 such homologies. The identical (predicted) conformations are listed in Table IX; there are also 26 such identities but the identities in Tables VIII and IX do not overlap completely. Only 19 of the 26 residues of Table VIII are shown in Table IX to have identical (predicted) conformations. On the other hand, residues 1, 6, 21, 22, 26, 41, and 62 have the same (predicted) conformations in all 19 homologous neurotoxins, despite the variations in species of amino acids.

Considering the data in Tables V, VIII, and IX, the (predicted) conformational features that appear in most of the homologous neurotoxins are the extended region at residues 1–4 and the helical region at residues 39–42. The other (predicted) helical and extended regions are indicated by underlining in Table V. The most structured of these neurotoxins is *Dendroaspis viridis;* toxin 4.11.3, where helical regions (three or more consecutive h states) are observed at residues 39–42, 50–53, and 58–60, and the extended region (four or more consecutive ϵ states) is found at residues 1–4. On the other hand, the most unstructured of these neurotoxins are naja haje haje; toxin α and hemachatus hemachatus; toxin IV, where no helical or extended regions are predicted.

V. Concluding Remarks

In section I, we have described a general predictive scheme that is based on short-range interactions and is applicable to models in which any number of states can be allowed. The method was applied to the three-state model (h, ϵ, and c) in section II, and numerical computations were carried out with this three-state model for BPTI and clostridial flavodoxin (in section III) and for homologous neurotoxins (in section IV).

The general predictive scheme of section I can also be applied to our earlier four-state[6] and multistate[7] models. In the four-state model,[6] the allowed conformational states are h, ϵ, c, and the chain reversal (R and S states); i.e., $\eta^* = 5$. In applying the method of section IA, it would be most convenient to again take $n = 3$. Then step (1) of section IA would be carried out by calculating $P^*_{\eta_1\eta_2\eta_3}$ for the $\eta^{*n} = 5^3 = 125$ possible conformational triads, where the η_i's can be h, ϵ, R, S, and c. [The number of possible conformational triads in the four-state model would be reduced to $(\eta^* - 1)[(\eta^* - 2)(\eta^* - 1) + 2]$, where $\eta^* = 5$, when applying the restriction[6] that R must precede S.] Step (2), for the calculation of $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$, would then be carried out repetitively for $\eta^* = 5$ possible states for η_{i+2}, with η_i and η_{i+1} maintained fixed at the conformations determined in the previous step. In the multistate model,[7] the same procedure would be carried out, but with $\eta^* = 7$ (and $n = 3$) since the allowed states[7] are h_R, ϵ, R, S, h_L, ζ_R, and c; here $\eta^{*n} = 7^3 = 343$. [The number of possible conformational triads in the multistate model would be reduced to $(\eta^* - 1)[(\eta^* - 2)(\eta^* - 1) + 2]$, where $\eta^* = 7$, when applying the restriction[6] that R must precede S.] Hence, it is practical (even though it would require much computer time) to apply the general predictive method of section I to the four-[6] and multistate[7] models to calculate the values of $P^*_{\eta_i\eta_{i+1}\eta_{i+2}}$. In this manner, we could eliminate the empirical rules of papers III[5] and IV,[6] as indicated in points (1) and (2) of the introductory section. Furthermore, the procedures of paper III can be improved by abandoning the restriction to regular sequences (eq 1) and considering arbitrary sequences (eq 2), as indicated in point (3) of the introductory section, and the procedures of papers IV and V can be improved by using nth-order a priori probabilities instead of first- and second-order probabilities, as discussed in point (4) of the introductory section.

In this series of papers (I–VI), we have presented a statistical mechanical treatment of protein conformation in terms of one-dimensional short-range interaction models and developed a predictive method in papers III–VI. The methods for evaluating the statistical weights, based on the x-ray data from protein structures, were described in paper I (where tentative numerical values were given) and in paper VI (to replace the tentative values presented in paper I) for the three-state model of paper II, and in papers IV and V (for the four-state and multistate models, respectively).

These statistical mechanical methods (based on short-range interaction models) are now being incorporated[31] into a pro-

cedure for determining the three-dimensional structures of proteins by introducing medium- and long-range interactions, as indicated in section VII of paper V and in ref 8, 32, and 33. The technique is a generalization of that used by Tanaka and Nakajima[34] for the two-state model of helix and coil states.

Acknowledgment. We are grateful to Dr. D. Gabel for providing us with the amino acid sequences of the homologous neurotoxins. The FACOM computer at Kyoto University was used for the numerical computations in this paper.

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 at Cornell University and 1975–1976 at Kyoto University and at the Weizmann Institute of Science; (b) to whom requests for reprints should be addressed.
(3) Paper I: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 142 (1976).
(4) Paper II: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 159 (1976).
(5) Paper III: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 168 (1976).
(6) Paper IV: S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 812 (1976).
(7) Paper V: S. Tanaka and H. A. Scheraga, *Macromolecules*, **10**, 9 (1977).
(8) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, **72**, 3802 (1975).
(9) As in our previous papers,[3-7] we use the symbol η_i to designate the conformational state of the ith residue. The symbol η_i can take on all conformational states allowed for in the model (e.g., h, ϵ, and c in the three-state model).
(10) R. Huber, D. Kukla, A. Ruelmann, and W. Steigemann, *Cold Spring Harbor Symp. Quant. Biol.*, **36**, 141 (1971).
(11) 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).
(12) The direction of a conformational transition in a polymer chain is usually taken from the N to the C terminus,[13] corresponding to the direction of matrix multiplication when computing the partition function and the averaged conformational properties of the polymer chain with statistical-weight matrices (see, e.g., eq II-20–22 and II-26).
(13) According to the IUPAC–IUB nomenclature, the initial and final ends of the polypeptide chain are the amino and carboxyl groups. Throughout this paper, we employ the recommendation proposed by the IUPAC–IUB Commission on Biochemical Nomenclature, *Biochemistry*, **9**, 3471 (1970).
(14) A set of statistical weights (w_{hJ}* and $v_{\epsilon J}$*) for the three-state model was also presented in Table I of paper III.[5] This was a tentative set because it was evaluated from conformational information reported by the x-ray crystallographers, rather than from the atomic coordinates; also, some of the older x-ray structures (used in paper III) have been revised. Thus, the set reported here in Table I replaces the earlier tentative set (see footnote 41 of paper IV for a discussion of the procedure to evaluate w_{hJ}* and v_{hJ}*). The statistical weights for the three-state model presented in Table I of this paper, those for the four-state model in Table III of paper IV,[6] and those for the multistate model in Table II of paper V[7] were all evaluated with the same numbers ($N_{h,j}$, $N_{\epsilon,j}$, etc.) from the same set of x-ray coordinates of 26 proteins. However, the statistical weights of Table I of this paper are expressed relative to that of the c state, whereas those in papers IV and V are relative to that of the ϵ state. The statistical weights of Table I (relative to the c state) can be converted to those relative to the ϵ state and, vice versa, those of papers IV and V (relative to the ϵ state) can be converted to those relative to the c state. For this purpose, we use the subscripts (c) and (ϵ) to indicate the basis on which the statistical weights are expressed, e.g., $w_{h(c)}$, $w_{h(\epsilon)}$, etc. The values in Table I can be converted into those relative to the ϵ state (in the three-state model) by means of the relations

$$w_{h(\epsilon)}^* = w_{h(c)}^*/v_{\epsilon(c)}^* \quad \text{(a)}$$

$$v_{h(\epsilon)}^* = v_{h(c)}^*/v_{\epsilon(c)}^* \quad \text{(b)}$$

$$u_{c(\epsilon)}^* = 1/v_{\epsilon(c)}^* \quad \text{(c)}$$

(15) IUPAC–IUB Commission on Biochemical Nomenclature, *J. Biol. Chem.*, **243**, 3557 (1968).
(16) C. C. Yang, H. J. Yang, and J. S. Huang, *Biochim. Biophys. Acta*, **188**, 65 (1969).
(17) S. Sato and N. Tamiya, *Biochem. J.*, **122**, 453 (1971).
(18) N. Tamiya and H. Abe, *Biochem. J.*, **130**, 547 (1972).
(19) S. Sato, unpublished; quoted by N. Maeda and N. Tamiya, *Biochem. J.*, **141**, 389 (1974).
(20) D. Eaker and J. Porath, *Proc. Plenary Sess., Int. Congr. Biochem., 7th*, 1087 (499).
(21) (a) D. P. Botes and D. J. Strydom, *J. Biol. Chem.*, **244**, 4147 (1969); (b) D. P. Botes, D. J. Strydom, C. G. Anderson, and P. A. Christensen, *J. Biol. Chem.*, **246**, 3132 (1971).
(22) A. J. C. Strydom and D. P. Botes, *J. Biol. Chem.*, **246**, 1341 (1971).
(23) D. P. Botes, *J. Biol. Chem.*, **246**, 7383 (1971).
(24) D. P. Botes, *J. Biol. Chem.*, **247**, 2866 (1972).
(25) (a) E. V. Grishin, A. P. Sukhikh, N. N. Lukyanchuk, L. N. Slobodyan, V. M. Lipkin, Yu. A. Ovchinnikov, and V. M. Sorokin, *FEBS Lett.*, **36**, 77 (1973); (b) H. Arnberg, D. Eaker, L. Fryklund, and E. Karlsson, *Biochim. Biophys. Acta*, **359**, 222 (1974).
(26) F. J. Joubert, *Hoppe-Seyler's Z. Physiol. Chem.*, **356**, 53 (1975).
(27) L. Fryklund, D. Eaker, and E. Karlsson, *Biochemistry*, 11, 4633 (1972).
(28) D. J. Strydom, *J. Biol. Chem.*, **247**, 4029 (1972).
(29) B. E. C. Banks, R. Miledi, and R. A. Shipolini, *Eur. J. Biochem.*, **45**, 457 (1974).
(30) D. J. Strydom and D. P. Botes, unpublished; quoted by D. J. Strydom, *Comp. Biochem. Physiol. B*, **44**, 269 (1973).
(31) S. Tanaka and H. A. Scheraga, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
(32) S. Tanaka and H. A. Scheraga, *Macromolecules*, **9**, 945 (1976).
(33) S. Tanaka and H. A. Scheraga, *Macromolecules*, preceding paper in this issue.
(34) S. Tanaka and A. Nakajima, *Macromolecules*, **5**, 714 (1972).