Macromolecules **1980**, *13*, 518–525

Stabilization of Regular Conformational Regions in Proteins by Intraregion Electrostatic Interactions†

R. L. Jernigan,* S. Miyazawa, and S. C. Szu‡

Laboratory of Theoretical Biology, DCBD, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205. *Received January 24, 1980*

ABSTRACT: In addition to the intrinsic proclivities of various types of amino acids for different backbone conformations, a detailed description of secondary conformations in globular proteins should account for the important effects of position within the regular secondary regions. These position effects are expected to be largest for ionized and polar amino acids. We have formulated a simple method to estimate the free energies of regular secondary regions for polypeptides of specified sequence. For all possible regions of regular conformation, intraregion electrostatic interactions are calculated explicitly, in the approximation of fixed side chain positions. The free energies of backbone–backbone and backbone–C^β atom interactions are treated as a single conformation-dependent parameter. These parameters are taken to be identical for all amino acids except glycine and proline. The principal parameter, the energy difference between α helix and β strand, is related in a simple way to the experimental ratio of the number of helical residues to the number of β strand residues. An energy minimum for the molecule is determined by a dynamic programming scheme which is rigorous if the total molecular energy is given in the form of a sum of energies of the independent secondary regions. Four standard backbone conformations are considered. The neglect of specific solvent effects and long-range interactions means that results will correspond to early stages of folding. For six proteins, results compared with reported X-ray conformations are correct, on average, for 65% of all residues and for 83% of residues within regular secondary regions. Backbone–backbone and backbone–side chain interactions are most important in determining secondary structures, with side chain–side chain interactions appearing to make a relatively small contribution. Two other methods utilized to select a best set of molecular conformers are: (1) "conformational stability" selection and (2) a priori conformational probabilities calculated with a partition function representing an equilibrium mixture of all combinations of independent secondary regions. All methods yield similar results. Inspection of conformational probabilities reveals that a single conformation is highly favored for many residues; for these residues that conformation is chosen, regardless of the selection method. Apparently many of these conformations are so stable as to be retained at later folding stages, after specific solvent and long-range interactions are imposed. The similar quality of these and other reported secondary structure prediction methods, in spite of their diversity, implies that the native conformations of globular proteins are most likely maintained through redundant interactions, with significant electrostatic intraregion stabilization.

In contrast to polymers composed of regularly repeating sequences,[1] detailed treatments of the configurations of large biological macromolecules of varying sequence are necessarily more complex. Also these monomers are usually composed of significantly larger numbers of atoms. These two complicating features, together with the present level of computing technology, dictate that treatments of free energies of such molecules must rely upon major approximations. Previous attempts at elucidating stable secondary conformers in globular proteins have, for the most part, directly utilized a variety of statistical compilations of X-ray crystallographic results. The efficacy of these methods has been questioned[2] in many quarters. Further developments and improvements in these methods require resolution of the conjugate issue of whether or not the reporting of additional experimental structures will provide significant improvements through an expansion of the statistical bases. A negative viewpoint was recently voiced by Maxfield and Scheraga.[3]

Two methods have attempted to overcome the restrictions imposed by these statistical origins. Lim[4] relied principally upon considerations of solvent interactions and packing to derive some simple rules to indicate the relative preferences of a given peptide sequence for various conformations. The method of Ptitsyn and Finkelstein[5] overcame the restrictions of its statistical origin by setting forth some rules with physical meaning. First in importance among these is the important role of backbone–side chain electrical interactions in α helix. There is a strong preference for positively charged side chains to be located near the carboxy terminus and for negatively charged side chains to occur in the vicinity of the amino terminus. These are based upon favorable side chain interactions with the helix backbone dipole. Perutz[6] has emphasized the broader importance of electrostatic interactions in proteins by reviewing their effects in folding, enzymatic activity, and denaturation kinetics.

Below we have calculated approximate free energies of regular secondary regions of proteins. The four standard conformations to be considered, together with their φ,ψ angles,[7] are: α, right-handed α helix at 122,133; β, right-handed β strand at 62,–68; α_L left-handed α helix at –122,–133; and β_L left-handed β strand at –62,68. Because of the very limited number of these states, each one must necessarily serve to represent a number of other nearby individual conformations. The free energies of these states are obtained by directly calculating electrostatic energies and by approximating the nonelectrostatic free energies in a simple manner, with the principal parameter being related to the fractions of secondary conformations occurring within a protein.

Free Energies

The total free energy of a molecule can be divided into contributions from pairwise interactions among backbone atoms, side chain atoms, and solvent molecules.

$$F_{tot} = F_{bb\text{-}bb} + F_{bb\text{-}sc} + F_{sc\text{-}sc} + F_{solv} \qquad (1)$$

For most regular regions, backbone–solvent interactions are rendered relatively unimportant by the interposition of side chain atoms. It is assumed here that the usually long range interregion side chain–side chain interactions and side chain–solvent interactions are more important in stabilizing tertiary interactions, whereas the first terms

‡Division of Bacterial Products, Bureau of Biologics, Food and Drug Administration, Bethesda, Maryland 20205.

†Dedicated to Paul Flory on his 70th birthday.

are more significant in determining secondary structures.

$$F_{tot} = F_{sec} + F_{tert} \quad (2a)$$

$$F_{sec} = F_{bb-bb} + F_{bb-sc} + F_{sc-sc}^{intraregion} \quad (2b)$$

$$F_{tert} = F_{sc-sc}^{interregion} + F_{solv} \quad (2c)$$

One test of the separation hypothesis in eq 2a would be to determine whether or not eq 2b yields the lowest energies for observed regular secondary regions. The extent of perturbation, by the two terms in eq 2c, of secondary conformations chosen with eq 2b remains to be determined. Such perturbations are to be expected because otherwise eq 2b alone would completely determine the total molecular conformation; this also serves to point out the arbitrary nature of the separation into secondary and tertiary conformations. Equation 2b typically includes short and medium range contributions, whereas eq 2c usually contains longer range interactions and solvent effects. The actual range of these interactions is variable because the lengths of the regions vary. Treating eq 2b as dominant, with eq 2c as a perturbation, is a particularly practical hypothesis since the greatest computational difficulties reside in estimating terms in the latter equation. In the present case, we are ignoring the terms in eq 2c. At early stages of folding, the conformations would be most properly represented by an average over many conformations. By this averaging, specific long-range and specific solvent interactions are likely rendered less important than in a final approach to the native conformation.

If the definition of backbone atoms is expanded to include β carbon atoms, then there must be two types of backbones: glycine and all others. By this device, the "backbone–backbone" interaction term will include the most common backbone–side chain steric interactions. The critical importance of electrostatic side chain–backbone interactions was indicated by the considerable success achieved with the secondary structure prediction method of Ptitsyn and Finkelstein.[5] Their method attributed relatively favorable weights to occurrences of negatively charged side chains near the amino terminus of a helix and of positively charged side chains near the carboxy terminus of a helix. Supporting evidence is offered in the terminal effect statistics provided by Crawford, Lipscomb, and Schellman.[8]

Equation 2b is simplified further by considering only the intraregion electrostatic side chain interactions

$$F_{sec} = F_{"bb-bb"} + F_{bb-sc}^{es} + F_{sc-sc}^{es,intraregion} \quad (3)$$

Energies in the first term of eq 3 for regions of regular conformation can be postulated to be approximately linear with the length of the region. Oobatake and Ooi[9] found this relationship for van der Waals' interactions in alanine helices and β strands. For a region comprising residues i to j in regular conformation ζ, the dimensionless "backbone–backbone" interaction energy is taken to be

$$F_{"bb-bb"}^{\zeta,i,j}/RT = n_{gly}F_{gly}^{\zeta} + (j - i + 1 - n_{gly})F^{\zeta} \quad (4)$$

where n_{gly} is the number of glycines included within the region i to j. We choose to treat F^{ζ} and F_{gly}^{ζ} as parameters; their evaluation will be considered later. Note that the parameters on the right side of eq 4 are dimensionless. If ζ corresponds to right-handed α helix, then two additional terms are included. The first is the helical hydrogen bond energy and the second represents the unfavorable energy[10] caused by a proline in the second or successive positions of a helix.

$$(j - i - 1 - n_{pro})F_H^{\alpha} + n_{xpro}F_{xpro}^{\alpha}$$

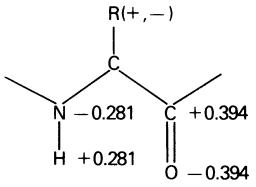

Figure 1. Diagram of peptide backbone charges.

Here n_{pro} is the number of prolines which prevent helical hydrogen bond formation, i.e., those prolines occurring within residues $i + 3$ to $j + 1$, and n_{xpro} is the number of residues, excluding glycine,[10] which precede prolines. This formulation has the major advantage that the only explicit dependence on interatomic distances is a simple inverse dependence; this obviates the sensitive higher power dependences on atomic positions which are encountered in a calculation of van der Waals' energies. As long as relatively few conformations are treated, it should be feasible to evaluate the required parameters. Also these parameters may absorb some of the error introduced by the fixed side chain approximation. Next, the details of calculating the electrostatic terms must be considered.

Electrostatic Energies

Energies of electrical interactions between polar or ionized atoms are taken to be Coulombic with a dielectric constant ϵ

$$F^{es} = Cq_iq_j/\epsilon r_{ij}$$

where r_{ij} is the distance between atoms i and j which have charges q_i and q_j, respectively. C is 332.0 for units of distance, charge, and energy taken as Å, electronic charge, and kcal mol^{-1}. Dipole moments are represented by partial charges. The backbone atom charges are those derived by Brant, Miller, and Flory[11] from small molecule dipole moments. These backbone charges are shown in Figure 1. The exception is proline; for that residue we have taken the nitrogen atom to have no charge and the carbon and oxygen charges to be the same as those for other residues. Backbone bond lengths and bond angles are those given in ref 11. The average positions of side chain atoms with respect to backbone atoms have been determined by averaging atomic positions for all occurrences of each amino acid within six proteins,[12] namely metmyoglobin, subtilisin BPN′, carboxypeptidase A, ribonuclease S, concanavalin A, and lysozyme. Alternative choices for the positions of side chain atoms could have been those side chain conformations determined to have minimum energies for an individual amino acid; however, such positions exhibit a dependence on the backbone conformation.[13]

In these calculations we have assumed a fixed mean position for each charged side chain atom. The effects of small errors in these positions on the side chain–backbone energies should be relatively small. Presumably the effects of this fixed side chain atom approximation would be greater for side chain–side chain interactions; however, as will be seen later, this class of interactions appears to exert relatively little influence in determining secondary forms. Mean positions for charged and polar atoms are tabulated in Table I for side chain atoms R in terms of r the distance between R and C^{α}, θ_R the effective bond angle N–C^{α}–R, and φ_R a rotation angle about the central bond of C–N–C^{α}–R, taken to be zero when it is in the trans conformation. Inspection of Table I reveals relatively little variation in the angle parameters for most amino acids. The largest variations occur among the α carbon to side chain atom distances. Charge assignments q for these atoms are given in the last column. They were obtained from the pK

Table I
Average Positions and Charges of Ionized and Polar Side Chain Atoms

amino acid	atom	r	θ_R	φ_R	q
arg	C^{ζ}	5.66	115	221	1.0
asn	O^{δ_1}	3.09	111	210	−0.5
	N^{δ_2}	3.50	105	226	−0.25
	C^{γ}	2.57	108	222	0.5
	H^{δ_2}	4.48	103	228	0.25
asp	$(O^{\delta_1},O^{\delta_2})_{av}$	3.21	110	215	−1.0
cys	S^{γ}	2.83	103	214	−0.31
	H^{γ}	3.15	102	211	0.31
gln	C^{δ}	3.63	120	211	0.5
	O^{ϵ_1}	4.17	117	205	−0.5
	N^{ϵ_2}	4.40	123	209	−0.25
	H^{ϵ_2}	5.37	125	207	0.25
glu	$(O^{\epsilon_1},O^{\epsilon_2})_{av}$	4.20	110	212	−0.998
his	$(N^{\delta_1},N^{\epsilon_2})_{av}$	3.96	107	214	0.091
lys	N^{ζ}	5.77	116	210	1.0
ser	O^{γ}	2.44	110	238	−0.35
	H^{γ}	2.78	110	238	0.35
thr	O^{γ_1}	2.35	99	233	−0.35
	H^{γ_1}	2.68	97	232	0.35
trp	N^{ϵ_1}	4.62	120	205	−0.40
	H^{ϵ_1}	5.58	119	202	0.40
tyr	O^{η}	6.44	107	207	−0.35
	H^{η}	7.43	107	206	0.35

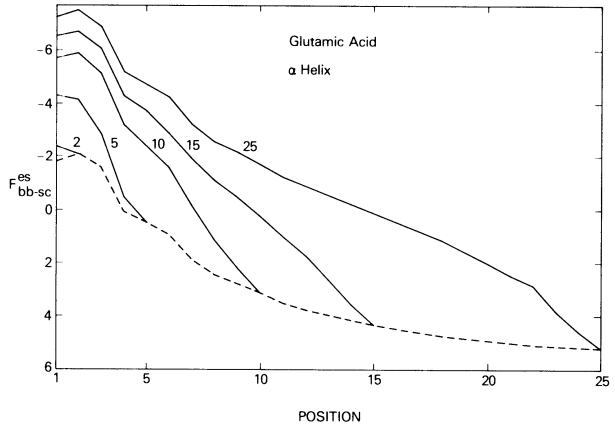

Figure 2. Backbone–side chain electrostatic interaction energies in kcal mol^{-1} for different lengths of α helix backbone fragments with one ionized glutamic acid side chain at the position indicated on the abscissa. Numbers on the curves are numbers of backbone residues in the chain. A complete peptide bond is always included on both ends of these fragments. The dashed curve corresponds to placement of the side chain at the carboxy terminus of various lengths of α helix backbone fragments. The dielectric constant is taken to be 3.5.

values[14] evaluated at pH 7 for ionized amino acids and for polar amino acids from dipole moments[15] of small molecules and standard bond lengths. The determination of the location of hydrogen atoms, which contribute to the dipole moments in asn, cys, gln, ser, thr, trp, and tyr, is somewhat arbitrary. We have chosen to assume that there is free rotation about the penultimate bonds; this reduces their effective dipole moments. For example, the O–H distance for serine calculated directly from the data in Table I is a highly unrealistic 0.34 Å. An alternative choice of charges could have been one of the complete sets of partial charges for all atoms such as those calculated by CNDO/2[16] or SCF–LCAO–MO[17] methods, but these would greatly increase the number of interacting atomic pairs in the calculations.

The relative importance of side chain–backbone interactions in determining conformations has been pointed out by Ralston and DeCoen,[13] and the importance of charged side chains in stabilizing helices were discussed by Blagdon and Goodman.[18] The effects of ionized side chains in determining helix termini can be critical. In Figures 2 and 3 we have presented the calculated electrostatic backbone–side chain energies for a single ionized glutamic acid side chain interacting with various α helical and β strand backbones. The energies in both cases favor placement of the glutamic acid near the amino terminus, but the effect is much larger for α helix. For the side chains with

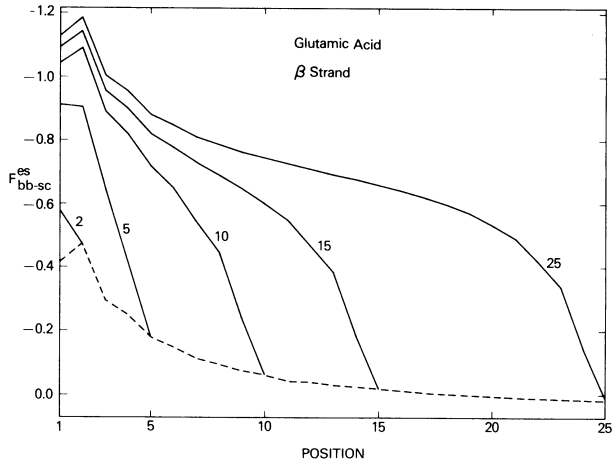

Figure 3. F_{bb-sc}^{es} for one ionized glutamic acid side chain at positions indicated on the abscissa for β strand backbone segments of lengths given on the curves. Units of energy are kcal mol^{-1}.

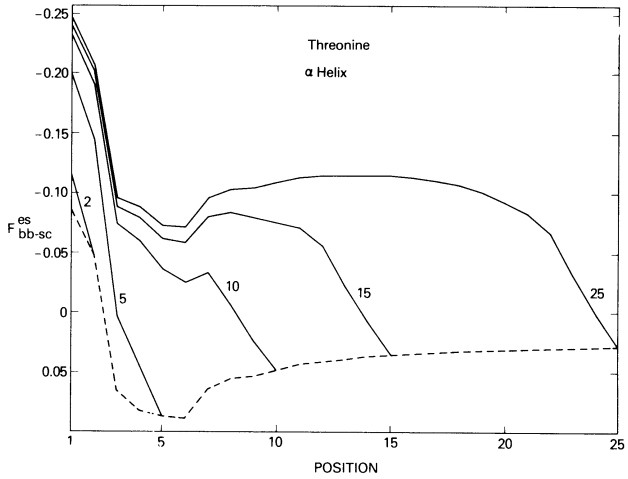

Figure 4. F_{bb-sc}^{es} for an α helical backbone with one threonine polar side chain.

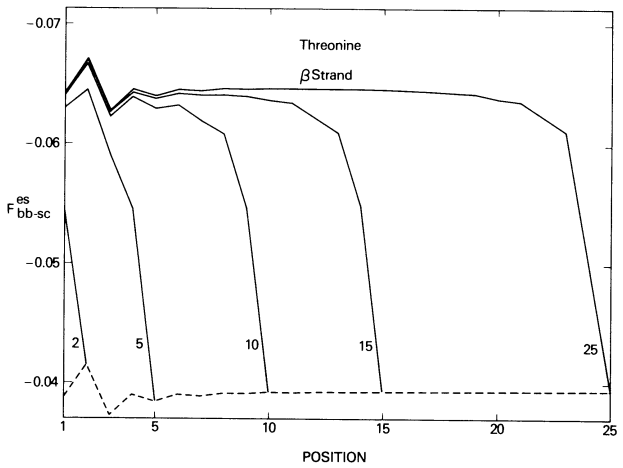

Figure 5. F_{bb-sc}^{es} for a β strand backbone with one polar threonine side chain.

dipoles, such energies are significantly smaller and characteristically are somewhat less uniform with successive relocation of the sole side chain along the fixed backbone. Examples for threonine are shown in Figures 4 and 5. The ordinate intercept of the dashed line is the electrostatic energy for a side chain and its two flanking peptide bonds. The intraresidue electrostatic interaction energies are lower for α helix than for β strand for all negatively charged ions and polar side chains except asn. The magnitudes of interresidue interactions for a given position in a given chain are almost always significantly larger for α helix than for β strand. The ionized side chains have the most favorable electrostatic energies when placed near the amino ends if they are negatively charged and near the carboxy termini if they are positively charged. The polar side chains cys, ser, thr, trp, and tyr behave similarly to the negatively charged ions and possess lower energies when located near amino termini. Asn and gln prefer carboxy termini. Slopes of these families of curves are usually larger for α helix than for β strand, but the signs of the slopes are usually the same.

Description of Conformations

In the present calculations we are selecting among *four standard backbone configurations*, namely right-handed α helix with φ,ψ values[7] of 122,133, right-handed β strand with φ,ψ values of 62,–68, left-handed α helix with φ,ψ of –122,–133, and left-handed β strand with φ,ψ of –62,68. Expansion of the number of states is possible but increases the number of parameters commensurately.

Energies of fragments formed by repeating these standard conformations are calculated. The maximum lengths of regular regions are designated by m_ζ. Regions beginning at every position with lengths from 1 to m_ζ are included. The total number of such independent regular regions in a molecule of n residues is

$$(n + \tfrac{1}{2})\sum_\zeta m_\zeta - \tfrac{1}{2}\sum_\zeta m_\zeta^2$$

if $n > m_\zeta$; otherwise the m_ζ's must be replaced with n's. In the present calculations, we have arbitrarily taken 26, 13, 3, and 3 as limits for m_α, m_β, m_{α_L} and m_{β_L}. With these values the total number of regions becomes $45n - 409$.

Fractions of the various secondary conformations are not reliably available from physical measurements in solution. However, we have chosen to proceed as if these were at hand and have utilized the X-ray structures to obtain fractions of a protein's residues in the standard conformations. We have assigned all residues strictly on the basis of their reported φ,ψ angles, regardless of their hydrogen bonding scheme.

The criteria we have used for assignments of conformations from the X-ray structures are based on determinations of distances in φ,ψ space from the standard angles, defined as

$$d_\zeta = [(\varphi - \varphi_\zeta)^2 + (\psi - \psi_\zeta)^2]^{1/2}$$

The smallest distance is used to determine whether it is right handed or left handed. If a right handed conformation is indicated, then the ratio of d_β/d_α is inspected. If it is greater than $3^{1/2}$, then right-handed α conformation is assumed, and if this ratio is less than or equal to $3^{-1/2}$, then it is taken as β. If calculated ratios fall outside either of these limits, they are assigned an intermediate conformation; some intermediate cases can be seen in Figure 7. Other methods for dividing φ,ψ space into discrete conformations have been espoused; this is among the simplest. For each molecule, the number of residues determined to be in the ζ conformation is specified by n_ζ.

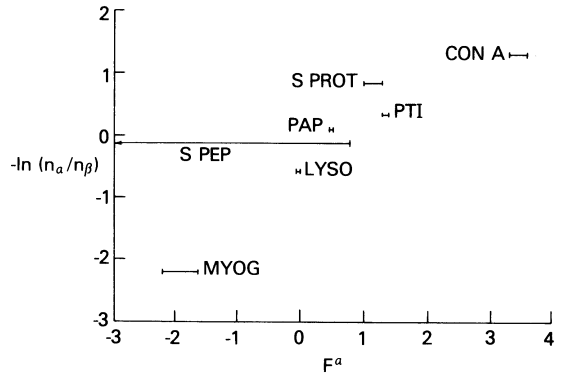

Figure 6. Comparison of the experimental value of $-\ln (n_\alpha/n_\beta)$ with the value of the parameter F^α which gives the best results with energy minimization for the secondary structures of the six molecules indicated. Values of the other parameters and the quality of results are given in Table II. The arrow on the left side of the S PEP bar indicates that any value of F^α in the indicated direction yields the same conformation.

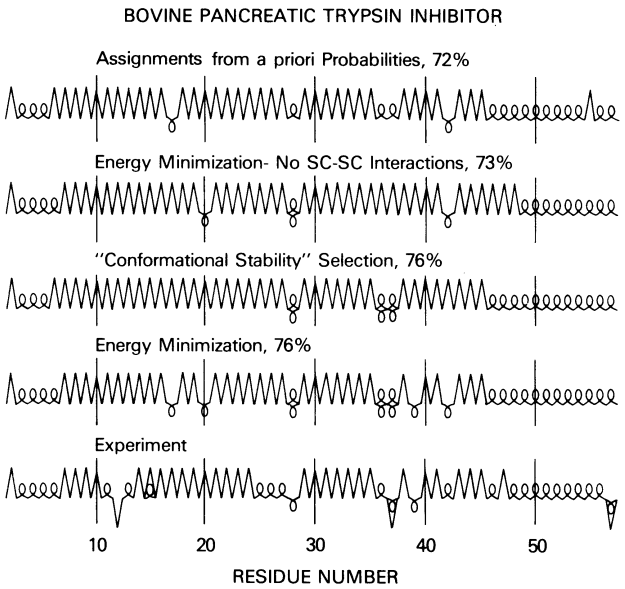

Figure 7. Comparison of results for several conformational selection methods applied to pancreatic trypsin inhibitor. Specific methods and quality of results are indicated above each curve. The diagrammatic symbols are: ⍝ right-handed α helix, ⍋ right-handed β strand, ⍒ left-handed α helix, and ⍕ left-handed β strand. Parameters used for the calculation were $F^\alpha = 1.5$, $F^{\alpha L} = 2.0$, $F_{gly}{}^\alpha = -0.4$, $F_{xpro}{}^\alpha = 5.0$, and $F_H{}^\alpha = -0.67$. For the case with no side chain–side chain interactions included, the parameters were $F^\alpha = 1.9$, $F^{\alpha L} = 2.4$, and $F_{gly}{}^\alpha = -0.2$.

Determination of Parameters

In the method as described so far there are a large number of parameters. Here we are trying to set forth a model with a minimum number of adjustable parameters. As we have mentioned above, we are considering only two types of principal parameters, for glycine and for all other amino acids.

Simple inspection of molecular models indicates that the smallest steric repulsion between side chain groups and backbone atoms is encountered in a β strand conformation. We have accordingly taken $F^\beta = 0$ for all amino acids. In addition, for glycine, the intraresidue interactions are the same for left-handed conformation as for right-handed ones; hence $F_{gly}{}^{\beta_L} = 0$ and $F_{gly}{}^\alpha = F_{gly}{}^{\alpha_L}$. The residue preceding a proline is disfavored to be in α helical form.

The two left-handed conformations for proline are *completely excluded* because of the constraint on the φ angle forced by the closing of the pyrrolidine ring. For simplification, the parameter for proline in α helix is assigned the same value as for the group of amino acids with β-methylene groups. A β-methylene group causes both of the left-handed conformations to be quite unfavorable; we have arbitrarily taken $F^{\alpha_L} = F^{\beta_L}$ for this group of amino acids. Furthermore, if these parameters are to represent simple steric repulsions, then their nature is such that we might expect the inequalities $F_{gly}{}^{\alpha} < F^{\alpha} < F^{\alpha_L}$ to hold. There are five molecular parameters: F_{H^α} helix hydrogen bond energy, $F_{xpro}{}^{\alpha}$ the energy of a residue preceding proline in an α helix, and the three parameters just described, namely F^{α}, F^{α_L}, and $F_{gly}{}^{\alpha}$. Calculations below have indicated an insensitivity of the final results to the value of the helical hydrogen bond energy, within the range $-1.33 \leq F_H{}^{\alpha} \leq -0.33$. We have chosen to present results for a value of $F_H{}^{\alpha} = -0.67$ but could have obtained nearly identical results for other values within the indicated range. However, the best values of the other parameters depend strongly on the exact value of $F_H{}^{\alpha}$ chosen. Because of the relatively low frequencies of occurrences of glycine and of all amino acids in left-handed conformations, results are not very sensitive to the values of F^{α_L} and $F_{gly}{}^{\alpha}$. Only one principal parameter, F^{α}, remains.

Although ϵ and RT have been formally and separately specified here, it is not necessary. Whenever free energies are minimized or placed in order, it is possible to scale all energies by the dielectric constant; in such cases, the factor ϵRT can be absorbed within each of the molecular parameters. Neither ϵ nor RT would appear explicitly in any expression. In order to calculate probabilities, it becomes necessary to identify unscaled free energies; this requires specification only of the combined single parameter ϵRT. In all calculations here this parameter ϵRT has been taken as 2.1 kcal mol^{-1}, corresponding to $RT = 0.6$ and a dielectric constant of 3.5 from the precedent set by Brant, Miller, and Flory.[11]

In order to attempt to relate the values of the parameters F^{α} to the molecules' overall conformational proclivities, we have plotted in Figure 6 the experimental values of $-\ln (n_\alpha/n_\beta)$ together with the range of values of the parameters F^{α} which give results with energy minimization closest to the native forms of the proteins. The bars indicate ranges of the parameter which yield identical results. If the model were properly formulated in complete detail, one would expect a linear relationship. For the six proteins included in this plot, there is some deviation from a single straight line. However, if one admits of variability caused by the numerous approximations in this simple model, there is clearly a linear trend with a turn toward a constant value at large F^{α}. Equality between F^{α} and $-\ln (n_\alpha/n_\beta)$ is observed for myoglobin. The largest deviation from this type of equality is found for concanavalin A. Deviations from equality for larger values can be attributed to our omission of an energy contribution from β strand hydrogen bonds. Accounting for these hydrogen bonds would require the difficult consideration of interregion interactions. This omission has the effect of requiring larger than expected values of the F^{α} parameter for molecules with large values of n_β/n_α. Although there appears to be a general trend for values of F^{α_L} to be correlated with the values of $-\ln (n_{\alpha_L}/n_\beta)$, this relationship cannot be critically evaluated for the few occurrences of the α_L state.

Relatively small changes in the conformations chosen were found upon varying the parameters F^{α_L}, $F_{gly}{}^{\alpha}$, and $F_{xpro}{}^{\alpha}$. The value of $F_{xpro}{}^{\alpha}$ has been fixed at 10.0. Usually $F_{gly}{}^{\alpha}$ has been taken to be 0.5 and F^{α_L} to be 2.0; however, negative values of $F_{gly}{}^{\alpha}$ were found to be slightly better for trypsin inhibitor and myoglobin and larger values of F^{α_L} were somewhat better for concanavalin A.

Three cases of cis peptide bonds, two in ribonuclease S and one in papain, are observed in the X-ray results for this group of six molecules. Cis–trans peptide bond isomerism has not been included in the present treatment. We have assumed that the locations of these rare cis peptides are known.

Methods for Choosing Sets of Secondary Conformations

Three methods have been utilized here: energy minimization, "conformational stability" selection, and assignments from probabilities calculated with an equilibrium partition function. A rigorous method for determining the total energy minimum of a set of independent molecular segments has been described in detail in ref 19. Interactions within each regular region are included but not interactions with atoms beyond the termini of the region. This approach is suitable only for molecules composed principally of long regions of regular conformations. For such a protein, this energy minimization scheme will yield the most favorable combination of secondary conformers. The assumed independence of the separate regions makes possible a rigorous energy minimization. Of the total number of possible combinations of regular regions, it is necessary to consider relatively few. In this dynamic programming scheme, one steps along the chain, one residue at a time. At each stage all regions terminating with that residue are considered; one best case for each of the standard conformations is determined. These conformations are later considered in combination with regions which begin at the adjacent exterior residues. In the version of this algorithm applied here, it is assumed that separate regions of the same conformation cannot be adjacent to one another. This method is remarkably simple because it yields a rigorous minimum energy form from the set of all possible combinations of a large number of independent regions.

"Conformational stability" selection is simpler and gives results which closely approximate the energy minimization results. In this method, all conformations are placed in order by their energies. The contents of this ordered conformational stewpot are examined by looking successively from bottom to top, from low energy forms to higher ones. Whenever a conformation is found which corresponds completely to unassigned residues, it is deposited in the corresponding residues' soup bowls. This is repeated until all residues are assigned a conformation. This operational method is a simple one and provides relatively good results. No attempt is made to prevent neighboring independent regions from assuming the same conformation. Both this and the energy minimization method may so position breaks between regions as to completely avoid highly unfavorable interactions. In reality, some of these interactions *cannot* be completely avoided. The artificial nature of these barriers to interactions between neighboring residues of adjacent regular conformational regions can be investigated.[20] Calculations which include, in addition, short range interactions between neighboring residues across these barriers do not appear to yield results significantly different from those presented here. Inclusion of such interactions, but of longer range than next nearest neighbors, might yield improved results.

Details of the calculation of the equilibrium conformational probabilities are given in the Appendix. It provides

significant contrast to the other two methods. Each residue's conformation is assigned independently to the conformation with the largest probability.

Results and Discussion

Molecules treated here are: concanavalin A (CON A); lysozyme (LYSO); myoglobin (MYOG); papain (PAP); ribonuclease S comprising two fragments, the S peptide (S PEP) and the S protein (S PROT); and trypsin inhibitor (PTI). Sequences and X-ray coordinates were taken from the compilation of Feldmann.[12] Neither of the terminal residues were included in the calculations; hence, the molecular terminal amine and carboxylic acid groups and their associated charges have also been neglected.

In Figure 7 results for trypsin inhibitor are compared for the three selection methods: energy minimization with and without inclusion of side chain–side chain interactions, "conformational stability" selection, and assignments according to the most favorable a priori probabilities. Parameters are slightly different than those for the more general cases in Table II. The quality of all of the results is similar, evidencing at most a 4% difference among the methods. This is remarkable since the selection methods are quite different in character. Also it can be seen that the neglect of the intraregion side chain–side chain electrostatic interactions in the energy minimization causes only a further 3% of the residues to be wrong. This implies that secondary conformations are determined predominantly by backbone–backbone and backbone–side chain interactions. The last term in eq 2b can possibly be neglected in good approximation. Inclusion of side chain–side chain interactions appears to lead to only small improvements for both PTI and MYOG which were the only two molecules so treated. This may or may not be consistent with the results of the examination of occurrences of charged pairs in helices by Maxfield and Scheraga.[21] The "conformational stability" selection method implies a mechanism for folding. It corresponds to the most energetically stable secondary conformational regions being successively formed. The differences among the results for the various methods are not large enough to indicate a best method; therefore it is not possible to infer here whether, at an early stage of folding, a definite folding mechanism or a simple equilibrium mixture of all conformations prevails.

Results for the group of six molecules are given in Table II. Especially good results are obtained for PTI and MYOG. The worst case is the S peptide of ribonuclease S; it corresponds to almost complete helix. There are three possible explanations for this poor result: the molecule is too small for these methods to be applicable, there are large errors in the calculations, or the interactions with the S protein significantly perturb the S peptide's intrinsically preferred conformation. Twenty residues may not be sufficient size for application of these approximate energy calculations. There may not be a single preferred conformation; small polypeptides often assume mixtures of conformations in solution. Arguments in favor of the importance of the S peptide–S protein interactions are compelling; however, there is some contrary experimental evidence[22] indicating that the S peptide conformation is unchanged upon its separation from the S protein.

As expected, results in the last column, counting only those residues which appear in regular regions, are significantly better in every case. The definition of regular regions has been taken as three or more consecutive residues of the same conformation for β strand and five or more consecutive residues for α helix. The two averages given on the last line of Table II are weighted by the number of residues and the number of residues occurring in the regular regions, respectively. The calculations presented here do not pretend to properly account for turns or regions of mixed conformations; therefore the results, which are significantly better than random, are surprisingly good. They are occasionally better than the numbers indicate because the definition of intermediate states and their counting means that completely correct results are unattainable. For example, in Figure 7, the experimental mixed conformations at residues 15, 37, and 57 cannot be obtained with the four standard conformational states. For this case the maximum achievable is 97% correct.

The importance of small molecules and ions included in the crystal remains to be determined. An inspection of such atoms and their locations with respect to residues with good or bad conformational choices was inconclusive. In a similar way disulfide bonds may perturb these results.

Let us consider some of the inadequacies in the present calculations. Charges on some of the residues may vary, depending on their environment. The most likely residues to have variable charges and dipoles are histidines and cysteines. The pKs of histidines are known to depend upon their neighboring residues, and sulfurs can be in the form of either S–H or S–S bonds with corresponding differences in charges. Details of van der Waals' side chain–side chain interactions have been completely ignored. Hydrogen bonds in α helices have been explicitly included but we have not attempted to consider the more difficult problem of counting hydrogen bonds in various types of turns or the long range hydrogen bonds in β strands. Electrical interactions have been considered in the simplest possible form. Dipoles have been approximated as point charges. A few cases of calculations including Debye–Hückel shielding did not yield results significantly different from those given here. Recently Skolnick and Fixman[23] have considered interacting charges on the surface of a cylinder. They observe large enhancements or decreases over simple Coulombic interactions, depending on the relative positions of the two charges. Unfortunately, their results are not in a form appropriate to permit direct application here.

The parameter F^{α} may reflect the compositional differences for the group of proteins considered. As formulated here this parameter includes a wide variety of interactions which have not been treated in detail. It could be related in further detail to protein composition by taking distinct parameters for more of the individual amino acids. A simple dependence of the parameter F^{α} on the protein's composition implies that a given residue may feel the influence of a mean field originating in all other residues; this is only possible for a relatively large molecule. Polypeptides are stiff enough so that large numbers of interresidue interactions are obtainable only for a protein of sufficient size. The parameter F^{α} may also reflect solvent conditions and temperature; such an identification would permit a simple description of denaturation and renaturation.

This simple model for interactions in globular proteins yields results which are good for 65% of all residues and for 83% of those residues in regular α helix and β strand regions. The present model can easily be extended to include additional conformations. Our original intention was to study the folding process by comparing results for different selection processes. It is surprising that all of the methods considered here appear to yield nearly identical results. However, inspection of the a priori probabilities reveals numerous residues with very large probabilities for

$\mathbf{R}_j$ =

	α					β					α_L			β_L		
α	0	y_2^{α}	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	0	0	y_3^{α}	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	0	0	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	0	0	0	y_{25}^{α}	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	0	0	0	0	y_{26}^{α}	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	0	0	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
β	y_1^{α}	0	0	0	0	0	y_2^{β}	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	0	0	y_3^{β}	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	0	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	0	0	0	y_{12}^{β}	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	0	0	0	0	y_{13}^{β}	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	0	0	0	0	0	$y_1^{\alpha_L}$	0	0	$y_1^{\beta_L}$	0	0
α_L	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	0	$y_2^{\alpha_L}$	0	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	0	0	$y_3^{\alpha_L}$	$y_1^{\beta_L}$	0	0
	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	0	0	0	$y_1^{\beta_L}$	0	0
β_L	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	0	$y_2^{\beta_L}$	0
	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	0	0	$y_3^{\beta_L}$
	y_1^{α}	0	0	0	0	y_1^{β}	0	0	0	0	$y_1^{\alpha_L}$	0	0	0	0	0

(A1)

a single conformation. The implication is that, even early in the folding process, many residues assume highly stable conformations. Any sensible conformational selection scheme will usually choose the same conformation for these residues. The extent of agreement of the present results with X-ray crystallographic results implies further that the subsequently imposed specific long-range interactions are not strong enough to change the conformational preferences of about $^2/_3$ of the residues.

The present methods and other more diverse methods of determining secondary conformations in proteins[2] meet with similar degrees of success. Here we have achieved some success by considering in detail only intraregion electrostatic interactions and the structure breaking features of glycine and proline. This leads to the postulate that globular proteins are most likely maintained in their native forms by an abundance of *redundant stabilizing* interactions, including electrostatic interactions.

Appendix. A Priori Conformational Probabilities for Independent Regular Secondary Regions

We have calculated approximate free energies of many regular regions within a protein. It is interesting to consider the equilibrium mixture of all possible combinations of these independent regular regions. Here we formulate a partition function in matrix form and use it to calculate the equilibrium probability for each residue in each of the standard conformations. The set of conformations we have considered in this paper is:

ζ_1	ζ_2	ζ_3	$\cdots$ ζ_{n-1}	ζ_n
$\zeta_{1,2}$	$\zeta_{2,3}$	$\zeta_{3,4}$	$\cdots$ $\zeta_{n-1,n}$	
.				
.				
.				
ζ_{1,m_ζ}	$\zeta_{2,m_\zeta+1}$	$\zeta_{3,m_\zeta+2}$	$\cdots$ $\zeta_{n-m_\zeta+1,n}$	

where ζ indicates the conformation which is taken in these calculations to be α, β, α_L, or β_L; the subscripts on ζ are indices to indicate the terminal residues of the region; and m_ζ is the maximum number of residues to be included within a single region of the conformation designated by ζ. The condition is imposed that neighboring independent regions cannot have the same conformation.

The free energies for the above set of conformations are converted into statistical weights, corresponding to the increments of energy for each additional single residue. The above group of conformations is described with the set of statistical weights:

w_1^{ζ}	w_2^{ζ}	w_3^{ζ}	$\cdots$ w_{n-1}^{ζ}	w_n^{ζ}
$w_{1,2}^{\zeta}$	$w_{2,3}^{\zeta}$	$w_{3,4}^{\zeta}$	$\cdots$ $w_{n-1,n}^{\zeta}$	
.				
.				
w_{1,m_ζ}^{ζ}	$w_{2,m_\zeta+1}^{\zeta}$	$w_{3,m_\zeta+2}^{\zeta}$	$\cdots$ $w_{n-m_\zeta+1,n}^{\zeta}$	

where $w_{i,j}^{\zeta}$ is the statistical weight formed as follows:

$$w_{i,j}^{\zeta} = \exp[-(F_{i,j}^{\zeta} - F_{i,j-1}^{\zeta})/RT]$$

A partition function is generated directly from these sets of statistical weights as a serial product

$$Z = \mathbf{R}_1 \prod_{j=2}^{n} \mathbf{R}_j \mathbf{J}$$

where $\mathbf{R}_j$ is a matrix composed of the above statistical weights. For the case at hand with four states α, β, α_L, and β_L and with maximum lengths m_ζ of 26, 13, 3, and 3, respectively, $\mathbf{R}_j$ is formed as indicated in eq A1. The nonzero elements of $\mathbf{R}_j$ are defined as $y_{j-i+1}^{\zeta} = w_{i,j}^{\zeta}$, with the right end of the region for y being specified by the index j of the matrix $\mathbf{R}_j$.

$\mathbf{R}_1$ is the row vector

$$\mathbf{R}_1 = (y_1^{\alpha}\ 0\ 0 \ldots 0\ 0\ y_1^{\beta}\ 0\ 0 \ldots 0\ 0\ y_1^{\alpha_L}\ 0\ 0\ y_1^{\beta_L}\ 0\ 0)$$

and $\mathbf{J}$ is a conforming column of ones. The a priori probability that residue i is in conformation ζ is[1,24]

$$p_i^{\zeta} = Z^{-1} \mathbf{R}_1 \prod_{k=2}^{i-1} \mathbf{R}_k \mathbf{R}_{\zeta,i}' \prod_{j=i+1}^{n} \mathbf{R}_j \mathbf{J}$$

The matrix $\mathbf{R}_{\zeta,i}'$ is derived from $\mathbf{R}_i$ by converting all elements to zero which are not located in columns indexed

Table II
Parameter Values and Results for Energy Minimization with $F_H^\alpha = -0.67$, $\epsilon = 3.5$, $RT = 0.6$ kcal mol^{-1} $F_{xpro}^\alpha = 10.0$, $F_{gly}^\alpha = 0.5$, and $F^{\alpha_L} = 2.0$

molecule	range of F^α	% residues correct for all res	% residues correct for res in reg regions
CON A	3.3 to 3.6	66	88
LYSO	−0.05 to 0	57	79
MYOG	−2.2 to −1.6	85	94
PAP	0.48 to 0.53	52	60
S PEP	≤0.8	44	54
S PROT	1.0 to 1.3	71	85
PTI	1.3 to 1.4	75	94
average for all above proteins		65	83

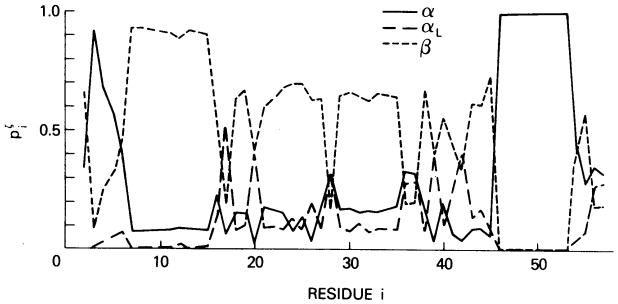

Figure 8. Probabilities of the three conformations α, β, and α_L for each residue of bovine pancreatic trypsin inhibitor. These were calculated with the equation in the Appendix and for the same parameters as for Figure 7. The most favorable conformation for each residue from this figure is displayed as the top line in Figure 7.

by ζ. Using this partition function, we can also apply methods in ref 1 and 24 to calculate mean square distances between all atom pairs and make a direct comparison with the reported crystal distances.

Results of the application of these equations to the free energies of trypsin inhibitor are shown in Figure 8 for the three conformations α, β, α_L. These probabilities are more informative about the relative probabilities of the unfavored conformations than are the results in Figure 7. For example, the low free energy helix centered about residue 50 has a helical probability which is essentially unity; there are no significant competing conformations. By contrast, residue 20 shows β to be only slightly more favorable than α_L. This is also subtly reflected in Figure 7 where this method and the "conformational stability" selection method indicated β for this residue but energy minimization gave α_L. In Figure 8, the two stable helices near the chain termini are clearly indicated together with the central portion that is dominantly β strand. Because of the neglect of β strand hydrogen bonds, it is typical to observe β probability maxima which are smaller than α helix ones. The probabilities for β_L are not given here; they are, with few exceptions, smaller than those for α_L.

References and Notes

(1) P. J. Flory, "Statistical Mechanics of Chain Molecules", Interscience, New York, 1969.
(2) T. T. Wu, S. C. Szu, R. L. Jernigan, H. Bilofsky, and E. A. Kabat, *Biopolymers*, **17**, 555 (1978).
(3) F. R. Maxfield and H. A. Scheraga, *Biochemistry*, **15**, 5138 (1976).
(4) V. I. Lim, *J. Mol. Biol.*, **88**, 857, 873 (1974).
(5) O. B. Ptitsyn and A. V. Finkelstein, *Biophysics*, **15**, 785 (1970).
(6) M. F. Perutz, *Science*, **201**, 1187 (1978).
(7) All rotational angles in this paper are expressed in the older convention with 0° taken to correspond to the trans position: J. T. Edsall, P. J. Flory, J. C. Kendrew, A. M. Liquori, C. Nemethy, G. N. Ramachandran, and H. A. Scheraga, *Biopolymers*, **4**, 130 (1966).
(8) J. L. Crawford, W. N. Lipscomb, and C. G. Schellman, *Proc. Natl. Acad. Sci. U.S.A.*, **70**, 538 (1973).
(9) M. Oobatake and T. Ooi, *J. Theor. Biol.*, **67**, 567 (1977).
(10) P. R. Schimmel and P. J. Flory, *J. Mol. Biol.*, **34**, 105 (1968).
(11) D. A. Brant, W. G. Miller, and P. J. Flory, *J. Mol. Biol.*, **23**, 47 (1967).
(12) R. J. Feldmann, "Atlas of Macromolecular Structure on Microfiche", Tracor Jitco Inc., Rockville, MD 1977. Molecules used, here and elsewhere in this paper, together with their reference numbers are: (a) sperm whale metmyoglobin, AM 1.4.1.1.1, (b) subtilisin BPN′ AM 3.1.1.1.1, (c) bovine carboxypeptidase A, AM 3.5.1.1.1., (d) bovine ribonuclease S–UpcA complex, AM 5.1.1.1.1., (e) jack bean concanavalin A, AM 9.1.1.1.2., (f) chicken lysozyme, AM 10.1.1.1.1., (g) papain, AM 3.2.1.1.1, and (h) bovine pancreatic trypsin inhibitor, AM 3.6.1.1.1.
(13) E. Ralston and J. L. DeCoen, *J. Mol. Biol.*, **83**, 393 (1974).
(14) E. J. Cohn and J. T. Edsall, "Proteins, Amino Acids and Peptides", Reinhold, New York, 1943, p 85.
(15) A. L. McClellan, "Tables of Experimental Dipole Moments", W. H. Freeman and Co., San Francisco, 1963. The dipole moments utilized were those for ethanethiol (1.48 D), ethanol (1.7 D), and indole (2.1 D).
(16) F. A. Momany, L. M. Carruthers, R. F. McGuire, and H. A. Scheraga, *J. Phys. Chem.*, **78**, 1595 (1974).
(17) E. Clementi, F. Cavallone, and R. Scordamaglia, *J. Am. Chem. Soc.*, **99**, 5531 (1977).
(18) D. E. Blagdon and M. Goodman, *Biopolymers*, **14**, 241 (1975).
(19) R. L. Jernigan and S. C. Szu, *Macromolecules*, **12**, 1156 (1979). Preliminary results given in this earlier paper were calculated with the present electrostatic energies and parameters in the form of eq 4 with $F_H^\alpha = 0$, $\epsilon = 1$, $F^\alpha = 3.8$ with only side chain–backbone electrostatic interactions.
(20) S. Miyazawa and R. L. Jernigan, to be published.
(21) F. R. Maxfield and H. A. Scheraga, *Macromolecules*, **8**, 491 (1975).
(22) J. E. Brown and W. A. Klee, *Biochemistry*, **8**, 2876 (1969).
(23) J. Skolnick and M. Fixman, *Macromolecules*, **11**, 867 (1978).
(24) R. L. Jernigan and P. J. Flory, *J. Chem. Phys.*, **50**, 4165 (1969).