Most Probable Intermediates in Protein Folding–Unfolding with a Noninteracting Globule-Coil Model[†]

Sanzo Miyazawa and Robert L. Jernigan*

ABSTRACT: Protein conformations are generated with a noninteracting globule-coil model in which each residue is assumed to take only the native or random-coil state, and a protein conformation is regarded to consist of alternating regions of random coil and globules of native conformation. Statistical weights are taken to have two parts, corresponding to intraresidue and interresidue interactions. The intraresidue statistical weight for a residue in its native state is assumed to be proportional to the empirical frequency of the native (ϕ,ψ), from tabulated statistics. Interresidue energies are taken to be proportional to the number of contacts within each native region. The principal adjustable parameter is the contact energy per contact pair. Proteins studied include trypsin inhibitor, ribonuclease A, lysozyme, and apomyoglobin. The number of native residues is employed as a simple one-dimensional representation of the folding–unfolding coordinate to describe probable folding pathways. When conformations are considered at each point on this coordinate, it is possible to obtain detailed descriptions of the conformational characteristics of relatively rare intermediates along the folding pathway. This technique of "trapping" conformational intermediates and statistically characterizing them appears to be a generally useful procedure for studying conformational transitions. Plausible equilibrium folding–unfolding pathways have been constructed by connecting most probable conformations in order according to the total number of native residues. The most probable conformations are determined from all possible combinations of native and random-coil regions. For each protein, free energies and the probabilities of each residue being in the native state are calculated at each stage of folding with a wide range of values of the contact energy parameter. Individual residues which have high probabilities of being native at each stage are indicated in diagrams. The free energies depend strongly on the value of the contact energy parameter; however, the most probable conformations are relatively insensitive to this parameter. Typically, helices and turns appear prior to formation of β sheets, and β-sheet formation coincides with a large maximum in the free energy because of the attendant loss of conformational entropy. In all cases, only a few separate native domains are observed at all stages of folding.

There have been many studies of protein folding mechanisms and folding pathways. Because it has often been considered impossible to obtain the native structures of proteins by means of an exhaustive random search of all of conformational space, there have been a variety of experimental and theoretical attempts to elucidate folding pathways. Detection of intermediates on folding pathways and descriptions of their conformational characteristics are rendered arduous by the usual two-step nature of the folding–unfolding transition. The population of partially folded or partially unfolded conformations is usually extremely small. Theoretical studies in this area can be categorized as (1) those in which energetic, geometrical, or packing considerations are explicitly taken into account, but entropic considerations are ignored, or (2) others in which both energetic and entropic contributions are explicitly included, but with simplifications to geometries and intramolecular interactions. In some of the latter studies, Monte Carlo methods have been used to simulate the kinetics of protein folding and unfolding (Go et al., 1980), as well as to generate a wide range of conformations from the native to the denatured state (Miyazawa & Jernigan, 1982). Because such Monte Carlo methods are difficult and time consuming, simpler methods are desirable. The one-dimensional lattice gas model of proteins proposed by Wako & Saito (1978a,b) is a simpler model. This two-state model, which is similar to helix–coil models, has been discussed in some detail as the "noninteracting local structure model" by Go et al. (Go et al., 1980; Go & Abe, 1981); here we are designating it as the noninteracting globule-coil model. Simplifications are possible because of the assumptions that each residue takes only native or random-coil states and that interresidue interactions can be neglected except within native conformational regions.

Reliance on knowledge of the native conformation is strong in the present model, as well as in all previous models of protein folding, including the three-step mechanism of Tanaka & Scheraga (1975) and the Lesk & Rose (1981) model of hierarchic organization of compact units in proteins. In all of these models, it appears to be almost impossible to treat folding which passes through nonnative intermediate states. In spite of a similar limitation, the present model is appealing because statistical mechanical formulations are possible with the noninteracting globule-coil model.

A detailed Monte Carlo generation of equilibrium conformations for trypsin inhibitor (Miyazawa & Jernigan, 1982) has indicated interactions between separated native globules to be improbable. Abe & Go (1981) have shown that their noninteracting local structure model substantially reproduces the interdependence of the entropies and energies as obtained from Monte Carlo simulations for two-dimensional lattice proteins. Both of these results indicate that the noninteracting globule-coil model may afford a useful description of the equilibrium properties of protein folding.

The principal purpose of this paper is to develop the noninteracting globule-coil model as a practical method for studying protein folding pathways in which one can account for significant conformational details. The major difficulty in describing folding intermediates comes from the all-or-none character of protein transitions; usually populations of folding intermediates are not abundant enough to be reflected in equilibrium averages. The latter are usually composed of

[†] From the Laboratory of Mathematical Biology, Building 10, Room 4B-56, DCBD, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20205. *Received February 10, 1982.*

Published in Biochemistry, October 12, 1982, pp. 5203–5213, by the American Chemical Society

significant contributions from only the two extreme states, corresponding to native and denatured forms. In an attempt to overcome this problem, the number of native residues will be employed in a simple one-dimensional representation of the folding coordinate. The use of a plausible folding coordinate in order to describe activated states has been employed in our Monte Carlo method (Miyazawa & Jernigan, 1982). Most probable conformations, at each point along the folding coordinate, will be described with the probabilities for each residue being in the native state. It then becomes possible to construct equilibrium folding–unfolding pathways by connecting most probable conformations in sensible order, according to the number of native residues. The method requires specification of statistical weights for all possible native conformational regions relative to their random-coil conformations. Rigorous estimates of such statistical weights are difficult because of unknown details of solvent interactions and volume exclusions. However, these weights will be estimated crudely, and the effects of a simple interaction parameter on the protein folding pathways will be examined. Each intraresidue statistical weight will be assumed to be proportional to the empirical frequency of that residue's native (ϕ,ψ). In addition, the interresidue energy of a native region will be assumed to be proportional simply to the number of C^β–C^β contacts. The principal parameters in these calculations will consist of this proportionality constant, which is the favorable energy to be associated with each interresidue close contact, and the distance used to define such contacts in the crystal. Results will be reported for trypsin inhibitor (PTI), ribonuclease A (RSA), lysozyme (LYZ), and apomyoglobin (MBN). Details of the free energy along the folding coordinate depend strongly on the values of these two parameters. However, the most probable conformations will be found to be relatively invariant. This indicates the pathways themselves to be less sensitive to parameter values. Folding pathways obtained for trypsin inhibitor (PTI) are similar to those obtained in the more detailed Monte Carlo generations (Miyazawa & Jernigan, 1982).

Noninteracting Globule-Coil Model. Each residue in a protein is assumed to exist in either of two conformational states, the native state or the random-coil state. Here, the native state of each residue is defined as the reported crystal structure, which includes the backbone conformations as well as the side-chain conformations. Properly, it should also include conformational fluctuations about the crystal conformation. Although each rotatable bond in a protein can be defined (Go & Abe, 1981) to take either the native or random conformation, we have chosen instead to treat each residue as a single unit. This simplification, that ϕ and ψ in a residue can be regarded as a unit, is based on the well-known result for *ideal* random coils (Flory, 1969) that ϕ and ψ within a residue are strongly interdependent but do not depend significantly on ϕ and ψ of neighboring residues. Thus the partially folded protein is to be regarded as consisting of alternating portions of random-coil and native conformational regions; the latter are designated here as globules and as local structures by Go et al. (Go et al., 1980; Go & Abe, 1981). Interresidue interactions are included only within each native conformational region; those within random-coil regions and between random-coil and native conformational regions are completely neglected. Some of these neglected interactions could perhaps be included, in the mean field sense. The goodness of this assumption depends on how random these nonnative regions are and may be less valid for residues adjacent to a native region (Flory, 1972). These simplifications permit a statistical mechanical formulation of a partition function and averages. Evaluation of the requisite statistical weights of random-coil and native conformational regions will be considered later.

Construction of Equilibrium Folding–Unfolding Pathways. Here, we are concerned with the folding and unfolding process, at equilibrium. Thermal fluctuations permit a molecule to change its conformation and pass through an activated state, in an all-or-none transition, between native and denatured states. Folding–unfolding pathways are defined as pathways along which a representative molecule passes in changing its conformations. Because conformational changes are intrinsically statistical, the concept of pathways is also statistical; pathways here are most probable ones. In equilibrium, these most probable pathways for both folding and unfolding must be identical except for direction.

One of the difficulties in describing folding–unfolding pathways arises from the all-or-none character of protein transitions. In an all-or-none transition, the equilibrium state near the melting point will consist of a mixture of completely folded and completely unfolded molecules; populations of folding intermediates would always be extremely rare. Changing external conditions would change only the relative populations of these two states but not provide useful information about intermediates. Therefore, methods (Wako & Saito, 1978b) of examining changes in the equilibrium distributions by changing environmental parameters, which would correspond to experimental methods of observing transitions by quasi-statically changing external conditions, are not so useful for studying folding–unfolding pathways. Instead, some means of "trapping" intermediates is needed.

Studying folding intermediates requires the introduction of measures of the extent of folding. For this purpose, Abe & Go (1981) employed two types of quantities, the intramolecular interaction energy and the number of random-coil residues. They showed that the most probable folding pathways for two-dimensional lattice proteins were roughly in the direction of decreases in both the conformational energy and the number of random-coil residues. This close relationship between these two quantities along the most probable pathway is expected because all components of the energy, both the short-range and long-range interaction energies, were taken to favor the native conformational state; therefore, the conformational energy tends to decrease as the number of random-coil residues decreases. The use of either conformational energy or the number of random-coil residues as a folding coordinate must be similar if intramolecular interactions are treated in such a manner. Here we choose the number of native residues as a simple one-dimensional folding coordinate.

Using the number of native residues as a folding coordinate is consistent with an intuitive concept of a folding process, that the overall trend in passing toward the native structures of proteins must generally be toward increasing numbers of native residues. As stated in the preceding section, folding through favored nonnative conformations cannot be treated with a noninteracting globule-coil model. Although it may impose a serious limitation on the applications of this model, the two-state, native and random coil, conformational description is conceptually consistent with using the total number of native residues as a folding coordinate.

The first step in constructing equilibrium folding–unfolding pathways consists of characterizing most probable conformations for each point along the folding coordinate. The most probable conformations are described in terms of the probability of each residue being native. A method for constructing a folding pathway is not obvious, even though the most probable conformations along the folding axis may be identifiable. If complete energy contours in the multidimensional conformational space were available, it would be possible to determine most probable folding pathways by determining the lowest energy barriers. In the present case, we have no information other than the most probable conformations along the specific folding coordinate. In a simple case in which most probable conformations change smoothly and continuously from the denatured state to the native state along the folding axis, a most probable pathway could be constructed by directly connecting most probable conformations in order along the folding axis. But there is the possibility that most probable conformations change discontinuously along the folding axis. For such cases, it is not clear whether or not a pathway constructed in this simple direct manner is sensible; later we will consider an alternative approach.

Partition Function. The total molecular partition function including all conformational states is

$$\Xi = \sum_{n=0}^{N} Z(n) \qquad (1)$$

where $Z(n)$ is a sum of the statistical weights for all states in which the total number of native residues is n. The number of native residues can vary from 0 to N, the latter being the total number of residues in a protein. $Z(n)$ and Ξ in this formulation correspond to expressions for a canonical partition function and a grand canonical partition function, respectively, where n is the number of particles. The direct calculation of $Z(n)$ is required in order to calculate free energies and to identify the most probable combination of native residues at each point on the folding coordinate. Although the matrix formulation of Ξ given by Wako & Saito (1978a,b) could have been extended to include the direct representation of $Z(n)$, we have instead utilized a recurrence relationship. The vector

$$\mathbf{Z}_N \equiv [\mathrm{Z}(0) \ldots \mathrm{Z}(N)]^t \qquad (2)$$

is calculated by a recurrence equation

$$\mathrm{Z}_j = \begin{bmatrix} \mathrm{Z}_{j-1} & 0 & \cdots & 0 & 0 & 0 \\ & \mathrm{Z}_{j-2} & 0 \cdots & 0 & 0 & 0 \\ & & \cdot\;\cdot & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot & \cdot \\ & & & \cdot & \cdot & \cdot \\ & & & \mathrm{Z}_1 & 0 & 0 \\ & & & & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ \mathrm{L}_{jj} \\ \cdot \\ \cdot \\ \cdot \\ \mathrm{L}_{3j} \\ \mathrm{L}_{2j} \\ \mathrm{L}_{1j} \end{bmatrix} \qquad (3)$$

where $\mathbf{Z}_i$ is a column vector of $i + 1$ dimension, which is embedded in the matrix so that its first element is located on the diagonal. $\mathbf{Z}_i$ is the partition function for a chain of length i. L_{ij} is the statistical weight of the native conformational region consisting of residues i through j, relative to the random-coil reference state for the same region; note that there is the further minor requirement that at least residues $i - 1$ and $j + 1$ must be in the random-coil state. This formulation is equivalent to the generating function method given by Go & Abe (1981) except for the trivial difference that each residue is treated as a single unit instead of each rotatable backbone bond. Here, it should be noted that this formulation of the partition function is based on the assumption of the noninteracting globule-coil model: *there are no long-range interactions included within random-coil regions or between random-coil and native conformational regions.*

The free energy of the state in which the number of native residues is equal to n is

$$\mathcal{F}(n)/(RT) = -\ln Z(n) \qquad (4)$$

where RT is the thermal energy. All energies are expressed on a molar basis. The random-coil state is the reference state; Z(0) is taken to be 1.

The probability that residues i through j, inclusive, are in the same single native conformational region and that they are the two termini of this native region is

$$Q_{ij}(n) = \frac{\partial \ln Z(n)}{\partial \ln L_{ij}} \qquad (5)$$

within the state in which all conformations have exactly n native residues. Then, the probability that the residues from i through j are all in the native conformational state, irrespective of the neighboring residues' conformations, is

$$P_{ij}(n) = \sum_{k=1}^{i} \sum_{l=j}^{N} Q_{kl}(n) \qquad (6)$$

Again this is for n native residues. The sum of $P_{ii}(n)$ over i is equal to n. Conversely, Q_{ij} can be expressed in terms of the P_{ij} as follows.

$$Q_{ij} = P_{ij} - P_{ij+1} + P_{i-1,j+1} - P_{i-1,j} \qquad (7)$$

These probabilities can be used to calculate averages of various quantities, e.g., the average number of contacts formed.

The entropy originating from the various arrangements of native residues for a fixed number of native residues, n, is

$$\begin{aligned} S^{\mathrm{comb}}(n)/R &= -\mathcal{F}(n)/(RT) + \sum_{i=1}^{N}\sum_{j=i}^{N} Q_{ij}(n)[-\ln (L_{ij})] \\ &= -\mathcal{F}(n)/(RT) + \\ &\quad \sum_{i=1}^{N}\sum_{j=i}^{N} P_{ij}(n)[-\ln [L_{ij}L_{i+1,j-1}/(L_{i,j-1}L_{i+1,j})]] \end{aligned} \qquad (8)$$

Sums in the second term of this equation correspond to averages of conformational free energies. This entropy is designated here as the combinatory entropy. The maximum value of $S^{\mathrm{comb}}(n)/R$ is achieved if all combinations of residues in the native conformational state have identical statistical weights; this corresponds to the number of ways of choosing n native residues from the N residues.

$$\ln \binom{N}{n} \qquad (9)$$

Statistical Weights of Random-Coil and Native Conformational States. Below we formulate the statistical weights of the random-coil state and the native conformational state, which are required for application of this method to specific proteins. As stated above, each residue is regarded as a unit rather than each rotatable bond; also interresidue interactions are neglected except within native conformational regions. Thus, the principal contributors to the statistical weight of the random-coil state are expected to be (1) intraresidue interactions and (2) interresidue interactions, solvent interactions, and volume exclusions between atoms, here to be limited to a mean field approximation. For the native conformational regions, principal contributions come from intra- and interresidue interactions, solvent effects, and conformational fluctuations around native conformations. Go & Abe (1981) gave a general formalism for evaluating such statistical weights; however, it is time consuming to calculate them semiempirically with account of side-chain conformations as well as backbone conformations. In addition, the difficulties of properly evaluating solvent effects and the crudeness of the approximations in this model might serve to vitiate such detailed calculations. Thus, it would be useful to assess this model by first employing a crude estimation of statistical weights. Later, more detailed calculations might be appropriate. In the present case, we use only empirical conformational frequencies and the contact maps of C^β–C^β atoms from crystal structures to represent contributions of intraresidue interactions and interresidue interactions, respectively. One parameter is used for contributions not included explicitly in this simple treatment.

The empirical intraresidue potential energies for three classes of amino acids, glycine, proline, and all other amino acids, are estimated (Miyazawa & Jernigan, 1982) from their frequency distributions in (ϕ,ψ) compiled from 20 protein crystal structures by Nemethy & Scheraga (1977). The empirical energy $f^{\mathrm{emp}}(\phi,\psi)$ for the conformation (ϕ,ψ) at each 10° grid point in (ϕ,ψ) space is obtained from their results as

$$f^{\mathrm{emp}}(\phi,\psi) = -0.6 \ln [g(\phi,\psi)] + \text{constant} \qquad (10)$$

where $g(\phi,\psi)$ is the probability with which (ϕ,ψ) of a residue is observed within the 10° square centered at (ϕ,ψ). Thermal energy, RT, is arbitrarily specified as 0.6 kcal mol^{-1} for converting probabilities to energies. A small arbitrary number has been added to the number of occurrences for all points on the (ϕ,ψ) grid, namely, 0.001 for proline and 0.01 for glycine and the others to assure 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^{-1} for glycine, proline, and the others, respectively, when the constant in eq 10 is adjusted so that the minimum value of f^{emp} is zero. The empirical potential energy for residues preceding proline, except glycine, is set to the maximum value of f^{emp} for $-140° \leq \psi \leq 40°$. These empirical energies crudely represent the dependence of intraresidue interaction energies on backbone conformations, as well as some contributions from neighboring residues.

Interresidue interaction energies within native conformational segments are greatly simplified by assuming attractive energies only for those residues in close contact in the crystal structure. Close contact between residues is defined on the basis of C^β–C^β distances rather than C^α–C^α distances, because the former are expected to be more sensitive to conformation. Attractive energies for contact are assumed to be the same regardless of residue pairs and are assigned for contacting residue pairs sequentially further apart than nearest-neighbor residues.

The total energy, relative to the random-coil state, of a native conformational region consisting of residues i through j, is assumed to have the form

$$F_{ij} = \sum_{k=i}^{j} f_k + n^c_{ij} f^c + (j - i + 1)\alpha \qquad (11)$$

where f_k is the difference in conformational free energy between the native and random-coil conformations of residue k. The conformational free energy of the native state has been calculated by preaveraging the Boltzmann factors over the four points on the 10° grid nearest to the (ϕ,ψ) value of the kth residue in the crystal structure, except for both terminal residues. The latter are assumed to be able to take any value of ϕ for the N-terminal residue and ψ for the C-terminal residue. The conformational free energy of the random-coil state has been obtained by subtracting the contribution of the native state from the sum of the Boltzmann factors over all 10° grid points of (ϕ,ψ). n^c_{ij} is the number of contacts in the native conformational part, defined as the number of residue pairs, except nearest-neighbor residues, whose distances are less than or equal to r^c. This cutoff distance r^c is used to define contacts, and f^c, the contact energy, is an attractive energy that favors the native conformation. The mean field parameter α may include contributions from interresidue interactions, volume exclusion, solvent interactions, and side-chain conformations as well as from conformational fluctuations of the native conformational regions. We assume, possibly simplistically, that these contributions are independent of the regions and sequence, i.e., α will be a constant.

The statistical weight for residues i through j within a single native conformational region is

$$L_{ij} = \exp[-F_{ij}/(RT)] \qquad (12)$$

In this formulation, three parameters, α, r^c, and f^c, are necessary to specify the statistical weights L_{ij}. The assumption that α is constant leads to probabilities and hence pathways independent of α. $Z(n)$ can be represented as

$$Z(n) = [Z(n)]_{\alpha=0} \exp[-n\alpha/(RT)] \qquad (13)$$

Free energies $\mathcal{F}(n)$ and probabilities $Q_{ij}(n)$ are given by

$$\mathcal{F}(n) = [\mathcal{F}(n)]_{\alpha=0} + n\alpha \qquad (14)$$

$$Q_{ij}(n) = \frac{\partial \ln Z(n)}{\partial \ln [L_{ij}]_{\alpha=0}} = \frac{\partial \ln [Z(n)]_{\alpha=0}}{\partial \ln [L_{ij}]_{\alpha=0}} \qquad (15)$$

The above equations show that the probabilities, $Q_{ij}(n)$, do not depend on α even though the free energies are linearly dependent. It is obvious from these equations and eq 6 and 8 that $P_{ij}(n)$, the combinatory entropy and other averaged quantities for fixed n, do not depend on α. Therefore, the effect of α on folding pathways is not expected to be significant, when they are constructed on the basis of probabilities for fixed n.

We will examine the dependences of free energy along the folding coordinate and the probable folding pathways on the two parameters r^c and f^c.

Results

The atomic coordinates of bovine pancreatic trypsin inhibitor (PTI), bovine pancreatic ribonuclease A (RSA), hen egg white lysozyme (LYZ), and sperm whale apometmyoglobin (MBN), are taken from the Brookhaven Protein Data Bank [coordinates used were those of 3PTI, 2RSA, 2LYZ, and 1MBN (Bernstein et al., 1977)]. We examine the effects of varying the contact energy f^c on $\mathcal{F}(n)$ and $P_{ij}(n)$ with a cutoff distance, r^c, of 6.5 Å. The results will be presented for several molecules and discussed in detail for pancreatic trypsin inhibitor. The effects of varying r^c have been examined only for PTI. In this case, use of 10 Å for r^c changed principally $\mathcal{F}(n)$, but the most probable conformations were little affected; consequently only the value of 6.5 Å is employed for r^c in calculations for all other proteins.

The value of f^c at $\alpha = 0$ has been determined for each protein so that the melting temperature falls within the range of thermal energy, $0.6 < RT < 0.7$. All energies in this paper are expressed in units of kilocalories per mole. The values of f^c found are -2.0 for PTI and RSA and -1.8 for LYZ and MBN. The effects of varying f^c are examined for the range of values up to -1.0 in increments of 0.2 with different *values of α chosen by fixing the melting temperature*. The melting point is taken approximately as the point of equality in the free energy minima on the two sides of the transition, which can be regarded as the folded and unfolded states. For an

Table I: Molecular Energy Parameters[a]

	no. of						
protein	residues	contacts	ratio	melting RT	α	$\overline{F_{ii}}$	f^c
PTI	58	99	1.7	0.665	0.0	3.48	-2.0
					-0.35	3.14	-1.8
					-0.70	2.78	-1.6
					-1.06	2.43	-1.4
					-1.42	2.06	-1.2
					-1.82	1.67	-1.0
RSA	124	223	1.8	0.639	0.0	3.61	-2.0
					-0.36	3.25	-1.8
					-0.73	2.88	-1.6
					-1.10	2.51	-1.4
					-1.48	2.13	-1.2
					-1.88	1.74	-1.0
LYZ	129	258	2.0	0.680	0.0	3.60	-1.8
					-0.40	3.20	-1.6
					-0.81	2.79	-1.4
					-1.22	2.38	-1.2
					-1.65	1.96	-1.0
MBN	153	287	1.9	0.695	0.0	3.40	-1.8
					-0.37	3.02	-1.6
					-0.76	2.64	-1.4
					-1.15	2.25	-1.2
					-1.55	1.85	-1.0

[a] Units are kilocalories per mole.

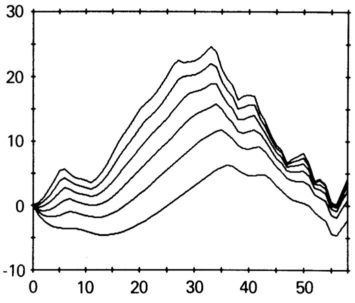

FIGURE 1: Free energies in kilocalories per mole for trypsin inhibitor (PTI) at the number of native residues, n, indicated on the abscissa. Curves are for the sets of parameters in Table I, in order from top to bottom, for f^c = -2.0, -1.8, -1.6, -1.4, -1.2, and -1.0.

all-or-none transition, the native and denatured states can be statistically defined as an average over the neighborhood of the respective free-energy minimum. At melting, α and f^c roughly satisfy

$$\sum_{i=1}^{N} f_i + N\alpha + N^c f^c \simeq 0 \qquad (16)$$

where N^c is the total number of contacts in the native structure. This equation is obtained by equating the free energies of the completely folded state and the completely unfolded state. The exact values of the parameters are given in Table I together with the number of contact pairs for r^c = 6.5 Å. Also it can be seen there that the ratio of the number of contacts to the total number of residues is relatively invariant. The $\overline{F_{ii}}$ is the average over all residues of the intraresidue energy, F_{ii}, given by eq 11. The range of values of f^c can seriously be questioned; however, it should be pointed out that it spans the range suggested for contact energies, -1.3 to -1.5, for hydrophobic energies per contact estimated from the equation (Janin & Chothia, 1979) $-3.9(N - 1.64N^{2/3})/N^c$ derived from calculated surface areas of protein crystal structures.

Characteristics of the Folding-Unfolding Transition. Free energies $\mathcal{F}(n)$ and combinatory entropies $S^{\text{comb}}(n)/R$ for each set of these parameter values of PTI are given for the total number of native residues in Figures 1 and 2. In Figure 3

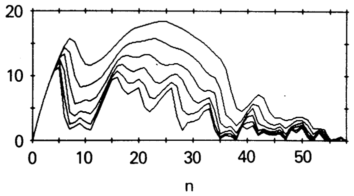

FIGURE 2: Combinatory entropies from eq 8 for various numbers of native residues, n, for PTI. Curves are for the sets of parameters in Table I, from top to bottom, for f^c = -1.0, -1.2, -1.4, -1.6, -1.8, and -2.0.

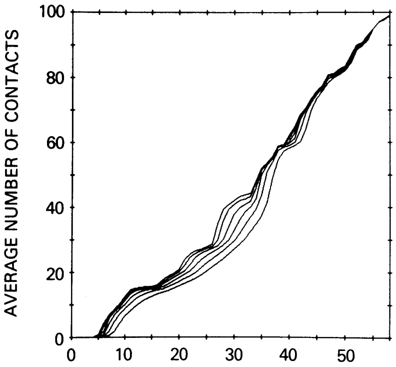

FIGURE 3: Average number of residue contact pairs for all numbers of native residues, n, for PTI. Curves are for the sets of parameters in Table I, from top to bottom, for f^c = -2.0, -1.8, -1.6, -1.4, -1.2, and -1.0.

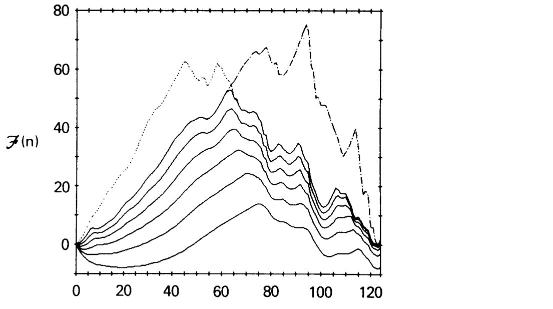

FIGURE 4: Free energies at all numbers of native residues, n, for ribonuclease A (RSA). Solid curves are for the sets of parameters in Table I, from top to bottom, for f^c = -2.0, -1.8, -1.6, -1.4, -1.2, and -1.0. The dotted curve and the dot-dash curves on the top are also for f^c = -2.0. In the dotted curve, residues 1-56 are fixed in their random-coil state, whereas in the dot-dash curve, residues 2-55 are fixed in their native state.

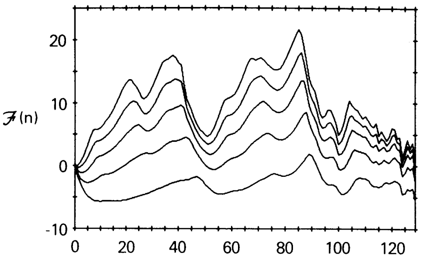

FIGURE 5: Dependence of free energies on the number of native residues, n, for lysozyme (LYZ). Curves are for the sets of parameters in Table I, from top to bottom, for f^c = -1.8, -1.6, -1.4, -1.2, and -1.0.

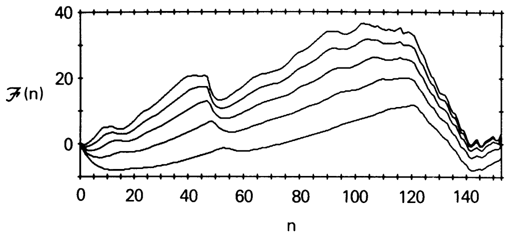

FIGURE 6: Free energies at each number of native residues, n, for apomyoglobin (MBN). Curves are for the sets of parameters in Table I, from top to bottom, for $f^{\mathrm{c}} = -1.8, -1.6, -1.4, -1.2$, and -1.0.

is shown the average number of contacts, calculated by using $P_{ij}(n)$, at each point of the folding, as specified by the number of native residues. Free-energy profiles for RSA, LYZ, and MBN are shown in Figures 4, 5, and 6, respectively.

In all proteins and for all values of f^{c}, an "all-or-none" character of the folding–unfolding transition is indicated by the appearance of large free-energy barriers which almost completely separate folded states from unfolded states. The major exception to this is lysozyme (Figure 5) which shows a free-energy profile with two deep intermediate minima; this would indicate the possibility of a three-step transition under some circumstances. The free-energy profiles of ribonuclease A (Figure 4) and apomyoglobin (Figure 6) stand in sharp contrast to one another. Ribonuclease A manifests multiple maxima on the native side of the principal maximum, whereas the highest maximum in apomyoglobin is shifted very far toward the native state with smaller maxima appearing on the denatured side.

Generally the activation energies increase, and the maximum free energy shifts toward the denatured side, as f^{c} decreases. The heights of the activation energies are intimately related to the cooperativity of the folding–unfolding transition; for higher activation energies, the all-or-none character of the transitions will be stronger and the transitions sharper. More negative values of f^{c} and less negative values of α correspond to strengthening the interresidue energies at the expense of introducing weaker intraresidue energies. Therefore, the folding–unfolding transitions are characteristically sharper for more negative values of f^{c}. A similar behavior would be observed in helix–coil transitions of finite chains if both s and σ were to be adjusted interdependently in a manner analogous to the interdependent adjustment of f^{c} and α: if s were increased and σ decreased to maintain the same melting temperature, then the transition would become sharper. One important distinction between helix–coil transitions and protein folding–unfolding transitions is that the range of interactions is significantly longer in the latter case and leads to sharper transitions, characteristically all-or-none in appearance. Indeed, it was shown with a one-dimensional lattice gas protein model (Wako & Saito, 1978a) that the longer the range of interactions, the sharper the transition. Although it would be possible to adjust the number of long-range interactions by choosing different r^{c}, we consider it more sensible to adjust f^{c}. The common feature of all free-energy curves, Figures 1, 4, 5, and 6, is that more favorable long-range interactions sharpen the transition. Let us consider the origin of this cooperativity.

As stated in a preceding section, if all protein conformations for fixed n have equal statistical weights, then the combinatory entropy $S^{\mathrm{comb}}(n)/R$ is given as $\ln \binom{N}{n}$. For the case of $L_{ij} = 1$, the free energy would be equal to $-TS^{\mathrm{comb}}(n)$, and the transition would be gradual and diffuse. Introduction of interresidue interactions leads to favoring definite conformations with specific combinations of residues in the native state which in turn reduces the combinatory entropy. Typically the high free energy of intermediate states originates in both the loss of combinatory entropy and the missing native interresidue interactions. When f^{c} and α are changed to maintain the melting condition in accordance with eq 16, the conformational free energy of any specific folding intermediate, with n native residues and n^{c} contacts, changes roughly by the amount

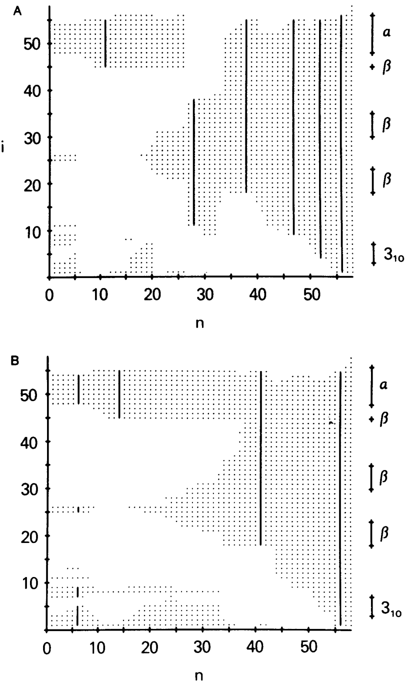

FIGURE 7: Locations of most probable native residues for PTI for all possible numbers of native residues, n. The presence of a dot indicates that the probability of residue i being native is greater than or equal to n/N, where N is the total number of residues. Free energy minima, shown in Figure 1, have been designated with solid lines instead of dots. Regular secondary regions are indicated by the labeled vertical bars on the right sides of the figures. Part A is for the set of parameters in Table I with the value of $f^{\mathrm{c}} = -2.0$ and part B is for $f^{\mathrm{c}} = -1.0$.

$$n\left(\frac{n^{\mathrm{c}}}{n} - \frac{N^{\mathrm{c}}}{N}\right)\Delta f^{\mathrm{c}} \qquad (17)$$

where Δf^{c} is the increment in f^{c}, n^{c}/n is the density of native contacts in the intermediate, and N^{c}/N is the density of contacts in the entire native molecule. For a folding intermediate that is less stable than the completely unfolded state, the intermediate's conformational free energy will increase as f^{c} becomes more negative, since $n^{\mathrm{c}}/n - N^{\mathrm{c}}/N$ is usually negative. Simultaneously, the combinatory entropy diminishes as interresidue interactions are more strongly favored relative to intraresidue interactions. An example is given in Figure 2. Conformations with more contacts at a given n become more probable as seen in Figure 3; that is, conformations with a single large native conformational region are usually more favorable than those consisting of several smaller native conformational regions. This effect can also be observed by comparing the dot plots in parts A and B of Figures 7–10. Consequently, making long-range interactions more favorable and short-range interactions less favorable increases the free energies $\mathcal{F}(n)$ of partially folded states by increasing the conformational free energy and decreasing the combinatory entropy. The small shift in the maximum toward the denatured side as f^{c} decreases also occurs because generally smaller

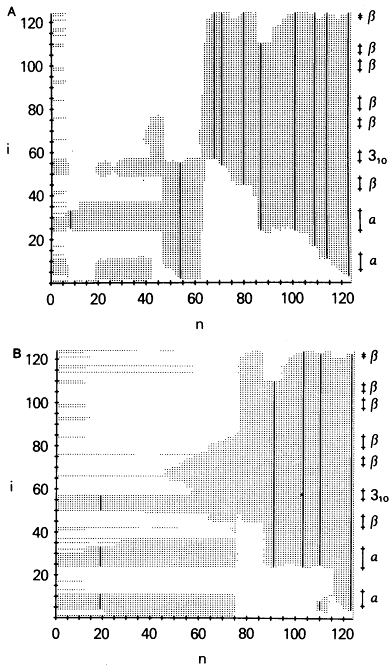

FIGURE 8: Locations, i, of most probable native residues for RSA for each number of native residues, n. Part A is for $f^c = -2.0$, and part B is for $f^c = -1.0$. See the legend to Figure 7 for details.

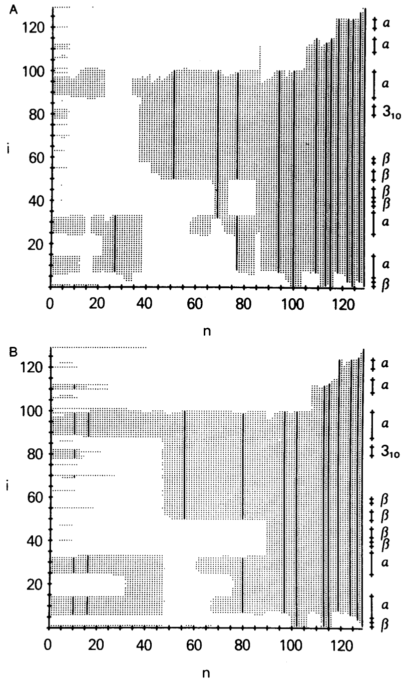

FIGURE 9: Locations, i, of most probable native residues for LYZ for all possible numbers of native residues, n. Part A is for $f^c = -1.8$, and part B is for $f^c = -1.0$.

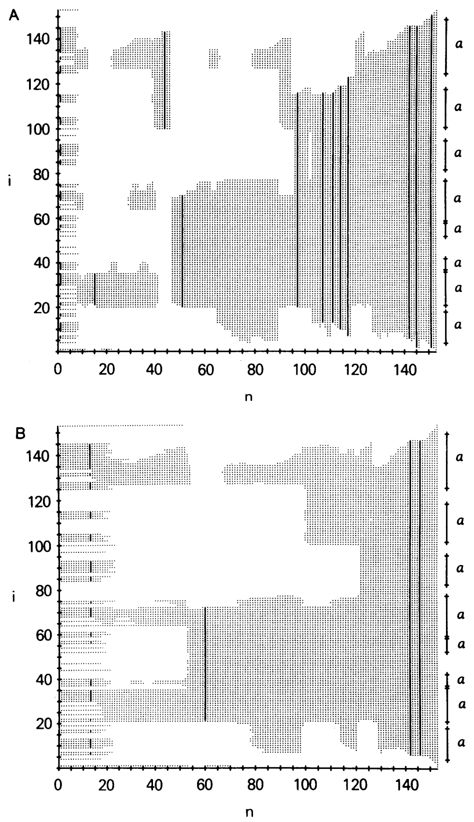

FIGURE 10: Locations, i, of most probable native residues for MBN for the number of native residues, n. Part A is for $f^c = -1.8$, and part B is for $f^c = -1.0$.

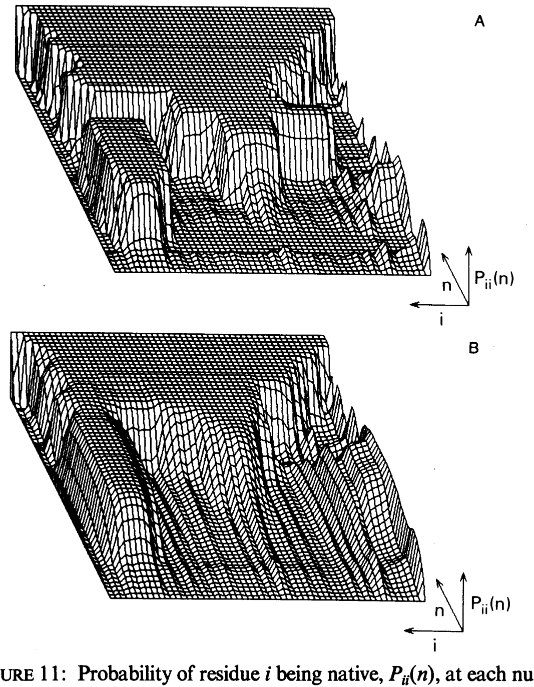

FIGURE 11: Probability of residue i being native, $P_{ii}(n)$, at each number of native residues, n, for PTI. Part A is for the set of parameters in Table I with the value of $f^c = -2.0$, and part B is for the set with $f^c = -1.0$. Parts A and B correspond to parts A and B of Figure 7, respectively.

values of n^c/n are found on the denatured side.

Another interesting feature is the existence of several minima on the free-energy profiles; this might indicate the existence of detectable intermediates, except that these minima are usually significantly higher than the two minima corresponding to the native and denatured states. It is noteworthy that these subsidiary minima on the free-energy profiles usually occur near minima in the combinatory entropies; see Figures 1 and 2 for PTI. This implies that fewer conformations contribute to these states. Most probable conformations at such subsidiary free-energy minima may consequently be useful for conformational characterization along the folding–unfolding pathways.

Most Probable Conformations. The most probable combination of residues in the native conformational state at each value of the number of native residues is examined by means of $P_{ii}(n)$, the probabilities of the ith residue being native. The dependences of $P_{ii}(n)$ on residue position, i, and the number of native residues, n, are given for PTI in parts A and B of Figure 11 for the cases of $f^c = -2$ and -1, respectively. The probability surfaces are generally less steep for $f^c = -1$ than for $f^c = -2$. The steepness of the surfaces reflects the cooperativity of the transition.

The dot representations in Figures 7–10 provide convenient indications of most probable native residues at each stage of folding. Each dot in these figures represents residues with probabilities for the native conformational state greater than or equal to the ratio of the number of native residues to the total number of residues, i.e., n/N. Locations of free-energy minima have been designated with solid lines instead of dots.

In the unfolded range of small numbers of native residues, intraresidue energies and relatively short-range interresidue energies determine most probable conformations because of limits to the maximum size of a single native conformational region; the maximum range of interresidue interactions is restricted by the total number of native residues. α helices and turns are likely candidates for most probable native conformational regions at small numbers of native residues, because they are characteristically stabilized by relatively short-range interactions. Generally, it is found in Figures 7–10 that free-energy minima in this domain of small numbers of native residues correspond to formation of helices or turns. The α-helix conformation is one of the intrinsically favored conformations for the empirical potential energies used and becomes even more stable as intraresidue energies are favored. In the case of $f^c = -1$, folding–unfolding transitions at melting occur between almost completely native conformations and denatured conformations in which small numbers of residues are still in their native conformational states (see Figures 1 and 4–6). Native conformational parts within such unfolded conformations are usually helical; however, it must be noted that all probable native residues indicated by dots cannot be accommodated simultaneously within a single molecule. As an example, see the incipient helices in apomyoglobin in Figure 10B; for the minimum at $n = 13$, the total number of dots is significantly larger than 13. All cannot be accommodated in a single molecule but instead are distributed among several molecules.

By contrast, β sheets are stabilized by significantly longer range interactions than α helices. Consequently, β sheets appear in a more cooperative step than α helices, because of the requisite longer range interactions. Formation of pairs of β strands becomes probable only at relatively large numbers of native residues, which can accommodate such complete structures. In Figures 7–9, long-range β-strand interactions appear suddenly and not gradually. It is noteworthy that, in contrast to helices and turns, β-sheet formation characteristically occurs near the maxima in the free-energy profiles.

Sharp changes in the most probable conformations along the axis of the number of native residues are also attributable to the strong cooperativity of native formation originating in the long-range interresidue interactions. A typical case is for RSA with $f^c = -2$ (Figure 8A). Ribonuclease A is composed of an N-terminal region consisting principally of helical parts and a C-terminal region of β strands. The N-terminal region assumes the native conformation with higher probability than the C-terminal for small numbers of native residues. A catastrophic event occurs at the number of native residues which is just large enough to accommodate the complete native conformation of the C-terminal region. Following this, as the number of native residues increases, the C-terminal native structure grows by adding helical parts on the N-terminal side, step by step. This sequence of events becomes most apparent whenever the long-range interresidue interactions are strongly favored as in Figure 8A.

As evidenced in Figure 5, the free-energy profiles of LYZ usually exhibit two deep intermediate minima in addition to those for the native and denatured states. Figure 9 shows that a middle part consisting of residues 50–100 is the most native region at the position of the first deep intermediate minimum located between 50 and 55 total native residues. An N-terminal native region consisting of residues 1–100 characterizes the other deep intermediate minumum. This can be understood easily by inspecting the C^β–C^β contact map of lysozyme. Except for the contacts near the diagonal, the map can be divided into three dense regions: contacts among residues 1–49, those among residues 50–100, and those of residues 1–40 with residues 80–129. Because the second region is the densest in contacts, it appears first. The second intermediate minimum in the free-energy profile corresponds to the appearance of the contacts of the first two regions. These two partially folded conformations at the intermediate minima are relatively stable intermediates on the folding–unfolding pathway for this equilibrium model of lysozyme.

Effects of r^c. Only two values, 6.5 and 10 Å, for r^c, the maximum separation for defining contact residue pairs, have been used to examine the effect of r^c on $\mathcal{F}(n)$ and $P_{ij}(n)$. These calculations have been performed only for PTI. For 10 Å, the number of contacts is 378 which is about 3.8 times larger than for $r^c = 6.5$ Å. The value of f^c for this case is determined by maintaining the same value of the total interresidue energy for the completely folded conformation; that is, f^c for $r^c = 10$ Å is given as its value for $r^c = 6.5$ Å multiplied by the ratio of the number of contacts with $r^c = 6.5$ Å to that for $r^c = 10$ Å, i.e., 99/378. Surprisingly, free energies are about 1.7 as large as those for $r^c = 6.5$ Å; however, the general characteristics of the free-energy profile are preserved with the maximum slightly shifted toward the denatured side. This is explained by the fact that the relative emphasis given to long-range effects is increased; the portion of the total contacts between residues greater than six residues apart is 58.6% for $r^c = 6.5$ Å and 64.3% for $r^c = 10$ Å, which indicates the characteristics of interresidue interactions to be substantially longer range in the latter case. Therefore, the activation energy increases, and the transition becomes sharper. However, the most probable conformations do not differ appreciably from those for $r^c = 6.5$ Å. Hence, the exact value of r^c does not appear to be so critical in the determination of most probable conformations or pathways.

Folding–Unfolding Pathways. In Figure 12, most probable native conformational fragments at representative points, usually free-energy minima, along the folding–unfolding coordinate between the native and denatured state minima are listed in order according to increasing numbers of native residues. The procedure consists of identifying the residue positions appearing as dots in Figures 7–10; regions including fewer than five contiguous residues are not shown. Punctuation in Figure 12 consists of (1) hyphens between pairs of numbers to specify terminal residues of single native conformational regions, (2) parentheses to designate those states not located at free-energy minima, (3) commas to separate native conformational regions which appear to be present in all molecules, and (4) semicolons to separate the native regions that cannot

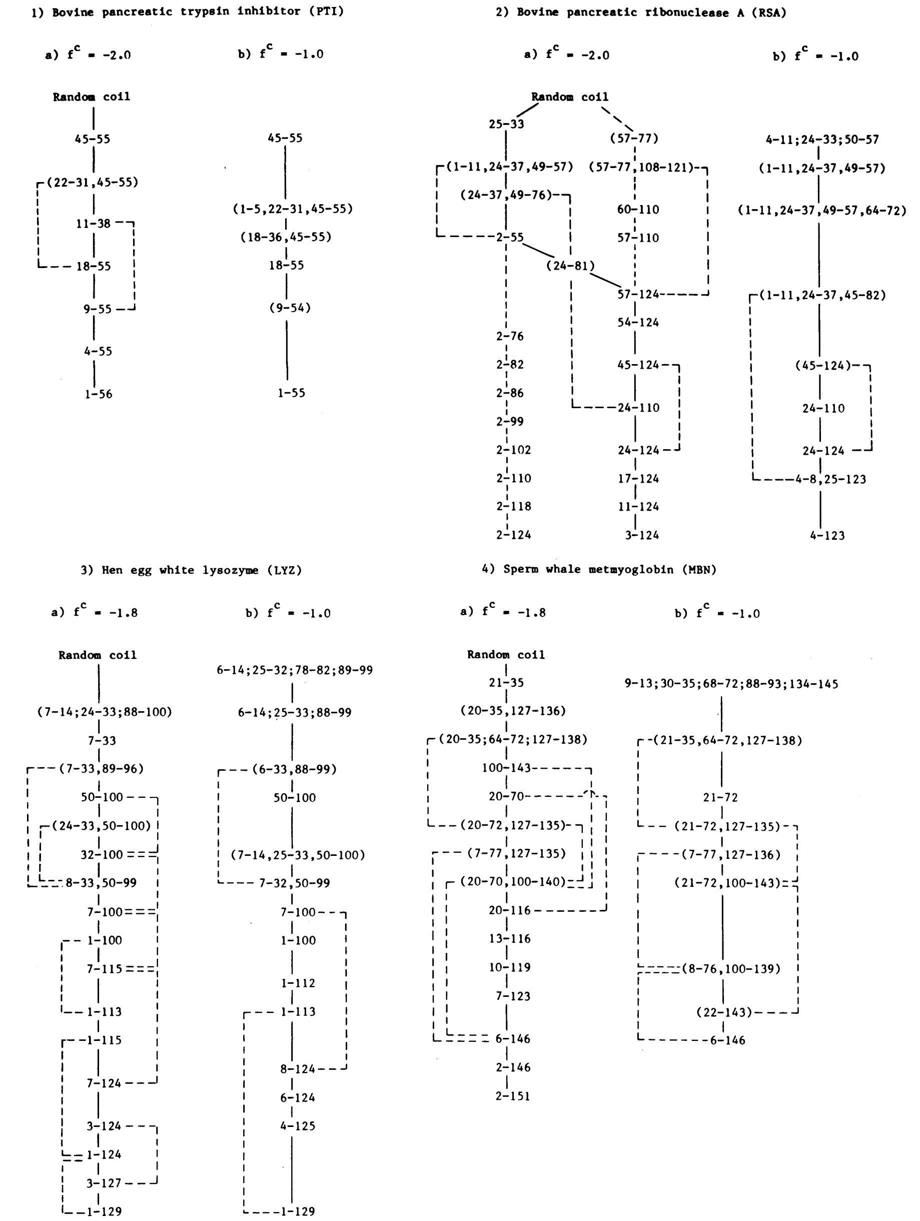

FIGURE 12: Probable folding pathways at equilibrium. Refer to Folding–Unfolding Pathways section for details. For RSA with $f^c = -2.0$, the lower left side corresponds to a path in which residues 2–55 are fixed in their native form, and the upper right side is for residues 1–56 fixed in random-coil form.

exist simultaneously within single molecules, corresponding to cases where the total number of dots significantly exceeds n.

If the number of native residues can serve as a simple one-dimensional folding–unfolding coordinate, then we might expect a pathway to consist of connections between the most probable conformations at each point along the axis. These probable pathways could be constructed simply by connecting most probable conformations in order of increasing numbers of native residues for folding, or decreasing numbers for unfolding. Because the present considerations correspond to equilibrium, such pathways must be reversible. Direct connections of probable conformations along the folding coordinate, to be designated pathways, are shown by solid lines in Figure 12. There are points where the most probable conformations are not related to their predecessors by simple growth, either by extending ends of regions or by merging separate regions while maintaining the previous result, but appear discontinuously. It is unclear whether such steps on pathways are probable or not, since large activation energies are often required for these discontinuous conformational transitions. These discontinuities appear as a manifestation of several pathways, for each of which the corresponding probable conformations appear at separated points along the folding coordinate. For such cases, the alternative pathways indicated by dotted lines in Figure 12 might be possible. These partially masked pathways are derived by choosing successive folding intermediates that include the previously native residues, with the occasional exception of a few terminal residues. It is, of course, possible to calculate free-energy changes accompanying the conformational changes which correspond to these dotted lines; they can be estimated crudely by fixing, in native form, the residues included in the less native state. The case of $f^c = -2$ for RSA is especially interesting because the most probable conformations change catastrophically near the maximum in free energy (see Figures 4 and 8A). Free energies and probabilities $P_{ii}(n)$ have been calculated with $f^c = -2$ for two cases in which specific regions are fixed in either the native or random-coil state. In the dotted curve in Figure 4, residues 1–56 are taken in the random-coil state, and in the dot–dash curve, the other residues 2–55 are fixed in their native state. These free-energy curves and $P_{ii}(n)$ indicate that if the folding is forced to begin in the C-terminal half, native conformational regions appear at 60–75 and 108–120; the activation energy of that process is lower than for the process in which the native conformational part in the N-terminal half grows toward the C-terminal. This method provides an alternative way to choose pathways for cases such as RSA that display large conformational discontinuities along the folding coordinate.

Discussion

The most probable folding–unfolding pathways have been presented with the noninteracting globule-coil model. These pathways characteristically represent growth followed by merging of separate native regions. One alternative is the diffusion–collision–coalescence model proposed by Karplus & Weaver (1976).

In that model, protein folding proceeds by random collisions of pairs of native nuclei, followed by their coalescence into a larger native structural entity. In other words, interactions between native conformational parts, which are completely neglected in the present noninteracting globule-coil model, are emphasized and are assumed to lead to a coalescence of native regions. It is possible that a collision–coalescence process may not be diffusion limited but instead behave like an activated process; this would be consistent with experimental observations of the relationship between relaxation times and viscosity (Tsong & Baldwin, 1978; Tsong, 1982). The activation energy in this collision–coalescence process originates principally from the entropy loss involved in bringing two small native parts into a specific relative position and orientation, i.e., the reduction in entropy accompanying the formation of the loop. Preliminary investigations with simple Gaussian models indicate that the growth–merge process is more probable, except for small native nuclei when the collision–coalescence process could become effective. The results of more detailed Monte Carlo generations (Miyazawa & Jernigan, 1982) of equilibrium conformations of PTI are consistent with that conclusion; in a general way, long-range native contacts form only following the appearance of shorter range native contacts. A minor exception occurs in the denatured state where some random contacts, both short and long range, appear; however, subsequently these usually disappear and do not lead directly to folded states. Also, Abe & Go (1981) demonstrated in their Monte Carlo simulations of a two-dimensional lattice protein that the interactions between native conformational parts are negligible except at an early stage of folding. Although it is not possible to arrive at a definitive conclusion here, the growth–merge process appears to be highly probable, especially through the later stages of folding.

In usual applications of both the collision–coalescence and the growth–merge models, protein folding is assumed to proceed with the growth of native conformational parts. The major evidence to oppose this viewpoint consists of the experimental results of Creighton (1978), who has reported that native S–S bonds form only after reshuffling of nonnative S–S bonds during the folding of PTI. These experiments may indicate that PTI folds by passing through intermediate nonnative conformations. An alternative interpretation might be that the nonnative forms detected are favored but do not lie directly on the folding pathway. If the nonnative conformers are actually on the direct folding pathway, then passing from such nonnative conformations to native conformations must be easier than the more direct process of gradually folding toward a native structure. It would appear to be possible for stable nonnative conformations to exist; however, it is difficult to say whether or not such nonnative conformations could be transformed directly into native conformations. Lim (1980) proposed the S-helix hypothesis of protein folding in which S helices could be candidates for such nonnative conformations. In his model, even β sheets would be formed in transitions from S helices; however, it is not clear whether such transitions could be easily realized. Ptitsyn & Finkelstein (1980) have proposed that single hairpinlike structures between two extended strands are precursors of β sheets. Even for β-sheet formation, these precursors could have substantial nonnative character. Further studies are required to provide further interpretation and understanding of Creighton's experiments. The present noninteracting globule-coil model can represent only nativelike intermediates; including nonnative conformations is significantly more difficult (Saito, 1981).

Although numerous details of folding pathways for the proteins considered are available in the figures, we choose not to discuss the results in further detail because of the absence of detailed information for comparison. Only for trypsin inhibitor (PTI) is information available from more detailed calculations. The folding pathways found here for PTI are consistent with those obtained in a Monte Carlo generation (Miyazawa & Jernigan, 1982): (1) A β sheet forms at the point of highest free energy. (2) The most probable folding pathway shown in Figure 12 is the same as the pathway with growth toward the C-terminal following β-sheet formation. However, the other folding pathway observed in the Monte Carlo samples, which corresponds to folding instead toward the N terminus subsequent to formation of the β sheet, has not been detected with the present noninteracting globule-coil model. The origin of this difference is not clear. There are sampling errors in the Monte Carlo method; however, several classes of interactions are completely neglected.

For all proteins, *the most general feature observed here is the coexistence of only a few dominant native conformational domains at all intermediate stages of folding.* This could be attributed to the contact energy parameters which strongly favor native conformations. Restricted numbers of conformations could arise for two reasons: (1) the intramolecular energies of a few conformers are strongly favored or (2) there are very few tightly packed forms, as required by solvent conditions. The exact contributions of long-range interactions to protein folding are unknown. However, the range of values used here for the contact energy parameters may be realistic since it spans the range of hydrophobic energy gains for protein folding estimated by Janin & Chothia (1979). More realistic treatments would require a detailed evaluation of the interaction energies within native globules.

It has been our intention to demonstrate procedures for cataloging all equilibrium conformations, from the native to the denatured. We have shown that it is possible to characterize all stages of folding and to construct plausible folding pathways.

Added in Proof

In a recent investigation of the effect of heme contacts in myoglobin (*Biopolymers*, in press), we have observed a strong sensitivity of intermediates to the details of the contact map.

References

Abe, H., & Go, N. (1981) *Biopolymers 20*, 1013–1031.

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.

Creighton, T. E. (1978) *Prog. Biophys. Mol. Biol. 33*, 231–297.

Flory, P. J. (1969) *Statistical Mechanics of Chain Molecules*, Chapter VII, Interscience, New York.

Flory, P. J. (1972) *Ciba Found. Symp. 7*, 109–124.

Go, N., & Abe, H. (1981) *Biopolymers 20*, 991–1011.

Go, N., Abe, H., Mizuno, H., & Taketomi, H. (1980) in *Protein Folding* (Jaenicke, R., Ed.) pp 167–181, Elsevier/North-Holland Biomedical Press, Amsterdam.

Janin, J., & Chothia, C. (1979) in *Protein: Structure, Function and Industrial Applications*, pp 227–238, Pergamon Press, Oxford.

Karplus, M., & Weaver, D. L. (1976) *Nature (London) 260*, 404–406.

Lesk, A. M., & Rose, G. D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 4304–4308.

Lim, V. (1980) in *Protein Folding* (Jaenicke, R., Ed.) pp 149–166, Elsevier/North-Holland Biomedical Press, Amsterdam.

Miyazawa, S., & Jernigan, R. L. (1982) *Biopolymers 21*, 1333–1363.

Nemethy, G., & Scheraga, H. A. (1977) *Q. Rev. Biophys. 10*, 239–352.

Ptitsyn, O. B., & Finkelstein, A. V. (1980) *Q. Rev. Biophys. 13*, 339–386.

Saito, N. (1981) Abstracts of the VII International Biophysics Congress, Mexico City, Aug 20–28.

Tanaka, S., & Scheraga, H. A. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 3802–3806.

Tsong, T. Y. (1982) *Biochemistry 21*, 1493–1498.

Tsong, T. Y., & Baldwin, R. L. (1978) *Biopolymers 17*, 1669–1678.

Wako, H., & Saito, N. (1978a) *J. Phys. Soc. Jpn. 44*, 1931–1938.

Wako, H., & Saito, N. (1978b) *J. Phys. Soc. Jpn. 44*, 1939–1945.