Sequential Collapse Model for Protein Folding Pathways

Fernando Bergasa-Caceres,[†] T. Andrew Ronneberg, and Herschel A. Rabitz*

Department of Chemistry, Princeton University, Princeton, New Jersey 08544

Received: January 6, 1999; In Final Form: June 28, 1999

A model of protein folding *pathways* is presented based on the entropy loss induced by loop closure and the protein−solvent interaction. The model is based on a simple physical picture that depends only on knowledge of the primary sequence to reproduce many experimental results. The sequential collapse model (SCM) predicts the sequence of folding events observed in the proton exchange experiments on the folding pathway of apomyoglobin, cytochrome *c*, and barnase. The SCM addresses the *mechanism* of the protein folding process by building a sequential picture of the stabilization free energy of folding and the activation barriers that the protein encounters along a sequential folding pathway towards the free energy minimum that prescribes the native structure. The paper (a) describes the theoretical foundations of the SCM, (b) discusses the intermediate states of the folding pathway in the model, and (c) presents detailed results for several globular proteins.

1. Introduction

Proteins spontaneously fold into compact and unique structures,[1] and all the information needed for the protein to fold into its native structure is contained in the primary sequence.[2] Since the pioneering work of Wu, Mirsky, and Pauling,[3,4] intense experimental and theoretical efforts have attempted to understand the mechanism by which proteins attain their native structure. Protein folding is generally believed to be under thermodynamic control because proteins often fold into the same native structure under a wide range of conditions.[5] However, Levinthal showed theoretically that proteins cannot fold by random search of conformation space within biologically relevant time scales,[6] thus implying the existence of preferred folding pathways. These two observations suggest the possibility of approaching the protein folding problem from two different and complementary perspectives: (a) understanding how the amino acid sequence determines the native folding pathway; (b) searching for the free energy ground state of the protein−solvent system, which is assumed to prescribe the native structure. A comprehensive theory of protein folding would be expected to include elements from both approaches. Such a theory is not yet available, and the model presented in this paper is almost exclusively concerned with approach a. The purpose of this effort is the determination of the sequence of events along the folding pathway of globular proteins of limited size at low resolution.

Proton exchange experiments have resolved the folding *pathways* of a number of proteins at millisecond time scales, including cytochrome *c*, apomyoglobin, and barnase.[7−11] Resolution of submicrosecond time scales has recently been reported using rapid mixing methods which have shorter dead times than conventional stopped and quenched flow techniques.[12,13] The photochemically induced ligand unbinding of cytochrome *c* as well as the folding of redox proteins upon electron injection have also been used to probe fast reaction kinetics.[14,15] Furthermore, temperature jump experiments coupled with infrared absorption monitoring have achieved submicrosecond resolution for RNAase A[16] unfolding, while temperature jump experiments monitored with tryptophan fluorescence have also been applied to the submicrosecond collapse of apomyoglobin.[17]

Although the folding pathway of cytochrome *c*, apomyoglobin, and barnase might not reflect the folding pathway of all globular proteins, they show three unexpected similarities: (1) the folding pathways are sequential; (2) long-range contacts are involved in the earliest resolved folding stages (milliseconds); while (3) later folding events occur on time scales up to seconds. Observation 1 seems at odds with the view that protein folding occurs by a large ensemble of coexisting intermediate pathways descending down a free energy funnel toward the minimum free energy state.[18−20] Instead, these results suggest that physical constraints exist on the possible folding intermediate states, such that perhaps only a very small number of folding pathways are traversed by the protein to attain its native structure.

A viable model of the folding pathway needs to explain the observed sequentiality, the observed initial long range contacts, and the experimental separation in time scales between early and later folding events. Models such as diffusion−collision[21] and framework[22] consider the folding pathway to be controlled to a significant degree by the propensities of localized regions to form secondary structure. Therefore, these models predict that small differences in the secondary structure propensities will alter the sequence of events along the folding pathway. Some success has been attained using such an approach for apomyoglobin in the diffusion−collision context,[23] but thus far, this result has not been extended to other proteins. This paper will show that the SCM provides a physical rationale for why the pathway should be sequential and include early long-range events. The SCM argues that the entropy loss associated with the restriction of the available torsional space for the amino acid side chains upon loop closure favors the formation of substantial contacts between residues at a few characteristic distances along the amino acid sequence.

The SCM relies on a simplifying theme based on the fact that proteins with vastly different primary sequences often fold into essentially the same tertiary structure.[24] On the basis of this observed robustness of the folding code, the SCM will only be concerned with the polarity of each residue rather than the

[†] Current address: ENDESA C/Príncipe de Vergara 187, 28002 Madrid, Spain. E-mail: FBergasa@Endesa.Es.

10.1021/jp990090y CCC: $18.00 © 1999 American Chemical Society
Published on Web 10/16/1999

residue's atomic structure. The present form of the SCM does not attempt to predict detailed atomic scale structure; the SCM just aims to deal with the coarse scale mechanistic folding steps on the way (i.e., the pathway) to the final structure. This limited goal permits the use of a simple model that is still predictive. In this regard, various degrees of detail are included in other folding models as the complexity of dealing with the full atomic resolution is computationally overwhelming. Significant success has been attained in determining tertiary structures using simplified representations of amino acids and interactions in the Hierarchical Condensation context.[25,26]

The goal of this paper is to present the physical basis of the SCM and then utilize the model to correlate with recent folding pathway experiments. Section 2 will present the foundations of the SCM. Section 3 will implement the SCM in an algorithm to predict the sequence of folding events in the pathway of several proteins and compare these results with laboratory data whenever available.

2. Sequential Collapse Model (SCM) for the Folding Pathway

In the SCM, the folding pathway is divided into two well-defined stages: (1) the early contact formation phase; (2) the cooperative collapse phase. This division will be shown below to arise naturally from the model. In the early contact formation phase, one or a few initial contacts are established governed by the thermodynamic consequences of loop closure and the hydrophobic effect. In the cooperative collapse phase, the native structure is attained cooperatively through a hydrophobic collapse process.

2.1. Early Contact Formation Phase. The protein folding process is likely to be initiated by the formation of one or a few early contacts between segments not necessarily in immediate vicinity along the sequence. This picture is fundamental to diffusion−collision. The notion of a loop formed by a stable intrachain contact is central to the SCM. The actual loop will likely be a highly fluctuational object by virtue of the contact's associated thermodynamic effects as well as chain dynamics.

Upon formation of a successful early contact there will be a free energy change ΔG_{cont} which may be decomposed as

$$\Delta G_{\text{cont}} = \Delta G_{\text{loop}} + \Delta G_{\text{hyd}} + \Delta G_{\text{sec}} + \Delta G_{\text{tert}} \qquad (1)$$

where ΔG_{loop} is the free energy change associated with loop closure and is expected to be positive because loop formation defines a state that has fewer conformational possibilities than an open protein chain, thus inducing a large entropic loss ΔS_{loop}. The term ΔG_{loop} could include an enthalpic contribution ΔH_{loop} due to nonspecific enthalpic interactions between neighboring side chains in the loop not involved in the contact that defines the loop. This contribution is expected to be small before water is excluded from the protein surface and native contacts are established, and thus, it will be neglected here. The free energy change ΔG_{hyd} is associated with the release of clathrate-like water structures from the surface of the hydrophobic residues forming the contact into the bulk and is expected to be negative,[27] ΔG_{sec} is the enthalpic free energy change associated with the formation of stable secondary structure in the contact region and is expected to be negative,[28] the entropic free energy change associated with the formation of secondary structure in the contact is included in ΔG_{loop}, and ΔG_{tert} is the free energy change induced by the establishment of tertiary enthalpic interactions such as salt bridges, disulfide bonds, and tertiary hydrogen bonds and is expected to be negative.[1] Most tertiary enthalpic interactions will likely be established in late stages of the folding process or when native-like contacts have already been established.[29] This is equivalent to the hydrophobic effect and secondary structure propensities being largely responsible for the stabilization of early contacts. This view in the SCM is common with the diffusion−collision picture. Thus, it is possible to write eq 1 approximately as

$$\Delta G_{\text{cont}} \approx \Delta G_{\text{loop}} + \Delta G_{\text{hyd}} + \Delta G_{\text{sec}} \qquad (2)$$

for the early contact formation phase.

Early successful contacts will form when $\Delta G_{\text{cont}} < 0$ and significantly greater than kT in magnitude. The only term that opposes the early contact formation process is ΔG_{loop}, and ΔG_{loop} should have only a limited dependence on the amino acid sequence of the loop and the contact segments because all amino acids, except for proline and glycine, have basically the same available torsional space.[30] Although the accessible ϕ and ψ angles for glycine and proline are significantly different from the other amino acids, this difference will be neglected in the present model. We will show that the approximation still allows for reliable predictions of the folding pathway for the proteins analyzed at the limited level of resolution sought for here. A more refined model of the torsional freedom would be necessary to arrive at a precise presentation of the native structure of the folded molecule. It is expected that ΔG_{loop} will be strongly dependent on the number of amino acids, n, forming the loop. The balance between the three terms of eq 2 generates two physically distinct situations depending on the sequence of the contact segments and the length of the loop that successful contact formation would define:

$$\text{situation 1: } \Delta G_{\text{loop}} < -(\Delta G_{\text{hyd}} + \Delta G_{\text{sec}})$$

$$\text{situation 2: } \Delta G_{\text{loop}} > -(\Delta G_{\text{hyd}} + \Delta G_{\text{sec}})$$

In situation 1, successful contact formation is possible depending on the absolute value of $(\Delta G_{\text{hyd}} + \Delta G_{\text{sec}})$; in situation 2, however, successful contact formation is not possible. The dependence of ΔG_{loop} on n is crucial in order to determine the early steps of the folding process.

In the SCM, the free energy ΔG_{loop} associated with formation of a loop will be shown to determine preferred distances along the protein chain for early contact formation. These preferred distances arise as a consequence of the dependence of ΔG_{loop} upon loop length n. ΔG_{loop} can be written as

$$\Delta G_{\text{loop}} \approx -T\Delta S_{\text{loop}} \qquad (3)$$

where ΔS_{loop} is the entropy change upon forming the loop

$$\Delta S_{\text{loop}} = k(\ln \text{g}_{\text{loop}} - \ln g_0) \qquad (4)$$

here g_{loop} is the number of accessible conformations in the configuration containing a loop of length n, and g_0 is the total number of accessible conformations of the open random protein chain. When the chain of N amino acids forms a loop, it will be composed of four distinct regions: c residues will be involved in the contact, n residues will form the loop, and m and l residues will form each of the tails such that $N = c + n + m + l$.

Quantitatively determining $\Delta S_{\text{loop}}(n)$ is difficult due to the self-avoiding nature of the protein backbone and the amino acid side chains and the fact that the dihedral angles ϕ and ψ determine both the relative orientation of the side chains and the backbone's overall configuration. Moreover, ϕ and ψ can take only a limited range of values due to steric hindrance between next neighbor side chains along the protein. In this paper, a simpler representation of the protein chain will be used: the protein backbone will be assumed to be a Jacobson−Stockmeyer polymer[31] with the side chains attached to each monomer and interacting sterically with each other.

The entropy loss upon loop closure in this model can approximately be split into two parts: the entropy loss due to steric hindrance between the side chains $\Delta S_{\mathrm{loop,s}}$, and the Jacobson−Stockmeyer entropy loss associated with the loss of backbone conformations upon loop closure $\Delta S_{\mathrm{loop,JS}}$:

$$\Delta S_{\mathrm{loop}} \approx \Delta S_{\mathrm{loop,s}} + \Delta S_{\mathrm{loop,JS}} \tag{5}$$

To obtain an expression for $\Delta S_{\mathrm{loop,s}}$, we define f to represent the number of accessible conformations per amino acid. Although the present model takes f to be independent of the type of amino acid, the value of f will depend on the location of the amino acid in relation to the protein loop structure. Amino acids that are part of the contact will have very limited conformational freedom denoted by the constant f_{c}. For amino acids in the tails, f is basically equivalent to the number of conformations accessible in the random coil, which is denoted by the constant f_0. For amino acids that are part of a loop, the conformational freedom f_n is a function of loop length n. This latter n dependence is due to the fact that amino acids in the loop have a restricted torsional space.

The loss of entropy upon forming a loop in eq 3 may now be written as

$$\Delta S_{\mathrm{loop,s}} = k(c \ln f_{\mathrm{c}} + n \ln f_n - (n + c) \ln f_0) \tag{6}$$

To determine $\Delta S_{\mathrm{loop,\ JS}}$, consider the number of conformations accessible to a loop of size n, g_n, estimated by a Jacobson−Stockmeyer treatment,

$$g_n = (\pi n l^2)^{-3/2} \nu \tag{7}$$

where l is the average monomer length and ν is the volume within which two monomers are said to be in contact. Thus, $\Delta S_{\mathrm{loop,JS}}$ can be written as

$$\Delta S_{\mathrm{loop,JS}} = k({}^{3}/_{2}(\ln n) - C) \tag{8}$$

where C is a constant identified from the factors in eq 7. Thus, the loss of configurational entropy ΔS_{loop} associated with forming a contact of fixed size c defining a loop of length n, up to a constant $C' = (c \ln f_{\mathrm{c}} - C)$, is

$$\Delta S_{\mathrm{loop}} \approx k(n \ln f_n - n \ln f_0 - {}^{3}/_{2} \ln n) \tag{9}$$

and the associated free energy loss ΔG_{loop} is given by substitution of eq 9 into eq 3.

The dependence of f_n on n is a key component of the SCM. The physical basis for this dependence is the restriction in the number of conformations accessible to the residues forming a loop. *Minimal* loops, defined as the smallest possible loop the protein chain can form, have very little conformational freedom in the torsional angles because of the strict restrictions on the dihedral angles necessary to reverse the direction of the protein chain under these conditions. An estimate of the minimal loop length, n_{min}, is $n_{\mathrm{min}} = n_{\mathrm{r}} + c$, where n_{R} is the minimal number of residues necessary to change the chain direction and c is the minimal number of residues necessary to establish a stable contact after the directional change is achieved. A reasonable observational value for n_{R}, based on the shortest observed turns in protein structures, is $n_{\mathrm{R}} = 4$ amino acids,[32] c should be sequence dependent but it is safe to say that a stable contact should involve at least a few residues. We assume that at least five amino acids on each segment are needed to form the contact, leading to $n_{\mathrm{min}} \approx 14$ residues.

The precise dependence of f_n on n as loop length increases from minimal loop length n_{min} is difficult to determine because of the self-avoiding nature of the protein backbone and the amino acid side chains. However, f_n is expected to be an increasing function starting at n_{min} and asymptotically approaching f_0 for large n. This behavior arises due to the fact that, as the loop gets larger, the accessible conformations for the amino acids in the loop increase and eventually approach those of the random coil.

The relationship between f_n and n is qualitatively depicted in Figure 1, while the resultant free energy loss ΔG_{loop} is represented in Figure 2. Although n is discrete, sufficient numbers of amino acids will be assumed to permit a continuous view for the purpose of model development. Considering the shape of f_n in Figure 1, a simple analysis of eq 9 shows that there will be a local minimum in ΔG_{loop}, at the shortest loop length for which $f_n \approx f_0$. The length where this condition is satisfied will be called the optimal length n_{op}. Loops formed by a contact at n_{op} are called primary loops. If f_{n} varies monotonically, there will be two other extrema in ΔG_{loop} vs n at smaller values of n, a minimum for $n = n_{\mathrm{min}}$ and a maximum between n_{min} and n_{op}, provided that $(\mathrm{d}^2 f_n/\mathrm{d}n^2) < 0$, as expected if f_n is monotonic and asymptotically reaches an equilibrium value. The minimum at n_{min} is given by the rapid rise in ΔG_{loop} at $n < n_{\mathrm{min}}$ due to the steric repulsion of the amino acids in the turn which would have to interpenetrate each other. This is the behavior depicted in Figure 2, which will be discussed in more detail below.

The minimum corresponding to region A of Figure 2 represents ΔG_{loop} for loops with length $n \approx n_{\mathrm{min}}$ where $f_n \ll f_0$. For loops slightly larger than n_{min}, the ratio f_n/f_0 will be small because only chain geometries in which the protein backbone is almost fully extended, except for the turn region, will exist that do not need the amino acid side chains to cross each other. The relatively small increase in the number of chain conformations as n increases causes ΔG_{loop} to rise for loops somewhat larger than minimal loops. This results in an increase in ΔG_{loop} represented by region B of Figure 2. As n continues to increase, the protein chain can deviate further from the open extended geometry because the constraints imposed by the steric interaction between the amino acid side chains are relaxed as more space becomes available. This domain is represented by region C in Figure 2. when n reaches the optimal length n_{op}, then $f_n \approx f_0$, and ΔG_{loop} is minimized. As n becomes larger than n_{op}, the Jacobson−Stockemeyer term dominates and ΔG_{loop} steadily increases. This domain is represented by region D of Figure 2.

Primary loops are favored over longer loops because longer loops have a larger loss of entropy due to the Jacobson−Stockmeyer term in eq 9 without a corresponding increase in f_n for the amino acids forming the loop. The absolute value of ΔG_{loop} at n_{op} cannot be easily estimated, but it is reasonable to expect it to be greater than the free energy loss associated with loop closure for minimal loops. This argument is based on the consideration that the number c of torsionally restricted amino acids in the contact should not depend on loop length, but a minimal loop will contain an extra n_{R} strongly constrained amino acids in the turn connecting the two segments forming the contact defining the minimal loop.

It is difficult to precisely estimate n_{op} without quantitative knowledge of the dependence of f_n on n. A first estimate of n_{op}

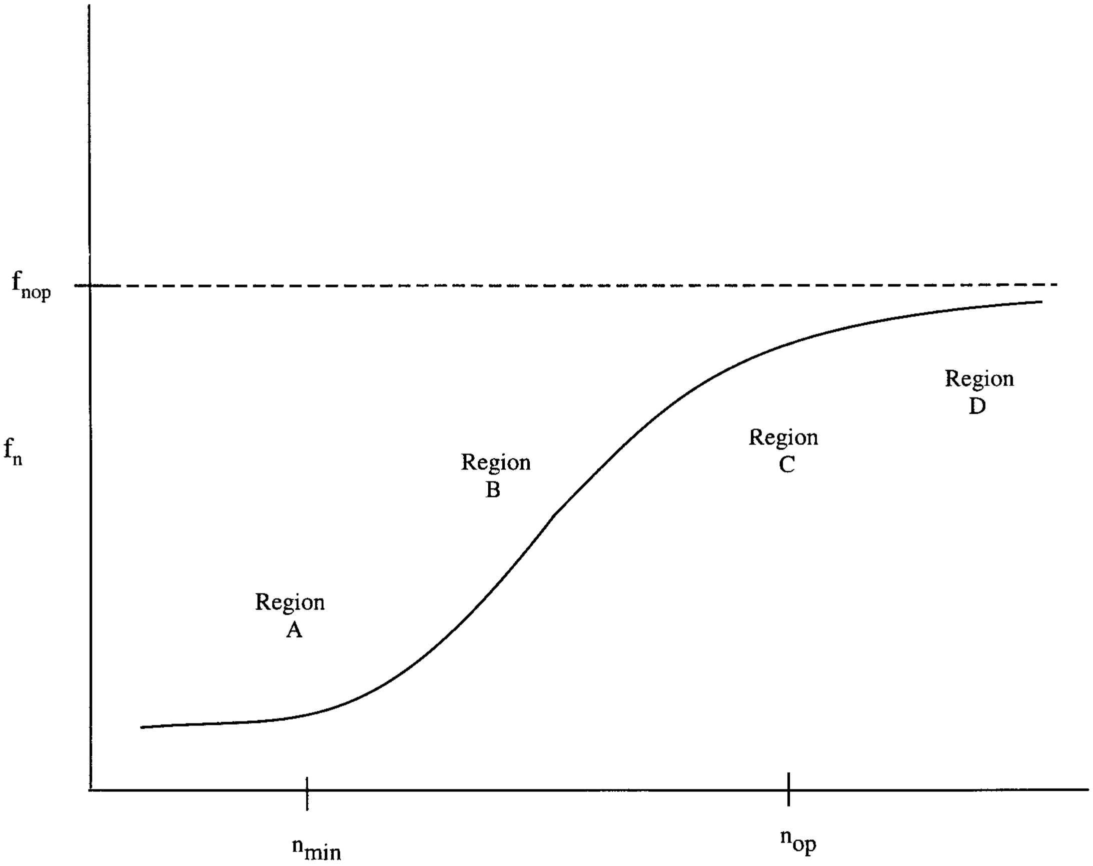

Figure 1. The conformational freedom f_n for residues in a loop is qualitatively depicted as a function of loop size n. The conformational freedom is expected to be smallest at minimal loop length n_{min}, and to increase monotonically as n increases. Region A corresponds to the conformational freedom of the minimal loops. Region B represents the conformational freedom accessible to loops somewhat larger than the minimal loop. Region C represents the conformational freedom for loops approaching the optimal length n_{op}. Region D is the domain in which the conformational freedom of the amino acids in the loop resembles that found in an open chain.

can be made by arguing that it should be the loop size for which all the amino acids in the loop could possibly take the most extreme torsional values of the Ramachandran plot without inducing unacceptable steric interactions between the side chains. It is expected that f_n for such a loop size should be approximately equivalent to f_0. This configuration is a helicoidal torus (more compact than α helical, closer to a π helix) with a rise per residue of ~1.1 Å and a radius of ~2.8 Å.[33] The maximum radius of the resulting geometry is the sum of ~8 Å for the average side chain length plus ~2.8 Å for the radius of the helix and ~2.8 Å to allow for a water layer attached to the surface of the hydrophobic residues. Thus, the number of residues needed to form a loop is $(\sim 27.2\pi)/1.1 \approx 79$ residues. This value is only a rough estimate, and the primary loop size n_{op} will be considered to have a range of at least 65−85 amino acids.

To summarize, the SCM suggests that, in the early folding stages, primary contacts are expected to form between segments of the protein chain not immediately adjacent along the sequence. These contacts tend to form at a generic distance, estimated here to lie between 65 and 85 amino acids, because of the need to minimize the free energy loss ΔG_{loop}, associated with loop closure. Early contacts are held together by the free energy gain associated with the release of the water on the surface of the contact forming segments into the bulk ΔG_{hyd} and the formation of secondary structure ΔG_{sec}.

In section 3, all these considerations result in a simple and practical algorithm. Implementation of the algorithm does not require precise values for the terms in eq 2, as the goal of the present model is to predict the sequence of folding events rather than the absolute time scale of the process or the thermodynamic stability of the intermediate states.

2.2. Later Folding Stages. *a. Cooperative Collapse.* After the primary contacts are established, the protein has a fluctuating open (i.e., not yet fully folded) primary loop. The folding of the open primary loop should happen on a larger time scale than the early contact formation phase because the folding of a primary loop must involve the temporary formation of loops shorter than the optimal length n_{op}, thereby generating intermediates with large $\Delta G_{\mathrm{loop}} > 0$. Thus, in general, $\Delta G_{\mathrm{loop}} \gg -(\Delta G_{\mathrm{hyd}} + \Delta G_{\mathrm{sec}})$ for the formation of individual contacts between protein segments included in the primary loop. This implies that the primary loop must fold through a mechanism of cooperative collapse in which most segments in the primary loop enter the folding process at once to maximize $-(\Delta G_{\mathrm{hyd}} + \Delta G_{\mathrm{sec}})$, most likely defining a collapsed intermediate held by a number of nonspecific contacts.

The tails of the protein will not be long enough for another primary loop to form in small proteins defined as having $(m + l) < n_{\mathrm{op}}$. Thus, they will fold through a short-range mechanism, involving formation of loops at the minimal length n_{min}.

b. Optimization of the Collapsed Primary Loop. In order for the primary loop to fold into a unique native conformation, there must be a sequence in which the protein segments form definitive native contacts within the cooperative collapse process. In the SCM, it is assumed that the need to release water from the protein surface for further folding generates activation barriers to the attainment of the native structure. These activation barriers induce a sequential ordering of folding events in the collapse of the primary loop. The rationale for there being

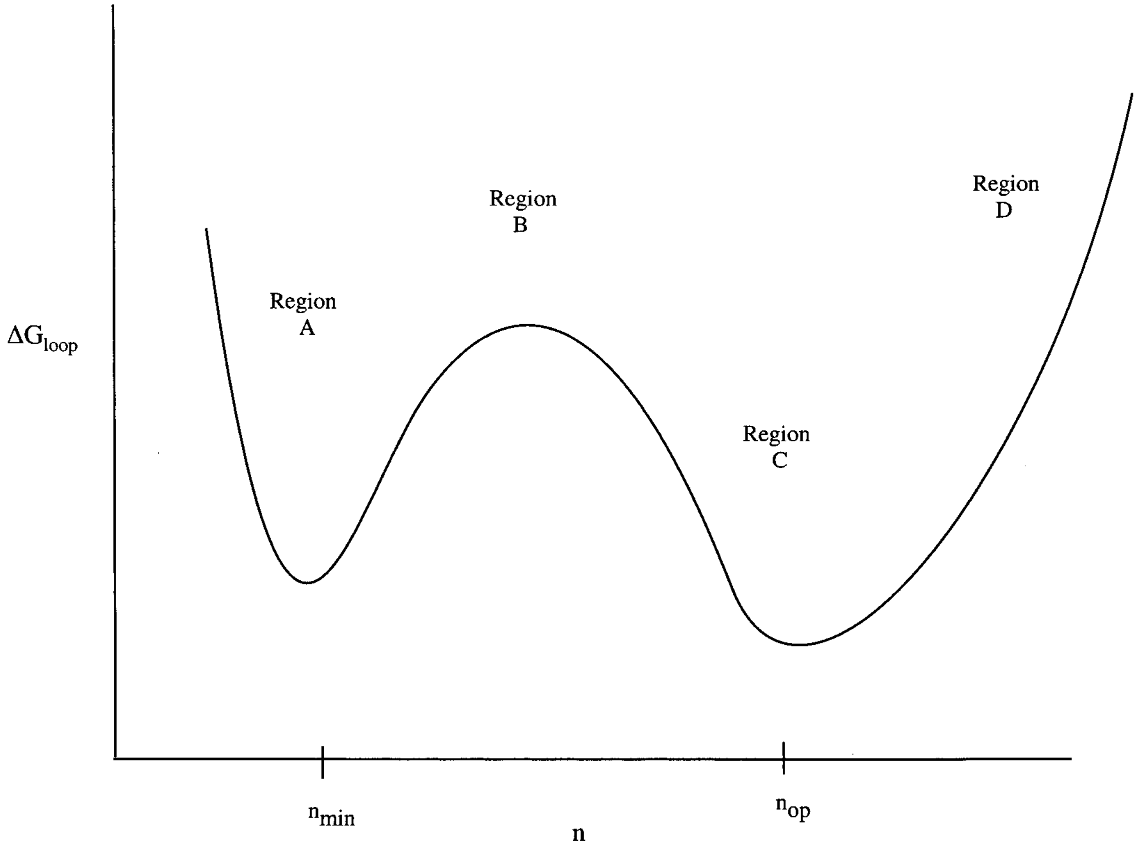

Figure 2. Qualitative form of the free energy loss for loop formation ΔG_{loop} as a function of loop size n. Region A corresponds to minimal loop length n_{min}, while region B corresponds to the region between n_{min} and the primary loop length n_{op}. Region C represents the primary loop. Region D corresponds to loops larger than the primary loop, $n > n_{\mathrm{op}}$.

activation barriers associated with the release of water from the protein surface can be understood as follows. The water surrounding hydrophobic residues is thought to be ordered into clathrate-like hydrogen bonding networks.[34] The water molecules forming the networks are released into the bulk upon formation of van der Waals' contacts between hydrophobic residues. It is the gain in entropy of the water molecules from the clathrate-like structure to the less ordered bulk structure that is thought to provide the protein with much of its stabilizing free energy.[35] Thermal fluctuations are required to exclude the water from the protein surface before sufficient van der Waals' contacts are established between hydrophobic side chains. Protein segments that have more hydrophobic residues, and as a result, more associated clathrate structure will require larger thermal fluctuations in order to fully break up the hydrogen bonding networks. Conversely, the activation barrier to native structure formation will be smaller for segments which are less hydrophobic. Thus, the activation energy to native contact formation in the collapsed primary loop, ΔE_{nat}, is expected to be proportional to the free energy of transfer of the protein segment between a polar and a nonpolar environment ΔG_{hyd},

$$E_{\mathrm{nat}} \propto \Delta G_{\mathrm{hyd}} \tag{10}$$

After the clathrate-like structure is released from the surfaces of a pair of protein segments and a contact established, the contact must be stable enough to remain in place. The stability of the contact will be dictated by the ratio $\Delta G_{\mathrm{cont}}/kT$ for the segments forming the contact. The term ΔG_{sec} should also have an influence in determining the sequence of formation of native contacts in the collapsed primary loop. In this paper, however, secondary structure will not be included in the simple form of the SCM, but this level of detail will be shown to be enough for a successful predictive algorithm, short of attempting atomic level detail.

It is to be expected that the same considerations above apply to the formation of native structure in the tails. In this case, however, it is reasonable to expect that, because the segments in the tails are less constrained than parts of the primary loop, the formation of native contacts will be faster than inside the primary loop. In addition, proteins that are shorter than 65−85 amino acids would have no primary contact. It is beyond the scope of this paper to treat this regime, but the SCM suggests that such short proteins likely fold through a series of minimal loops.

3. Determining Primary Loops, the Cooperative Collapse Sequence, and Folding of the Tails

3.1. Primary Loop. The identification of the primary contact is determined by the minimum value of eq 2 over the protein sequence for segments 65−85 amino acids apart. Only the terms in eq 2 that are sequence dependent will determine the specific location of the primary contact. The ΔG_{loop} term is sequence independent in the approximation followed here and thus provides no predictive capabilities to decide the location of the primary contact along the sequence. The hydrophobic term ΔG_{hyd} and the secondary structure term ΔG_{sec} are strongly sequence dependent. Furthermore, experimental evidence shows that hydrophobic interactions are energetically dominant with respect to secondary structure propensities.[36] These considerations suggest that determining the minimum value of ΔG_{hyd} over the sequence for segments located 65−85 amino acids apart might be sufficient to determine the location along the sequence of the primary contact. Although this approximation is simplistic, results will show that it has the predictive capability to reproduce the observed folding pathway of several globular proteins at low resolution.

To computationally minimize ΔG_{hyd}, an extra consideration must be made. Primary loops will have two contact segments whose physical extent is expected to be greater than one amino acid. Although the precise physical extent of the segments might vary for different proteins, a length of five amino acids for each segment was chosen here for computational purposes.

Since the identification of the primary contact is determined by the hydrophobicity of the segments forming the contact, polarity values obtained from the Fauchere–Pliska scale[37] were assigned to each residue. The results were observed to be robust, as other cruder hydrophobicity scales[38] gave the same conclusions regarding the predicted contact for primary loop formation. The hydrophobicity P_k of each residue is added over a contact window of five amino acids, resulting in a polarity P_i of a potential contact segment centered at residue i. To determine the best contact, the P_i value of a segment centered at residue i is added to the P_j value of a segment centered at residue j that is 65–85 residues apart, to give a contact propensity $P_{ij} = P_i + P_j$. The ij pair along the sequence separated by 65–85 residues which produces the highest value of P_{ij} is selected as the primary contact. Differences in P_{ij} larger than ∼0.45 reflect differences in ΔG_{hyd} larger than kT.[37]

3.2. Cooperative Collapse Sequence and Tails. The activation barriers E_{nat} governing the formation of native contacts in the cooperative collapse of the primary loop and the tails in the SCM are determined by the amount of water attached to the surface of the hydrophobic residues contained in each segment. Thus, only hydrophobic residues will be considered to contribute to E_{nat}. These residues are taken to be I, L, W, F, V, and M and were assigned hydrophobicity values from the Fauchere–Pliska scale while other amino acids were considered to be nonhydrophobic and assigned a hydrophobicity of zero. For the calculations, a segment size of 15 amino acids similar to n_{min} was chosen and the results were seen to be robust for windows between 13 and 19 amino acids. This length is long enough that even the largest possible secondary structure elements, the ω loops[39] could be detected in the cooperative collapse sequence. A more refined model would have to consider different segment sizes for the cooperative collapse sequence, in consideration of the secondary structure propensities of the amino acids.

For computational purposes, the hydrophobicities P_k of 15 consecutive residues centered at residue j are summed, resulting in a polarity value M_j. The M_j's are calculated for all possible segments of 15 amino acids along the protein sequence. To determine the sequence of folding events, the 15 amino acids with the lowest M_j which do not overlap with each other are sequentially chosen. These segments are assumed to reach their native structure in increasing order of M_j, because E_{nat} for each protein segment is assumed to be directly proportional to its polarity represented by M_j (i.e., a large M_j value means that more water needs to be excluded upon cooperative collapse).

4. Results and Discussion

Because the SCM implies that the protein folds by a series of well-defined steps, the model makes several predictions about the nature of the intermediates along the folding pathway. The pathway for the proteins studied in this paper will be composed of the primary contact and the three first steps of the cooperative collapse and the folding of the tails. These predictions will be compared with laboratory data whenever available. For each protein, the optimal pathway and the second best primary contacts are presented. Some molecules may also fold through a suboptimal pathway depending on the relative values of ΔG_{cont} in the early contact formation phase and the relative size of the activation barriers in the cooperative collapse phase.

Although there are experimental results for the folding pathway of ribonuclease A,[40] no attempt will be made to compare those with the predictions here. These experiments were carried out without reducing the disulfide bridges upon unfolding, thus constraining the initial configuration space. The presence of preformed loops due to the disulfide bridges prevents making a comparison of these results with the predictions of the present form of the SCM.

4.1. Properties of the SCM Molten-Globule-Like Intermediate State. The SCM predicts a molten-globule-like intermediate state (MGLIS) in which the protein has formed a primary loop and perhaps a few minimal loops with the following properties.

(a) The MGLIS is expected to have a primary loop in a fluctuating, but generally open conformation, due to the need to minimize ΔG_{loop} in the early folding stages.

(b) The MGLIS is expected to show structural fluctuations on a larger scale than the native state because of the unfolded primary loop.

(c) The overall dimensions of the MGLIS are expected to be larger than those of the native protein because of the open primary loop.

(d) The extra volume of the MGLIS with respect to the native state is expected to be water due to the open conformation of the loop.

(e) Side chains in the primary loop of the MGLIS are expected to retain much of their torsional freedom, because the loop is not a fully folded structure.

(f) The MGLIS is expected to show structural fluctuations on a larger scale than the native state because of the unfolded primary loop.

All of these features are consistent with observed properties of the molten globule state,[41] which have been shown to be part of the native folding pathway of a few proteins including apomyoglobin. The SCM in its present form is not intended for addressing the stability and kinetics of the MGLIS.

4.2. Folding Pathway of Cytochrome *c*. The preceding primary loop analysis was applied to the amino acid sequence of horse heart cytochrome *c* (1HRC). The primary contact analysis identified the segments comprising residues 9–13 and 94–98, with $P_{ij} = 10.1$. The corresponding contact in the native structure is shown in Figure 3. The second highest P_{ij} obtained corresponds to segments 9–13 and 81–85, with $P_{ij} = 9.3$. The residues in the predicted primary contact are in the terminal regions of the protein chain and correlate well with the proton exchange experiments for horse heart cytochrome *c*.[7,8] The proton exchange experiments on horse heart cytochrome *c* show that amide protons in these terminal regions are the first detected to be protected in folding and the last to be exposed in the unfolding experiments.[8] The predicted primary contact is part of a helix–helix contact in the native structure, while the second best is not, thus suggesting that consideration of the ΔG_{sec} term in a more refined model for P_{ij} would have produced a larger difference between the two highest values.

The M_j values for the cooperative collapse sequence for 1HRC are plotted in Figure 6 against the residue number corresponding to the center of the 15 amino acids long protein segment considered in the calculations. This plot is proportional to the activation barriers to formation of native contacts E_{nat} and is

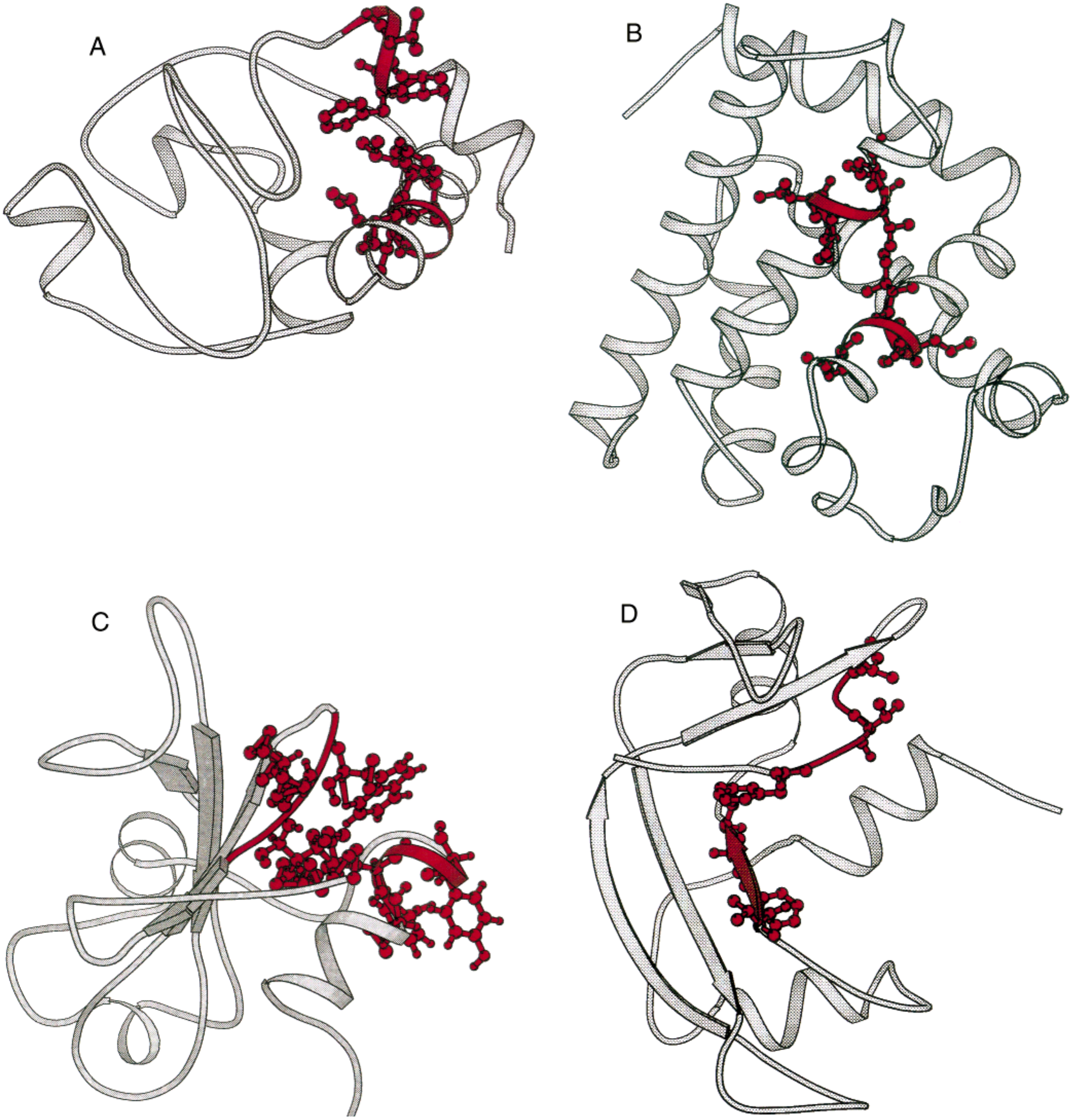

Figure 3. (A−D) The figures are ribbon representations of the proteins from their coordinate Protein Data Bank (PDB) files. Primary contacts are shown in red. The hydrophobic residues of the primary contact are also shown in a ball-and-stick representation. The primary contact is predicted to form first on the protein's folding pathway: (A) cytochrome *c* (1HRC); (B) myoglobin (5MNB); (C) barnase (1BNR); (D) ribonuclease A (3RNR).

shown in Figure 4. The smallest activation barriers to native contact formation which do not overlap with previously formed elements correspond to the protein segments including residues 17−31, 39−53, and 61−75. Each of these protein segments is similar to one of the three ω loops seen to constitute autonomous folding elements in the proton exchange experiment.[8] Furthermore, the predicted order of formation of native contacts in the SCM is the inverse of the order observed for the unfolding of omega loops in cytochrome *c*.[8] The SCM in its present form does not include prosthetic groups, and it therefore cannot deal with folding events associated with the heme group in 1HRC.

4.3. Folding Pathway of Apomyoglobin. The predicted primary contact for apomyoglobin (5MNB) is established between the segments comprising residues 28−32 and 111−115, with $P_{ij} = 12.6$. The primary contact in the native structure is shown in Figure 3. The predicted primary contact correlates well with the proton exchange result, the regions of the protein that include the primary contact segments are observed to be involved in the earliest experimentally observed folding events.[9,10] The second best possible primary contact includes segments 7−11 and 72−76, with $P_{ij} = 12.2$. The difference between the predicted primary contact and the second best possibility is

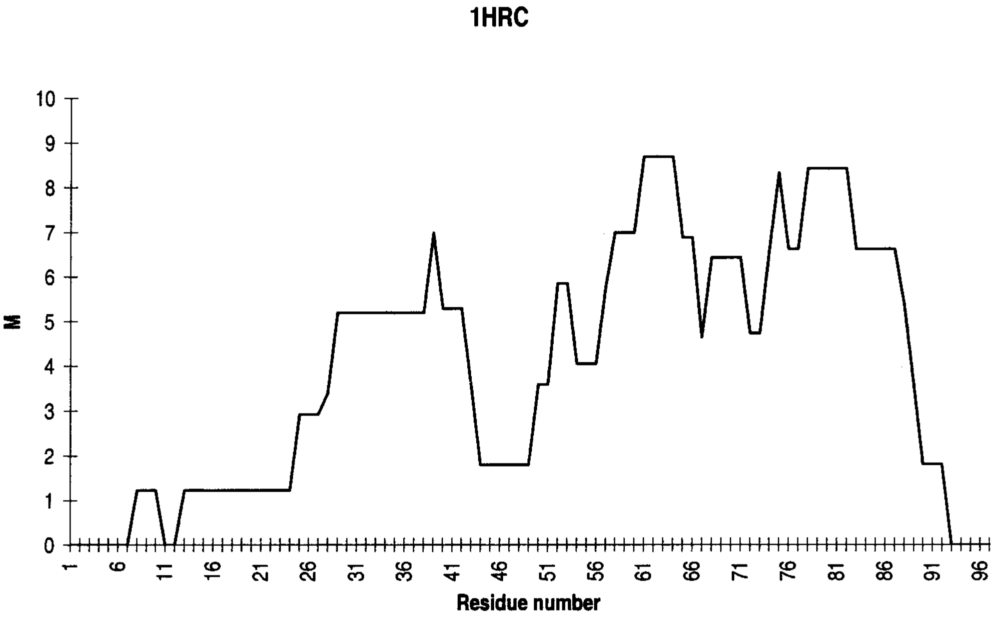

Figure 4. Relative polarity M_j versus the center of the 15 amino acid segments for cytochrome *c* (1HRC). The plot shows that the sequence of formation of native contacts should start in segment 17−31, followed by segments 39−53 and 61−75.

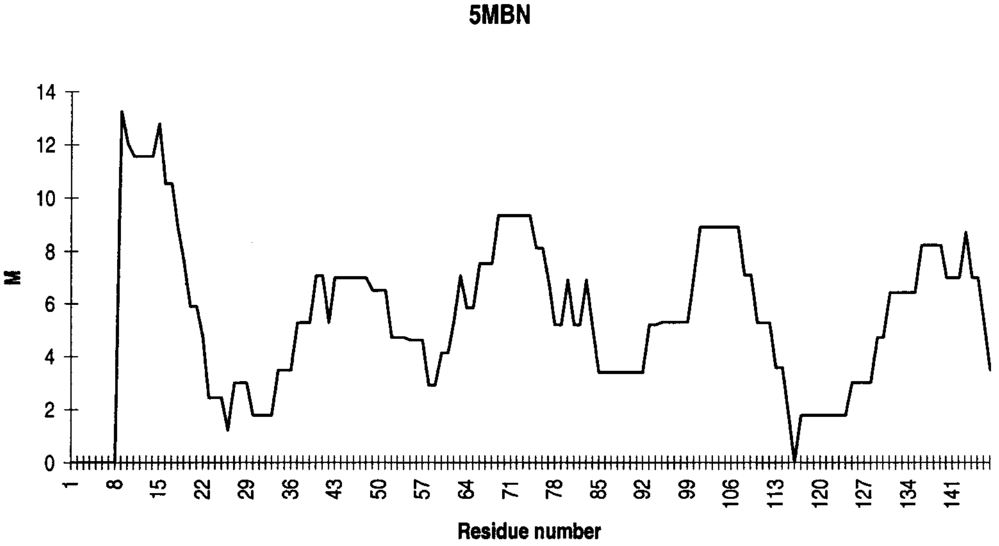

Figure 5. Relative polarity M_j versus the center of the 15 amino acid segments for apomyoglobin (5MNB). The plot shows that the sequence of formation of native contacts should start in segment 109−123, outside the primary loop, followed sequentially by segments 51−65 and 81−96 inside.

small. As in the case of cytochrome *c*, secondary structure propensities embodied in the ΔG_{sec} term might help discriminate between the two possibilities. There are recent results suggesting that helix E of apomyoglobin, which includes residues 72−76, is less stable than helices B and G that include the predicted primary contact in the native structure.[42]

The M_j values for the cooperative collapse sequence for 5MNB are shown in Figure 5 against the residue number corresponding to the center of the 15 amino acid segments. The smallest activation barriers to native contact formation which do not overlap with previously formed contacts correspond to structural elements including residues 109−123, outside the primary loop and 51−65 and 81−96 inside. These results correlate well with the results of the proton exchange experiment,[9] where the 109−123 segment in the tails is observed to fold first in the experiment, while the other segments, all inside the primary loop, are seen to form native contacts significantly later than the segments involved in the primary contact. The folding of the tails would imply an MGLIS that was composed of portions of helix A, B, G, and H. This type of intermediate correlates well with characterized molten globule states in apomyoglobin.[43]

4.4. The Folding Pathway of Barnase. The best predicted primary contact for barnase (1BNR) is established between the segments 13−17 and 93−97 with $P_{ij} = 12.6$ as shown in the native structure in Figure 3. The second highest includes segments 7−11 and 72−76, and $P_{ij} = 12.2$. Once again, the difference is small, suggesting the need to include ΔG_{sec} in a more refined model. The predicted primary contact correlates reasonably well with the proton exchange experiment for barnase.[11] The experimentally folded population at 0.1 s of the proton exchange probed amino acids of barnase shows peaks corresponding to amino acids 16 and 98. The experiment does not provide probes in the 93−97 segment other than I96 which is significantly populated in 0.1 s, while Y97 is observed to fold fast with not enough reliable data for a quantitative estimate of its folded population.

The values of M_j for the formation of native contacts in the cooperative collapse sequence are shown in Figure 6. The

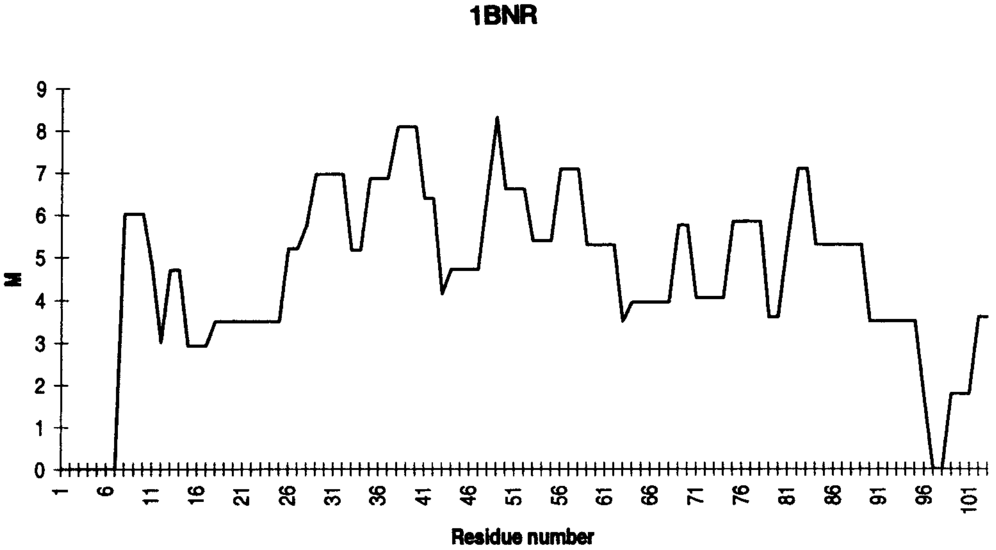

Figure 6. Relative polarity M_j versus the center of the 15 amino acid segments for barnase (1BNR). The plot shows that the sequence of formation of native contacts should start in segment 17−31, followed sequentially by segments 56−70 and 73−87 all inside the primary loop. The plot does not however show clear differences in M_j between different regions of the protein chain.

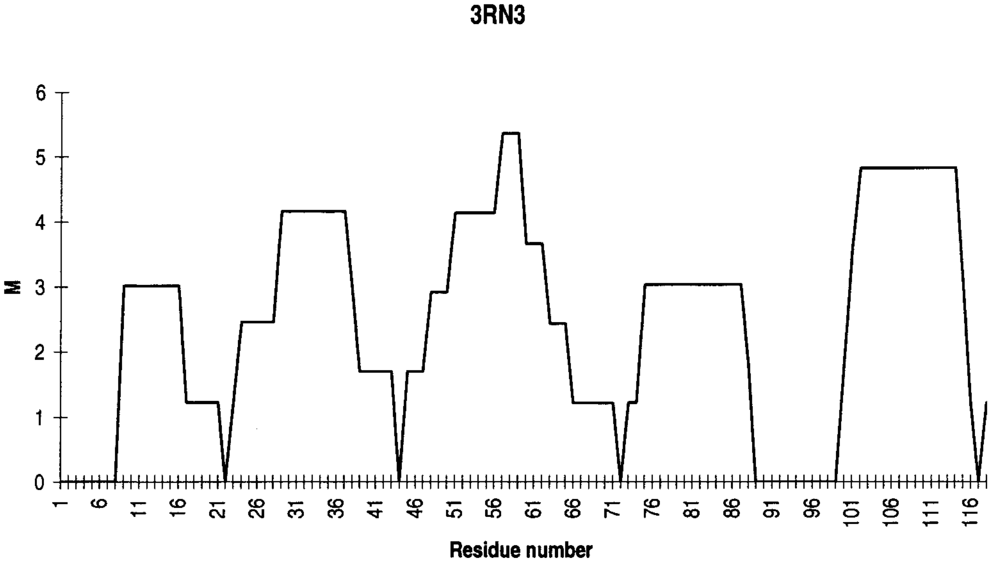

Figure 7. Relative polarity M_j versus the center of the 15 amino acid segments for ribonuclease A (3RNR). The plot shows that the sequence of formation of native contacts should start in segment 15−29, outside the primary loop, followed sequentially by segments 87−101 and 65−79 inside the primary loop.

smallest activation barriers to native contact formation correspond to structural elements including residues 18−32, 57−70, and 73−87. However, all activation barriers are roughly the same size thus implying the absence of a significant separation of time scales between the different folding events in the cooperative collapse. In accord with this behavior, no discernible time scale separation was found in the proton exchange experiment.[11]

4.5. Folding Pathway of Ribonuclease A. The best predicted contact for ribonuclease A (3RNR) is established between the segments comprising residues 43−47 and 116−120 with P_{ij} = 9.0. The primary contact is shown in Figure 3. The second best primary contact includes segments 26−30 and 106−110, with P_{ij} = 8.7. The predicted primary contact should be expected to be involved in the earliest folding stages of 3RNR.

The values of M_j for the cooperative collapse sequence of 3RNR are shown in Figure 7. The smallest activation barriers to native contact formation correspond to segments including residues 15−29, outside the primary loop and 65−79 and 87−101 inside. The activation barrier for the 87−101 segment is lower than that for the 65−79; thus, it is expected that it enters the folding process before. In the native structure the 87−101 segment is entirely included in a β-sheet. If the experimental sequence of folding events deviated from the one predicted here, it would probably suggest that the ΔG_{sec} term should be included in a more refined version of the SCM in order to deal successfully with large differences in stability between different secondary structure motifs.

5. Conclusions

The SCM provides a conceptually and operationally simple framework to explain protein folding pathways. In the SCM, the entropic consequences of loop closure and the protein/solvent interaction determine the folding pathway in two stages: an early phase in which initial contacts are established driven by thermodynamic optimization, and a later cooperative collapse phase in which the need to exclude water from the protein surface governs the sequential attainment of the native structure.

Although the present paper does not rigorously prove the validity of the SCM, the physical basis of the model was presented and its predictions are consistent with recent protein folding data. The model explains many of the features observed in proton exchange experiments on cytochrome *c*, apomyoglobin, and barnase and predicts an intermediate state that has many of the properties of a molten globule.

The next step in the development of the model will be to treat larger proteins and incorporate secondary structure in the calculations in order to increase the resolution attainable and eventually attempt detailed structure predictions. It is possible that additional complexity will enter the picture when larger proteins are considered. For example, it is not clear whether the formation of a second primary loop may overlap with the first primary loop or whether they occur in different segments of the protein chain. Such issues must be addressed in order for the SCM to make more detailed predictions of the folding pathways of broad classes of proteins.

Acknowledgment. One of the authors (F.B.C.) acknowledges Dr. Dirk Dittmer at Princeton and Prof. Juan J. Saenz at the Universidad Autónoma de Madrid for insightful comments and discussions. We also acknowledge support from the National Science Foundation.

References and Notes

(1) Dill, K. A. *Biochemistry* **1989**, *29*, 7133.

(2) Anfinsen, C. B.; Haber, E.; Sela, M.; Whiter, F. H. *Proc. Natl. Acad. Sci. U.S.A.* **1961**, *47*, 1309.

(3) Wu, H. *Am. J. Physiol.* **1929**, *90*, 562.

(4) Mirsky, A. E.; Pauling, L. *Proc. Natl. Acad. Sci.* **1936**, *22*, 439.

(5) Anfinsen, C. B. *Science* **1973**, *181*, 223.

(6) Levinthal, C. *J. Chem. Phys.* **1968**, *65*, 44.

(7) Roder, H.; Elove, G. A.; Englander, S. W. *Nature* **1988**, *335*, 700.

(8) Bai, Y.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192.

(9) Jennings, P. A.; Wright, P. E. *Science* **1993**, *262*, 892.

(10) Loh, S. N.; Kay, M. S.; Baldwin, R. L. *Proc. Natl. Acad. Sci.* **1995**, *92*, 5446.

(11) Matouschek, A.; Serrano, L.; Meiering, E. M.; Bycroft, M.; Fersht, A. R. *J. Mol. Biol.* **1992**, *224*, 837.

(12) Chan, C. K.; Hu, Y.; Takahashi, S.; Rousseau, D. L.; Eaton, W. A.; Hofrichter, J. *Biophys. J.* **1996**, *70*, A177 (abstr.).

(13) Callender, R. H.; Dyer, R. B.; Gilmanshin, R.; Woodruff, W. H. *Annu. Rev. Phys. Chem.* **1998**, *49*, 173−202.

(14) Jones, C. M.; Henry, E. R.; Hu, Y.; Chan, C. K.; Luck, S. D.; Bhuyan, A.; Roder, H.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 11860−11864.

(15) Pascher, T.; Chesnick, J. P.; Winker, J. R.; Gray, H. B. *Science* **1996**, *271*, 1558−1560.

(16) Phillips, C. M.; Mizutani, Y.; Hochstrasser, R. M. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 7292−7296.

(17) Ballew, R. M.; Sabelko, J.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5759−6764.

(18) Dobson, C. M.; Sali, A.; Karplus, M. *Angew Chem. Int. Ed. Engl.* **1998**, *37*, 868−893.

(19) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.

(20) Bryngelson, J. D.; Onuchic, J. N.; Socci, N. D.; Wolynes, P. G. *Proteins: Struct., Funct., Genet.* **1995**, *21*, 167.

(21) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404.

(22) Ptisin, O. B.; Rashin, A. A. *Biophys. Chem.* **1975**, *3*, 1.

(23) Pappu, R. V.; Weaver, D. L. *Protein Sci.* **1998**, *7*, 480.

(24) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*; Springer-Verlag: New York, 1979; pp 166−168.

(25) Rose, G. D. *J. Mol. Biol.* **1979**, *134*, 447.

(26) Srinivasan, R.; Rose, G. D. *Proteins* **1995**, *22* (2), 81.

(27) Kuntz, I. D.; Kauzmann, W. *Adv. Protein Chem.* **1974**, *28*, 239.

(28) Pauling, L.; Corey, R. B.; Branson, H. R. *Proc. Natl. Acad. Sci. U.S.A.* **1951**, *37*, 205.

(29) Waldburger, C. D.; Jonnson, T.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93* (7), 2629.

(30) Ramachandran, G. N.; Sasisekharan, V. *Adv. Prot. Chem.* **1968**, *23*, 283.

(31) Jacobson, H.; Stockmeyer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600.

(32) Schultz, G. E.; Schirmer, R. H. *Principles of Protein Structure*; Springer-Verlag: New York, 1979.

(33) Richardson, J. S.; Richardson, D. C. In *Prediction of Protein Structure and Protein Conformation*; Fasman, G., Ed.; Plenum Press: New York, 1989; pp 1−88..

(34) Teeter, M. M. In *Protein Folding: Deciphering the second half of the Genetic Code*; Gierasch, L. M., King, J., Eds.; AAAS Press: Washington, DC, 1991; pp 44−54.

(35) Kauzmann, W, *Adv. Protein Chem.* **1959**, *14*, 1.

(36) O'Neil, K. T.; Degrado, W. F. *Science* **1990**, *250*, 646.

(37) Fauchère, J. L.; Pliska, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.

(38) Nozaki, Y.; Tanford, C. *J. Biol. Chem.* **1971**, *246*, 2211.

(39) Leszczynski, J.; Rose, G. D. *Science* **1986**, *234*, 849.

(40) Ugdaonkar, J. B.; Baldwin, R. L. *Proc. Natl. Acad. Sci.* **1990**, *87*, 8197.

(41) Kuwajima, K. *Proteins: Struct., Funct., Genet.* **1989**, *6*, 87.

(42) Hirst, J. D.; Brooks, C. L., III *Biochemistry* **1995**, *34*, 7614.

(43) Eliezer, D.; Yao, J.; Dyson, H. J.; Wright, P. *Nat. Struct. Biol.* **1998**, *5* (2), 148−155.