Ising Model Treatment of Short-Range Interactions in Polypeptides and Its Application to the Structure of Bovine Pancreatic Trypsin Inhibitor[1]

Lawrence G. Dunfield[2a] and Harold A. Scheraga*[2b]

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853, and Biophysics Department, The Weizmann Institute of Science, Rehovoth, Israel.
Received December 6, 1979

ABSTRACT: A nearest-neighbor Ising model based on the empirical conformational energies of two successive residues has been derived for polypeptide chains in the "random-coil" state and used to compute contour maps that represent the free energy of the entire (constrained) denatured molecule as a function of the backbone dihedral angles (ϕ_i, ψ_i) of each individual residue (having averaged over the conformations of all other residues). The positions of minimum free energy in each contour map indicate the most likely conformations for the backbone of each residue (and hence of the whole molecule) in the denatured state. The results may then be compared to the backbone dihedral angles of the *native* structure of the protein. For a specific protein (bovine pancreatic trypsin inhibitor) this comparison indicates that the results of the Ising model agree with the experimental dihedral angles of the native protein with an accuracy (41–29%) as good as that of the best prediction methods in the literature (46–29%), when the latter are recast in a form to yield predicted values of (ϕ_i, ψ_i) for each residue. When the residues in helical and bend regions are omitted, so as to provide a proper test of this nearest-neighbor interaction model, the backbone dihedral angles of the remaining residues of the denatured protein agree with those of the native structure with a much higher degree of accuracy (60–43%) than that obtained with any other (prediction) method in the literature (47–30%). The major advantages of the model are that it is independent of X-ray crystallographic data on proteins and that, unlike any other method, it describes the conformational space of the protein in terms of *continuous* values of the backbone dihedral angles (ϕ_i, ψ_i); it also provides information about the heretofore ambiguous "coil region". Analysis of the results reveals that certain sections of the amino acid sequence of a protein are well represented by this model and, as a consequence, interactions in such sections must be predominantly short range. (An assessment of the relative importance of side-chain to side-chain, side-chain to backbone, and backbone to backbone contributions to the nearest-neighbor short-range interactions is provided.) Other sections of the chain involve strong interactions along the chain (e.g., those that promote helix formation) and are by their nature beyond the scope of a nearest-neighbor model. The implications of this model for protein folding are discussed briefly. Also, a formal matrix representation of the theory is presented for application to the amino acid sequence of any protein. Finally, the nearest-neighbor Ising model can serve as the basis of an efficient Monte Carlo calculation for Ising models with short- and medium-range interactions and can be extended systematically until the entire conformational energy of the protein is accounted for, thereby avoiding the multiple-minima problem.

I. Introduction

In attempting to deduce the unique three-dimensional structure of a native protein from its amino acid sequence, one encounters the multiple-minima problem if one minimizes the empirical conformational energy or a paucity of experimental protein crystallographic data if one applies a statistical analysis that correlates the conformational behavior of different types of amino acid residues in different positions of the chain.[3] Thus, the *initial* stages of a protein-folding algorithm need to make use of an approximation based on the observation[4] that short-range interactions dominate and hence proceed by initially neglecting medium- and long-range interactions.[3] Such an assumption is the basis of various statistical prediction procedures[5–14] and of those empirical conformational energy calculations[15–18] that have made use of the information obtained from terminally blocked residues,[19] i.e., the *N*-acetyl-*N'*-methylamides of single amino acid residues.

There are several different definitions of short- and long-range interactions in use.[3,4,20] We shall use the one[3] in which interactions between all atoms in the structural unit composed of the backbone and side chain of a given residue and the two adjacent amide groups are termed "intraresidue" interactions; those between a given residue and others within four residues in each direction along the sequence are "short-range" interactions, those involving a given residue and residues 5–20 along the sequence are "medium-range" interactions, and those involving residues further removed in the sequence are "long-range" interactions. As defined here, short-range interactions will stabilize β bends and α-helical sequences, and possibly extended structures. Short-range interactions will dominate in "random-coil" structures, including that of a denatured protein. Medium- and long-range interactions will give the native structure a lower conformational energy compared to that of the denatured protein. Since the short-range interactions are expected to dominate in *both* the native and denatured states[4] (though their relative contribution to the total energy of the native structure will probably be *slightly* less than in the denatured structure), the lowest energy conformation in the ensemble of "random-coil" structures will be close to the native structure. Such a premise is consistent with much of the experimental results found for protein folding.[3,21,22]

An important question is how detailed the model for short-range interactions must be in order to be useful, i.e., how many neighbors must be considered in order to predict the conformational state of a given residue.[17] For example, α helices derive some stability from interactions between residues i and $i + 4$. A model that would treat such interactions in a direct manner is, unfortunately, quite complicated. The following questions then arise: Can a useful model for predicting the conformational states of a protein be derived from triplets of residues, or pairs, or even single residues? How important are the degrees of freedom of the side chains? Do they have a strong influence on the conformation of the backbone, and are their conformations correlated from one residue to the next?

This paper deals with these questions. The 58-residue amino acid sequence of bovine pancreatic trypsin inhibitor (BPTI) is represented by an Ising model for the interactions between *nearest neighbors* in the amino acid se-

0024-9297/80/2213-1415$01.00/0 © 1980 American Chemical Society

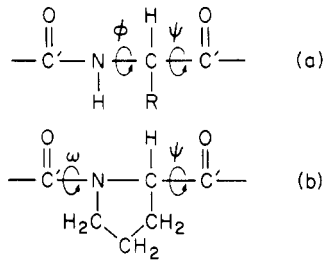

Figure 1. Fragment to which intraresidue interactions apply for computation of E_i for (a) nonproline and (b) proline residues.

quence; i.e., medium- and long-range (and some short-range) interactions are neglected. Thus, the results derived here would be expected to pertain only to those sections of a protein that are stabilized by short-range interactions *between nearest neighbors* and not by hydrogen bonds or *other* strong interactions that span several residues, such as those occurring in α helices or β bends. Hence, the applicability of the model should, and will, be tested by examining how well it reproduces the experimental backbone dihedral angles of those residues of BPTI that do *not* occur in α helices or β bends.

This model then corresponds to the "random-coil" or denatured state. It will be seen, however, that the results for the nearest-neighbor Ising model treatment of the denatured protein show a strong similarity to the *native* structure (of BPTI) in the nonhelical and nonbend regions; i.e., the Ising model treatment of the denatured protein appears to be applicable to these regions of the native molecule. Hence, their conformations in the native and denatured states must be similar.

The model derived here differs substantially from any previous short-range prediction algorithm for proteins[5-14] in that conformations are not described in terms of large general regions such as helix, coil, extended, etc. but are actually assigned specific values of the backbone dihedral angles (the procedure of Wu and Kabat,[6] in which backbone dihedral angles are predicted, is applicable only to families of homologous protein sequences). Further, the treatment is a rigorous statistical thermodynamic description of a model protein containing only *nearest-neighbor* short-range interactions and is solvable exactly. Since use is made of empirical conformational energies, computed with the program ECEPP[23] (Empirical Conformational Energy Program for Peptides), no parameterization from experimental protein crystallographic data is required, and the associated problem of the paucity of such data is avoided.

The present study presents an alternative to an earlier one[13] that was based on conditional probabilities derived from experimental protein crystallographic data. This earlier study[13] also used a nearest-neighbor Ising model but employed a small discrete number of conformational states that were defined by rather large regions of the (ϕ, ψ) conformational space; in contrast, a continuous range of values of (ϕ, ψ) is considered here. Also, the earlier study[13] suffered from the paucity of crystallographic data needed for its parameterization and required the use of ad hoc prediction rules.

II. The Model

The intraresidue interactions arise within the structural units shown in Figure 1, a and b, for nonproline and proline residues, respectively. The energy of such a unit is represented by E_i (eq A-1 of Appendix A). Interactions between neighboring residues arise within the unit depicted in Figure 2, a and b, for residue $i + 1$ being a nonproline

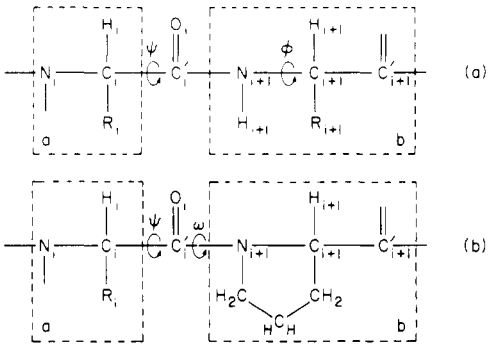

Figure 2. Portion of polypeptide chain for calculation of interresidue interaction energy, $E_{i,i+1}$, when residue $i + 1$ is (a) a nonproline and (b) a proline residue. The dashed lines enclose blocks a and b between which the interaction energies are computed. It should be noted that H_i is missing in block a and O_{i+1} in block *b* because interblock interactions involving these atoms are of longer range than the nearest-neighbor ones included in the present model. As is usual, (1-3)-type interactions are omitted.

or a proline residue, respectively. The interaction energy between residues i and $i + 1$ is represented by $E_{i,i+1}$ (eq A-2 of Appendix A). By choosing the structural units as in Figures 1 and 2, E_i and $E_{i,i+1}$ each become functions of exactly two backbone dihedral angles and depend only on the type of residue or residue pair, respectively.

Inclusion of any longer range interactions than those shown in Figures 1 and 2 would result in a much larger and more complicated matrix representation, a model that would not yield to present-day computational efforts. The restriction to those interactions shown in Figures 1 and 2 does represent an approximation to the energy of interaction between neighboring residues. In particular, the dipole–dipole interactions of neighboring peptide groups are incomplete because of the omission of the atoms H_{N_i} and O_{i+1} in Figure 2. However, in the context of the nearest-neighbor Ising model, this is a valid approximation and, in fact, is a good one. The effectiveness of this approximation was studied in preliminary calculations on terminally blocked alanine. The conformational energies for high-probability residue conformations have a nearly constant contribution from the omitted interactions between the H_{N_i} (or the O_{i+1}) atom and the atoms in the neighboring residue. These interactions can thus be neglected in the present model, since only relative energies are required. Despite the neglect of these interactions (and also the hydrogen bond between O_{i-1} and H_{i+1}, which is a *next*-nearest-neighbor interaction) the remaining interactions are sufficient to lead to the C_7^{eq} conformation for many of the residues. While the residues could have been redefined to include this hydrogen bond, this alternative definition would not have allowed us to treat the proline residue by such a simple model (because proline influences the conformation of the previous residue). The influence of the approximation of Figure 2 on the results will be discussed further in section IV.

The mathematical theory for evaluating the partition function for a specific-sequence polypeptide chain with nearest-neighbor interactions (Ising model) and for computing the probability distribution $P_i(\phi_i, \psi_i)$ for residue i to have backbone dihedral angles (ϕ_i, ψ_i) is given in Appendix A. A simple matrix representation of the partition function employing a self-consistent effective potential approximation for $P_i(\phi_i, \psi_i)$ is given in Appendix B.

Specifically, the value of $P_i(\phi_i, \psi_i)$ is given by eq A-9 of Appendix A, where $F_i(\phi_i, \tilde{\chi}_i)$ and $G_i(\psi, \tilde{\chi}_i)$ are given by eq A-7 and A-8, respectively. These quantities pertain to the ith residue, having averaged over the properties of residues

1, 2, ..., $i-1$ to obtain $F_i(\phi_i, \tilde{\chi}_i)$ and over residues N, $N-1$, ..., $i+1$ to obtain $G_i(\psi_i, \tilde{\chi}_i)$; i.e., eq A-7 and A-8 are recursion formulas. Thus, one has to compute E_i and $E_{i,i+1}$ for every reside and residue pair, respectively; then, F_1, F_2, ..., F_N and G_N, G_{N-1}, ..., G_1 are evaluated in the order of the amino acid sequence of the protein under investigation.

In this paper, this procedure is carried out with the side-chain dihedral angles held fixed, in order to test the theory with a minimum expenditure of computer time. In one calculation, the side-chain dihedral angles are preset to the values of the native protein; this represents the most favorable test of the model and must yield good results if the nearest-neighbor interactions are the most important. In a second, separate, calculation, the side-chain dihedral angles are deliberately preset to values different from those of the native structure. This represents the most difficult test of the model; also, it will provide an indication of the importance of the side-chain conformations. If the Ising model yields good information about the protein structure with a poor choice of side-chain dihedral angles, then it can be expected to be even more useful when the side-chain dihedral angles are accounted for properly.

On the other hand, if we were to *average* over the side-chain conformations, then, instead of simply *testing* the theory, we could use the Ising model to *predict* the most probable random-coil conformation from protein-sequence data; this could then be used as a starting point for the determination of the native structure of the protein. For this purpose, we introduce an approximation which averages over the side-chain dihedral angle in E_i and $E_{i,i+1}$ (Appendix B), thereby enabling $P_i(\phi_i, \psi_i)$ to be obtained by simple matrix multiplication, using eq B-31. Each matrix depends only on the type of dimer to which the matrix applies, and the amino acid sequence gives the correct order of matrix multiplication. Though not used in this paper (which focuses only on the aforementioned *test* of the theory), the matrix treatment of Appendix B is a general solution to the problem. It allows one to define a matrix for each of the 400 possible dimer pairs, and this matrix has to be evaluated only once. Thus, *any* protein sequence can be treated easily. Since the present test treats only a single side-chain conformation for each residue, thereby avoiding the *variation* of E_i and $E_{i,i+1}$ with side-chain dihedral angles, it is not necessary to use the approximation of Appendix B here.

With either method (i.e., with or without the approximation of Appendix B), the most probable values of (ϕ_i, ψ_i) for each residue are those that correspond to the maximum value of each $P_i(\phi_i, \psi_i)$. Equation A-9 or B-31 can be used to produce a contour map of $P_i(\phi_i, \psi_i)$, and the maximum value of $P_i(\phi_i, \psi_i)$ for each residue can be observed. The most probable conformation of the chain is then considered to be the one with all residues in their most probable individual conformations.

III. Details of Calculation

Four different (and unrelated) levels of short-range approximations were used to obtain the *backbone* dihedral angles of native BPTI. These four levels of approximation are the following:

(1) The dihedral angles of all of the side chains were fixed at the values observed experimentally for BPTI[24-26] but fitted to ECEPP geometries,[23] and a nearest-neighbor Ising model calculation (based on dipeptide interactions) was carried out.

(2) The dihedral angles of all of the side chains were fixed at the values found by Zimmerman et al.[19] for the lowest energy conformation of each type of residue with *N*-acetyl- and *N'*-methylamide terminal blocking groups, and a nearest-neighbor Ising model calculation (based on dipeptide interactions) was carried out.

(3) *Without* performing any calculations, the backbone dihedral angles were simply assigned (independently for each residue) the values for the lowest energy conformation (*global* minimum) of each type of residue, with terminal blocking groups, as determined by Zimmerman et al.[19]

(4) *Without* performing any calculations, the backbone dihedral angles were simply assigned (independently for each residue) the values for the most probable *local* energy minimum of each type of residue, with terminal blocking groups,[19] for which all side-chain dihedral angles were within ±30° of the ECEPP-fitted native conformation.[24,25]

Approximations 1 and 2 are nearest-neighbor Ising models based on dipeptide interactions, while approximations 3 and 4 are based on a single-residue model and contain no correlation between the neighboring residues. This single-residue model is introduced to provide a basis for measuring the improvement that the nearest-neighbor Ising model provides over calculations on terminally blocked single residues. The side-chain dihedral angles in approximations 1 and 4 were very different from those in approximations 2 and 3; for 35 out of 42 residues (excluding Gly, Ala, and Pro), the differences corresponded to different rotational isomeric minima, i.e., differences in side-chain dihedral angles of up to 120°.

Approximations 1 and 2 enable us to assess the influence of the conformation of the side chain on that of the backbone. A proper evaluation of the partition function would have included an averaging over all conformations of each side chain. This would have required the computation of a large number of conformational energies (e.g., in the case of the arginine residue, there are 324 minimum-energy side-chain conformations for each backbone conformation[19]). Thus, we have used approximations 1 and 2, each of which fixes the side-chain conformations at different sets of values. If the computed backbone dihedral angles of any residue differ in these two approximations, then the backbone conformation of such a residue would be considered to be sensitive to the conformation of its side chain.

A comparison of the results of approximations 1 and 2, on the one hand, and approximations 3 and 4, on the other, enables us to assess the importance of nearest-neighbor side-chain to side-chain interactions, for these are included in approximations 1 and 2 but not in approximations 3 and 4. (All four approximations, however, contain the important interactions of a side chain with the peptide group on each side, since the terminal blocking groups in approximations 3 and 4 are connected to the central residue by peptide bonds.) In the absence of side-chain to side-chain interactions, the conformational energy of each residue is essentially the same as that of the terminally blocked amino acid residue,[19] except for the negligibly small contributions of the end groups. A comparison of the results of approximations 3 and 4 (like that for approximations 1 and 2) provides information about the influence of the conformation of the side chain on that of the backbone.

In using the Ising model, in approximations 1 and 2 only, the conformational energy of each residue E_i and the nearest-neighbor interaction energy $E_{i,i+1}$ were evaluated for each residue and residue pair of BPTI, using the ECEPP geometry and energy algorithm.[23] The conformations of the side chains were held fixed, and E_i and $E_{i,i+1}$ were evaluated (for the structures of Figures 1 and 2, respectively) at 15° intervals of the backbone dihedral angles. For nonproline residues, ω was held fixed at 180° (planar

trans conformation); for the peptide bond preceding proline, however, ω was taken as a variable at 15° intervals over the range of −180 to +180°. The value of $E_{i,i+1}$ in ECEPP arises from the interaction of each atom in block a with all atoms in block b which are separated from it by three or more bonds (see Figure 2).

The integrals in eq A-7 and A-8 were evaluated numerically (with the temperature taken as 300 K) to give the functions $F_i(\phi_i, \tilde{\chi}_i)$ and $G_i(\psi_i, \tilde{\chi}_i)$, respectively, which in turn were used to evaluate the partition function (eq A-10) and $P_i(\phi_i, \psi_i)$ (eq A-9) for each residue of BPTI in approximations 1 and 2. These probability distributions were represented as *free-energy* $[-kT \ln P_i(\phi_i, \psi_i)]$ contour maps in (ϕ_i, ψ_i) for each residue; they differ from the usual (ϕ_i, ψ_i) contour maps in which potential energy is represented. Each map represents the (nearest-neighbor) short-range contribution of that residue to the free energy *of the entire protein as a function of* (ϕ_i, ψ_i) *of that residue*; this is in reality the free energy of a protein with residue *i constrained* to have a given (ϕ_i, ψ_i). The position of minimum free energy [and hence maximum $P_i(\phi_i, \psi_i)$] in each of these free-energy contour maps indicated the most likely conformation for the backbone of each residue; hence, the sequence of minimum free-energy positions taken from the 58 contour maps indicated the most likely conformation for the entire protein molecule. If the ECEPP-fitted crystallographic backbone dihedral angles[27] of a given residue fell within a circle of radius 47° about the global minimum of the free-energy map for that residue [this 47° circle representing a *mean* error in the *pair* (ϕ_i, ψ_i) of less than 30°], then the model was considered to have computed the conformation of that residue correctly. The effectiveness of each of the four levels of approximation was assessed in terms of the number of correct computations for the whole molecule.

The present work does not take account either of the effect of solvent or of charges on ionizable residues. It has been shown previously[28] that the use of uncharged side chains for Asp, Glu, Lys, and Arg simulates the effect of ion shielding by the solvent and provides good agreement with experiment. It appears better, then, to take all residues in their uncharged states. The protein–solvent interactions could influence the structure. Recent work by Hodes et al.[29] and Némethy et al.,[30] however, showed that, although the inclusion of the effect of water in the calculations of the conformational energies of *N*-acetyl-*N'*-methylamides of amino acids and dipeptides did cause the relative energies of these conformations to shift, it did not lead to a large change in the shape of the energy contour maps or in the positions of the minima. Thus, for the nearest-neighbor interactions considered in this model, water introduces only slight perturbations. The existence of α helices and bends, however, may depend strongly on solvent. Such long-range (solvent) effects cannot be included in the nearest-neighbor Ising model of the present paper, and the results of the model should be interpreted only for residues in extended or coil states.

Since the purpose of the present work was to investigate the suitability of the Ising approximation, the calculations were greatly simplified by considering each residue to have only one side-chain conformation. A general application of the theory would require the evaluation of the matrices of eq B-22 and B-23. This would be done by fitting the dipeptide energies with the Fourier series defined by eq B-18 and B-19.

IV. Results and Discussion

Computed Conformational Space. The calculations with approximations 1 and 2 produced contour maps showing the variation in free energy of the whole molecule with variation in the backbone dihedral angles of any given residue (with the side-chain dihedral angles held fixed). Representative maps for approximations 1 and 2 are shown in Figure 3. The conformation of lowest free energy was taken as the predicted one for each residue in each approximation.

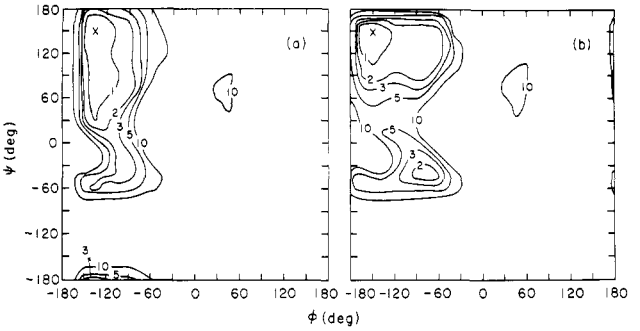

Figure 3. Free energy of BPTI at 300 K as a function of (ϕ, ψ) of residue 45 (Phe) for (a) approximation 1 with all χ's assigned the values of the native protein[24] (χ^1, g⁻) and (b) approximation 2 with all χ's assigned the values found for blocked single residues[19] (χ^1, t). The free energies are in kcal/mol, relative to zero at the conformation of lowest free energy (indicated by ×).

From an examination of the free-energy contour maps from approximations 1 and 2, it was possible to deduce the sensitivity of the map to the side-chain conformation for each type of residue. In comparing maps such as those of Figure 3, it was first necessary to establish that the differences are not due solely to the use of two different (fixed) side-chain conformations in approximations 1 and 2. For this purpose, we refer to the supplementary material in the paper of Paterson and Leach[31] and to the data of Zimmerman et al.[19] For β-methylalanine,[31] a change of the side-chain dihedral angle χ^1 from that of the t to the g⁻ conformation raised the energies at values of ϕ near −180° (as also seen in our Figure 3a, as compared with Figure 3b). On the other hand, however, whereas the α-helix minimum exists for both the t and g⁻ conformations of β-methylalanine[31] and of phenylalanine,[19] its energy is raised significantly for the g⁻ conformation of residue 45 (Phe) in the Ising model calculations (e.g., when going from approximation 2 to approximation 1 in Figure 3) but not in the single-residue map. Therefore, we may conclude that differences, such as those between Figures 3a and 3b, are due both to the differences in side-chain conformations (i.e., intraresidue interactions) and to nearest-neighbor interactions.

In addition, the similarity of Figures 3a and 3b to the g⁻ and t energy contour maps, respectively, presented by Paterson and Leach[31] provides support for our earlier assertion that the interactions omitted in the definition of $E_{i,i+1}$ (Figure 2) provide a relatively constant contribution to the conformational energy and hence are unimportant for the relative energies required in this study.

We now consider the results of the various approximations. First, we compare the contour maps from approximations 1 and 2 with that of a blocked single residue, viz., blocked alanine.[19] It is possible to compare these free-energy maps with energy maps such as those of ref 19 since the latter are equivalent to free-energy maps in the present sense because there is an entropy contribution from averaging over the conformations of the blocking end groups; in free-energy maps such as those of Figure 3, an entropy from such averaging over the rest of the chain appears. In general, it was observed that local minima shift their positions, constrained free-energy wells broaden or narrow, and regions of low energy disappear.

In both Figure 3a and Figure 3b, the broadness of the low-free-energy regions indicates that there is significant conformational freedom for each residue, regardless of the side-chain conformation. The existence of such broad minima is consistent with the results of Ponnuswamy et al.,[17] who found that the conformational space of a residue is broad but becomes more restricted as more residues are added to the chain; i.e., while the intraresidue and nearest-neighbor interactions dominate, it is the interactions with at least four residues on each side that are responsible for the uniqueness of the native conformation.[17] This is shown both by the large conformational space of low free energy in the maps of both approximations 1 and 2 (which implies that longer range interactions are required to define a specific conformation uniquely) and by our failure to predict the bends and helices which are stabilized by interactions over more than two residues.

Next, we compare the two Ising models, approximations 1 and 2, with each other. All residues (except Gly, Ala, Pro, and Val) were found to show some variation in the free-energy contour map with the different side-chain conformations used in the two approximations. The local minima were shifted and the low-energy regions changed in shape and size between the two approximations. This effect was particularly noticeable for Asp, Glu, Ile, and Leu, residues in which the β carbon is well screened by the rest of the side chain. While Val also has a well-screened β carbon, the side-chain conformation of Val in BPTI is the same in approximations 1 and 2; hence, there is no difference in the maps for this residue in BPTI.

Approximation 2 can be used to determine whether there are *large* effects of nearest-neighbor interactions on the conformation of a given type of residue when such a residue occurs in several different places in the chain but with different neighbors, i.e., when there are different nearest-neighbor interactions for a given type of residue. Since the side-chain conformation is the same for all residues of a given type (in this approximation), any changes in backbone conformation with position in the chain will be due to differences in the types of neighbors. A marked effect is observed only if one of the neighbors is a proline, as found for blocked dipeptides.[32] The contour map for a residue followed in the sequence by proline was very different from the map for the same residue type followed by any residue other than proline. In these maps, the range of ψ for residues preceding proline was distributed narrowly (within the range 120–160°), and conformations outside this range corresponded to high energies.

In approximation 1, where the side-chain conformations of a given type of residue differ in different positions in the sequence, the same types of restricted contour maps (as described in the previous paragraph) were observed for residues X in X–Pro sequences. Thus, a change of the side-chain conformation of residue X does not relieve the strain of its high-energy conformations,[32] and the interactions of its backbone with proline must represent the most important contribution to the nearest-neighbor interaction.

In a Pro–X sequence, a narrowing of the contour maps was also observed in approximation 2. Here ϕ of any residue X became narrowly distributed in the range −170 to −130°. When the side-chains were assigned the native conformations, however, as in approximation 1, the effect of a preceding proline disappeared and the contour map resembled that at any other location. The side-chain conformation in the native structure appears to be such as to minimize the interactions of the side chain with the preceding proline. In general, it appears that the type of neighbors can have a *major* influence on the allowed conformational space only if one of the neighbors is a proline residue.

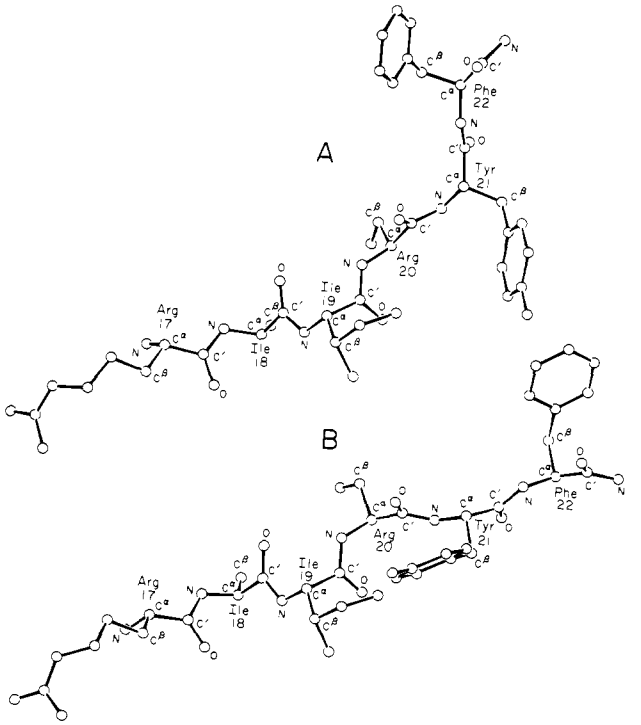

Figure 4. Experimental (A) and calculated (B) backbone conformations (including C^β atoms) for segment 17–22 of BPTI.

In summary, side-chain to backbone nearest-neighbor interactions produce *large* changes in backbone conformation of the X residue only in Pro–X sequences, and likewise for backbone–backbone nearest-neighbor interactions in X–Pro sequences. Side-chain to side-chain interactions between nonproline residues produce much smaller changes in backbone conformation, with the relative energies of different regions in the energy contour map being more sensitive to the neighbor toward the C terminus. While these changes are small, they are nevertheless important and must be taken into account (e.g., by an Ising model) in the prediction of protein conformation.

Comparison of Computed and Experimental Structures. Table I shows the experimental values of (ϕ_i, ψ_i) for each residue of BPTI and the values determined within each of the four approximations. Figures 4 and 5 compare the experimental[27] dihedral angles with those from approximation 1 for residues 17–22 and 29–35, respectively, as examples.

The segment from residues 17–22 is one for which all backbone dihedral angles have been computed correctly, and Figure 4 shows that the experimental and computed conformations of this segment appear similar.

The comparison in Figure 5 is interesting because the backbone dihedral angles of residues 29–31 and 33–35 were computed correctly, but those of residue 32 were computed incorrectly. As a result, the computed chain conformation undergoes a change of direction at residue 32. Therefore, for a proper comparison of computed and experimental conformations, residues 29–31 and 33–35 should be compared separately, in Figure 5A,B and Figure 5C,D, respectively. The experimental and calculated conformations of these two segments compare favorably. Conceivably, if our model had included *next*-nearest-neighbor interactions, we might have predicted the dihedral angles of

Table I
Determination of Conformation of BPTI at Four Levels of Approximation

residue posn	residue type	values of (ϕ_i, ψ_i),[a] deg: exptl[27]	approximation 1	2	3	4
1	Arg	−60, 171	−90, 150 (*)	−60, 150 (*)[b]	−75, 90	−165, 165
2	Pro	−75, 150	−75, 90	−75, 150 (*)[b]	−75, 75	−75, 75
3	Asp	−70, −36	−75, 90	−180, 150	−165, 150	−90, 105
4	Phe	−69, 3	−150, 150	−150, 150	−150, 150	−150, 150
5	Cys	−104, −8	−90, 90	−90, 90	−90, 90	−90, 90
6	Leu	−89, −27	−90, 90	−90, 90	−90, 90	−90, 75
7	Glu	−58, 151	−150, 150	−150, 90	−150, 135	−150, 150
8	Pro	−75, 170	−75, 165 (*)	−75, 165 (*)	−75, 75	−75, 75
9	Pro	−75, 131	−75, 90 (*)	−75, 165 (*)	−75, 75	−75, 75
10	Tyr	−108, 108	−150, 150 (*)	−150, 165	−150, 165	−150, 150 (*)
11	Thr	−61, −57	−145, 120	−90, 90	−90, 90	−90, 75
12	Gly	113, 171	−180, 180	−180, 180	−90, 75	−90, 75
13	Pro	−75, −28	−75, 75	−75, 75	−75, 75	−75, 75
14	Cys	−60, 157	−75, 120 (*)	−165, 150	−90, 90	−90, 90
15	Lys	−120, 32	−135, 150	−90, 90	−75, 90	−90, 75
16	Ala	−83, 169	−90, 90	−90, 75	−90, 75	−90, 75
17	Arg	−116, 77	−90, 90 (*)	−90, 90 (*)	−75, 90 (*)	−90, 90 (*)
18	Ile	−111, 121	−90, 105 (*)	−150, 135 (*)	−150, 135 (*)	−75, 105 (*)
19	Ile	−80, 113	−75, 105 (*)	−150, 135	−150, 135	−90, 90 (*)
20	Arg	−120, 173	−90, 135 (*)	−90, 75	−75, 90	−75, 90
21	Tyr	−116, 135	−150, 150 (*)	−150, 150 (*)	−150, 165 (*)	−150, 150 (*)
22	Phe	−138, 134	−150, 150 (*)	−150, 165 (*)	−150, 150 (*)	−150, 150 (*)
23	Tyr	−73, 143	−150, 150	−150, 150	−150, 165	−150, 150
24	Asn	−118, 102	−150, 120 (*)[b]	−150, 135 (*)	−165, 135	−75, 105 (*)
25	Ala	−54, −30	−105, 75	−90, 90	−90, 75	−90, 75
26	Lys	−69, −69	−135, 150	−90, 90	−75, 90	−90, 75
27	Ala	−53, −40	−90, 90	−150, 150	−90, 75	−90, 75
28	Gly	98, 10	90, 90	90, −75	90, −75	90, −75
29	Leu	−142, −179	−150, 165 (*)	−90, 90	−90, 90	−150, 150 (*)
30	Cys	−97, 144	−90, 105 (*)	−90, 120 (*)	−90, 90	−90, 90
31	Gln	−132, 168	−135, 150 (*)	−150, 150 (*)	−150, 135 (*)	−90, 75
32	Thr	−91, 144	−90, 90	−90, 90	−90, 90	−90, 75
33	Phe	−155, 141	−150, 150 (*)	−150, 165 (*)	−150, 150 (*)	−150, 150 (*)
34	Val	−66, 128	−90, 120 (*)	−90, 120 (*)	−90, 105 (*)	−90, 105 (*)
35	Tyr	−111, 132	−135, 135 (*)	−150, 150 (*)	−150, 165	−150, 150 (*)
36	Gly	−47, −51	90, −75	90, −75	−90, 75	−90, 75
37	Gly	116, 9	−90, 75	−90, 75	90, −75	90, −75
38	Cys	−158, 151	−90, 90	−90, 75	−90, 90	−90, 90
39	Arg	56, 56	−90, 135	−90, 90	−75, 90	−75, 90
40	Ala	−80, 149	−90, 120 (*)	−90, 75	−90, 75	−90, 75
41	Lys	−92, −167	−135, 150 (*)[b]	−90, 90	−75, 90	−90, 75
42	Arg	−76, −46	−135, 135	−90, 90	−75, 90	−75, 90
43	Asn	−78, 77	−165, 150	−165, 135	−165, 135	−75, 105 (*)
44	Asn	−167, 113	−165, 105 (*)	−165, 135 (*)	−165, 135 (*)	−165, 120 (*)
45	Phe	−123, 172	−135, 150 (*)	−150, 150 (*)	−150, 150 (*)	−150, 150 (*)
46	Lys	−98, −18	−90, 75	−90, 90	−75, 90	−90, 75
47	Ser	−143, 158	−150, 165 (*)	−90, 75	−75, 75	−165, 150 (*)
48	Ala	−77, −35	−90, 150	−90, 75	−90, 75	−90, 75
49	Glu	−57, −46	−60, 90	−150, 135	−150, 135	−90, 75
50	Asp	−76, −32	−135, 135	−165, 150	−165, 150	−165, 150
51	Cys	−64, −55	−75, 90	−75, 90	−90, 90	−90, 90
52	Met	−67, −28	−75, 105	−75, 105	−75, 105	−90, 75
53	Arg	−66, −38	−165, 120	−90, 90	−75, 90	−75, 105
54	Thr	−47, −56	−75, 105	−90, 90	−90, 90	−150, 150
55	Cys	−68, −56	−90, 90	−90, 90	−90, 90	−90, 90
56	Gly	−39, −47	−90, 75	−90, 75	−90, 75	−90, 75
57	Gly	161, 64	−90, 75	−90, 75	−90, 75	−90, 75
58	Ala	−168, −164	−150, 150 (*)	−150, 150 (*)	−90, 75	−90, 75

[a] Correct determinations are indicated by (*) [i.e., the experimental values of (ϕ_i, ψ_i) lie within a circle of radius 47° about the computed point; thus the *mean* error in the *pair* (ϕ_i, ψ_i) is less than ±30°]. [b] The free energy map shows two minima with <0.1 kcal difference between them; the one closest to the experimental value is listed here. The circles of radius 47° about both minima, however, included the experimental value; hence, either computed value would be correct. In all other cases, there was either only one minimum or a second minimum >0.5 kcal above the global minimum, and the choice of (ϕ_i, ψ_i) was clear.

residue 32 correctly and thereby would have avoided the erroneous change of direction of the chain at residue 32.

Comparison of Approximations. The dihedral angles computed within each of the four approximations may be compared with the experimental values (Table I). In particular, a comparison of the percentage obtained correctly in approximations 1 and 2, on the one hand, with that obtained correctly in approximations 3 and 4, on the other, illustrates how the backbone dihedral angles computed with the Ising model compare to those for blocked single residues. Columns 3–6 in the first row of Table II show the percent of the residues with correctly assigned

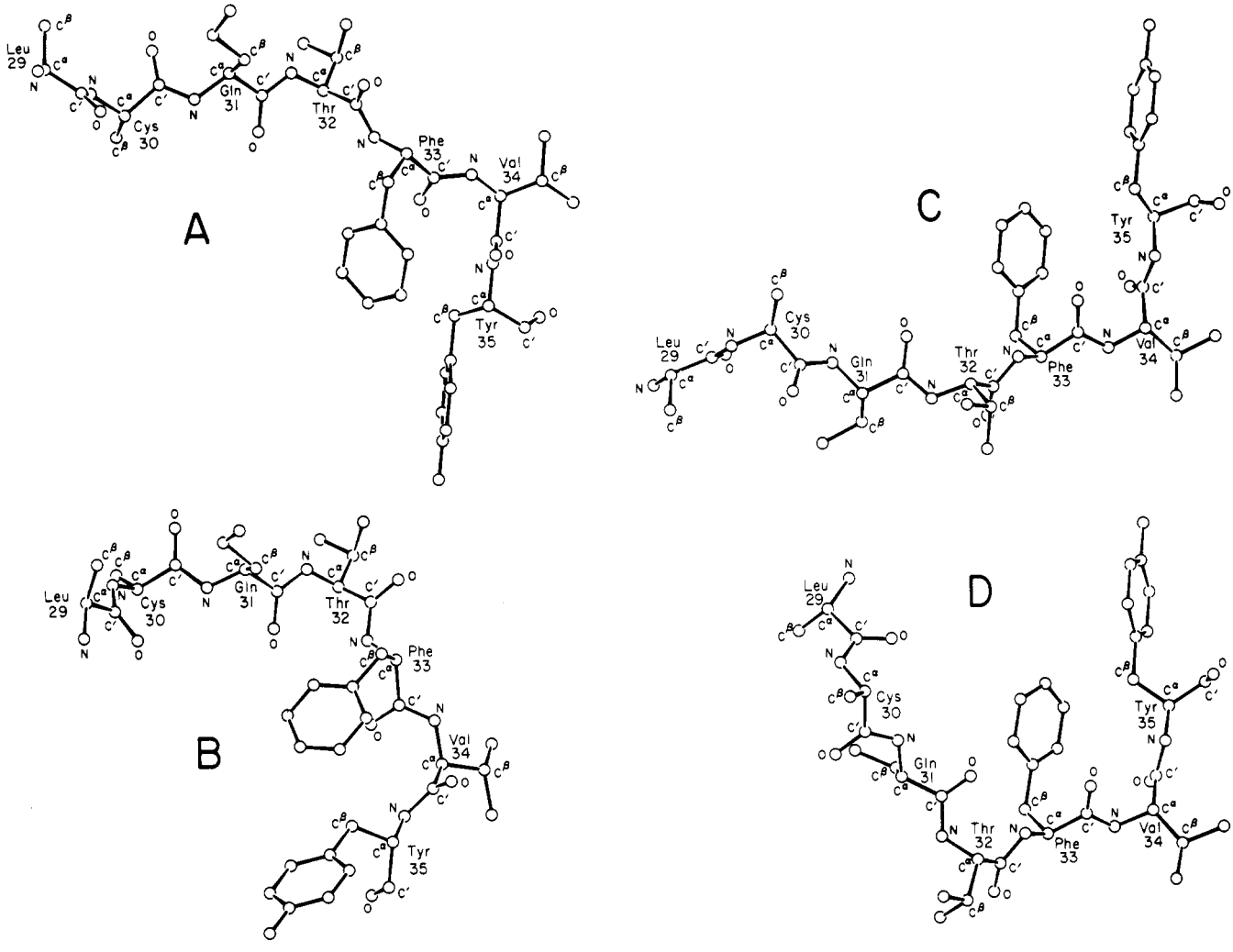

Figure 5. Experimental (A and C) and calculated (B and D) backbone conformations (including C^{β} atoms) for segment 29–35 of BPTI. The comparison of A and B pertains only to residues 29–31 and that of C and D only to residues 33–35.

(ϕ, ψ)'s in each approximation. Approximations 1 and 2 are statistically better (at the 95% confidence level, using a χ^2 test[33]) than approximations 4 and 3, respectively, but none is better than 41% correct. Also, the assignment of the native conformations of the side chains improves approximation 1 over approximation 2, and approximation 4 (in which the side-chain dihedral angles are within 30° of the native ones) over appoximation 3, but neither improvement is significant at the 95% confidence level.

Now, as pointed out in the Introduction, while α helices and β bends arise primarily because of short-range interactions, they also involve the interactions with neighbors that are removed by four and three residues, respectively. Since these longer range interactions were not included in any of the four approximations, such structures should not be included when judging how well the approximations represent nearest-neighbor interactions. Thus, the four approximations should be evaluated by considering only those residues of BPTI that are *not* in α-helical (or β-bend) conformations (this can also be justified on energetic grounds[34]).

The results of such an evaluation are shown in the second row of Table II, in which the entries are considerably higher than those in the first row, but the difference between the entries in the first and second rows is statistically significant within the 90% confidence level only for approximation 1. Further, the data of the second row show that the percent correct for approximations 1 and 2 is considerably higher than for approximations 4 and 3, respectively, within the 95 and 90% confidence levels, respectively. Also, the value of 60% correct for approximation 1 is much higher at the 95% confidence level (for regions of the chain *not* involved in α helices or β bends) than the majority of those based on short-range algorithms[7-13] that rely on statistical analyses of experimental protein crystallographic data (see next section for further comparison with other prediction procedures).

It is not quite fair to compare approximations 1 and 2 with approximations 3 and 4 by including the residues preceding proline because it is known[32] that the conformational space of an X residue in an X–Pro sequence is greatly restricted by the following proline residue, compared to its properties as a blocked single residue.[19] The use of $E_{i,i+1}$ in approximations 1 and 2 would include this effect, which would not appear in approximations 3 and 4. Indeed, as seen in Table I, approximations 3 and 4 fail to predict the conformations of any residues preceding proline, whereas approximations 1 and 2 do as well (2 out of 4) for these residues as for any other residue. Therefore, to compare these approximations properly, we should eliminate the four residues preceding prolines (third row of Table II). With this elimination, the results of approximation 4, while still not as good as approximation 1, improve to the point where they are not statistically distinguishable at the 95% confidence level from approximation 1. For approximation 4 to succeed so well, it must mean that the side-chain to nearest-neighbor backbone interactions (present in both models 1 and 4) play an im-

Table II
Summary of Determinations at the Four Levels of Approximation and of Various Other Procedures in the Literature

no. of residues of BPTI considered	% with correctly assigned (ϕ,ψ)'s												
		approximation							Tanaka and Scheraga				
	perfect[18]	1	2	3	4	Burgess et al.[7]	Lim[8]	Chou and Fasman[9]	Robson and Pain[10]	3-state[13]	4-state[13]	multi-state[13]	Maxfield and Scheraga[a]	random
58[b]	79[c]	41	29	16	26	29	38	45	41	40	39	21	46[f]	10
40[d]	73	60	43	23	35	35	35	38	30	33	33	23	47[f]	15
36[e]	78	61	42	25	39	38	36	42	31	36	33	25	50[f]	17

[a] These predictions were made by the method of ref 11, with $P_{max} = 0.1$ and $S = 40$, using the data set (27 protein structures) of ref 12. [b] This is the total number of residues in BPTI. [c] The "perfect" prediction is <100% accurate because, while the conformational *region* is predicted with 100% accuracy, the values of (ϕ_i, ψ_i) correspond to the mean values within each region and hence do not always meet our criterion for a correct prediction (see footnote *a* of Table I). [d] Excluding residues in α helices (3–6, 25–27, 48–56) and β bends (42–43). [e] Excluding residues in α helices and β bends and those preceding proline (1, 7, 8, 12). [f] Residues 1 and 58 are not included in this prediction algorithm.

portant role in those conformations occurring in the globular protein. The nearest-neighbor side-chain to side-chain interactions (absent in approximation 4) will be of less importance, although the better results of approximation 1 over approximation 4 (with the results for prolines eliminated) are due partly to the inclusion of side-chain to side-chain interactions and partly to the correlation between neighbors inherent in the Ising model. The poor results of approximation 3 strongly indicate that a short-range Ising model must include the variability of side-chain conformations.

Finally, it must be emphasized that, in predicting the backbone dihedral angles of an unknown protein structure, no prior information is available as to the location of helices and bends or as to the values of the side-chain dihedral angles. The data presented here represent, on the one hand, the most favorable assignment of side-chain dihedral angles (approximations 1 and 4) and, hence, an upper bound on the success rate, and, on the other hand, an incorrect assignment of side-chain dihedral angles (approximations 2 and 3) and, hence, a lower bound on the success rate. These calculations are meant as tests, and it is the procedure of Appendix B that must be followed if the Ising model is to be used for the *prediction* of protein structure. Thus, the success rates of line 1 of Table II can indicate only the range of successful prediction possible with the nearest-neighbor Ising and single-residue models. The remaining lines of Table II serve only as tests of the importance of *nearest-neighbor* interactions.

Comparison with Prediction Procedures in the Literature. High success rates have been claimed for various prediction procedures,[7-13] but it must be realized that these methods assign residues to *regions* of conformational space, leaving a large degree of uncertainty in the actual values of ϕ and ψ. Indeed, Burgess and Scheraga[18] have demonstrated that a "perfect" prediction algorithm [assignment of correct *region* for *every* residue, with the (ϕ_i, ψ_i)'s taken as the mean value for each residue in each region] results in a structure that differs considerably from the native one. If we apply our criterion that the *mean* error in the *pair* (ϕ_i, ψ_i) be less than 30°, then even the "perfect" prediction[18] leads to incorrect dihedral angles for residues 1, 7, 12, 14, 17, 29, 31, 33, 38, 43, 44, and 58 (see footnote *c* of Table II); i.e., the conformational regions, as usually defined, are too large for meaningful assignments of the backbone dihedral angles.

In order to compare our procedure with those in the literature, we have used the following method to deduce dihedral angles from the predictions reported earlier.[7-13] If a procedure[7-13] assigns a region correctly, as defined above, and if the "perfect" prediction[18] can assign the dihedral angles correctly, then we consider that the conformation of that residue has been predicted correctly. There still remain some ambiguities, however, that must be dealt with. Many algorithms[7-9,13] for, say, a two-state model assign the nonhelical residues to a coil region that includes the entire conformational map; i.e., the conformations of such residues are unpredicted. If a residue is unpredicted,[7-9,13] i.e., assigned to the "coil" region, we give it a 25% chance that its conformation would be assigned correctly to one of the four regions, viz., $(\alpha_R, \alpha_L, \zeta_R, \epsilon)$,[18] which are the only ones that occur in BPTI. But, even if assigned to the correct region, their dihedral angles are considered to be predicted correctly only if they meet our criterion that the *mean* error in the *pair* (ϕ_i, ψ_i) be less than 30°. Thus, for example, if residues 6 and 7 were predicted to be in "coil" states, we would assign $^1/_4$ of a correct prediction to residue 6 but zero to residue 7 since the latter

is not predicted correctly by even the "perfect" prediction algorithm. A similar problem arises in the prediction of bends,[7-9,13] since there are several combinations of dihedral angles for a bend and the prediction methods[7-9,13] do not specify the type of bend. In considering bends in the Gly–Gly sequence, we have assumed that there is an equal chance ($^1/_8$) for the occurrence of each of eight types of bends,[32] I (α_R, ζ_R), I′ (α_L, ζ_L), II (ϵ, α_L), II′ (ϵ^*, ζ_R), III (α_R, α_R), III′ (α_L, α_L), V (ϵ, ϵ^*), and V′ (ϵ^*, ϵ), where ϵ^* is the mirror-image equivalent of the ϵ region and is possible only for Gly.[35] Hence, $^1/_8$ of a correct residue is assigned to each residue in a pair that is predicted to be in a bend, provided that (again) the 30° criterion is met; otherwise, zero is assigned. Other pairs of residues in bends are treated similarly. For Ala-type–Ala-type[36] and Pro–Ala-type sequences, we assign equal chance ($^1/_3$) for bend types I, II, and III; for Ala-type–Pro and Pro–Pro sequences, we assign equal chance ($^1/_2$) for bend types I and III; for the Gly–Pro sequence, we assign equal chance ($^1/_4$) for bend types I, II′, III, and V′; for Ala-type–Gly sequences, we assign equal chance ($^1/_6$) for bend types I, I′, II, III, III′, and V; and for Gly–Ala-type sequences, we assign equal probability ($^1/_7$) for all bend types except V. In all cases, the 30° criterion is also applied.

The first line of Table II shows the assigments of various prediction methods.[7-13] All of them, including those of this paper, assign (ϕ, ψ)'s correctly with significantly less success (99.9% confidence level) than the "perfect" prediction. Approximations 1 and 2 are *tests* of the Ising model. The latter can be used for prediction only if one averages over the side-chain conformations (Appendix B); the expected success rate of the Ising model would lie between those of approximations 1 and 2, and that of the single-residue model (with *averaged* side-chain conformations) would lie between those of approximations 3 and 4. The success rates of the Ising model, Maxfield and Scheraga,[11,12] Lim,[8] Robson and Pain,[10] Chou and Fasman,[9] and Tanaka and Scheraga (three- and four-state models)[13] lie grouped around 40%, and these methods are signficantly better (95% confidence level) than a random assignment to the four regions of the (ϕ, ψ) map of BPTI. The success rates of the single-residue model, Burgess et al.,[7] and Tanaka and Scheraga (multistate model)[13] are grouped around 25% and are not statistically different (95% confidence level) from a random-choice success rate of 10%.

When the helical and bend regions are removed from the sample (line 2 of Table II), the Ising model is statistically indistinguishable (95% confidence level) from the "perfect" prediction and is significantly better (95% confidence level) than all but Chou and Fasman[9] and Maxfield and Scheraga.[11,12] The single-residue model and the remaining prediction methods[7,8,10,13] are not significantly better (95% confidence level) than a random assignment. It thus appears that the nearest-neighbor Ising model predicts the backbone dihedral angles of BPTI with an accuracy as good as that of the best methods in the literature; when the helical and bend regions are omitted, the accuracy of the Ising model is best.

Ising Model and Native Structure. One of the problems in applying short-range interaction models to proteins is that, as is well-known, there are different types of local structure. α helices and β bends display a narrow range of ϕ and ψ and involve interactions that span several residues. All other conformational states (e.g., nonregular or extended structure) exist over a wide range of ϕ and ψ and arise largely as a consequence of the intraresidue interactions (which include those of the residue with its adjacent peptide groups) and, to a lesser extent, nearest-neighbor interactions. The Ising model treatment in this paper was suitable only for this latter type of structure. Prediction of α helices and bends is beyond the scope of an exact nearest-neighbor model.

The short-range model used in the calculations reported here is a good one for a random-coil protein, and the contour maps can be used to indicate the available conformations for such a protein in the absence of long-range interactions, i.e., in the unfolded state. (It should be noted that the term "random coil", as used here, refers to a Boltzmann average over the dihedral angles of the entire protein, this average being dominated by low-energy conformations in which short-range interactions dominate.) In other words, the calculations carried out here pertain to Θ conditions.[4,37] The fact that the results compare favorably with the nonhelical and nonbend regions of the *native* protein indicates that these regions are not very sensitive to long-range interactions. When the external conditions (e.g., pH, temperature, solvent) are changed from Θ conditions to those favoring the native structure, changes arise in the conformation of each residue and, hence, in the overall conformation. A large fraction of backbone dihedral angles in the native structure lie close to the global minimum of the short-range model,[38] as shown here. Therefore, when the external conditions are changed from native ones, the backbone dihedral angles of many of the residues will be expected to fluctuate about values that are displaced by *small amounts* (certainly within the 30° criterion used here) from the native structure; i.e., the intraresidue and nearest-neighbor interactions are the most important in a large fraction of residues, both for the random-coil structure and for the native structure. It must be emphasized that even very small changes in the backbone dihedral angles (10° or less) will cause gross changes in the overall conformation of the protein.[18] In other words, a denatured protein may, in fact, have the majority of its residues in individual conformations near their native values and still have a conformation very different from that of the native protein.

In light of this, the short time required for the protein to refold into its native conformation can be better understood in that many residues will require only minor changes in their dihedral angles in order to reach their individual native conformations. Of course, since refolding is a cooperative process, the more residues that are in or near their native conformation, the faster will the remaining residues approach their native conformations.

Applications of the Ising Model. In the past, protein structure has been predicted by restricting the statistical analysis of the crystallographic data to residue-pair frequencies[5-14] and by formulating ad hoc rules for the prediction of conformational regions. This approach suffers from deficiencies such as the limitation of the statistical analysis to pairs of residues [which does not distinguish isolated helical (α_R) states from helical sequences and isolated ζ_R states from bends, many of which contain ζ_R states], with a resulting overemphasis of isolated α_R and ζ_R regions. Also, not enough protein crystallographic data exist to carry out the analysis of triplets and quadruplets needed for proper statistical prediction of the correct amounts of all types of local structures. The paper by Maxfield and Scheraga[11] contains a critique of this approach. (Currently, work is under way[39] to optimize the existing protein data and to improve some of the assumptions made in the Tanaka–Scheraga approach.[13]) As indicated earlier, however, even a perfect prediction of *regions*[18] leaves a large uncertainty in the values of the dihedral angles.

It would thus be very desirable to have a detailed short-range alogorithm, such as a second- and/or third-nearest-neighbor Ising model, which included such interactions as the α-helix hydrogen bond. Such Ising models are extremely difficult to solve in a direct manner and would involve extensive conformational energy calculations to parameterize. On the other hand, techniques such as Monte Carlo provide a powerful method for extending the Ising model presented here, so as to include interactions with residues twice removed as neighbors. Such an extension would allow our Ising model to treat helices and bends properly.

The Ising model of this paper, as presented in Appendix B, can be used as the sampling basis for an efficient Monte Carlo calculation of the relative probabilities of occurrence of conformations in an extended short-range model (including interactions with up to five adjacent residues in each direction along the chain). These probabilities in turn can be used as the sampling basis for a Monte Carlo calculation in a medium-range model, which in turn can be used as the sampling basis for a Monte Carlo calculation of the most probable conformations for the entire sequence with all interactions present. These then serve as the starting points for full energy minimizations. In this way, the multiple-minima problem is dealt with in a computationally feasible manner. The results of this paper, and of recent experiments,[22] suggest that such a model may be sufficiently complete and accurate to generate many conformations of a protein that lie close to (and in equilibrium with) the native structure; i.e., it is reasonable to expect that, because of the success of the nearest-neighbor Ising model compared to other methods (see Table II), its extension (with variable side-chain dihedral angles) by this Monte Carlo procedure should lead to values of the (ϕ_i, ψ_i)'s that are more accurate than the mean values assigned to predicted *regions*,[18] even if such regions are predicted with 100% accuracy.

V. Conclusions

The nearest-neighbor Ising model has been found to reproduce the native backbone dihedral angles of a large fraction of the residues in BPTI. Since the model properly corresponds to random-coil conditions and hence to a denatured protein, we are led to conclude that a denatured protein may, in fact, have the majority of its residues in individual conformations very close to their native ones. This would greatly reduce the time needed for the refolding process.

This study has also provided further assessment of the relative importance of various contributions to the short-range interactions. The following were found: (1) side chain–side chain interactions are less important than side chain–backbone interactions; (2) residues preceding proline and, to a lesser extent, following proline have a smaller allowed conformational space than do those with other types of neighboring residues; (3) from a comparison of approximations 1 and 2, the effect of side-chain dihedral angles on backbone conformations is due mainly to intraresidue interactions; (4) while the blocked single residue is a good representation of short-range interactions, the Ising model represents an improvement; (5) the better results of approximation 1 compared to approximation 2, as well as the success of approximation 4, show that variation of the side-chain conformation influences the predicted backbone conformation and is a necessary feature of any short-range model.

A very important feature of the present treatment is its use of backbone dihedral angles that are continuous, rather than the assignments of the backbone conformations to large regions of conformational space (with a consequent large uncertainty in the values of the dihedral angles). In particular, this treatment circumvents the ambiguity present in conformations that lie in the heretofore-designated "coil" region.

In summary, the present method of calculation of the conformation of globular proteins has avoided many of the weaknesses of previous efforts; i.e., the use of continuous backbone dihedral angles allows more precise specification of the protein conformation than the conformational states previously employed. Also by parameterizing the Ising model with conformational energy calculations, we have avoided the problem of insufficient experimental data. The model has been formulated in a mathematically rigorous manner without ad hoc rules and in such a way that it can be expanded in the future to include interactions more distant along the chain. It has been successful in predicting the conformations of a large percentage of residues correctly and has given rise to further useful insights as to the nature of local structures in globular proteins. Further, it can serve as a basis for more complete calculations in the future.

Acknowledgment. We are indebted to Drs. H. Meirovitch, G. Némethy, M. Oka, Y. Paterson, S. Rackovsky and H. Wako for helpful discussions and to S. Rumsey for preparing the ORTEP drawings of Figures 4 and 5.

Appendix A. Partition Function of a Specific-Sequence Polypeptide Chain with Nearest-Neighbor Interactions

Consider a polypeptide chain having a specific amino acid sequence. The chain can be considered to be constructed from fragments of the type represented in Figure 1a (or in Figure 1b if the residue is proline). If E_i represents the energy of interaction between all pairs of atoms *within* residue i, i.e., within the fragments of Figure 1a and 1b, respectively, then

$$E_i = E_i(\phi_i, \psi_i, \chi_i^1, ..., \chi_i^{\nu_i}) \text{ for Figure 1a}$$

and (A-1)

$$E_i = E_i(\omega_{i-1}, \psi_i) \text{ for Figure 1b}$$

where ν_i is the number of dihedral angles in the side chain of residue i. The theory will be formulated in terms of two backbone dihedral angles per residue [(ϕ, ψ) and (ω, ψ) for nonproline and proline residues, respectively]; the equations will be written, however, only in terms of (ϕ, ψ), but it should be understood that (ω, ψ) replaces (ϕ, ψ) when the residue is proline.

Interactions between neighboring residues are depicted in Figure 2a (or in Figure 2b if the second residue is proline). If $E_{i,i+1}$ is the energy of interaction between nearest neighbors, i and $i + 1$, i.e., between each atom of block a with all atoms of block b, which are separated by three or more bonds, then

$$E_{i,i+1} = E_{i,i+1}(\psi_i, \phi_{i+1}, \chi_i^1, ..., \chi_i^{\nu_i}, \chi_{i+1}^1, ..., \chi_{i+1}^{\nu_{i+1}}) \text{ for Figure 2a}$$

and (A-2)

$$E_{i,i+1} = E_{i,i+1}(\psi_i, \omega_i, \chi_i^1, ..., \chi_i^{\nu_i}) \text{ for Figure 2b}$$

The partition function for a polypeptide chain of N residues, taking into account interactions only within residue i and between residues i and $i + 1$, may be written as

$$Z = \int \cdots \int \left\{ \prod_{i=1}^{N-1} \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \right\} \times \exp(-\beta E_N)\, d\phi_1 \cdots d\phi_N\, d\psi_1 \cdots d\psi_N\, d\chi_1^1 \cdots d\chi_N^N \quad \text{(A-3)}$$

Hereafter, we introduce the shorthand $\tilde{\chi}_i$ for $\chi_i^1, \ldots, \chi_i^{p_i}$ and $d\tilde{\chi}_i$ for $d\chi_i^1 \cdots d\chi_i^{p_i}$. The probability distribution of the backbone dihedral angles (ϕ_i, ψ_i) for residue i is

$$P_i(\phi_i, \psi_i) = \frac{1}{Z} \int \cdots \int \left\{ \left[\prod_{k=1}^{i-1} \exp(-\beta E_k) \exp(-\beta E_{k,k+1})\, d\phi_k\, d\psi_k\, d\tilde{\chi}_k \right] \times \left[\prod_{k=i+1}^{N} \exp(-\beta E_k) \exp(-\beta E_{k-1,k})\, d\phi_k\, d\psi_k\, d\tilde{\chi}_k \right] \times \exp(-\beta E_i)\, d\tilde{\chi}_i \right\} \quad \text{(A-4)}$$

This expression for $P_i(\phi_i, \psi_i)$ may be represented in terms of other functions as follows. Define F_i as

$$F_i(\phi_i, \tilde{\chi}_i) = \int \cdots \int \prod_{k=1}^{i-1} [\exp(-\beta E_k) \exp(-\beta E_{k,k+1})\, d\phi_k\, d\psi_k\, d\tilde{\chi}_k] \quad \text{(A-5)}$$

$$\text{for } i = 2, N$$

(since $E_{k,k+1}$ depends on ϕ_{k+1} and χ_{k+1}'s and not ψ_{k+1}; see eq A-2) and F_{i-1} similarly. Then

$$F_i = \int \cdots \int \exp(-\beta E_{i-1}) \exp(-\beta E_{i-1,i}) \prod_{k=1}^{i-2} [\exp(-\beta E_k) \times \exp(-\beta E_{k,k+1})\, d\phi_k\, d\psi_k\, d\tilde{\chi}_k]\, d\phi_{i-1}\, d\psi_{i-1}\, d\tilde{\chi}_{i-1} \quad \text{(A-6)}$$

$$\text{for } i = 3, N$$

i.e.

$$F_i(\phi_i, \tilde{\chi}_i) = \int \cdots \int F_{i-1}(\phi_{i-1}, \tilde{\chi}_{i-1}) \times \exp(-\beta E_{i-1}) \exp(-\beta E_{i-1,i})\, d\phi_{i-1}\, d\psi_{i-1}\, d\tilde{\chi}_{i-1} \quad \text{(A-7)}$$

$$\text{for } i = 2, N$$

where $F_1 \equiv 1$. Likewise, define G_i as

$$G_i(\psi_i, \tilde{\chi}_i) = \int \cdots \int \prod_{k=i+1}^{N} [\exp(-\beta E_{k-1,k}) \exp(-\beta E_k)\, d\phi_k\, d\psi_k\, d\tilde{\chi}_k]$$

$$= \int \cdots \int G_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) \times \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1})\, d\phi_{i+1}\, d\psi_{i+1}\, d\tilde{\chi}_{i+1} \quad \text{(A-8)}$$

$$\text{for } i = 1, N-1$$

where $G_N \equiv 1$. Then we may write $P_i(\phi_i, \psi_i)$ as

$$P_i(\phi_i, \psi_i) = \frac{1}{Z} \int \cdots \int F_i(\phi_i, \tilde{\chi}_i) \exp(-\beta E_i) G_i(\psi_i, \tilde{\chi}_i)\, d\tilde{\chi}_i \quad \text{(A-9)}$$

and the partition function as

$$Z = \int \cdots \int F_N(\phi_N, \tilde{\chi}_N) \exp(-\beta E_N)\, d\phi_N\, d\psi_N\, d\tilde{\chi}_N \quad \text{(A-10)}$$

or as

$$Z = \int \cdots \int G_1(\phi_1, \tilde{\chi}_1) \exp(-\beta E_1)\, d\phi_1\, d\psi_1\, d\tilde{\chi}_1 \quad \text{(A-11)}$$

Since eq A-7 and A-8 are recursion formulas such that F_N has all the information of the preceding $N-1$ residues and G_1 all the information of the succeeding $N-1$ residues, eq A-10 or A-11 gives the partition function [and hence eq A-9 gives $P_i(\phi_i, \psi_i)$] by a simple numerical algorithm. Whereas E_i can be evaluated for each of the 20 amino acids and $E_{i,i+1}$ for each of the 400 dipeptide pairs, as indicated in section II, the quantities $F_1, F_2, F_3, \ldots, F_N$ and $G_N, G_{N-1}, G_{N-2}, \ldots, G_1$ must be evaluated separately for each protein since the F's and G's depend on the amino acid sequence.

Appendix B. Matrix Representation of the Partition Function

It is useful to introduce approximations that allow the partition function to be decoupled into products of matrices, each matrix depending only on the type of dipeptide. These matrices can then be evaluated for all 400 dipeptides and, since the amino acid sequence of any particular protein provides the correct order of multiplication of the matrices, the probability distributions $P(\phi, \psi)$ can be calculated for any amino acid sequence. It is important to average properly over the χ's as we decouple the partition function. This can be done by formulating the partition function in terms of a self-consistent effective potential, as in N-body theory. In the latter theory, the average potential energy of the jth particle, $\phi(\mathbf{r}_j)$, whose vector position is $\mathbf{r}_j$, is obtained by averaging the pair interaction potential $U(\mathbf{r}_j - \mathbf{r}_i)$ over the positions $\mathbf{r}_i$ of all the other particles,[40] i.e.

$$\phi(\mathbf{r}_j) = \sum_{i \neq j} \int U(\mathbf{r}_j - \mathbf{r}_i) P(\mathbf{r}_i | \mathbf{r}_j)\, d\mathbf{r}_i \quad \text{(B-1)}$$

where $P(\mathbf{r}_i|\mathbf{r}_j)$ is the conditional probability that the ith particle is at $\mathbf{r}_i$, given that the jth particle is at $\mathbf{r}_j$. In a similar manner, we make use of a self-consistent effective potential, U_{SCP}, for the nearest-neighbor interactions in the Ising model treatment of a polymer chain. We define U_{SCP} in this case as

$$U_{SCP}(\psi_i, \phi_{i+1}) \equiv \int \cdots \int E_{i,i+1}(\psi_i, \tilde{\chi}_i, \phi_{i+1}, \tilde{\chi}_{i+1}) \times P(\tilde{\chi}_i, \tilde{\chi}_{i+1} | \psi_i, \phi_{i+1})\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1} \quad \text{(B-2)}$$

where $P(\tilde{\chi}_i, \tilde{\chi}_{i+1}|\psi_i, \phi_{i+1})$ is the probability distribution for the two side chains having the conformation $(\tilde{\chi}_i, \tilde{\chi}_{i+1})$ when the backbone has the conformation (ψ_i, ϕ_{i+1}), and

$$P(\tilde{\chi}_i, \tilde{\chi}_{i+1} | \psi_i, \phi_{i+1}) = P(\tilde{\chi}_i, \tilde{\chi}_{i+1}, \psi_i, \phi_{i+1}) / P_i(\psi_i, \phi_{i+1}) \quad \text{(B-3)}$$

$$= \left\{ \int \cdots \int \left[\prod_{j=1}^{i-1} \exp(-\beta E_j) \exp(-\beta E_{j,j+1})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \right] \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) \times \left[\prod_{j=i+2}^{N} \exp(-\beta E_j) \exp(-\beta E_{j-1,j})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \right] d\phi_i\, d\psi_{i+1} \right\} \times \left\{ \int \cdots \int \left[\prod_{j=1}^{i-1} \exp(-\beta E_j) \exp(-\beta E_{j,j+1})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \right] \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) \times \left[\prod_{j=i+2}^{N} \exp(-\beta E_j) \exp(-\beta E_{j-1,j})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \right] d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1} \right\}^{-1} \quad \text{(B-4)}$$

$$= \left\{ \int \int f_i(\phi_i, \tilde{\chi}_i) \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) \times g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1})\, d\phi_i\, d\psi_{i+1} \right\} \Big/ \left\{ \int \cdots \int f_i(\phi_i, \tilde{\chi}_i) \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) \times d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1} \right\} \quad \text{(B-5)}$$

where

$$f_i(\phi_i, \tilde{\chi}_i) = \int \cdots \int \prod_{j=1}^{i-1} \exp(-\beta E_j) \exp(-\beta E_{j,j+1})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \quad \text{(B-6)}$$

and

$$g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) = \int \cdots \int \prod_{j=i+2}^{N} \exp(-\beta E_j) \exp(-\beta E_{j-1,j})\, d\phi_j\, d\psi_j\, d\tilde{\chi}_j \quad \text{(B-7)}$$

While f_i and g_i formally look like F_i and G_i of eq A-5 and the first of eq A-8, respectively, they are different quantities; because of the self-consistent potential approximation, f_i and g_i cannot obey the recursion relations of eq A-7 and the second of eq A-8, respectively.

By inserting eq B-5 into eq B-2, we obtain a self-consistent interaction potential

$$U_{SCP}^{(i,i+1)}(\psi_i, \phi_{i+1}) = \left\{\int \cdots \int E_{i,i+1}(\psi_i, \tilde{\chi}_i, \phi_{i+1}, \tilde{\chi}_{i+1}) f_i(\phi_i, \tilde{\chi}_i) \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) \times d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\}\left\{\int \cdots \int f_i(\phi_i, \tilde{\chi}_i) \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) \times d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\}^{-1} \quad \text{(B-8)}$$

At this stage, we introduce two different approximations. In the first-order approximation, we take all f's and g's as unity. This is equivalent to ignoring all residues in the chain except residues i and $i + 1$ when considering the interactions between the side chains of these two residues. In this approximation, eq B-8 becomes

$$U_{SCP}^{(i,i+1)}(\psi_i, \phi_{i+1}) = \left\{\int \cdots \int E_{i,i+1}(\psi_i, \tilde{\chi}_i, \phi_{i+1}, \tilde{\chi}_{i+1}) \times \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1})\, d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\} \times \left\{\int \cdots \int \exp(-\beta E_i) \exp(-\beta E_{i,i+1}) \exp(-\beta E_{i+1}) \times d\phi_i\, d\psi_{i+1}\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\}^{-1} \quad \text{(B-9)}$$

$$= \left\{\int \cdots \int E_{i,i+1}(\psi_i, \tilde{\chi}_i, \phi_{i+1}, \tilde{\chi}_{i+1}) \exp(-\beta E_i^*) \times \exp(-\beta E_{i,i+1}) \exp(-\beta \tilde{E}_{i+1})\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\} \times \left\{\int \cdots \int \exp(-\beta E_i^*) \times \exp(-\beta E_{i,i+1}) \exp(-\beta \tilde{E}_{i+1})\, d\tilde{\chi}_i\, d\tilde{\chi}_{i+1}\right\}^{-1} \quad \text{(B-10)}$$

where

$$\exp[-\beta E_i^*(\psi_i, \tilde{\chi}_i)] = \int \exp[-\beta E_i(\phi_i, \psi_i, \tilde{\chi}_i)]\, d\phi_i \quad \text{(B-11)}$$

and

$$\exp[-\beta \tilde{E}_{i+1}(\phi_{i+1}, \tilde{\chi}_{i+1})] = \int \exp[-\beta E_{i+1}(\phi_{i+1}, \psi_{i+1}, \tilde{\chi}_{i+1})]\, d\psi_{i+1} \quad \text{(B-12)}$$

A better approximation would be to use the f's and g's from the N-acetyl-N'-methyl dipeptide amides in the calculation of U_{SCP}. In this approximation, instead of ignoring all residues in the chain except residues i and $i + 1$, we simulate the borders of these two residues by the CH_3 of the acetyl group (preceding residue i) and the $NHCH_3$ of the methylamide (following residue $i + 1$). Thus, we approximate eq B-6 and B-7 by

$$f_i(\phi_i, \tilde{\chi}_i) = \int \exp[-\beta E_{CH_3}(\chi_{CH_3})] \exp[-\beta E_{CH_3,i}(\chi_{CH_3}, \phi_i, \tilde{\chi}_i)]\, d\chi_{CH_3} \quad \text{(B-13)}$$

and

$$g_{i+1}(\psi_{i+1}, \tilde{\chi}_{i+1}) = \int \exp[-\beta E_{NHCH_3}(\chi_{CH_3}')] \times \exp[-\beta E_{i+1,NHCH_3}(\psi_{i+1}, \tilde{\chi}_{i+1}, \chi_{CH_3}')]\, d\chi_{CH_3}' \quad \text{(B-14)}$$

where χ_{CH_3} is the dihedral angle for rotation of the CH_3 of the acetyl group and χ_{CH_3}' is the dihedral angle for rotation of the CH_3 of the $NHCH_3$ group. In practice, these dihedral angles are almost invariable. This gives

$$\exp[-\beta E_i^*(\psi_i, \tilde{\chi}_i)] = \int \int \exp(-\beta E_{CH_3}) \exp(-\beta E_{CH_3,i}) \exp(-\beta E_i)\, d\chi_{CH_3}\, d\phi_i \quad \text{(B-15)}$$

and

$$\exp[-\beta \tilde{E}_{i+1}(\phi_{i+1}, \tilde{\chi}_{i+1})] = \int \int \exp(-\beta E_{NHCH_3}) \exp(-\beta E_{i+1,NHCH_3}) \exp(-\beta E_{i+1}) \times d\psi_{i+1}\, d\chi_{CH_3}' \quad \text{(B-16)}$$

This approximation may be used for evaluating $U_{SCP}^{(i,i+1)}$ of eq B-10, which may then be inserted in place of $E_{i,i+1}$ in eq A-3 to obtain the partition function in the form

$$Z_{SCP} = \int \int \left[\prod_{i=1}^{N-1} \left\{\int \int \left[\int \cdots \int \exp(-\beta E_i)\, d\tilde{\chi}_i\right] \times \exp[-\beta U_{SCP}^{(i,i+1)}]\, d\phi_i\, d\psi_i\right\}\right] \int \cdots \int \exp(-\beta E_N)\, d\tilde{\chi}_N \times d\phi_N\, d\psi_N \quad \text{(B-17)}$$

When the Boltzmann factors in eq B-17 are evaluated numerically (see section III), they may be expressed as a Fourier series of order l, where l is an adjustable parameter (usually about 7) to achieve an adequate Fourier series representation of these Boltzmann factors. The purpose of using a Fourier series is to enable these integrals to be expressed as matrix products which can be evaluated numerically very easily. (Expansion in a Fourier series, so as to arrive at a matrix representation, had been carried out earlier by Allegra et al.[41] in the calculation of characteristic ratios of alternating copolymers.) For this purpose, we introduce the following matrix notation:

$$\int \cdots \int \exp[-\beta E_i(\phi_i, \psi_i, \tilde{\chi}_i)]\, d\tilde{\chi}_i = \mathbf{\Phi}_i^* \mathbf{A}_i \mathbf{\Psi}_i \quad \text{(B-18)}$$

and

$$\exp[-\beta U_{SCP}^{(i,i+1)}(\psi_i, \phi_{i+1})] = \mathbf{\Psi}_i^* \mathbf{B}_{i,i+1} \mathbf{\Phi}_{i+1} \quad \text{(B-19)}$$

where $\mathbf{\Phi}_i$ and $\mathbf{\Psi}_i$ are the column vectors

$$\mathbf{\Phi}_i \equiv \begin{bmatrix} 1/2^{1/2} \\ \cos\phi_i \\ \vdots \\ \cos(l\phi_i) \\ \sin\phi_i \\ \vdots \\ \sin(l\phi_i) \end{bmatrix} \quad \text{(B-20)}$$

$$\mathbf{\Psi}_i \equiv \begin{bmatrix} 1/2^{1/2} \\ \cos\psi_i \\ \vdots \\ \cos(l\psi_i) \\ \sin\psi_i \\ \vdots \\ \sin(l\psi_i) \end{bmatrix} \quad \text{(B-21)}$$

and the asterisk indicates a transpose. Also

$$\mathbf{A}_i \equiv \begin{bmatrix} 2a_{00} & 2^{1/2}a_{01} & \cdots & 2^{1/2}a_{0l} & 2^{1/2}b_{01} & \cdots & 2^{1/2}b_{0l} \\ 2^{1/2}a_{10} & a_{11} & \cdots & a_{1l} & b_{11} & \cdots & b_{1l} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 2^{1/2}a_{l0} & a_{l1} & \cdots & a_{ll} & b_{l1} & \cdots & b_{ll} \\ 2^{1/2}c_{10} & c_{11} & \cdots & c_{1l} & d_{11} & \cdots & d_{1l} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 2^{1/2}c_{l0} & c_{l1} & \cdots & c_{ll} & d_{l1} & \cdots & d_{ll} \end{bmatrix} \quad \text{(B-22)}$$

$$\mathbf{B}_{i,i+1} \equiv \begin{bmatrix} 2\alpha_{00} & 2^{1/2}\alpha_{01} & \cdots & 2^{1/2}\alpha_{0l} & 2^{1/2}\beta_{01} & \cdots & 2^{1/2}\beta_{0l} \\ 2^{1/2}\alpha_{10} & \alpha_{11} & \cdots & \alpha_{1l} & \beta_{11} & \cdots & \beta_{1l} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 2^{1/2}\alpha_{l0} & \alpha_{l1} & \cdots & \alpha_{ll} & \beta_{l1} & \cdots & \beta_{ll} \\ 2^{1/2}\gamma_{10} & \gamma_{11} & \cdots & \gamma_{1l} & \delta_{11} & \cdots & \delta_{1l} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 2^{1/2}\gamma_{l0} & \gamma_{l1} & \cdots & \gamma_{ll} & \delta_{l1} & \cdots & \delta_{ll} \end{bmatrix} \quad \text{(B-23)}$$

The quantities a_{mn}, b_{mn}, c_{mn}, and d_{mn} are to be understood as having superscripts (i) and are the constants of the terms of the Fourier expansion in cos $(m\phi_i)$ cos $(n\psi_i)$, cos $(m\phi_i)$ sin $(n\psi_i)$, sin $(m\phi_i)$ cos $(n\psi_i)$, and sin $(m\phi_i)$ sin $(n\psi_i)$, respectively, for $\int \ldots \int \exp[-\beta E_i(\phi_i, \psi_i, \tilde{\chi}_i)]\, d\tilde{\chi}_i$. The quantities α_{mn}, β_{mn}, γ_{mn}, and δ_{mn} are to be understood as having superscripts $(i, i+1)$ and are the similar terms of the Fourier expansion for $\exp[-\beta U_{\mathrm{SCF}}^{(i,i+1)}(\psi_i, \phi_{i+1})]$.

With this notation, the partition function becomes

$$\begin{aligned} Z_{\mathrm{SCP}} &= \int \ldots \int \left[\prod_{i=1}^{N-1} \mathbf{\Phi}_i^* \mathbf{A}_i \mathbf{\Psi}_i \mathbf{\Psi}_i^* \mathbf{B}_{i,i+1} \mathbf{\Phi}_{i+1}\right] \times \left[\mathbf{\Phi}_N^* \mathbf{A}_N \mathbf{\Psi}_N\right] d\phi_1 \ldots d\phi_N\, d\psi_1 \ldots d\psi_N \\ &= \int \ldots \int \mathbf{\Phi}_1^* \mathbf{A}_1 \left[\prod_{i=1}^{N-1} \mathbf{\Psi}_i \mathbf{\Psi}_i^* \mathbf{B}_{i,i+1} \mathbf{\Phi}_{i+1} \mathbf{\Phi}_{i+1}^* \mathbf{A}_{i+1}\right] \times \left[\mathbf{\Psi}_N\right] d\phi_1 \ldots d\phi_N\, d\psi_1 \ldots d\psi_N \\ &= \left[\int \mathbf{\Phi}_1^*\, d\phi_1\, \mathbf{A}_1\right]\left[\prod_{i=1}^{N-1}\left\{\int \mathbf{\Psi}_i \mathbf{\Psi}_i^*\, d\psi_i\, \mathbf{B}_{i,i+1} \times \int \mathbf{\Phi}_{i+1} \mathbf{\Phi}_{i+1}^*\, d\phi_{i+1}\, \mathbf{A}_{i+1}\right\}\right]\left[\int \mathbf{\Psi}_N\, d\psi_N\right] \end{aligned} \quad \text{(B-24)}$$

Now

$$\begin{aligned} \int_0^{2\pi} \mathbf{\Phi}_1^*\, d\phi_1 &= \left(\frac{1}{2^{1/2}}\int_0^{2\pi} d\phi, \int_0^{2\pi} \cos\phi\, d\phi, \ldots, \int_0^{2\pi} \cos(l\phi)\, d\phi, \int_0^{2\pi} \sin\phi\, d\phi, \ldots, \int_0^{2\pi} \sin(l\phi)\, d\phi\right) \\ &= 2^{1/2}\pi(1, 0, 0, \ldots, 0) = 2^{1/2}\pi \mathbf{e}^* \end{aligned} \quad \text{(B-25)}$$

where $\mathbf{e}^*$ is a row vector with the components indicated in eq B-25. Similarly

$$\int_0^{2\pi} \mathbf{\Psi}_N\, d\psi_N = 2^{1/2}\pi \mathbf{e} \quad \text{(B-26)}$$

where $\mathbf{e}$ is a corresponding column vector, and

$$\int_0^{2\pi} \mathbf{\Phi}_i \mathbf{\Phi}_i^*\, d\phi_i = \int_0^{2\pi} \mathbf{\Psi}_i \mathbf{\Psi}_i^*\, d\psi_i = \pi \mathbf{I}_{2l+1} \quad \text{(B-27)}$$

where $\mathbf{I}_{2l+1}$ is the identity matrix of order $(2l + 1)$. Hence, the partition function of eq B-24 becomes

$$Z_{\mathrm{SCP}} = 2\pi^{2N} \mathbf{e}^* \left[\prod_{i=1}^{N-1} \mathbf{A}_i \mathbf{B}_{i,i+1}\right] \mathbf{A}_N \mathbf{e} \quad \text{(B-28)}$$

Thus, the evaluation of Z_{SCP} involves the evaluation of the matrices of eq B-22 and B-23 and hence the integrals in (B-10) and (B-18).

Finally, the probability distribution functions for each pair of dihedral angles (ϕ_i, ψ_i) are given by

$$\begin{aligned} P_{\mathrm{SCP}}(\phi_i, \psi_i) &= \frac{1}{Z_{\mathrm{SCP}}} \int \ldots \int \left\{\left[\prod_{k=1}^{i-1} \exp(-\beta E_k) \exp(-\beta U_{\mathrm{SCP}}^{(k,k+1)}) \times d\phi_k\, d\psi_k\, d\tilde{\chi}_k\right]\left[\prod_{k=i+1}^{N} \exp(-\beta E_k) \exp(-\beta U_{\mathrm{SCP}}^{(k-1,k)}) \times d\phi_k\, d\psi_k\, d\tilde{\chi}_k\right]\right\}\left\{\exp(-\beta E_i)\, d\tilde{\chi}_i\right\} \quad \text{(B-29)} \\ &= \frac{2\pi^{2N-2}}{Z_{\mathrm{SCP}}} \mathbf{e}^* \left[\prod_{k=1}^{i-1} \mathbf{A}_k \mathbf{B}_{k,k+1}\right]\left[\mathbf{\Phi}_i \mathbf{\Phi}_i^* \mathbf{A}_i \mathbf{\Psi}_i \mathbf{\Psi}_i^*\right] \times \left[\prod_{k=i+1}^{N} \mathbf{B}_{k-1,k} \mathbf{A}_k\right] \mathbf{e} \quad \text{(B-30)} \\ &= \mathbf{t}_i^* \mathbf{\Phi}_i \mathbf{\Phi}_i^* \mathbf{A}_i \mathbf{\Psi}_i \mathbf{\Psi}_i^* \mathbf{h}_i \quad \text{(B-31)} \end{aligned}$$

where

$$\mathbf{t}_i^* = (2/Z_{\mathrm{SCP}})^{1/2} \pi^{N-1} \mathbf{e}^* \left[\prod_{k=1}^{i-1} \mathbf{A}_k \mathbf{B}_{k,k+1}\right] \quad \text{(B-32)}$$

and

$$\mathbf{h}_i = (2/Z_{\mathrm{SCP}})^{1/2} \pi^{N-1} \left[\prod_{k=i+1}^{N} \mathbf{B}_{k-1,k} \mathbf{A}_k\right] \mathbf{e} \quad \text{(B-33)}$$

References and Notes

(1) This work was supported by research grants from the National Institute of General Medical Sciences, National Institutes of Health, U.S. Public Health Service (GM-14312), and from the National Science Foundation (PCM75-08691).
(2) (a) National Research Council of Canada Postdoctoral Fellow, 1976–1977. (b) To whom requests for reprints should be addressed at Cornell University.
(3) Némethy, G.; Scheraga, H. A. *Q. Rev. Biophys.* **1977**, *10*, 239.
(4) Scheraga, H. A. *Pure Appl. Chem.* **1973**, *36*, 1.
(5) Kotelchuck, D.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1969**, *62*, 14.
(6) Wu, T. T.; Kabat, E. A. *J. Mol. Biol.* **1973**, *75*, 13.
(7) Burgess, A. W.; Ponnuswamy, P. K.; Scheraga, H. A. *Isr. J. Chem.* **1974**, *12*, 239.
(8) Lim, V. I. *J. Mol. Biol.* **1974**, *88*, 873.
(9) Chou, P. Y.; Fasman, G. D. *Biochemistry* **1974**, *13*, 211, 222.
(10) Robson, B.; Pain, R. H. *Biochem. J.* **1974**, *141*, 869, 883, 899.
(11) Maxfield, F. R.; Scheraga, H. A. *Biochemistry* **1976**, *15*, 5138.
(12) Maxfield, F. R.; Scheraga, H. A. *Biochemistry* **1979**, *18*, 697.
(13) Tanaka, S.; Scheraga, H. A. *Macromolecules* **1976**, *9*, 142, 159, 168, 812; **1977**, *10*, 9, 305.
(14) Nagano, K. *J. Mol. Biol.* **1977**, *109*, 251.
(15) Kotelchuck, D.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1968**, *61*, 1163.
(16) Finkelstein, A. V.; Ptitsyn, O. B. *J. Mol. Biol.* **1976**, *103*, 15.
(17) Ponnuswamy, P. K.; Warme, P. K.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1973**, *70*, 830.
(18) Burgess, A. W.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 1221.
(19) Zimmerman, S. S.; Pottle, M. S.; Némethy, G.; Scheraga, H. A. *Macromolecules* **1977**, *10*, 1.
(20) Tanaka, S.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1975**, *72*, 3802.
(21) Anfinsen, C. B.; Scheraga, H. A. *Adv. Protein Chem.* **1975**, *29*, 205.

(22) Chavez, L. G.; Scheraga, H. A. *Biochemistry* **1980**, *19*, 1005.
(23) Momany, F. A.; McGuire, R. F.; Burgess, A. W.; Scheraga, H. A. *J. Phys. Chem.* **1975**, *79*, 2361.
(24) The *side-chain* dihedral angles were the *final* values computed by Swenson et al.[25] from the experimental coordinates of Deisenhofer and Steigemann.[26] These coordinates[26] are referred to as X-ray 2 by Swenson et al.[25]
(25) Swenson, M. K.; Burgess, A. W.; Scheraga, H. A. "Frontiers in Physicochemical Biology"; Pullman, B., Ed.; Academic Press: New York, 1978; p 115.
(26) Deisenhofer, J.; Steigemann, W. *Acta Crystallogr., Sect. B* **1975**, *31*, 238.
(27) The values of the *backbone* dihedral angles were again the *final* X-ray 2 computed ones of Swenson et al.[25]
(28) Howard, J. C.; Momany, F. A.; Andreatta, R. H.; Scheraga, H. A. *Macromolecules* **1973**, *6*, 535.
(29) Hodes, Z. I.; Némethy, G.; Scheraga, H. A. *Biopolymers* **1979**, *18*, 1565, 1611.
(30) Némethy, G.; Hodes, Z. I.; Scheraga, H. A. *Proc. Natl. Acad. Sci. U.S.A.* **1978**, *75*, 5760.
(31) Paterson, Y.; Leach, S. J. *Macromolecules* **1978**, *11*, 409 (see supplementary material).
(32) Zimmerman, S. S.; Scheraga, H. A. *Biopolymers* **1977**, *16*, 811.
(33) Crow, E. L.; Davis, F. A.; Maxfield, M. W. "Statistics Manual"; Dover: New York, 1960; p 100.
(34) Such an omission can also be justified on energetic grounds. For residues *not* in the helical conformation in the native molecule (see footnote *d* of Table II), the calculated (Ising model) free energy of the helical conformation was always several kcal/mol (and sometimes more) above that at the (nonhelical) global minimum. On the other hand, for residues that *are* in the helical conformation in the native molecule, the calculated (Ising model) free energy (at the minimum corresponding to the helical conformation) was always only 1–1.5 kcal/mol above that calculated for the global minimum; hence, the helical conformation was (incorrectly) not assigned. This small difference in free energy (and the accompanying entropy loss), however, would be more than offset if the stabilizing hydrogen bonds (and the dipole–dipole interactions between the first and fourth residues) were included.
(35) The definition of the regions used in ref 32 is slightly different from the definition used in ref 18, but the method described for the "perfect" prediction leads to the same dihedral angles.
(36) "Ala-type" pertains to all types of residues except Gly and Pro.
(37) Flory, P. J. "Principles of Polymer Chemistry"; Cornell University Press: Ithaca, N.Y., 1953; Chapter 14.
(38) Actually, we find that the native structure lies within 1 kcal/mol of the short-range model.
(39) Ueda, Y.; Scheraga, H. A., work in progress.
(40) Reiss, H. *J. Chem. Phys.* **1967**, *47*, 186.
(41) Allegra, G.; Calligaris, M.; Randaccio, L. *Macromolecules* **1973**, *6*, 390.

Helix–Coil Stability Constants for the Naturally Occurring Amino Acids in Water. 18. Tryptophan Parameters from Random Poly[(hydroxypropyl)glutamine-*co*-L-tryptophan][1]

J. A. Nagy,[2a] S. P. Powers, B. O. Zweifel,[2b] and H. A. Scheraga*[2c]

Baker Laboratory of Chemistry, Cornell University, Ithaca, New York 14853.
Received May 5, 1980

ABSTRACT: Water-soluble, random copolymers containing L-tryptophan and N^5-(3-hydroxypropyl)-L-glutamine have been synthesized, fractionated, and characterized, and the thermally induced helix–coil transitions of these copolymers in water have been investigated. The incorporation of L-tryptophan was found to increase the helix content of the polymers at all temperatures in the range 0–60 °C. The Zimm–Bragg parameters σ and s for the helix–coil transition in poly(L-tryptophan) in water were deduced from an analysis of the melting curves of the copolymers in the manner described in earlier papers. The computed values of s indicate that L-tryptophan enhances helix growth at low temperatures and reduces it at high temperatures; the large value of σ indicates that, in water, this residue has a tendency to promote helix–coil boundaries at all temperatures.

I. Introduction

This paper is concerned with the determination of the helix–coil stability constants of L-tryptophan in water and is a continuation of the series of papers[3-19] in which the conformational preferences of the naturally occurring amino acids in water have been investigated by use of the "host–guest" technique. In this technique, a water-soluble, α-helical host homopolymer with nonionizable side chains is selected, and various amounts of a guest residue are incorporated into it to form random copolymers. By studying the thermally induced helix–coil transitions in these copolymers, it is possible to calculate the Zimm–Bragg[20] helix–coil parameters σ and s for the *guest* residues from an examination of their influence on the helix–coil transition properties of the *host* homopolymer. L-Tryptophan residues are incorporated into copolymers with an N^5-(3-hydroxypropyl)-L-glutamine host, and the thermally induced helix–coil transitions in these copolymers in water are studied.

Although no experimental determination of the helix–coil stability constants for L-tryptophan in aqueous solution has been reported, several studies on the relative preference of L-tryptophan residues for helix and coil conformations have been carried out. Poly(L-tryptophan) has been investigated in organic solvents,[21-28] and both block and random copolymers of tryptophan and amino acids with ionized as well as nonionized side chains have been studied in water[29-32] and in organic solvents.[22,23,26-28,32] The results of the present study indicate that, in water, L-tryptophan has a pronounced ability to promote helix–coil boundaries at all temperatures and that it can either enhance or reduce helix growth, depending on the temperature.

The synthesis of water-soluble random copolymers of L-tryptophan with N^5-(3-hydroxypropyl)-L-glutamine is described in section II, and the experimental characterization of these copolymers and their melting behavior in aqueous solution are presented in section III. Finally, in section IV, the data are analyzed by means of an appropriate form of the theory[3] to determine the helix–coil stability parameters of L-tryptophan in water. The theory is based on evidence[33,34] that short-range interactions dominate in determining the local conformation of a polypeptide or protein. The parameters for L-tryptophan are compared with empirical observations on the behavior of this residue in proteins and with a theoretical analysis of these quantities.

II. Experimental Section

A. Preparation and Characterization of the Copolymers. The synthesis of the copolymers was achieved by first co-

0024-9297/80/2213-1428$01.00/0 © 1980 American Chemical Society