# Equilibrium Folding and Unfolding Pathways for a Model Protein

S. MIYAZAWA and R. L. JERNIGAN, *Laboratory of Mathematical Biology, DCBD, NCI, National Institutes of Health, Bethesda, Maryland 20205*

## Synopsis

The folding–unfolding process of reduced bovine pancreatic trypsin inhibitor was investigated with an idealized model employing approximate free energies. The protein is regarded to consist of only $C^{\alpha}$ and $C^{\beta}$ atoms. The backbone dihedral angles are the only conformational variables and are permitted to take discrete values at every 10°. Intraresidue energies consist of two terms: an empirical part taken from the observed frequency distributions of $(\phi,\psi)$ and an additional favorable energy assigned to the native conformation of each residue. Interresidue interactions are simplified by assuming that there is an attractive energy operative only between residue pairs in close contact in the native structure. A total of 230,000 molecular conformations, with no atomic overlaps, ranging from the native state to the denatured state, are randomly generated by changing the sampling bias. Each conformation is classified according to its conformational energy, $F$; a conformational entropy, $S(F)$ is estimated for each value of $F$ from the number of samples. The dependence of $S(F)$ on energy reveals that the folding–unfolding transition for this idealized model is an "all-or-none" type; this is attributable to the specific long-range interactions. Interresidue contact probabilities, averaged over samples representing various stages of folding, serve to characterize folding intermediates. Most probable equilibrium pathways for the folding–unfolding transition are constructed by connecting conformationally similar intermediates. The specific details obtained for bovine pancreatic trypsin inhibitor are as follows: (1) Folding begins with the appearance of nativelike medium-range contacts at a $\beta$-turn and at the $\alpha$-helix. (2) These grow to include the native pair of interacting $\beta$-strands. This state includes intact regular secondary conformations, as well as the interstrand sheet contacts, and corresponds to an activated state with the highest free energy on the pathway. (3) Additional native long-range contacts are completely formed either toward the amino terminus or toward the carboxyl terminus. (4) In a final step, the missing contacts appear. Although these folding pathways for this model are not consistent with experimental reports, it does indicate multiple folding pathways. The method is general and can be applied to any set of calculated conformational energies and furthermore permits investigation of gross folding features.

## INTRODUCTION

Since Anfinsen et al.[1] succeeded in experimentally refolding ribonuclease, the theoretical prediction of a protein's native structure from its amino acid sequence has provided a fundamental challenge. The most direct approach of globally optimizing the free energy is rendered arduous, if not impossible, by the existence of numerous local minima on the free-energy surface. Simplifications both to the geometry of proteins[2,3] and to the energy functions can be incorporated in an attempt to shorten these search times.

Biopolymers, Vol. 21, 1333–1363 (1982)

CCC 0006-3525/82/071333-31$04.10

If, contrary to expectation, all conformations were equally probable, it would be impossible for the protein to reach its native structure through a complete random search of all possible conformations.[4–6] One possible reason why folding occurs within moderate periods of time is that there are a limited number of preferred folding pathways. Recent attempts at obtaining protein native structures have been based on either searches of all conformations or on assumed folding pathways. It should be pointed out here that a simplification of conformational energy functions and an assumption of a folding pathway are intimately related: simplified conformational energy functions, together with a proper evaluation of entropies, would usually yield limited numbers of highly preferred conformers, which might correspond to intermediates on specific folding pathways; conversely, an assumed folding pathway can be explained in terms of a limited number of conformers with favorable free energies. In the study of protein folding, it may be useful to elaborate on the relationship between the potential functions utilized and the resulting folding processes. This paper represents such a study strictly from an equilibrium point of view.

Gō and coworkers investigated the roles in the folding process of various classes of interactions, hydrophobic, short- and long-range, both specific and nonspecific but for two-dimensional lattice models of proteins[7–11] and a highly simplified three-dimensional lattice model.[12,13] They simulated the kinetic folding processes of the model proteins. Kinetic simulations may be a desirable method in the study of the folding process of proteins, but as their result for the three-dimensional lattice model indicates, it is not easy to obtain refolding steps, even when the protein's geometry and potential energy are very simple. Rather than pursue a kinetic simulation, we will employ a simpler equilibrium method.

Bovine pancreatic trypsin inhibitor (PTI) is advantageous because it is small and yet its secondary structures are representative, since they include both $\beta$-strands and an $\alpha$-helix. Its principal disadvantage is the presence of three disulfide bonds in the native protein; the present calculations correspond only to the case of completely reduced S-S pairs. In the present treatment, the protein molecule is regarded as a chain with volume exclusion consisting only of $C^\alpha$ and $C^\beta$ atoms, and only $C^\alpha$ atoms for glycine residues. The only conformational variables are the pairs of backbone dihedral angles flanking each $C^\alpha$; here they are permitted to take discrete values only at increments of 10°. Intraresidue interactions are estimated from the empirical frequency distributions of $(\phi,\psi)$ observed in the crystal structures of 20 proteins, as tabulated by Nemethy and Scheraga.[14] Also, an additional favorable energy is assigned to the native conformation of each residue as a part of the intraresidue interaction. Interresidue interactions are greatly simplified by including attractive energies only for contacting pairs of residues identical to those found in the crystal structure; these include cysteine pairs, which are accorded no further special treatment. The treatment of interresidue interactions in our model is similar to that of Ueda et al.[12]; however, the present approximations for the intraresidue interactions and the molecular geometry are more detailed.

Large numbers of molecular conformations are generated, and the entropies for different ranges of conformational energies are evaluated. A Monte Carlo method has been devised to generate various conformations ranging from the native state to the denatured state in a scheme that reduces sample attrition problems caused by excluded volume effects. In this Monte Carlo method, the native crystal structure is assumed to be at the global minimum in the conformational energy. The folding process of this model protein is examined in terms of the statistical averages of quantities that characterize the extent of folding for different ranges of conformational energies. Specifically, the probability of forming each contact pair is calculated. By taking the energy as a folding coordinate and comparing the conformational characteristics at different points, it is possible to construct most probable pathways for the folding–unfolding transition. This method provides a simple treatment of most probable folding pathways. It is a general method and offers an equilibrium description of protein folding.

## IDEALIZED MODEL OF PROTEIN

### Geometry

The protein molecule is treated as a chain consisting only of hard sphere $C^{\alpha}$ and $C^{\beta}$ atoms, except for glycine. The backbone dihedral angles $(\phi,\psi)$ are taken as the only conformation variables and are permitted to take discrete values only at every 10°, i.e., at 0°, 10°, . . . , 350°. We have employed this approximation for dihedral angles in order to be able to reproduce most of the close contacts in the native $C^{\beta}$-$C^{\beta}$ distance map.[15] These rather fine divisions on the $(\phi,\psi)$ map are needed to give the protein molecule sufficient flexibility to achieve contacts required in the native form. Such fine divisions in $\phi$ and $\psi$ should be sufficient to permit an estimation of conformational entropy. All other geometric quantities are fixed at their crystal values; these include $C^{\alpha}$-$C^{\beta}$ distances, the relative orientation of $C^{\beta}$ atoms with respect to the N-$C^{\alpha}$-C′ plane, and the peptide-bond conformations. We have avoided the use of a standard, uniform backbone geometry for bond angles and lengths, because it would cause some overlaps between pairs of $C^{\alpha}$ and $C^{\beta}$ atoms for the native dihedral angles and also would not permit as good a reproduction of the native $C^{\beta}$-$C^{\beta}$ distance map; these effects have previously been reported.[16] The atomic coordinates of PTI[17] are taken from the Brookhaven protein data bank.[18]

From the crystal structure, the number of $C^{\beta}$-$C^{\beta}$ pairs whose distances are less than or equal to 6.5 Å is 99, excluding nearest-neighbors; see the number of black squares in Fig. 1. $\alpha$-Carbon atoms alone represent glycines. The distance map representation using $C^{\beta}$ atoms, with this distance, displays the usual tertiary conformational features found using other atoms. For the assignment of dihedral angles to the single nearest point on the 10°

dihedral angle grid, 80 of the 99 $C^\beta$-$C^\beta$ pairs can be reproduced. But, as we shall see further, a slight relaxation of the restriction to the nearest point will add significantly to the number of nativelike contacts achieved. For the present case, the minimum distance between all pairs of $C^\alpha$ and $C^\beta$ atoms is 2.47 Å for the $C^\beta$-$C^\beta$ pair between arginine-20 and tyrosine-35 residues. On the basis of this value, we have chosen 1.2 Å as the van der Waals radii of $C^\alpha$ and $C^\beta$ atoms; that is, distances between centers of all $C^\alpha$ and $C^\beta$ atoms will not be permitted to approach one another more closely than 2.4 Å. This value, although smaller than usual values for van der Waals radii of $C^\alpha$ and side chains, depends to some extent on the fineness of the dihedral angle grid; this particular value is suitable only for 10° divisions of $\phi$ and $\psi$ for this molecule. Also, choice of a small radius may serve to partially compensate for effects of neglected solvent interaction energies.

## Intra- and Interresidue Interaction Energies

For convenience, we separate interaction energies into two categories, those for intraresidue interactions and for interresidue interactions. Intraresidue energies could have been calculated with semiempirical energy functions and utilized directly in the present calculations; however, explicit calculation of interresidue interaction energies is not possible with the present simple model because each residue has been replaced with $C^\alpha$ and $C^\beta$ atoms. We choose instead to use completely empirical conformational probabilities. The empirical potential energies for three classes of amino acids—glycine, proline, and all other amino acids—are estimated from their frequency distributions in $(\phi,\psi)$ compiled from 20 protein crystal structures by Nemethy and Scheraga.[14] The empirical energy $F^{emp}(\phi_\mu,\psi_\nu)$ for the conformation $(\phi_\mu,\psi_\nu)$ is obtained from their results as follows:

$$F^{emp}(\phi_\mu,\psi_\nu) = -0.6 \ln[q(\phi_\mu,\psi_\nu)] + \text{const} \quad \text{(kcal/mol)} \qquad (1)$$

where $q(\phi_\mu,\psi_\nu)$ is the probability with which $(\phi,\psi)$ of a residue is observed within a 10° square centered at $(\phi_\mu,\psi_\nu)$. Thermal energy, $RT$, is assumed to be 0.6 kcal/mol in converting probability to energy. The constant in Eq. (1) is chosen so that the minimum value of $F^{emp}$ is zero. These probabilities, for the group of amino acids excluding glycine and proline, are calculated from the results reported by Nemethy and Scheraga[14] by subtracting the numbers of occurrences of glycine and proline from those for all amino acids. In their figures, the interval of dihedral angles is 10° centered at 5°, 15°, . . . ; here, however, each $\phi_\mu$ and $\psi_\nu$ corresponds to values at 0°, 10°, . . . . The probability $q(\phi_\mu,\psi_\nu)$ in Eq. (1) has been smoothed by averaging over the four nearest neighbors to $(\phi_\mu,\psi_\nu)$. A small arbitrary number has been added to the number of occurrences for all points on the $(\phi,\psi)$ grid, namely, 0.001 for proline and 0.01 for glycine and the others; this assures that all conformations possess a nonzero probability of occurrence. Use of these values yields maximum values of $F^{emp}$ of 3.6, 4.9, and 5.4 kcal/mol

for glycine, proline, and the others, respectively. There are notable differences between these empirical energy surfaces and ones calculated semiempirically.[2] A bridge region connecting the regions of $\alpha$- and $\beta$-conformations has lower energy in the former than in the latter; the shape of the $\alpha$-helix region is somewhat different. These differences may arise either from inadequacies in the sample or because the empirical energies implicitly include averages of longer-range interactions.

An additional intraresidue energy is introduced in order to include some effects of various residue types and their nearest-neighbor interactions. It is assumed to be a square-well potential energy with a favorable value only for $(\phi_\mu,\psi_\nu)$'s within 10° of the crystal dihedral angles $(\phi_n,\psi_n)$. Thus, four points on the dihedral angle grid are favored. The energy favoring the native conformation is designated here as the short-range energy:

$$F_i^{\text{short}}(\phi_\mu,\psi_\nu) = \begin{cases} -1 \text{ (kcal/mol)}, & \text{if } |\phi_\mu - \phi_n^i| < 10^\circ \quad \text{and} \quad |\psi_\nu - \psi_n^i| < 10^\circ \\ 0, & \text{for others} \end{cases} \tag{2}$$

where $i$ is an index of residue number. This value of $-1$ has been chosen arbitrarily. The total intraresidue interaction energy is simply the sum of $F_i^{\text{emp}}$ and $F_i^{\text{short}}$.

Interresidue interactions between residues further apart than nearest-neighbor residues along the primary structure are introduced in the strongest limit of specificity. If, and only if, the two residues that make contact with each other are a pair in contact in the crystal structure, then a negative energy, which is the same for all native contacting residue pairs, is assigned to the conformation. Close contact between residues is defined on the basis of $C^\beta$-$C^\beta$ distances rather than $C^\alpha$-$C^\alpha$ distances, because the former is expected to be more sensitive to conformation. If the distance between a pair of $C^\beta$ atoms is less than or equal to 6.5 Å, then this pair of residues is taken to be in close contact. The contact map for the crystal structure is shown in Fig. 1. Of the hydrogen bonds identified by Deisenhofer and Steigemann,[17] most also correspond to close contact between $C^\beta$ atoms; among these, all residues with side chain–side chain hydrogen bonds are in close contact, most residues with backbone–side chain hydrogen bonds and somewhat fewer with backbone–backbone hydrogen bonds. Whenever hydrogen-bonded residue pairs are not found in close contact, usually some of the neighboring residues will be observed in close proximity.

The interresidue contact energy $F_{ij}^{\text{cont}}$ between the $i$th and $j$th residues is given as

$$F_{ij}^{\text{cont}} = \begin{cases} -2 \text{ (kcal/mol)}, & \text{if } 2.4 \text{ Å} \leqq r_{ij} \leqq 6.5 \text{ Å, for } |i-j| > 1 \text{ in both the crystal structure and the given conformation} \\ \infty, & \text{for } r_{ij} < 2.4 \text{ Å} \\ 0, & \text{for } r_{ij} > 6.5 \text{ Å} \end{cases} \tag{3}$$

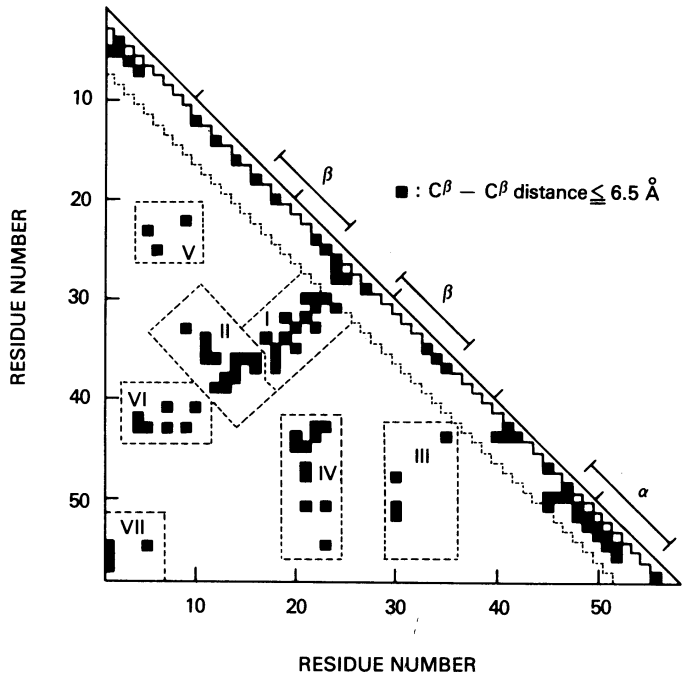


Fig. 1. $C^\beta$-$C^\beta$ contact map of the crystalline structure of PTI taken from data in Ref. 18. Ordinate and abscissa designate residue numbers. Solid squares indicate that the $C^\beta$-$C^\beta$ atomic distances are less than or equal to 6.5 Å. For glycine residues, $C^\alpha$ atoms are employed. Nearest-neighbor contacts are omitted in this figure. Close residue pairs are termed contacts and have been grouped into the seven regions, I–VII, enclosed by dotted lines in this figure. Medium-range contacts are also separated by the dotted line running parallel to the diagonal. Bars above the diagonal indicate the locations of the regular $\alpha$-helix and $\beta$-strands.

Such a contact energy serves as a simplest representation of all favorable interactions, including hydrogen bonds, electrostatic interactions, and hydrophobic interactions. This value of $-2$ has been chosen so that this model protein melts within a reasonable temperature range; this choice causes the model PTI to melt at $RT = 0.67$.

The total approximate energy of a conformation is taken to be the sum of three contributions:

$$F(\{\phi,\psi\}) = \sum_i (F_i^{\text{emp}} + F_i^{\text{short}}) + \sum_{j>i+1} \sum F_{ij}^{\text{cont}} \tag{4}$$

The use of this artificial energy function involves the assumption that the lowest energy forms closely resemble the crystal structure.

## CONFORMATIONAL ENTROPY, $S(F)$

In a statistical-mechanical study of folding–unfolding transitions in proteins, Gō[11,19] established the usefulness of the $S(H,T)$ function, the conformational entropy of a molecule in solution for a given value of enthalpy $H$ at temperature $T$. Here, we use a somewhat different formulation of the partition function.

A configuration partition function[20–22] is usually expressed as an integral over all configuration coordinates $Q$, which will be taken here to be the dihedral angles $\phi$ and $\psi$ for each residue:

$$Z = \int \exp[-F(Q,T)/RT]\, dQ \tag{5}$$

where $R$ is the gas constant and $T$ is temperature. $F(Q,T)$ is a conformational energy that represents the Gibbs free energy of a single conformation specified by the set of conformational coordinates $Q$. This $F(Q,T)$ is taken to include the effects of a potential of mean force arising from solvent molecules.[20,21] This integration is broken into two steps, first over coordinates $Q$ for a fixed value of $F$ and then over all values of $F$ by means of a Dirac delta function of $F$.

$$Z = \iint \exp[-F(Q,T)/RT]\, \delta[(F(Q,T) - F)/RT]\, dQd(F/RT) \tag{6}$$

By defining entropy as

$$\exp[S(F,T)/R] = \int \delta[(F(Q,T) - F)/RT]\, dQ \tag{7}$$

it becomes possible to express the partition function as

$$Z = \int \exp[-(F - TS(F,T))/RT]\, d(F/RT) \tag{8}$$

$S(F,T)$ is the conformational entropy of protein molecules for conformational states with a given value of conformational energy $F$ at temperature $T$. Equations (7) and (8) are similar to expressions for microcanonical ensemble and canonical ensemble partition functions, respectively. This last partition function can be transformed into the following form useful for numerical calculations[7]:

$$S(F,T)/R = \ln[\mathrm{P}(F,T)] + F/RT + \text{const} \tag{9}$$

where $P(F,T)$ is the probability of conformations with energy $F$ at temperature $T$.

Here, it should be noted that the integration in the first stage of Eq. (6) is performed at a fixed value of free energy, rather than enthalpy as in Gō's formalism.[11,19] Although there is a slight difference in the formalisms, Gō's interpretation of the characteristics of an $S(H,T)$ function can be directly applied without modification to the present $S(F,T)$ function. The $S(F,T)$ function directly describes the folding and unfolding transition of a protein. From Eq. (8), the most populous state would have the minimum value of $\mathcal{F} = F - TS(F,T)$. It follows from Gō's considerations that the curve of conformational entropy $S(F,T)$ versus conformational energy $F$, at the melting temperature, must have at least a concave part if it is to reproduce the observed "all-or-none" character of the protein-folding transition. The present formalism emphasizes that a proper description of the folding–unfolding equilibrium in proteins requires specification of both the conformational energy $F$ and the conformational entropy $S(F,T)$.

The native state will have low conformational energy and low conformational entropy, whereas the denatured state is distinguished by both

high conformational energy and high conformational entropy. It should be noted that we are assuming that the free energy $F(Q,T)$ of an individual conformation consists of energy terms only; $F(Q,T)$ and consequently $S(Q,T)$ are assumed to be independent of temperature. This allows the function $\mathcal{F} = F - TS(F)$ to satisfy the customary relationship for free energies $\partial\mathcal{F}/\partial T = -S$.

## CONFORMATIONAL AVERAGES

The averaging of any conformational quantity can be separated into the same two integrations that were utilized in the formulation of the partition function. In the first step, the statistical average over conformational states, of any function $f(Q)$ of conformational coordinates, with a fixed value of conformational energy $F$ and temperature $T$, is

$$\langle f(F)\rangle = [\exp(S(F,T)/R)]^{-1} \int f(Q)\delta[(F(Q,T) - F)/RT]\, dQ \qquad (10)$$

In the second step, of averaging over $F$, it is our intention to perform the integration only over a limited range of energy, $x$, say from $F_a$ to $F_b$. The statistical averge of $f(Q)$ over this limited region of conformational energy is

$$\langle f\rangle_x = \frac{\int_x \langle f(F)\rangle \exp[-(F - TS(F,T))/RT]\, d(F/RT)}{\int_x \exp[-(F - TS(F,T))/RT]\, d(F/RT)} \qquad (11)$$

This limited averaging will be used to provide statistical conformational averages at intermediate stages in the folding process.

## MONTE CARLO METHOD FOR EVALUATING CONFORMATIONAL ENTROPY AND AVERAGES

Quantities related to macromolecular conformations are complicated functions of conformational variables such as dihedral angles. In these cases, Monte Carlo methods provide powerful tools for evaluating various average conformational properties of large macromolecules. These methods, in which conformations of a macromolecule are generated randomly, have often been used in polymer studies, for example, to estimate the effect of volume exclusion on the mean-square radius of gyration. Here, a Monte Carlo method is used to estimate the entropy $S(F)$ in Eq. (7) and then the averages in Eqs. (10) and (11), with excluded volume. There is a specific common problem with the efficiency of Monte Carlo methods for excluded-volume applications. The number of successful conformations generated with proper account of excluded volume diminishes exponentially with chain length.[23] Short chains are easy to generate, but it is exceedingly difficult to generate sufficient samples of long molecules. A special sampling method can be employed to reduce this inefficiency of

generating conformations. A small group of conformations are explicitly considered for each added residue; sampling is then performed only for those with permissible excluded volume, rather than over the entire conformational space of each residue. A similar approach has been used elsewhere.[24,25]

The continuous conformational space $(\phi,\psi)$ for each pair of backbone dihedral angles is approximated here with points on discrete 10° grids. Because the number of points on a single residue's grid, $36 \times 36$, is quite large, it would require substantial computation time to check at each step to see whether or not each grid point is permitted on the basis of atomic overlaps. Therefore, we employ a two-stage sampling method. The total set of $36 \times 36$ points is divided into independent subsets in such a manner that all intersections between the subsets are empty. In the first stage, a particular subset is selected on the basis of a random number; each point in that subset is checked for its volume exclusion. If there is at least one allowed $(\phi_\mu,\psi_\nu)$, then a second sampling is performed to choose a single $(\phi_\xi,\psi_\eta)$ from among the permissible $(\phi_\mu,\psi_\nu)$'s in the subset. Otherwise, the molecular generation ends unsuccessfully, and conformation generation starts again from the first residue. The actual computation time required to generate a given number of successful molecular conformations depends on the way in which the total number of points is distributed into subsets.

Here, the total set of $36 \times 36$ points is divided into $12 \times 12$ subsets, each consisting of $3 \times 3$ points from the 10° grids. First, $(\phi,\psi)$ space is divided into 30° grids. A separate subset is constructed for each of these 30° grid points to include it and eight other points formed by adding either 0°, 70° or −70° to each of the $\phi$ and $\psi$ angles. Preliminary tests indicate that this placement of nonadjacent lattice points in subsets is relatively efficient for computations on a small protein. If desired, a third stage could be added in order to sample conformations between grid points.

Another major problem in sampling protein conformations is the necessity of providing a representative sample of conformational space that includes all important conformations. Such a sample is required in order to construct a representative partition function for calculation of average properties. The native structures of proteins are relatively unique conformations, in contrast to the thermodynamic equilibrium mixture of large numbers of random-coil conformations. The conformational energy of the native conformation must be at least low enough to overcome the conformational entropy loss accompanying the transition from the denatured state to the native state. The entire conformational space accessible to a protein is vast. These two facts conspire to prevent a good estimation of the partition function at temperatures below melting by uniform random sampling of all of conformational space; for reasonable numbers of samples, there is almost no chance of obtaining conformations with substantial similarity to the native form, which would correspond to a major term in the partition function. This problem has been widely recognized[4–6] in the

sense that it is not generally expected that the native structure of proteins can be realized by a complete random search of all possible conformations.

To avoid this pitfall, we have taken the expedient measure of imposing a strong bias toward the native conformation. To sample a full range of conformations, from native to denatured, we employ a random sampling of conformations with varying bias toward the native conformation. A method for biased generation of self-avoiding random walks was developed by Wall et al.[26] and subsequently used by others[27,28] in such a manner that important conformations of chain molecules are generated with greater than uniform frequency. It would be ideal to generate conformations with probabilities based on the Boltzmann factors of their energies, but, of course, that is impossible, since before a conformation is generated, its actual energy is unknown. For sample generation we have utilized Boltzmann factors for the intraresidue energies plus a bias factor.

The probability $p_i(\phi_\mu,\psi_\nu)$ used in the random sampling of the dihedral angles $(\phi_\mu,\psi_\nu)$ of the $i$th residue is

$$p_i(\phi_\mu,\psi_\nu) = C^{-1} \exp[-(F_i^{\text{emp}}(\phi_\mu,\psi_\nu) + F_i^{\text{short}}(\phi_\mu,\psi_\nu) + \epsilon_i(\phi_\mu,\psi_\nu))/RT] \quad (12)$$

where

$$\epsilon_i(\phi_\mu,\psi_\nu) = \epsilon\delta_{\phi_\mu\phi_0^i}\delta_{\psi_\mu\psi_0^i}$$

and $\phi_0^i$ and $\psi_0^i$ are the dihedral angles of the $i$th residue in the lowest energy conformation. $C$ in Eq. (12) is simply a normalization factor. $\epsilon_i$ is a parameter that may correspond to the mean contribution of unknown interresidue interactions; for practical purposes, it is used to change the extent of native character. Here, it has been taken to be identical for all residues, although perhaps it should vary along the chain. The $\delta$'s are Kronecker deltas. Various conformations, from the native conformation to the denatured state, can be generated by changing both $\epsilon$ and $RT$. Temperature in Eq. (12) is not to be construed to be an external physical variable, but rather as a parameter in the sample generation. We have chosen a minimum value for $\epsilon$ of about $-3$, which is approximately equal to the interresidue interaction energy per residue in the lowest energy conformation. For such biased sampling, each generated sample must be weighted with the inverse of the sampling weights to remove the bias in the sampling; for large samples, results then become equivalent to unbiased ones.

The weight $W^\omega$ to remove the sampling bias, for sample $\omega$, is given by

$$W^\omega = \left[\prod_i p_i^a(A)\, p_i^b(\phi_\zeta,\psi_\eta)\right]^{-1} \quad (13)$$

where the probability for each residue $i$ is composed of a product of two probabilities corresponding to the two-stage selection described above; $p_i^a(A)$ corresponds to the selection of one particular subset $A$ composed of nine conformations, and $p_i^b(\phi_\zeta,\psi_\eta)$ to the second choice of one specific conformation $(\phi_\zeta,\psi_\eta)$:

$$p_i^a(A) = \sum_{(\phi_\mu,\psi_\nu)\in A} p_i(\phi_\mu,\psi_\nu)$$

and

$$p_i^b(\phi_\zeta,\psi_\eta) = K^{-1}p_i(\phi_\zeta,\psi_\eta) \tag{14}$$

with

$$K = \sum p_i(\phi_\mu,\psi_\nu)$$

The last sum is over the grid points $(\phi_\mu,\psi_\nu)$, which are in the set $A$ and also do not overlap previous atoms. This renormalization is necessary because the samples with bad volume exclusion have been removed.

By using this weight $W^\omega$, the required $P(F,T)$ and $\langle f(F)\rangle$ as defined in Eqs. (9) and (10) are calculated as follows:

$$P(F,T) = \sum_\omega W^\omega \exp(-F^\omega/RT)\,\Delta[(F^\omega - F)/RT] \Big/ \sum_\omega W^\omega \exp(-F^\omega/RT) \tag{15}$$

where

$$\Delta[(F^\omega - F)/RT] = d^{-1}\int_{F-d/2}^{F+d/2} \delta[(F^\omega - F)/RT]\,dF$$

$$\langle f(F)\rangle = \sum_\omega f^\omega W^\omega \exp(-F^\omega/RT)\,\Delta[(F^\omega - F)/RT] \Big/ \sum_\omega W^\omega \exp(-F^\omega/RT)\,\Delta[(F^\omega - F)/RT] \tag{16}$$

The previous integrals have been replaced by sums over samples $\omega$. Here, the function for a continuous variable has been replaced with its average over a small interval $d$. By using Eqs. (15) and (9), $S(F,T)$ can be calculated, except for a constant.

## MOST PROBABLE FOLDING–UNFOLDING PATHWAYS

The net overall trend in folding toward the native conformation must be toward conformations of lower energy. Because kinetics are not considered here, intermediate reversals in energy are not permitted. Here, the conformational energy is proportional to both the number of correct long-range contacts and to the number of correct dihedral angle pairs; consequently, this conformational energy can be identified intuitively as an extent of folding. The conformational energy serves as a useful one-dimensional representation of folding coordinates.

Averages at points along the folding coordinate can yield details about most probable conformations, as well as provide the means of characterizing groups of conformations. We have chosen to characterize protein conformations in terms of the probabilities of close approaches of favored residues. These characteristics are determined for various regions of en-

ergy, corresponding to different stages of folding. These regions of energy include many individual conformations and sometimes will include groups that belong to distinguishable classes of conformations. Folding pathways can be constructed by connecting those classes of conformations with the greatest conformational similarities, in order of decreasing energy. These are equilibrium considerations; consequently, it is not possible to describe the kinetics or to ascribe a direction to the pathways.

## RESULTS

In preliminary samples, conformations near the crystal structure were generated by using Eq. (12) with $\epsilon = -3$ and $RT = 0.45$ kcal/mol. Among the conformations generated, the lowest energy conformation obtained reproduces 94 of the 99 $C^\beta$-$C^\beta$ contacts observed in the crystal structure. This reproducibility of about 95% of the native $C^\beta$-$C^\beta$ contacts confirms the 10° divisions of $(\phi,\psi)$ space as fine enough to permit formation of a nativelike conformation for PTI. Dihedral angles of all residues, except for glycine-57 and alanine-58, in this lowest energy conformation, are within 10° of their crystal structure values. They furthermore correspond to the single values on the 10° $(\phi,\psi)$ grid closest to the crystal form for all cases except for glutamine-31, arginine-42, asparagine-44, methionine-52, threonine-54, and cysteine-55. However, these few small deviations are responsible for achieving 14 of these contacts; only 80 contacts were obtained when all dihedral angles were fixed at the points nearest their crystal values. Glycine-57 and alanine-58 appear to be flexible in the present idealized model. These terminal residues' conformations are not critical for achieving other contacts; furthermore, their only native contacts are glycine-57 with arginine-1 and alanine-58 with glycine-56. The intraresidue interactions alone play a conformation-determining role for those two residues, but they yield a nonnative conformation. This lowest energy conformation with 94 of the native contact pairs has been used in the estimate of $\epsilon_i$ in Eq. (12).

Various conformations from the native state to the denatured state are randomly generated by changing the parameter $\epsilon$ from $-3$ to 0 and the thermal energy $RT$ from 0.45 to 0.95 in Eq. (12). The total number of molecular samples generated is 230,000. $P(F,T)$ and $\langle f(F)\rangle$ in Eqs. (15) and (16) are estimated for increments in $d$ of 2 kcal/mol in energy $F$. Units of energy are taken as kcal/mol throughout this paper.

### "All-or-None" Folding–Unfolding Transition for This Model Protein

The dependence of entropy, $S(F)$, on energy $F$ is shown in Fig. 2. The values of entropy are taken relative to a reference state of the lowest energy conformation with an entropy of zero; this serves to specify the constant in Eq. (9). Sufficient numbers of samples have been generated to yield a

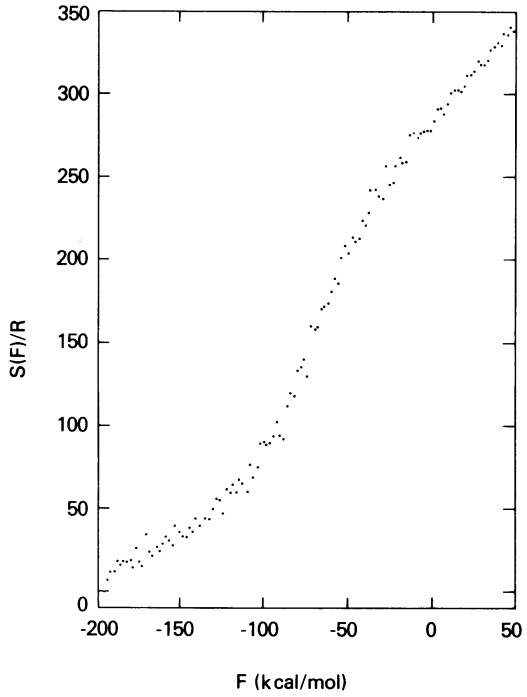


Fig. 2. Dependence of conformational entropy $S(F)/R$ on conformational energy $F$. The values of entropy represent accumulations of samples at 2 kcal/mol intervals in energy. A total of 230,000 molecular conformations have been generated; approximately 25,000 of these possess energies higher than 50 kcal/mol.

moderately smooth curve. It is not clear how much of the nonsmoothness in this figure is statistical in origin, i.e., from the random sampling, or how much it reflects genuine variations in entropy. Even so, this curve clearly consists of a concave part in the region of energy from about $-160$ to $-80$ and a convex part for energies above about $-80$. For very large energies, the $S(F)$ function is still an increasing function of $F$, even at an energy of 100. Gō has concluded that a concave shape in this function indicates that the folding–unfolding transition is an "all-or-none" type. We would like to understand the origin, in terms of ranges of interactions, of the concave portion of this curve.

The transition is quite sharp and yields a melting temperature near $RT = 0.67$. Figure 3(a) depicts probability distributions of conformations at various energies, in the vicinity of the melting temperature. This figure clearly demonstrates an "all-or-none" behavior in the folding–unfolding transition of the model PTI; there is no appreciable population of intermediate energy conformations through the transition region. It can be seen that the native and denatured sides of the transition correspond to energies near $-190$ and $-30$, respectively. The reduction of conformational entropy on folding is large, about 240 for $S/R$ or about 4 per residue. All further results depend strongly on the $S(F)$ function. Because there are some statistical fluctuations arising from the random sampling, results can be

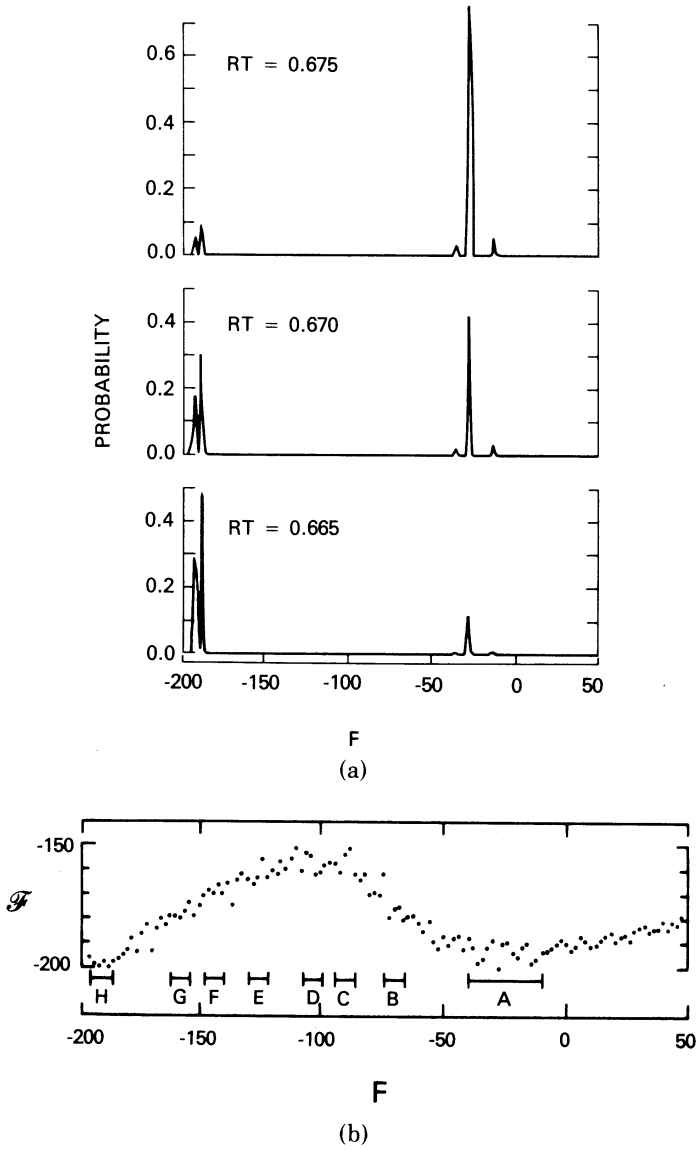


Fig. 3. (a) Probability distributions of conformational energy in the vicinity of the melting temperature at $RT$ = 0.67 kcal/mol. These probability distributions are calculated from the results in Fig. 2 and include statistical errors originating in the random sampling. This figure is most meaningful to demonstrate the nature of the transition as two-step or an "all-or-none" type. Throughout this paper the units of energy are kcal/mol. (b) The dependence of free energy, $\mathcal{F} = F - TS(F)$, at the melting temperature, on conformational energy $F$. Bars at the bottom, designated as A–H, are the regions of energy for which contact energies and short-range energies are averaged in Figs. 6 and 7.

interpreted only semiquantitatively. The free energy, $\mathcal{F} = F - TS(F)$, at the melting temperature is calculated and displayed in Fig. 3(b). The details of most probable conformations at different points along this curve will be described next.

## Conformational Characteristics of Partially Folded Intermediates and Construction of Pathways

First, the various components of the energy are examined. Figure 4 depicts the averages of the sums over all residues of empirical energies, short-range energies, and interresidue contact energies for the full range of the transition. The interresidue interactions are divided into two categories: medium-range interactions, defined as the interresidue interactions between a residue and its six nearest-neighbor residues on each side; and longer-range interactions. In the denatured energy region above −30, the short-range energy does not change significantly. The most significant changes of energy components in that range are in the empirical energy and the medium-range energy. The long-range energy is, on average, nearly constant in this region. In that range of energy, most residues can be seen, from the values of their short-range energies, not to be in their

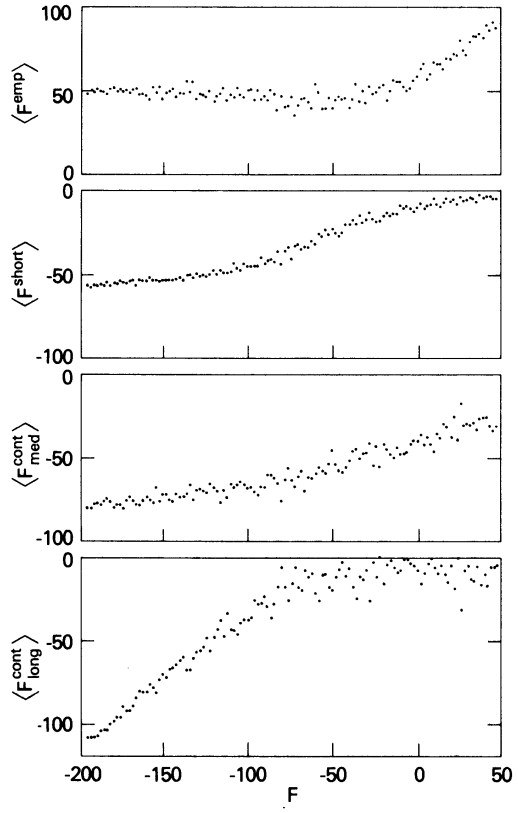


Fig. 4. Statistical averages of various energy components for values of the total energy indicated on the abscissa. These are averages for fixed energy $F$ as given in Eq. (16). These components comprise, from top to bottom, intraresidue empirical energy, short-range intraresidue energy, medium-range interresidue contact energy, and long-range interresidue contact energy. Medium-range contacts are taken to be those between residues less than seven residues apart along the primary sequence; those longer in range are termed long-range contacts.

native conformations. Below an energy value of −30, the short-range energy begins to decrease, which indicates that increasing numbers of residues are taking nativelike conformations. The empirical energy increases slightly from this point until the native state. This small increase indicates that the native state does not correspond to a minimum in the empirical potential energy alone. The significant decrease in the short-range energy continues until an energy value of about −80, where the short-range energy continues to decrease, but very slowly. The medium-range energy changes similarly; it gradually decreases until a total energy of −80, where its convergence slows. In the region of total energy below −80, the decreases in energy arise principally from the long-range energy component. It is noteworthy that this value of −80 corresponds closely to the inflection point in the $S$ vs $F$ curve (see Fig. 2). This indicates that the concave part of the $S$ vs $F$ curve, which is responsible for the "all-or-none" behavior of the folding–unfolding transition, arises mainly from long-range interactions and that the convex part originates in short-range and medium-range interactions. These results indicate that intraresidue and medium-range interresidue interactions dominate during early stages of folding and that long-range interresidue interactions are important only during the later stages of folding. Medium-range interactions appear to serve to stabilize secondary structures even in the denatured state. Such medium-range interactions include hydrogen-bond interactions within an $\alpha$-helix and in turns.

Long-range contacts are somewhat arbitrarily classified into seven regions, specified as I–VII in Fig. 1, which is a conventional contact map for all $C^{\beta}$-$C^{\beta}$ distances less than or equal to 6.5 Å. In Fig. 5, the average contact

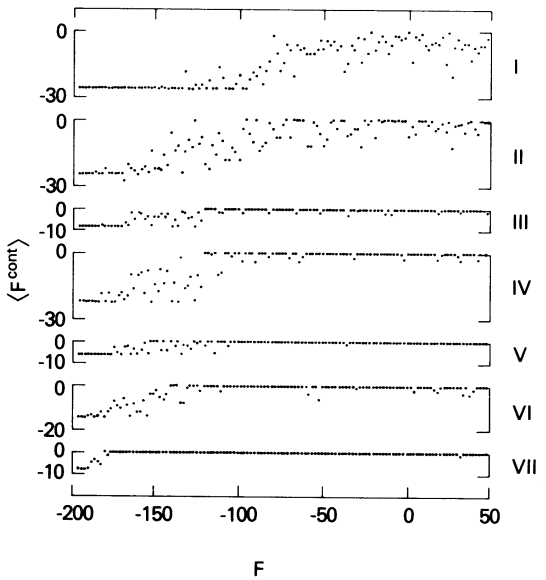


Fig. 5. The statistical averages for fixed energy $F$ of interresidue contact energies in each contact region of the crystal structure of PTI from I–VII as designated in Fig. 1.

energy for each region is depicted. The general order of formation of contact regions can usually be distinguished, but there is significant scatter that obscures details. A more detailed and informative representation of the process is given in Fig. 6(A–H). The contact interaction energy for each

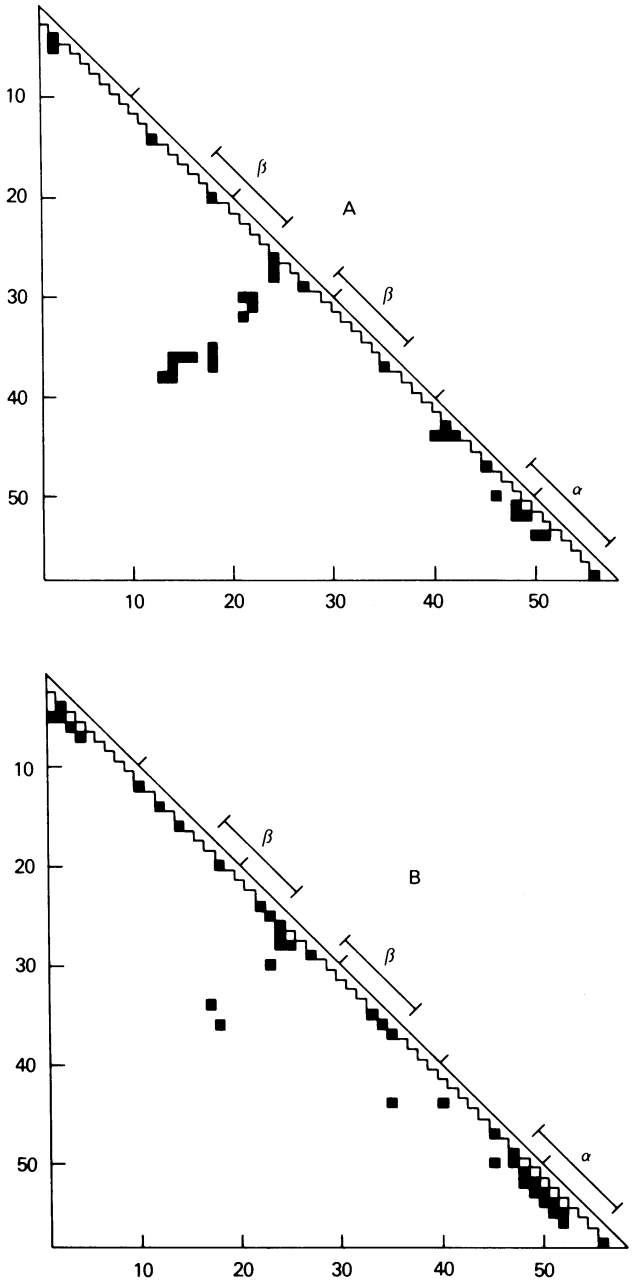


Fig. 6. Probabilities of contact formation for those residue pairs in close contact in the crystal structure. Figures labeled A–H represent averages over groups of randomly generated conformations corresponding to the energy regions A–H demarcated by the bars at the bottom of Fig. 3(b). Solid squares represent a probability of contact formation greater than or equal to ¾; open squares are for probabilities less than ¾ but greater than or equal to ¼.

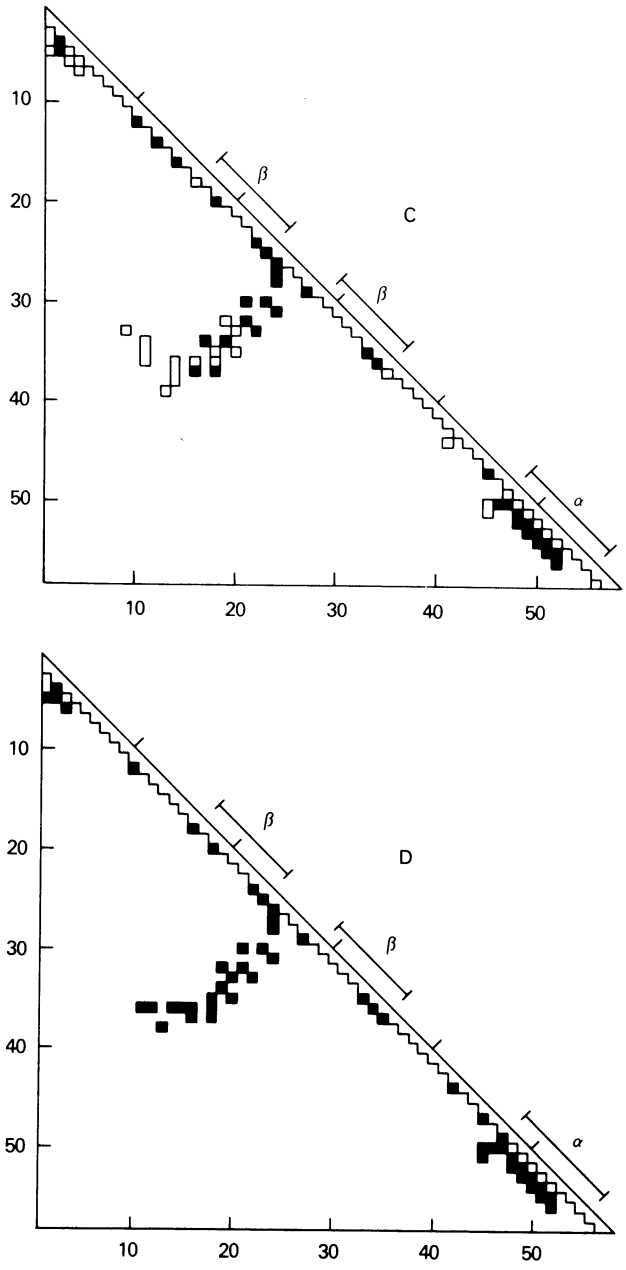


Fig. 6. (*continued from the previous page*)

native contact pair is statistically averaged over selected regions of energy at the melting temperature. This averaging somewhat reduces the statistical variations originating in the random sampling. The specific regions of energy utilized are those designated by the bars at the bottom of Fig. 3(b). Also, the statistical averages of the short-range energies of each residue for these same energy regions are shown in Fig. 7. We have employed the values $-1$ and $-2$ for the short-range energy of each residue and the interresidue contact energy, respectively; see Eqs. (2) and (3). The average

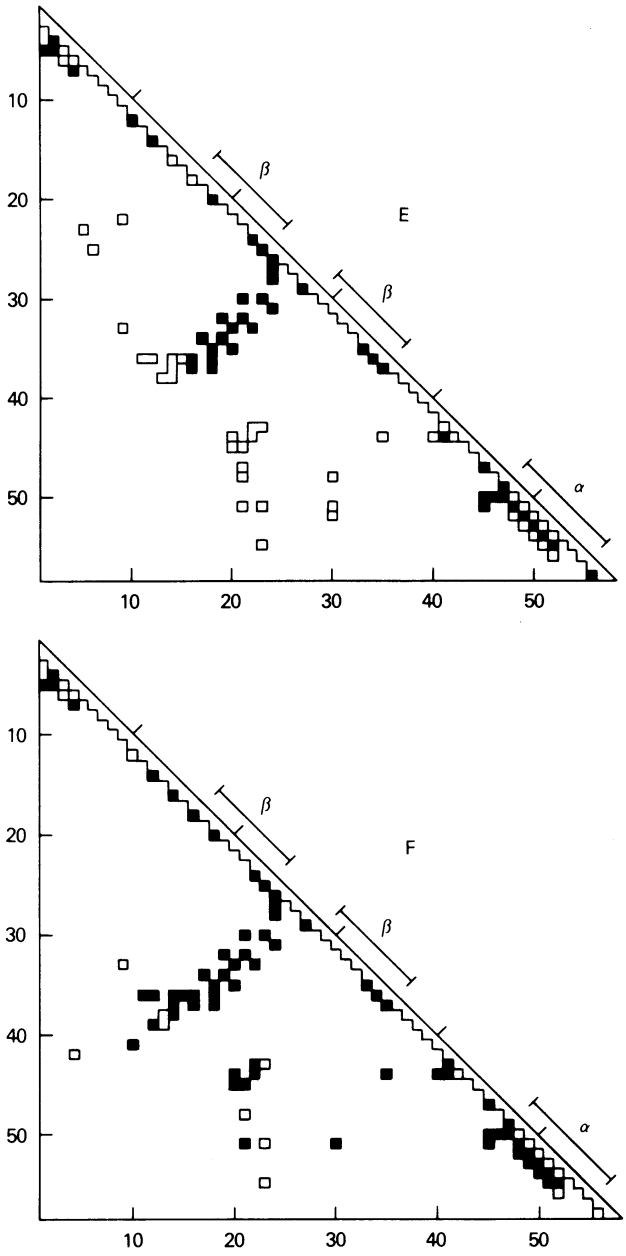


Fig. 6. (*continued from the previous page*)

contact energy in Fig. 5 divided by −2 yields the average number of contacts formed for each contact region. The average of the short-range energy in Fig. 7 divided by −1 gives the probability that the dihedral angles of each residue are within 10° of their crystal structure values.

As pointed out before, the equilibrium folding of this model protein begins with the appearance of intraresidue and medium-range interresidue interactions. At a very early stage of folding, nativelike medium-range contacts appear at the turn between $\beta$-strands and in the $\alpha$-helix [see Figs.

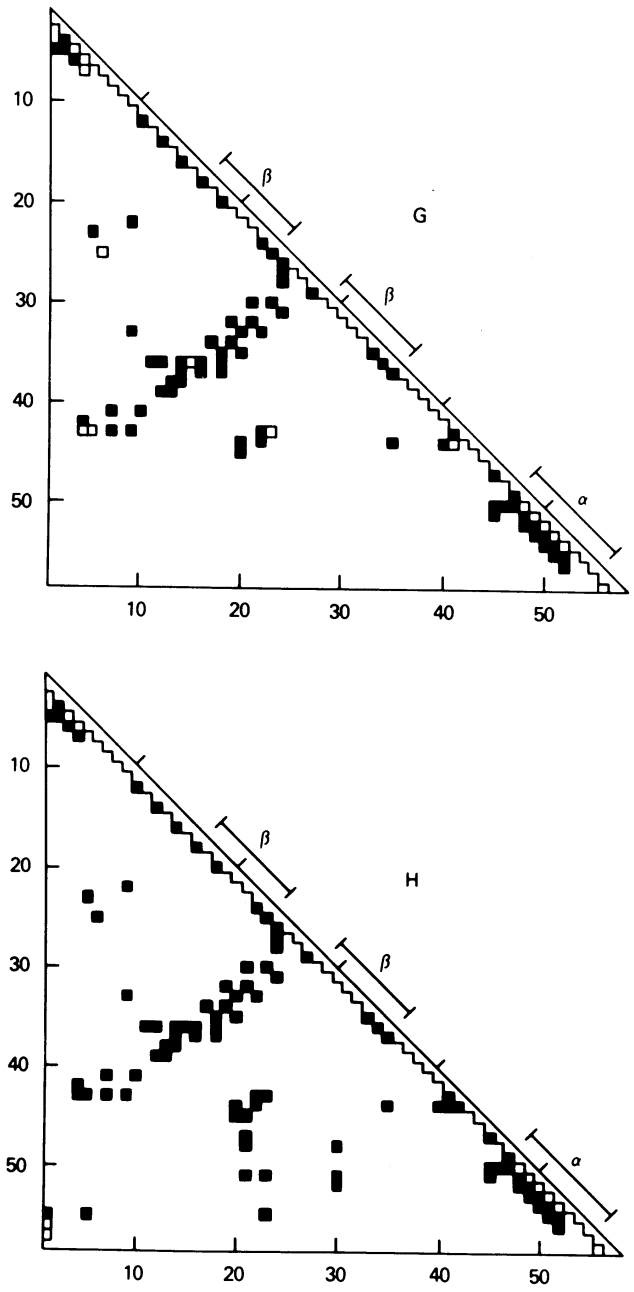


Fig. 6. (*continued from the previous page*)

6(B) and 7(B)]. Through the folding process, some individual residues appear to be nearly native at one stage but become nonnative before passing back to native form. The denatured regions of Figs. 6 and 7 indicate that most residues do not take their native conformation, but that some may form loose secondary structures, i.e., the $\alpha$-helix, $\beta$-strands, and the turn. Those small nuclei grow locally toward both chain ends, as the total con-

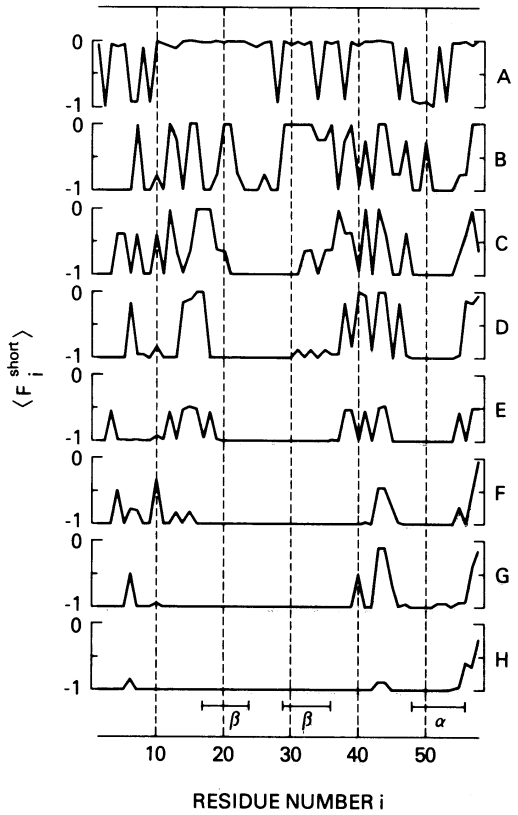


Fig. 7. The statistical average of the short-range intraresidue energy for each residue in PTI, averaged for each of the energy regions A–H from Fig. 3(b). The ordinate values divided by −1 correspond to the probabilities of the dihedral angles being within 10° of the native ones. As a single measure of the extent of native character, the ordinate values can be averaged over all residues for each of the figures A–H. These averages, for individual curves, from top to bottom, are −0.23, −0.58, −0.70, −0.76, −0.88, −0.91, −0.91, and −0.96.

formational energy decreases. These appear to be formed because of favorable intraresidue interactions and medium-range interresidue interactions. Figure 8 shows the total intraresidue energy for the lowest energy conformation. Note that the turn region (residues 25–28) and the $\alpha$-helix are relatively stable, compared to all other regions, on this basis alone. The formation of a turn is particularly useful because it introduces the possibility of interactions between the flanking residues on each side of the turn. Turns should precede long-range $\beta$-sheet interactions; they can be stabilized by favorable intraresidue and medium-range interresidue interactions in the initial stage of folding.

The present results permit a detailed examination of the order of appearance of long-range contacts at different stages of folding. A small nucleus at the turn is followed by the formation of a $\beta$-sheet, comprising the two $\beta$-strands flanking the turn, at residues 25–28 [see Figs. 6(B), 6(C), 7(B), and 7(C)]. The contact region I corresponding to the $\beta$-sheet is almost

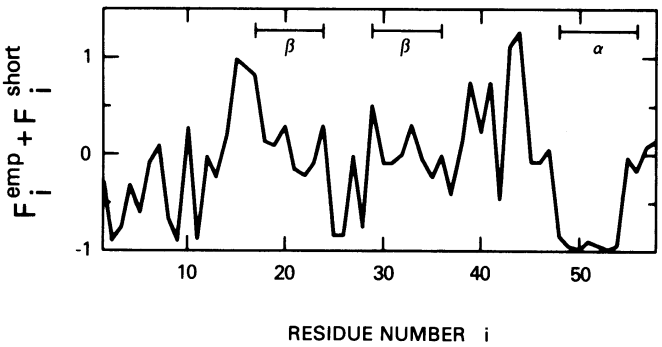


Fig. 8. Intraresidue energies of each residue for the single lowest energy conformation. Intraresidue energies consist of the sum of the empirical energy and the short-range energy.

completely formed for conformations with energies less than −100 [see Fig. 5(I)]. Formation of the β-sheet appears to be a limiting process, because conformations with energies of about −100 are located near the maximum in the free energy between the denatured state and the native state [see Fig. 3(b)]. This is consistent with the fact that the two-strand β-sheet formation would cause a large conformational entropy reduction. After this β-sheet is formed, the folding could proceed very rapidly, descending the free energy curve by forming favorable long-range interactions. The order of formation of contact regions in this latter stage of folding is not so simple. A contact region II appears in conformations with energies between about −110 and −100 [see Fig. 6(D)]. Figure 6(E) shows that contact regions II, III, IV, and V are formed with intermediate probability for energies from about −130 to −120. Examination of the conformational characteristics of single points in Fig. 3(b) indicates that region E consists of a mixture of conformations in which either contact regions I, II, and V or contact regions I, III, and IV are formed. Contact regions I, II, III, and IV are formed in conformations with energies between about −150 and −140; and contact regions I, II, V, VI, and parts of III and IV are formed at energies of about −165 to −155 [see Figs. 6(F) and 6(G)]. The last step in the folding is the formation of contact region VII, which corresponds to contacts between the chain termini [see Figs. 5(VII) and 6(H)]. It is observed that contact regions do not form uniformly in order of increasing distance from the diagonal of the contact map. The formation of contact region II precedes that of contact region V. Also, contact regions III and IV appear to be formed nearly simultaneously, rather than formation of contact region III prior to contact region IV. We infer the equilibrium folding pathways shown in Fig. 9 by connecting similar conformations in order of decreasing energy.

It should be pointed out that Figs. 6 and 7 display changes that could not be observed if properties were averaged over all molecular energies; regions of intermediate energy corresponding to high free energies would not

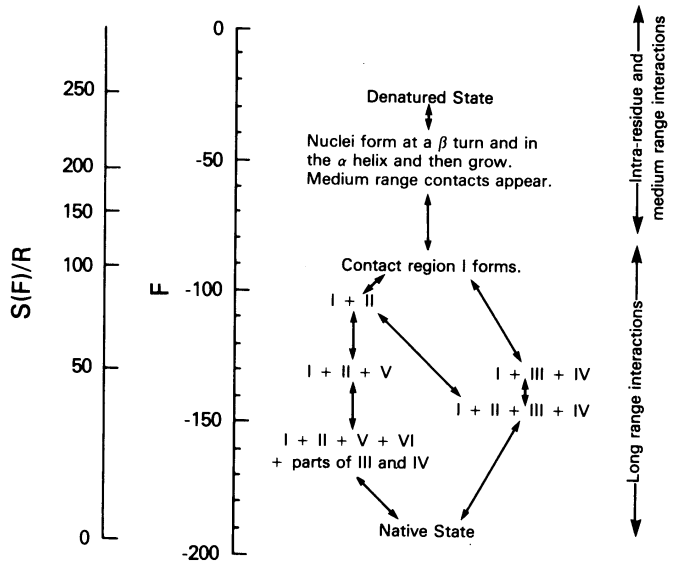


Fig. 9. Most probable folding pathways in this idealized model of PTI. These qualitative results are obtained from the Monte Carlo generation of 230,000 molecular conformations, with approximate treatment of excluded volume. The characteristic features of the most probable conformations at fixed conformational energies are summarized in this figure. By assuming smooth growth, arrows connect the intermediates to indicate the folding–unfolding pathways. Scales of conformational entropy and conformational energy are given on the left side of the figure to assist in the characterization of intermediates.

contribute significantly to equilibrium averages. Partially folded conformers are too high in free energy to appear in significant amounts in equilibrium mixtures at the melting temperature. The equilibrium state near the melting point consists principally of a mixture of completely folded and completely unfolded molecules; small changes in temperature affect only the relative populations of these two states. Therefore, use of properties averaged over all free energies is inappropriate for detecting folding intermediates.

## SUMMARY

The calculated entropy-versus-energy curve is convex in the vicinity of the denatured state and concave near the native state; therefore, the folding–unfolding transition of this model protein is an "all-or-none" type. The convex part of the entropy-versus-energy curve originates in the intraresidue and medium-range interresidue interactions; the concave part is due to the specific long-range interresidue interactions. This indicates that in the context of the present calculation, the specific long-range interactions are responsible for the "all-or-none" behavior of the transition. However, the folding of the protein begins with the appearance of intra-

residue and medium-range interresidue interactions. An unfortunate fact about "all-or-none" transitions is that equilibrium averages over the full range of free energies do not offer useful information about intermediates in the folding process; this has led us to attempt to infer a most probable folding pathway by statistically describing individual contacts for relatively small ranges of conformational energy. The resulting equilibrium folding pathway of this model protein is shown in Fig. 9. Arrows represent pathways among the native conformation, intermediates, and denatured forms. Folding is found to consist of four stages: (1) Nativelike medium-range contacts appear at a $\beta$-turn and in the $\alpha$-helix. (2) These grow toward the completion of the $\beta$-sheet and $\alpha$-helix. Conformations located near the maximum in the free energy correspond to the formation of the two-strand $\beta$-sheet. (3) After formation of this $\beta$-sheet, the folding proceeds with the growth of the nativelike structure toward either the amino terminus or the carboxyl terminus. (4) The final event is the formation of native contacts between N-terminal residues and C-terminal residues. This corresponds to a crude picture of four strands folding, for which there are three steps in the appearance of long-range interactions: (1) coalescence of the two central strands, (2) addition of either the first or fourth strand to the nucleus, and (3) finally folding together the remaining free strand.

## DISCUSSION

Evidence of the important role of short- and medium-range interactions in protein conformations is afforded by the considerable successes in predictions of protein secondary structures with schemes in which long-range interactions are ignored. Such predictions are usually correct for about 60% of the residues, but this depends on the molecule and the actual method applied.[29] These results indicate, however, that native protein conformations are not determined by short- and medium-range interactions alone. It has been observed that transitions in real proteins are not gradual, as in one-dimensional systems,[30] but are instead well represented as a two-state transition.[31] Both of these facts hint at the critical role of long-range tertiary interactions in determining protein conformations. Furthermore, in lattice models of proteins, it was demonstrated that completely nonspecific interunit interactions cannot usually cause an "all-or-none" transition on either two-dimensional[7] or three-dimensional lattices.[32] Therefore, it is quite likely[12] that long-range interresidue interactions, specific in character, are responsible for the "all-or-none" transitions observed in real proteins and play an important role in determining native conformations. The origin of this apparent specificity may reside either in genuinely specific, individually favored interactions or in less specific solvent and excluded-volume effects.

It is difficult to determine the relative importance of various classes of these interresidue interactions. Hydrophobic interactions are relatively less specific than electric interactions between polar residues, with or

without salt bridges. At present, the relative contributions of each class of interactions—electric interactions, hydrogen bonds, and solvent effects including hydrophobic interactions—to the folding process of proteins are unknown. Finney et al.[33] estimated the contribution of each of these classes to the free energy change between the native state and the denatured state for a few proteins. Although that limited analysis is useful in presenting a picture in which the native structures of proteins are realized through a delicate balance of several contributions, more detailed studies of the roles of each class of interactions in the folding process should be pursued.

In protein native structures, amino acid residues are packed like tightly fitting jigsaw puzzles.[34–36] Therefore, there is a possibility that the number of compact structures, such as the native conformation, is extremely limited by the heterogeneous amino acid sequence. In this way, packing might lead to interactions that could appear to be specific. It is useful to examine effects of specificity, regardless of its origin; the present model serves this purpose well. The present approximation of completely specific interactions is one of the simplest models to yield an "all-or-none" type of transition.

Some models of folding pathways have been proposed on the basis of contact maps of crystal structures,[37–39] the contacts between chain segments in crystal structures,[40] and the compactness of structures formed by chain segments in crystal structures.[41] All of these models are based on characteristics of native structures. In other words, nonspecific interactions are completely neglected, and only specific interresidue interactions—as represented by the contact map, by the extent of contact, or by the extent of compactness—are taken into account. In none of these previous models has account been taken of conformational freedom of nonnative forms. This can lead to an incorrect assessment of the importance of entropy for the process. In the present model, conformational entropies are explicitly evaluated. The largest errors in the present calculation are likely to occur in the evaluation of the energies of nonnative conformations.

The importance of nonnative forms has clearly been established in the experimentally determined folding pathways[42–46]: most intermediates were determined to be nonnative conformations, as detected by both the existence of incorrect disulfide bonds and immunochemical methods. But those results do not offer evidence to permit distinguishing between the specific or nonspecific origin of interactions. The present calculation does not favor any nonnative contact pairs; in latter calculations we plan to attempt to include other favorable interresidue interactions in addition to the native ones.

Contacts near the diagonal of the contact map, which here are designated as medium-range contacts, are formed at the initial stages of folding; these correspond roughly to formation of secondary structures. The principle that regular secondary conformations precede the appearance of longer-range contacts during folding is strongly supported by the results in Fig. 6. Tanaka and Scheraga[37] proposed a three-step folding hypothesis in which contacts are formed in the order of increasing distance from the di-

agonal of the contact map. This hypothesis was employed in predicting folding pathways for several proteins.[38] The limitations in the applications of this hypothesis were discussed in detail by Nemethy and Scheraga.[39] Their stated assumption is that contact regions of native structures correspond to fairly stable structures, simply because of the presence of many contacts. This assumption is similar to our treatment of the interresidue interactions because we have assigned a negative interaction energy only to those contacting residue pairs reported in the native conformation. However, the order of formation of contact regions in our model protein does not coincide with strict application of their hypothesis. Specifically, conformation contact groups (III), (V), (I,III), (I,V), (I,III,V), or (I,II,III,IV,V), which would be expected with the three-step hypothesis, have not been detected. This can be understood intuitively: most intervening residues between a long-range contact must be nativelike; consequently, most shorter-range native contacts would also appear simultaneously. For example, the formation of contact regions I and III necessarily accompanies the appearance of contact region IV. Intermediate conformations shown in Fig. 9 with the following groups of contact regions have been detected: (I), (I,II), (I,III,IV), (I,II,V), (I,II,III,IV), and (I,II,V,VI, parts of III and IV). These can be seen in the contact maps (Fig. 6). It should be noted that Fig. 6(E) corresponds to a mixture of conformations of the contact regions (I,III,IV) and (I,II,V).

Conformations including contact regions III or V alone have not been detected, although they are as close to the diagonal of the contact map as contact reigon I. It should be realized that regions III and V possess fewer long-range contacts than the other regions. The following simple scheme may be useful for understanding qualitatively why conformations forming contact regions III or V alone are not favorable. Conformational entropy loss and conformational energy gain accompanying the formation of a contact region may be roughly proportional to the number of residues participating and the number of long-range contacts, respectively. Values of these quantities for the intermediates from Fig. 9 are plotted in Fig. 10. This may be regarded as a crude approximation to the entropy–energy curve in Fig. 2. Because most of the short- and medium-range contacts form prior to the appearance of long-range contacts, they are not included in this rough description. The number of residues participating in the formation of contact regions III or V is almost the same as for contact region I. This indicates that the conformational entropy loss accompanying the formation of contact regions III or V is comparable to that for the formation of contact region I. However, the conformational energy gain accompanying their formation would be much less than that obtained by the formation of contact region I, as inferred from the difference in the number of long-range contacts within these contact regions. Therefore, the formation of contact region III or V alone is less favorable than that of contact region I. Also, this simple consideration explains why the formations of contact regions (I,III,IV) in residues 17–55 and (I,II,V) in residues 5–39 are

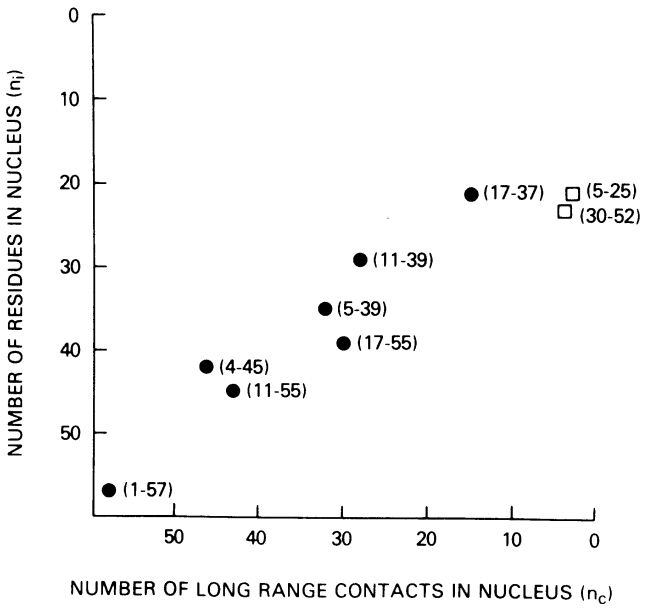


Fig. 10. Relationship between the numbers of residues $n_i$ and the corresponding number of long-range contacts $n_c$ for the folding intermediates indicated in Fig. 9. Open squares represent contact regions III and V. The numbers in parentheses are terminal residue numbers of the native conformational regions. $n_i$ and $n_c$ provide a rough approximation to the entropy loss and the energy gain accompanying the formation of such local structures.

accompanied by almost the same conformational energy and entropy changes, as indicted in Fig. 6(E); both the numbers of participating residues and the numbers of contacts are almost the same for the formation of these groups of contact regions. Similar changes also occur for the formation of either of the two contact groups (I,II,III,IV) in residues 11–55 or (I,II,V,VI, parts of III and IV) in residues 4–45.

These simple considerations are similar to what has been termed a noninteracting local structure model of protein conformations.[10,13,47–50] In that model, a protein conformation is depicted as an alternating series of random-coil and local nativelike structures; interactions between random parts and local structures are neglected. In the case of specific nativelike interresidue interactions, as employed in the present model, the approximation of a noninteracting local structure model appears reasonable (to be published). A model of hierarchical condensation proposed by Lesk and Rose[41] can also be regarded as similar to the noninteracting local structure model. They obtained hierarchical structures by choosing most compact local conformations of a given size and then connecting them in order of increasing size. Their proposed folding pathway for PTI (G. D. Rose, personal communication) is similar to one of the folding pathways shown in Fig. 9, specifically, the formation of contact regions in the order of (I), (I,II), (I,II,V), (I,II,V,VI) and (I,II,III,IV,V,VI,VII). A similar intermediate (I,II,V,VI) was reported[13] for a lattice kinetic simulation of the folding of PTI. However, neither obtained the other pathways indicated

in Fig. 9. This coincidence in pathways for the three models occurs because these models all consider only specific native interresidue interactions. Also, the obvious similarity between the nature of compact structures and favorable weighting accorded large numbers of long-range native contact pairs should be noted.

The folding pathways proposed here are consistent with an important structural aspect of PTI, namely, the two loop–thread structures in this molecule. One of them consists of the loop linked by the S-S bond between the 30th and 51st cysteines and of the thread composed of the 19th to 23rd residues; the other consists of the loop linked by the formation of contacts between the 5th and 23rd residues and of the thread composed of the 34th to 37th residues.[51] The contact regions III and V lead to the formation of these loops. The threads consisting of the 19th to 23rd residues and of 34th to 37th residues are fixed at one side of each loop by the $\beta$-sheet formation, i.e., the formation of contact region I, and by the formation of contact region II, respectively. Connecting threads on one side of each loop is preferable for formation of the loop–thread structure; otherwise, the loop–thread structure is likely to be missed. Therefore, it would be more favorable to form the contact region I prior to all other long-range contact regions. Also, it is reasonable by the same simple geometric considerations that the formation of the contact region II should precede that of the contact region V.

By trapping and identifying intermediates, Creighton[42–45] studied the folding pathway of PTI from the fully reduced, unfolded form to the native, folded state with three disulfide bonds. Single disulfide intermediates formed in the initial stage of the refolding had an S-S bond between the 30th and 51th cysteines or 5th and 30th cysteines, rather than that between 14th and 38th cysteines; this is not consistent with the present results. The S-S bond between the 14th and 38th cysteines was formed as the last S-S bond toward the native conformation from a two-disulfide intermediate (30–51,5–55). The two-disulfide intermediate (30–51,5–55) was not formed directly from a single-disulfide intermediate (30–51), but by rearranging a wrong S-S bond in two other two-disulfide intermediates, (30–51,5–14) and (30–51,5–38). Although a two-disulfide intermediate (30–51,14–38) was detected as a product from the single-disulfide intermediate (30–51), it did not lead directly to the same intact disulfide bonds as in the native form, i.e., (14–38,30–51,5–55). This indicates that this two-disulfide intermediate (30–51,14–38) may not have the native loop–thread structure. In other words, the N-terminal part of the chain is not threaded in the loop linked by the S-S bond between the 30th and 51st cysteines. This threading is probably missing in a single-disulfide intermediate (30–51) and in two other two-disulfide intermediates, except (30–51,5–55). The threading would appear to be performed in a transition to the two-disulfide intermediate (30–51,5–55). This is also consistent with this transition being the slowest step. However, this observed order of disulfide bond formation does not coincide with any that could be anticipated from the present re-

sults. The experimental conditions may not correspond to the equilibrium conditions of the present calculations.

The nmr results of States et al.[52] indicate the coexistence of the native conformer and a nativelike metastable form. Even though the spectra differ only for residues tyrosine-21, tyrosine-23, glutamine-31, threonine-32, phenylalanine-45, alanine-48, and methionine-52, they ascribe this additional form to an alternative folding pathway. The concept that major differences between folding pathways lead to relatively minor, but observable, differences in the ultimate folded conformation is intriguing. Unfortunately, detailed experimental descriptions of such an alternative pathway are not available. In the absence of such details, it is not possible to consider whether or not there is any connection between this alternative pathway and the folding schemes outlined here. It is only possible to say that our intermediates are most similar to the two-disulfide intermediate (30–51,14–38), which was also determined to be most nativelike in the immunochemical studies of Creighton et al.[46] However, they have determined that this form is not on the direct pathway to the native structure.[42–45]

The folding pathway in this simple model of reduced PTI has been examined in an attempt to understand general features of folding pathways. It is necessary to consider the propriety of the approximations in the present calculations. Oversimplifications in the present treatment of the interresidue interactions may prevent a valid determination of folding pathways. However, these simplifications are useful to indicate the effects of specific interresidue interactions on the folding pathway. It would be useful to consider in detail how the above results depend on the potential energy functions used. The formation of local regions of nativelike conformation at an early stage of folding depends on both the intraresidue interaction potential and the medium-range interresidue interactions. The empirical energy and the additional energy used for intraresidue interactions are somewhat arbitrary. However, the latter value is not so strong as to produce single dominant conformations. Effects on the pathway of varying the contact energy and the distance used for defining the contact map have been examined by employing a noninteracting local structure model. Although the free energies manifest large changes, the most probable conformations at various stages of the transition appear to be relatively invariant over large ranges of the parameters. The effects of neglecting the side chain energies and entropies are unknown. The assumption of specific nativelike interresidue interactions is much less sound. Neglecting competing nonnative interresidue interactions is a serious omission; these include both those specific and nonspecific in character. Effects of such competing conformations on the folding pathways remain to be determined. Hydrophobic residue–water interactions may reduce the energy for compact forms and increase it for expanded forms in which hydrophobic residues are exposed to water. Choice of a larger, more realistic value for the radius of the side chains would reduce the entropies. There can be compensation

between these two effects as manifested in the net change of free energy between the denatured to native state. However, the extent of compensation at intermediate stages in the transition is uncertain; consequently, the nature of favored intermediate conformations is more susceptible to modification. Energies of interresidue interactions have been taken to be favorable for close contact residue pairs, as observed in the crystal structure, and zero for all other contact pairs. Models in which interresidue interactions are treated more realistically remain to be studied.

We thank N. Gō for useful discussions and suggestions about the manuscript, P. J. Flory for valuable comments, George Rose for kindly sending us a copy of his manuscript prior to its publication, and H. Mizuno for his program to plot distance maps.

## References

1. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H., Jr. (1961) *Proc. Natl. Acad. Sci. USA* **47,** 1309–1314.
2. Brant, D. A. & Flory, P. J. (1965) *J. Am. Chem. Soc.* **87,** 2791–2800.
3. Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Interscience, New York, pp. 248–255.
4. Wetlaufer, D. B. (1973) *Proc. Natl. Acad. Sci. USA* **70,** 697–701.
5. Anfinsen, C. B. & Scheraga, H. A. (1975) *Adv. Protein Chem.* **29,** 205–300.
6. Levinthal, C. (1968) *J. Chim. Phys.* **65,** 44–45.
7. Taketomi, H., Ueda, Y. & Gō, N. (1975) *Int. J. Pept. Protein Res.* **7,** 445–459.
8. Gō, N. & Taketomi, H. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 559–563.
9. Gō, N. & Taketomi, H. (1979) *Int. J. Pept. Protein Res.* **13,** 235–252.
10. Gō, N. & Taketomi, H. (1979) *Int. J. Pept. Protein Res.* **13,** 447–461.
11. Gō, N. (1976) *Adv. Biophys.* **9,** 65–113.
12. Ueda, Y., Taketomi, H. & Gō, N. (1978) *Biopolymers* **17,** 1531–1548.
13. Gō, N., Abe, H., Mizuno, H. & Taketomi, H. (1980) in *Protein Folding*, Jaenicke, R., Ed., Elsevier/North Holland, Amsterdam, pp. 167–181.
14. Nemethy, G. & Scheraga, H. A. (1977) *Q. Rev. Biophys.* **10,** 239–352.
15. Nishikawa, K., Ooi, T., Isogai, Y. & Saito, N. (1972) *J. Phys. Soc. Jpn.* **32,** 1331–1337.
16. Ooi, T., Nishikawa, K., Oobatake, M. & Scheraga, H. A. (1978) *Biochem. Biophys. Acta* **536,** 390–405.
17. Deisenhofer, J. & Steigemann, W. (1975) *Acta Crystallogr.* **31,** 238–250.
18. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542; The Brookhaven Protein Data Bank.
19. Gō, N. (1975) *Int. J. Pept. Protein Res.* **7,** 313–323.
20. Gō, N. & Scheraga, H. A. (1969) *J. Chem. Phys.* **51,** 4751–4767.
21. Flory, P. J. (1974) *Macromolecules* **7,** 381–392.
22. Gō, N. & Scheraga, H. A. (1976) *Macromolecules* **9,** 535–542.
23. Hammersley, J. M. & Handscomb, D. C. (1964) *Monte Carlo Methods*, Methuen, London.
24. Crippen, G. M. (1977) *Macromolecules* **10,** 21–25.
25. Crippen, G. M. (1977) *Macromolecules* **10,** 25–28.
26. Wall, F. T., Windwer, S. & Gans, P. J. (1962) *J. Chem. Phys.* **37,** 1461–1465.
27. Premilat, S. & Hermans, J., Jr. (1973) *J. Chem. Phys.* **59,** 2602–2612.
28. Premilat, S. & Maigret, B. (1977) *J. Chem. Phys.* **66,** 3418–3425.
29. Jernigan, R. L., Miyazawa, S. & Szu, S. C. (1980) *Macromolecules* **13,** 518–525.
30. Landau, L. D. & Lifshitz, E. M. (1958) *Statistical Physics*, Pergamon, London, p. 478.

31. Baldwin, R. L. (1975) *Annu. Rev. Biochem.* **44,** 453–475.
32. Kron, A. K., Ptitsyn, O. B., Skvortsov, A. M. & Fedrov, A. K. (1967) *Mol. Biol.* **1,** 576–582.
33. Finney, J. L., Gellatly, B. J., Golton, I. C. & Goodfellow, J. (1980) *Biophys. J.* **32,** 17–33.
34. Richards, F. M. (1974) *J. Mol. Biol.* **82,** 1–14.
35. Finney, J. L. (1975) *J. Mol. Biol.* **96,** 721–732.
36. Richards, F. M. (1977) *Annu. Rev. Biophys. Bioeng.* **6,** 151–176.
37. Tanaka, S. & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 3802–3806.
38. Tanaka, S. & Scheraga, H. A. (1977) *Macromolecules* **10,** 291–304.
39. Nemethy, G. & Scheraga, H. A. (1979) *Proc. Natl. Acad. Sci. USA* **76,** 6050–6054.
40. Crippen, G. M. (1978) *J. Mol. Biol.* **126,** 315–332.
41. Lesk, A. M. & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. USA* **78,** 4304–4308.
42. Creighton, T. E. (1977) *J. Mol. Biol.* **113,** 275–293.
43. Creighton, T. E. (1977) *J. Mol. Biol.* **113,** 295–312.
44. Creighton, T. E. (1977) *J. Mol. Biol.* **113,** 313–328.
45. Creighton, T. E. (1978) *Prog. Biophys. Mol. Biol.* **33,** 231–297.
46. Creighton, T. E., Kalef, E. & Arnon, R. (1978) *J. Mol. Biol.* **123,** 129–147.
47. Wako, H. & Saito, N. (1978) *J. Phys. Soc. Jpn.* **44,** 1931–1938.
48. Wako, H. & Saito, N. (1978) *J. Phys. Soc. Jpn.* **44,** 1939–1945.
49. Gō, N. & Abe, H. (1981) *Biopolymers* **20,** 991–1011.
50. Abe, H. & Gō, N. (1981) *Biopolymers* **20,** 1013–1031.
51. Connolly, M. L., Kuntz, I. D. & Crippen, G. M. (1980) *Biopolymers* **19,** 1167–1182.
52. States, D. J., Dobson, C. M., Karplus, M. & Creighton, T. E. (1980) *Nature* **286,** 630–632.

Received April 30, 1971
Accepted January 6, 1982