Role of Topology in the Cooperative Collapse of the Protein Core in the Sequential Collapse Model. Folding Pathway of α-Lactalbumin and Hen Lysozyme

Fernando Bergasa-Caceres*,† and Herschel A. Rabitz

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

Received: December 7, 2000

This paper further explores the recently proposed sequential collapse model (SCM) for protein folding pathways. Two specific items are considered within the SCM: (a) the cooperative collapse phase for protein folding pathways and (b) applications of the model to suggest the folding pathways of α-lactalbumin and hen lysozyme. With regard to the first goal, it is shown that major topological rearrangements of the protein core after the hydrophobic collapse are entropically unfavorable, suggesting that the final native structure is attained sequentially through a limited number of nativelike optimization steps. It is also shown that cross-links between protein loops are entropically unfavorable. This result suggests that long proteins fold in independent folding units, with each one of them corresponding to initially forming a single loop defined by a hydrophobic contact. The second objective of this paper is to illustrate further the predictive capabilities of the SCM through its application to the folding pathways of α-lactalbumin and hen lysozyme in comparison with available experimental data and previous theoretical results. It is found that the predicted folding pathways are similar for both proteins.

1. Introduction

The recently proposed sequential collapse model (SCM)[1] strives to explain how the primary sequence of a protein governs its folding pathway. In the SCM, the entropic consequences of forming protein loops (i.e., the topology of the intermediate states) and the hydrophobic effect guide the protein through a small number of intermediate states along a sequential folding pathway. The SCM has been successfully applied to analyze the multistate folding pathways of proteins of relatively small size (∼100−150 amino acids) at low resolution, including cytochrome *c* and apomyoglobin.[1]

In the SCM, the folding pathway is divided into two distinct phases: (1) an early phase defined by the formation of an initial intramolecular protein contact. This initial contact is established between segments located at a characteristic distance along the protein sequence, thus defining a relatively open fluctional protein loop containing the amino acids between the two contact-forming segments (i.e., the primary loop). (2) A cooperative phase in which the primary loop undergoes hydrophobic collapse defining a compact intermediate held together by a number of contacts. The collapsed primary loop subsequently undergoes an optimization subphase in which the native structure is finally reached.

The purpose of this paper is to (a) investigate the cooperative collapse of the primary loop and (b) to further illustrate the capabilities of the SCM by exploring the folding pathways of α-lactalbumin and hen lysozyme. In connection with the cooperative collapse analysis, it will be shown that (i) the topology of the collapsed intermediate tends to be nativelike and (ii) cross-links between primary loops are entropically unfavorable, suggesting that proteins long enough to form more than one primary loop fold in independent units. With respect to the study of α-lactalbumin and hen lysozyme, it is found that their predicted folding pathways are similar. This conclusion is consistent with theoretical and experimental results [2−7] which suggest that there is often a link between structural homology and similar folding pathways, although this is not a universal rule.[8,9] The results are also discussed in connection with existing experimental data and theoretical results.

In section 2, the SCM is briefly reviewed. In section 3, the study of the cooperative collapse phase is presented. Section 4 presents results for the sequential folding pathway of α-lactalbumin and hen lysozyme. Finally, section 5 presents conclusions and directions for further research.

2. The Model

The theoretical basis of the SCM has been explained elsewhere.[1] A brief review is presented below, with emphasis on the issues that will be further investigated here.

2.1. Early Contact Formation Phase. In the SCM, the folding process starts by the formation of one or a few initial contacts between segments not necessarily adjacent along the amino acid sequence. These initial contacts define open protein loops. Formation of an early successful contact will imply a free energy change ΔG_{cont} that can be decomposed as

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

where ΔG_{conf} is the free energy change associated with the loss of configurational freedom upon formation of the contact and loop closure. ΔG_{conf} 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_{conf}. The term ΔG_{conf} could include a contribution due to nonspecific enthalpic interactions between neighboring side chains in the loop not involved in the contact that defines the open 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.

* Corresponding author.
† Current address: Endesa, Príncipe de Vergara 187, 28003 Madrid, Spain. E-mail: FBergasa@Endesa.Es.

10.1021/jp004421e CCC: $20.00
Published on Web 03/16/2001

ΔG_{hyd} is the free energy change associated with the release of water into the bulk from the surface of the hydrophobic residues forming the contact and is expected to be negative.[10] ΔG_{sec} is the enthalpic free energy change associated with the formation of secondary structure in the contact and is expected to be negative.[11] The entropic free energy change associated with formation of secondary structure is included in ΔG_{conf}. Finally, ΔG_{tert} is the free energy change due to the establishment of tertiary enthalpic interactions such as salt bridges, tertiary hydrogen bonds, and disulfide bonds and is expected to be negative.[12]

When the chain of N amino acids forms a loop, it will be composed of four distinct regions: c residues will form the contact, n residues will form the open loop, and m and l residues will form each of the tails such that $N = c + n + m + l$. Up to a constant, ΔG_{conf} can be written approximately as[1]

$$\Delta G_{conf} = -kT[c \ln f_c + n \ln f_n - (n+c) \ln f_0 - 3/2 \ln n] \quad (2)$$

where f is the number of accessible conformations per amino acid. 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. The term $^3/_2 \ln n$ represents the classical Jacobson−Stockmeyer contribution to the entropy loss from loop closure.[13] A more refined model could replace the Jacobson−Stockmeyer term with a contribution that fully included excluded volume considerations.[14]

ΔG_{conf} can be shown to have two minima[1] with respect to n: a deeper one at a distance of about 65−85 amino acids, called the optimal loop length n_{op}, and a shallower one at a distance of about $n \approx 10-15$ amino acids, called the minimal loop length, n_{min}. The minimal loop length represents the minimal loop size that allows for successful contact formation. In the early contact formation phase of the SCM, the protein will tend to form a first contact at a distance n_{op} along the sequence, and this contact is referred to as the primary contact. Formation of the primary contact requires that $\Delta G_{conf} < -(\Delta G_{hyd} + \Delta G_{sec} + \Delta G_{tert})$. If the protein is shorter than n_{op} or sufficiently hydrophobic segments do not exist at the optimal distance to create a stable contact, then the SCM predicts that the protein will fold through a short-range pathway by formation of loops of length as close as possible to the minimal length n_{min}. Because of the absence of initial long-range contacts, these proteins will tend to show an average distance along the sequence between amino acids involved in contacts (i.e., lower contact order) shorter than that of proteins forming a primary loop. Contact order has been shown to be an important determinant of protein folding dynamics by experimental and theoretical studies probing the influence of topology on the folding of globular proteins.[15,16] In general, there is a rising entropic cost to forming protein contacts depending on loop length up to n_{op}. Then it is to be expected that most protein contacts, other than primary contacts, tend to form close to the minimal loop range.

The entropy loss associated with loop closure has been generally assumed to favor the formation of short-range contacts over long-range ones along the folding pathway.[15,16] In the SCM, the situation is slightly more complex: most protein contacts will form at very short range, but the entropy loss associated with loop closure is minimized at the relatively long-range distance n_{op}, thus promoting the formation of an initial contact at this distance in proteins containing segments hydrophobic enough to form a stable contact at n_{op}.

2.2. Molten Globule-Like Intermediate State (MGLIS). In the SCM, folding of the tails follows formation of the primary contact. The intermediate thus defined has a compact region comprising the amino acids in the primary contact and the tails and an open, fluctuating primary loop. This intermediate is referred to as a molten globule-like intermediate, and it shares many of the properties of a molten globule.[1]

2.3. Cooperative Collapse Phase. After the primary contact is established, the protein is expected to have a fluctuating open (i.e., not yet fully folded) primary loop. In general, $\Delta G_{conf} > -(\Delta G_{hyd} + \Delta G_{sec} + \Delta G_{tert})$ for the formation of individual contacts between protein segments included in the primary loop because the partial folding of a primary loop must involve the temporary formation of loops shorter than the optimal loop length n_{op}, thereby generating intermediates with large $\Delta G_{conf} > 0$. This implies that the primary loop should fold through a mechanism of cooperative collapse in which most segments in the primary loop enter the folding process simultaneously to maximize $-(\Delta G_{hyd} + \Delta G_{sec} + \Delta G_{tert})$ and thereby overcome the entropic barrier generated by ΔG_{conf}.

After the primary loop undergoes hydrophobic collapse, the SCM hypothesizes that the core optimization subphase is controlled by the need to fully exclude water from the interacting surface of any protein segment in order to attain its native conformation. Thus, the need to break up the remaining clathrate-like water structures expected to exist[17] around hydrophobic residues is assumed to produce activation barriers to structure formation. The activation barrier E^i to native structure formation in the segment i can then be written as $E^i \propto \Delta G^i_{hyd}$, where ΔG^i_{hyd} is the free energy change associated with the exclusion of water from the interacting surface of segment i upon attainment of the native structure.

The SCM is in the same spirit as that of recent theoretical efforts to develop simple models able to describe the intermediates along the folding pathway in the context of the so-called "funnel" view of the folding process.[18−21] The SCM, however, shares much of the "old" view of the folding process, in which the protein descends toward the free energy minimum through few intermediate steps.[22] Although the SCM does not preclude the possibility that a protein might fold through several parallel pathways,[1] it suggests that there is a strong preference for those pathways that minimize the conformational entropy loss upon the formation of primary contacts. Also, because n_{op} is relatively large, the number of possible primary contacts (i.e., the nucleation event that initiates the folding pathway) for a given sequence is likely to be small. The "new" view embodied in the funnel picture postulates instead a large number of intermediates, especially in the early folding stages. These two views need not, however, be antagonistic because recent theoretical results show that some folding pathways might be strongly statistically preferred in a free energy landscape that allows for a multiplicity of folding pathways.[23]

3. Theoretical Investigation of the Cooperative Collapse Phase

In this section, two issues are investigated: (a) the possibility of major topological rearrangements of the primary loop and (b) the implications of the cooperative collapse for the folding of long proteins. The physical conclusions presented in this section are consequences of the SCM that has been shown to successfully reproduce the folding pathway of a number of relatively small proteins (proteins of length $n_{op} < N < 2\, n_{op}$) at

low resolution.[1] It is impossible to fully validate the conclusions presented here at this stage due to (i) the absence of sufficiently detailed experimental data concerning core rearrangements along the folding pathway in relevant proteins (i.e., proteins longer that ~100 amino acids) and (ii) the lack of detailed experimental pathway data for long proteins (i.e., proteins with $N > 2n_{op}$).

3.1. Topological Rearrangements of the Collapsed Primary Loop. Once the primary loop has collapsed, an interesting question is whether major rearrangements are likely to occur before the folding process proceeds toward the native structure. As major rearrangements, those topological changes involving at least a region of the primary loop of the same size as the one that remains topologically fixed are considered. If major rearrangements can take place, the overall topology of the collapsed primary loop need not be nativelike. If, instead, major rearrangements are not possible, then the overall topology of the collapsed primary loop should be nativelike.

A significant rearrangement of an already compact structure must involve the unfolding of large portions of the primary loop in order to refold into the correct final structure. This partial unfolding of the already collapsed structure involves the formation of an intermediate, including at least one unfolded region k of length $n_k < n_{op}$ and a region m that remains compact of length $n_m = n_{op} - n_k$. Region k is equivalent to an open loop of length n_k. Major topological rearrangement as defined above, moreover, implies that $n_k \geq n_m$; that is, the minimum length considered here for n_k is ~30−40 amino acids.

Before the unfolding of region k, the free energy of stabilization of the collapsed primary loop ΔG_{coll} can be written as

$$\Delta G_{coll} = \Delta G^{m}_{coll,conf} + \Delta G^{m}_{coll,int} + \Delta G^{k}_{coll,conf} + \Delta G^{k}_{coll,int} + \Delta G^{km}_{coll,int} \quad (3)$$

where $\Delta G^{i}_{coll,conf} > 0$ represents the configurational free energy change upon collapse of region i, the term $\Delta G^{i}_{coll,int} = \Delta G^{i}_{hyd} + \Delta G^{i}_{sec} + \Delta G^{i}_{tert} < 0$ describes the free energy change upon collapse from the hydrophobic contacts and enthalpic interactions established within regions $i = k$, m, and $\Delta G^{km}_{coll,int}$ represents the free energy of stabilization arising from the interactions established between regions k and m and should be negative. For region k to successfully collapse (i.e., $\Delta G_{coll} < 0$), $\Delta G^{k}_{coll,conf} < -(\Delta G^{k}_{coll,int} + \Delta G^{km}_{coll,int})$.

Upon unfolding of region k, $\Delta G^{k}_{coll,int}$ and $\Delta G^{km}_{coll,int}$ are lost. Thus, the free energy change of the collapsed primary loop upon unfolding of region k, ΔG^{k}_{un}, can be written as

$$\Delta G^{k}_{un} = -\Delta G^{km}_{coll,int} - \Delta G^{k}_{coll,int} - \Delta G^{k}_{un,conf} \quad (4)$$

where $\Delta G^{k}_{un,conf}$ represents the free energy change due to the restricted conformational space of the amino acids in region k with respect to the open primary loop and is positive. The amino acids in region k gain conformational freedom upon the unfolding of region k, and it is expected also that $\Delta G^{k}_{un,conf} < -(\Delta G^{k}_{coll,int} + \Delta G^{km}_{coll,int})$, so in general, $\Delta G^{k}_{un} > 0$. Partial unfolding of the collapsed primary loop is in general an unfavorable process. The free energy of unfolding of a given collapsed primary loop, ΔG^{k}_{un}, will vary from protein to protein and between different segments of the same protein because $\Delta G^{k}_{coll,int} + \Delta G^{km}_{coll,int}$ are strongly sequence dependent. If $\Delta G^{k}_{un} < kT$, topological rearrangements could still be possible, even if they imply the unfolding of large portions of the collapsed region. In general, however, the collapsed intermediate must be stable with respect to fluctuations in order to undergo the secondary structure optimization subphase, so it is to be expected that $\Delta G^{k}_{coll} \leq -k\mathrm{T}$ and $\Delta G^{k}_{un} \geq k\mathrm{T}$. Also, the collapsed protein core is held together by a number of nonspecific contacts, and its stabilization energy must be much smaller than that of the fully folded protein. As a consequence, ΔG_{coll} is expected to be on the order of at most a few kT, and we can conclude that topological rearrangements of the collapsed protein core are energetically unfavorable. Topological rearrangement could still occur if the stabilization energy of region m, ΔG^{m}_{coll}, is significantly larger than ΔG^{k}_{un}; whether this happens depends on the amino acid sequence making $\Delta G^{m}_{coll} \ll 0$. This situation is unlikely to arise because hydrophobic amino acids in globular proteins tend to be located in clusters of a few residues placed more or less evenly along the amino acid sequence, and the size of each individual cluster is much less than n_k.[24] This behavior suggests that the optimization subphase of globular proteins implies the formation of a series of nativelike contacts, once nativelike topology is established in the hydrophobic collapse of the primary loop.

Recent experiments point to the importance of topology in the folding pathway of globular proteins.[4−6,15,16] Mutations that severely affect the contacts determining the topology of the protein core also severely affect the folding rates and the stability of the resulting structures. The theoretical result obtained here is consistent with these observations because the early establishment of the native topology is shown to be an important factor in determining the folding pathway leading to the native structure. Theoretical studies have also shown the importance of topological considerations in the folding process.[25,26]

3.2. Cross-Links in Long Proteins. If a protein is long enough that it can form more than one primary loop (i.e., $N > 2n_{op}$, referred to as "long proteins"), a relevant question is whether two or more primary loops can share a protein segment (i.e., whether they form cross-links). In this section, we will show that cross-links are entropically unfavorable in the SCM. This entropic cost suggests that long proteins will tend to form independent primary loops, and each of them can be naturally viewed as an independent folding unit.

Consider a long protein that has formed two primary loops. Each primary loop is defined upon formation of a contact between two protein segments at the optimal distance n_{op}. Primary loop 1 is defined by a contact established between segments i and j, while primary loop 2 is defined by a contact established between segments l and k, with lengths satisfying $n_i < n_l < n_j$. There is a segment S of length $n_s = (n_j - n_l)$, which is shared by the two primary loops.

If primary loop 1 undergoes a cooperative hydrophobic collapse, then S would be involved in the process as it is part of primary loop 1. This means that f_s, the conformational freedom of the amino acids in S after the hydrophobic collapse, is smaller than f_0 of the same amino acids in the open primary loop, $f_s < f_0$. After the hydrophobic collapse of primary loop 1, the still open fluctuating primary loop 2 now includes a segment of length n_s, with a conformational space more restricted than that of the open loop. Because f is fundamentally a function of the interactions between the amino acid side chains, the constraints imposed on f_s by the hydrophobic collapse of primary loop 1 reduces the conformational freedom available to the rest of the amino acids included in primary loop 2.

After collapse of the primary loop 1, the intermediate thus formed includes an open primary loop with restricted conformational possibilities, with an associated entropic cost which is not compensated by any extra entropy gain by the solvent. The free energy change $\Delta G_{coll,1}$ associated with the collapse of

primary loop 1 can be written as

$$\Delta G_{\mathrm{coll},1} = \Delta G^{0}_{\mathrm{coll},1} + \Delta G_{\mathrm{s},2} \tag{5}$$

where $\Delta G^{0}_{\mathrm{coll},1}$ is the free energy change for the collapse of primary loop 1 in the absence of cross links with other primary loops and is expected to be negative, $\Delta G_{\mathrm{s},2}$ is the free energy change associated with the restriction of the conformational space available to the amino acids in primary loop 2 and is expected to be positive, and $\Delta G_{\mathrm{coll},1} > \Delta G^{0}_{\mathrm{coll},1}$. In general, then, it is less energetically favorable to form cross-links between different primary loops.

It is difficult to rigorously determine the dependence of $\Delta G_{\mathrm{s},2}$ on n_s. This is so because upon the collapse of primary loop 1, the constraints on the conformational space available to the amino acids in primary loop 2 will depend both on n_s and the overall conformation adopted by segment *s*. It is expected however, that the magnitude of $\Delta G_{\mathrm{s},2}$ depends directly on n_s because the larger the values of n_s are, the more the conformational space of the amino acids in primary loop 2 will be constrained. If n_s is small compared to n_{op}, then $\Delta G_{\mathrm{s},2}$ could be relatively small. For n_s values smaller but comparable to n_{op} values, it is to be expected that $\Delta G_{\mathrm{s},2}$ is large because the conformational freedom left to the amino acids in primary loop 2 approaches that of the amino acids in a loop of intermediate size between n_{min} and n_{op}. The protein should tend to minimize $\Delta G_{\mathrm{s},2}$ and form separate primary loops that undergo independent hydrophobic collapse or, at most, primary loops that share just a few amino acids. This result suggests a correspondence between separate primary loops in long proteins and the observed independent folding domains seen in many proteins.[27] A single primary loop in a long protein would correspond to an independent folding unit. The absence of detailed experimental data on the folding pathway of long proteins makes it impossible to firmly confirm this conclusion at this stage.

4. Folding Pathway of α-Lactalbumin and Hen Lysozyme

We shall now consider the second aspect of this paper, further demonstrating the capabilities of the SCM in predicting folding pathways. Hen lysozyme and α-lactalbumin are small proteins of 129 and 123 amino acids, respectively. Both proteins have very similar native structures even though their sequences show little homology. The structures include two distinct domains: a helical one, including the C- and N-termini, called the α domain, and a domain including significant β sheet content, the β domain. There are four disulfide bridges in each protein. In this section, results for the folding pathway of α-lactalbumin and hen lysozyme are presented. Comparison with available experimental results is done. The detailed results are only compared in their broader features with experiments in which disulfide bridges were not reduced[28−35] because disulfide bridges do not enter the calculation and they induce topological constraints in the unfolded state and the intermediates along the folding pathway. It has been shown recently that a mutant form of α-lactalbumin is able to reach a compact state in the absence of disulfide bridges, similar to the molten globule of the native protein,[36] including a fully folded α domain. This experiment suggests that disulfide bridges help stabilize the native structure of the α domain but do not determine the folding pathway. The experiment also suggests that the same conclusion likely holds for the β domain. The results shown below for the β domain should be taken with caution, however, because the β domain does not fully fold into its native conformation in the absence of disulfides in the mutant form of α-lactalbumin.[36] These experimental findings for α-lactalbumin are consistent with experimental evidence that shows that hen lysozyme is able to fold into a nativelike conformation when the disulfide bridge established between residues 6 and 127, included in the α domain, is cleaved.[37−39] Finally, the results are also compared to the theoretical predictions for the same proteins obtained in the context of the funnel picture.[20]

4.1. Calculational Method. In this paper, we follow closely the method introduced previously[1] to determine the SCM folding pathway of apomyoglobin, cytochrome *c*, barnase, and ribonuclease A. A summary of the procedure is presented below. The primary contact is determined by the minimum value of ΔG_{hyd} for segments of five amino acids located 65−85 residues apart along the sequence. 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[40] were assigned to each residue. In summary of the procedure, the hydrophobicity P_k of each residue is added over a contact window of five amino acids segment centered at residue *i*, resulting in a polarity P_i. To determine the best contact, we added the P_i value of a segment centered at residue *i* to the P_j value of a segment centered at residue *j* that is 65−85 residues away from *i* 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 ∼0.5 reflect differences in ΔG_{hyd} larger than kT.[40]

For the purpose of determining the activation barriers E^j governing the formation of native contacts in the cooperative collapse, the amino acids L, W, F, V, M, and I were assigned hydrophobicity values from the Fauchere−Pliska scale. All other amino acids were considered to be nonhydrophobic and assigned a hydrophobicity of zero. For the calculations, a segment size of 15 amino acids was chosen, and the results were seen to be robust for windows between 13 and 17 amino acids. This length is long enough that even the largest possible secondary structure elements, the omega loops,[41] could be detected in the cooperative collapse sequence.

The hydrophobicities P_k of 15 consecutive residues centered at residue *j* are summed, resulting in a hydrophobicity value H_j. The H_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 H_j which do not overlap with each other are sequentially chosen. These segments are assumed to reach their native structure in increasing order of H_j because E^j for each protein segment is assumed to be directly proportional to its hydrophobicity represented by H_j (i.e., a large H_j value means that more water needs to be excluded upon structure optimization).

4.2. Primary Contact of α-Lactalbumin. The best predicted primary contact in α-lactalbumin is established between residues 27−31 and 101−105, with $P_{ij} = 12.7$. The second best primary contact is established between residues 51−55 and 116−120, with $P_{ij} = 9.7$. The predicted primary contact occurs between residues 27−31 in helix B and a short piece of 3^{10} helix comprising residues 101−103, residue 104 in a turn, and residue 105 at the beginning of helix D in the crystal structure.[42] This is consistent with NMR unfolding experiments that show that the 3^{10} helix and the D helix are the regions of α-lactalbumin most resistant to unfolding by guanidinium chloride and temperature.[30,31] They are the only regions that do not show resonances in 8 M guanidinium chloride and 50 °C.[30,31] Under strongly denaturing conditions less biased toward the unfolded state, 10 M urea and 50 °C, half of the residues from helices A

and B are also protected.[30,31] A few residues in the region between amino acids 60 and 80 in the β domain also show significant NMR protection under denaturing conditions.[30,31] It is interesting to note, however, that in the disulfide-less mutant form of α-lactalbumin, the 60−80 region becomes much less resistant to unfolding after elimination of the 61−77 and 73−91 disulfide bridges.[36] The SCM prediction for the primary contact differs from theoretical predictions obtained within the funnel context in the high capillarity regime,[20] in which helix C forms first, followed by helices A and B, with helix D being the last of the helical regions to form. It must be kept in mind, however, that the present form of the SCM does not include the disulfide bridges. It is possible that inclusion of the disulfide bridges in the calculations might alter the predicted folding pathway. In this regard, experimental studies with mutant forms of α-lactalbumin show that out of the six possible disulfide bridges that could form in the α domain, only the native 28−111disulfide, including a residue in the primary contact, shows a higher propensity to form than a random walk model would predict.[43] The 6−120 disulfide also included fully in the α-domain remains labile in the same experiment.[43]

4.3. Primary Contact of Hen Lysozyme. The best predicted primary contact in hen lysozyme is established between residues 28−32 and 107−111, with $P_{ij} = 11.4$. The second best primary contact is established between residues 54−58 and 120−124, with $P_{ij} = 10.4$. The predicted primary contact occurs between residues 28−32 in helix B and residue 107 in a 3^{10} helical segment, residue 108 in a bend, and residues 109−111 in helix D in the crystal structure.[44] The location of the predicted primary contact in the native structure of hen lysozyme is equivalent to that of the predicted primary contact of α-lactalbumin in its structure. The regions included in the predicted primary contact are observed to fold within the shortest observable time scale (i.e., milliseconds) in proton exchange refolding experiments together with most of the α domain of the native protein.[33,34] They are however, less protected at this early stage than other regions in the α domain, including a few residues in the vicinity of residue 95 (helix C) and around residue 124 (helix D). This difference could be related to the fact that refolding is initiated from a random coil state that includes the four disulfide bridges seen in the native structure. Two disulfide bonds are established between amino acids included in the α domain (28−111 and 6−127), and a disulfide bridge is established between the α and β domains (76−94). As in α-lactalbumin, a few residues in the 60−80 region in the β domain are also highly protected within the shortest observable time scale. These residues do not form any well-defined region of secondary structure in the native protein. There are two disulfide bridges (64−80 and 76−94) included in the region between amino acids 60 and 80. Unfolding experimental results show that residues 108−115 included in the predicted primary contact are the most resistant to pressure-assisted cold denaturation.[35] The segment 28−32 is also included in a region highly resistant to denaturation in the same experiment, second only to the 108−115 segment in the α domain. Residues 61−65, 76, and 78 in the β domain, however, show higher protection factors than the 28−32 region. The SCM predicts that the residues included in the primary contact fold first, followed by the tails, defining most of the α-domain and that residues 61−65, 76, and 78 in the β domain should fold in a later phase. Whether this prediction is correct will probably be established when refolding experiments in the submillisecond range become available for hen lysozyme as they have for other proteins (45). It is also possible that the inclusion of disulfide bridges in the calculation might significantly alters the predicted folding pathway, as discussed above. The observation that in the disulfideless mutant form of α-lactalbumin the 60−80 region becomes much less resistant to unfolding after elimination of the 61−77 and 73−91 disulfides[36] suggests that the same might apply for hen lysozyme if the 64−80 and 76−94 disulfides were eliminated by directed mutagenesis. In summary, the location of the SCM predicted primary contact for hen lysozyme is in general agreement with experimental data. There is better agreement with unfolding experiments than with refolding ones. The observed discrepancies are hypothethized to arise because of the influence of the disulfide bridges in the early refolding pathway. It is a prediction of the SCM that if refolding experiments were performed on a disulfideless mutant form of hen lysozyme, the agreement between the SCM and experiment should improve. Calculations within the funnel context, including a more complete account of the interactions than the one considered here, are able to reproduce the experimentally observed features of the folding pathway of hen lysozyme to a significant degree.[20]

4.4. MGLIS of α-Lactalbumin. The primary contact between helix B and the 3^{10} helical segment and the beginning of helix D implies that the compact region of the MGLIS includes most of the α-domain of the native structure, including helices A, B, and D and the short piece of the 3^{10} helix at the beginning of helix D. There is an open fluctuating primary loop, including most of the amino acids between helix B and the 3^{10} helical segment before helix D. The open primary loop includes the totality of the β domain of the native protein. α-Lactalbumin forms a well-defined molten globule with significant nativelike structure, especially in the regions included in the α domain of the native protein.[28,29] As explained above, the compact region of the MGLIS predicted by the SCM coincides with the regions of the α-lactalbumin molten globule experimentally most resistant to unfolding.[28−32] The main difference is the few residues in the 60−80 region inside the primary loop that also show significant NMR protection factors under denaturing conditions.[28−32] As explained above, this difference is probably due to the absence of the disulfide bridges in the SCM calculation. The molten globule of α-lactalbumin has also been studied theoretically employing molecular dynamics[46] and through the use of simplified free energy functionals within the funnel context[20] in greater detail than here. The results of these studies are in general agreement with the experimental data.

4.5. MGLIS of Hen Lysozyme. Formation of the primary contact between helix B and the 3^{10} helical segment and the beginning of helix D implies that the compact region of the MGLIS of hen lysozyme is very similar to that of α-lactalbumin. The compact region includes most of the α domain, including helices A, B, and D and a short piece of the 3^{10} helix at the beginning of helix D. There is an open fluctuating primary loop, including most of the amino acids between helices B and the 3^{10} helical segment before helix D. The open primary loop includes the totality of the β domain in the native structure. This result is consistent with the experimental observation that the α and β domains are independent folding domains in the protein with disulfide bridges.[32−35] The main difference between the experimentally observed molten globule and the MGLIS predicted by the SCM is the presence of a few folded residues in the region between residues 60 and 80 inside the primary loop. These residues are equivalent to those observed to be folded in the 60−80 region of the molten globule of α-lactalbumin. As explained above, the discrepancy is possibly due to the presence of disulfide bridges between residues 64 and 80 and 76 and 94, and it would disappear if experiments were

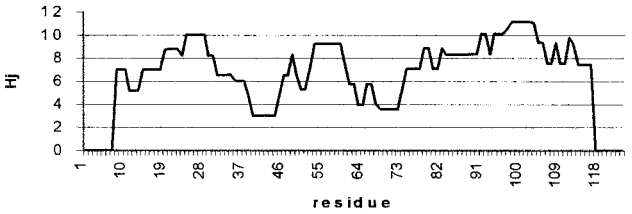

Figure 1. Relative hydrophobicity H_j vs the center of the 15 amino acid segments for α-lactalbumin. The plot shows that the sequence of formation of native contacts should start in the compact region of the MGLIS in the segment centered in residue 12. Inside the primary loop, the native structure should be attained sequentially by segments 35−49 and 63−77. The H_j values for windows in the C-and N-termini shorter than 15 amino acids were set to zero.

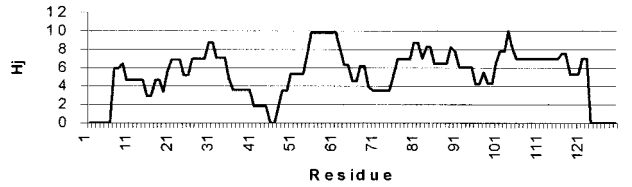

Figure 2. Relative hydrophobicity H_j vs the center of the 15 amino acid segments for hen lysozyme. The plot shows that the sequence of formation of native contacts should start in the compact region of the MGLIS in the segment centered in residue 16. Inside the primary loop, the native structure should be attained sequentially by segments 39−53 and 66−80. The H_j values for windows in the C-and N-termini shorter than 15 amino acids were set to zero.

carried out for a disulfide-less mutant form of hen lysozyme. Calculations within the funnel picture, including a consistent treatment of the disulfide bridges, are able to semiquantitatively reproduce the main features of the hen lysozyme molten globule, including the presence of nativelike structure in the 60−80 region.[20]

4.6. Optimization Subphase of α-Lactalbumin. According to the calculated H_j shown in Figure 1, within the compact region of the MGLIS, the region around residue 12 should be the fastest to reach its native structure. The regions around residues 42 and 70 inside the primary loop should attain their native structure before the rest of the primary loop. As commented before, the region around residue 70 is compact and nativelike in the molten globule of α-lactalbumin.[30,31] The possible connection of this observation with the 64−80 disulfide bridge was discussed above. Unfolding experiments in general, however, do not represent an ideal test for the suboptimization sequence because the native topology (i.e., the native contacts) of the protein core is established earlier in the cooperative collapse, and unfolding probably implies the undoing of the nativelike topology of the unfolded regions.[43]

4.7. Optimization Subphase of Hen Lysozyme. According to the calculated H_j shown in Figure 2, within the compact region of the MGLIS, the region around residues 16 should be the fastest to reach its native structure. The regions around residues 46 and 73 inside the primary loop should attain their native structure before the rest of the primary loop. There is once again a striking similarity between the predicted optimization subphase sequence of hen lysozyme and that of α-lactalbumin.

4.8. Comparison of the SCM Folding Pathways of α-Lactalbumin and Hen Lysozyme. The SCM folding pathways of α-lactalbumin and hen lysozyme are very similar. In both cases, the primary contact is established between amino acids corresponding to helix B, a 3^{10} helical segment at the beginning of helix D, and a piece of helix D in the tertiary structure. Formation of this primary contact leads to a MGLIS state, including helices A, B, D, and a strand of the 3^{10} helix. This result means that most of the α domain folds first, followed by the cooperative collapse of the primary loop that consists mostly of the amino acids that form the β domain in the native structure. The optimization subphase sequence is also similar for both proteins. This observation is consistent with theoretical and experimental results that suggest that there is often a link between structural homology and similar folding pathways in globular proteins.[2−7]

5. Conclusions

This paper explored the cooperative collapse of the protein primary loop in the SCM, and the model was applied to the folding pathways of α-lactalbumin and hen lysozyme. It was shown that the cooperative collapse process is expected to be quite specific. This conclusion arose because topological rearrangements of the folded core are unfavorable. It was also shown within the SCM that formation of cross-links between primary loops in long proteins is entropically unfavorable. This result suggests that long proteins will tend to form independent primary loops, with each one producing a distinct folding unit. The folding pathway of α-lactalbumin and hen lysozyme were studied and compared with relevant experimental data. The folding pathways of both proteins are similar, consistent with the suggested link observed between homologous structures and similar folding pathways. The findings are in general agreement with previous experimental and theoretical results; the remaining differences found probably reflect the need for a more complete account of the interactions determining the structure of the intermediate states along the folding pathway, especially the disulfide bridges. This is an important issue that deserves further investigation.

Acknowledgment. The authors would like to thank T. A. Ronneberg for useful discussions and comments. F.B.C. would like to thank the Departamento de Física de la Materia Condensada at the Universidad Autónoma de Madrid and Dr. Manuel Rico's group at the Consejo Superior de Investigaciones Científicas (Spain) for their hospitality during part of the period in which this work was prepared. H.R. acknowledges support from the Petroleum Research Fund of the American Chemical Society.

References and Notes

(1) Bergasa-Caceres, F.; Ronneberg, T. A.; Rabitz, H. A. *J. Phys. Chem. B* **1999**, *103* (44), 9749.
(2) Krebs, M.; Schmid, F. X.; Jaenicke, R. *J. Mol. Biol.* **1983**, *163*, 619.
(3) Hollecker, M.; Creighton, T. E. *J. Mol. Biol.* **1983**, *168*, 409.
(4) Chiti, F.; Taddei, N.; White, P. M.; Bucciantini, M.; Magherini, F.; Stefani, M.; Dobson, C. M. *Nat. Struct. Biol.* **1999**, *6*, 1005.
(5) Martínez, J. C.; Serrano, L. *Nat. Struct. Biol.* **1999**, *6*, 1010.
(6) Riddle, D. S.; Grantcharova, V. P.; Santiago, J. V.; Alm, E.; Ruczinski, I.; Baker, D. *Nat. Struct. Biol.* **1999**, *6*, 1016.
(7) Clarke, J.; Cota, E.; Fowler, S. B.; Hamill, S. J. *Fold. Design* **1999**, *7*, 1145.
(8) Ternstrom, T.; Mayor, U.; Akke, M.; Oliveberg, M. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 14854.
(9) Nishimura, C.; Prytulla, S.; Dyson, M. J.; Wright, P. E. *Nat. Struct. Biol.* **2000**, *7*, 679.
(10) Kuntz I. D.; Kauzmann, W. *Adv. Protein Chem.* **1978**, *28*, 239.
(11) Pauling, L.; Corey, R. B.; Branson, H. R. *Proc. Nat. Acad. Sci. U.S.A.* **1951**, *37*, 205.
(12) Dill, K. A. *Biochemistry* **1990**, *29*(31), 7133.
(13) Jacobson, H.; Stockmeyer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600.
(14) Edwards, S. F. *Proc. Phys. Soc.* **1965**, *85*, 613.
(15) Baker, D. *Nature* **2000**, *405*, 39.
(16) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. *Biochemistry* **2000**, *39*, 11177.

(17) 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
(18) Sali, A.; Shakhnovich, E.; Karplus, M.; *Nature* **1994**, *369*, 248.
(19) Bryngelson, J. D.; Onuchic J. N.; Socci, N. D.; Wolynes, P. G., *Proteins: Struct. Funct. Genet.* **1995**, *21*, 167.
(20) Shoemaker, B. A.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 657.
(21) Shoemaker, B. A.; Wang, J.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 675.
(22) Levinthal, C. J. *J. Chim. Phys.* **1968**, *65*, 44.
(23) Lazaridis, T.; Karplus, M. *Science* **1999**, *278*, 1928.
(24) White, S. H.; Jacobs, R. E.; *Biophys. J.* **1990**, *57*, 911.
(25) Go, N. *Annu. Rev. Biophys. Bioeng.* **1983**, *12*, 183.
(26) Micheletti, C.; Banavar, J. R.; Maritan, A.; Seno, F. *Phys. Rev. Lett.* **1999**, *82*, 3372.
(27) Schulz, G. E.; Schirmer, R. H. *Principles of Protein Structure*, Springer-Verlag: New York, 1979.
(28) Baum, J.; Dobson, C. M.; Evans, P. A.; Hanley, C. *Biochemistry* **1989**, *28*, 7.
(29) Peng, Z. Y.; Wu, L. C.; Kim, P. S. *Biochemistry* **1994**, *33*, 2136.
(30) Schulman, B. A.; Redfield, C.; Peng, Z. Y.; Dobson, C. M. *J. Mol. Biol.* **1995**, *253*, 651.
(31) Schulman, B. A.; Kim, P. S.; Dobson, C. M.; Redfield, C. *Nat. Struct. Biol.* **1997**, *4*, 630.
(32) Miranker, A.; Radford, S. E.; Karplus, M.; Dobson, C. M. *Nature* **1991**, *349*, 633.
(33) Radford, S. E.; Dobson, C. M.; Evans, P. A. *Nature* **1992**, *358*, 302.
(34) Gladwin, S. T.; Evans, P. A. *Fold. Design* **1996**, *1*, 407.
(35) Nash, D.; Jonas, J. *Biochemistry* **1997**, *36*, 14375.
(36) Redfield, C.; Schulman, B. A.; Milhollen, M. A.; Kim, P. S.; Dobson, C. M. *Nat. Struct. Biol.* **1999**, *6*, 948.
(37) Buck, M.; Radford, S. E.; Dobson, C. M. *J. Mol. Biol.* **1994**, *237*, 247.
(38) Eyles, S. J.; Radford, S. E.; Robinson, C. V.; Dobson, C. M. *Biochemistry* **1994**, *33*, 13038.
(39) Chung, E. W.; Nettleton, E. J.; Morgan, C. J.; Gross, M.; Miranker, A.; Radford, S. E.; Dobson, C. M.; Robinson, C. V. *Protein Sci.* **1997**, *6*, 1316.
(40) Fauchère, J. L.; Pliska, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.
(41) Leszcynski, J.; Rose, G. D. *Science* **1986**, *234*, 849.
(42) Acharya, K. R.; Stuart, D. I.; Walker, N. P.; Lewis, M.; Phillips, D. C. *J. Mol. Biol.* **1989**, *208*, 99.
(43) Wu, L. C.; Peng, Z. Y.; Kim, P. S. *Nat. Struct. Biol.* **1995**, *2*, 281.
(44) Motoshima, H.; Mine, S.; Masumoto, K.; Abe, Y.; Iwashita, H.; Hasimoto, Y.; Chijiiwa, Y.; Ueda, T.; Imoto, T. J. *J. Biochem. (Tokyo)* **1997**, *121*, 1076.
(45) Ballew, R. M.; Sabelko, J.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5759.
(46) Smith, L. J.; Dobson, C. M.; Van Gunsteren, W. F. *J. Mol. Biol.* **1999**, *286*, 1567.