Sequential Collapse Folding Pathway of Staphylococcal Nuclease: Entropic Activation Barriers to Hydrophobic Collapse of the Protein Core

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

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

Received: August 27, 2003; In Final Form: February 5, 2004

In this paper the sequential collapse model (SCM) is applied to reveal the folding pathway of staphylococcal nuclease. It is found that there are two energetically equivalent dominant primary contacts leading potentially to two distinct folding pathways. A third weaker contact is likely to initiate an additional less populated pathway. The findings are compared with previous theoretical and experimental results, including laboratory data suggesting that the later stages of the folding pathway of staphylococcal nuclease might be kinetically controlled. The activation barriers observed to control the intermediate stages of the folding pathway are postulated to correspond to the configurational activation barriers governing the cooperative collapse phase along each of the three predicted folding pathways.

1. Introduction

Since Levinthal's demonstration that there must be folding pathways,[1] the precise mechanism by which a protein's primary sequence controls its folding pathway toward the native structure has been the subject of intense experimental and theoretical study. In recent years, a wealth of experimental results and theoretical developments have allowed for the development of a number of predictive models for protein folding dynamics.[2−7] Many of these models rely on relatively simple thermodynamic and entropic considerations to investigate the main features of the protein-folding pathways.[2−7] The sequential collapse model (SCM) shares these simplifying assumptions. In the SCM the entropic consequences of forming protein loops and the hydrophobic effect are assumed to guide proteins of length ∼100−150 amino acids through a multistate folding pathway.[8−11] The purpose of this paper is to apply the SCM to reveal the folding pathway of staphylococcal nuclease. The SCM suggests that there might be two different nucleation events which are thermodynamically equivalent, for the first intermediate along the folding pathway of staphylococcal nuclease. Each of these nucleation events leads potentially to a distinct folding pathway. Additionally, a third weaker contact is likely to initiate a much less populated folding pathway parallel to the two dominant ones. This behavior is consistent with recent results within the SCM showing that there might be several possible pathways leading to the same native structure within a given protein family and for each single protein within the family.[10,11] The existence in the SCM of multiple pathways leading to the same native structure is consistent with experimental results,[12−14] and common to most current theoretical models of protein folding.[15−20] The predicted folding pathways for staphylococcal nuclease are compared and shown to be generally consistent with existing laboratory data and theoretical work,[21−51] including experimental results suggesting that the later stages of the folding pathway might be kinetically controlled.[34] The SCM suggests that the observed activation barriers have a configurational entropic origin related to the hydrophobic collapse process of the three predicted parallel folding pathways.

2. The Model

The SCM has been outlined in full detail elsewhere.[8] Here, only a brief review appropriate for the specific goals of this paper is presented. In the SCM the free energy of formation of a successful contact is written as:

$$\Delta G_{\mathrm{cont}} = \Delta G_{\mathrm{loop}} + \Delta G_{\mathrm{int}} \tag{1}$$

ΔG_{loop} is the free energy change associated with formation of protein loops and is expected to be positive because loop formation defines a state in which the backbone and the amino acid side chains have fewer conformational possibilities than in an open protein chain thus inducing a large entropic loss ΔS_{loop}. ΔG_{int} represents all the interactions that help stabilize the contact, including hydrophobic interactions, van der Waals interactions, hydrogen bonds, disulfide bonds, and salt bridges,[52] $\Delta G_{\mathrm{int}} < 0$. Hydrophobic interactions have been postulated to constitute the main driving force of the folding process.[53] Also, the hydrophobic effect has been observed to be dominant with respect to secondary structure formation propensity.[54] Thus, it is reasonable to write eq 1 as

$$\Delta G_{\mathrm{cont}} \approx \Delta G_{\mathrm{loop}} + \Delta G_{\mathrm{hyd}} \tag{2}$$

ΔG_{hyd} is expected to be negative as it is the free energy change associated with the burial of hydrophobic groups and the gain of entropy in the solvent.[52,53]

Upon formation of a contact in a protein of N amino acids, three regions are distinguished within the SCM: the contact region c of length n_c, the open connecting loop l of length n_l, and the free ends or tails of combined length $n_0 = N - (n_c + n_l)$. Up to a constant, ΔG_{loop} can be written as:[8]

$$\Delta G_{\mathrm{loop}} \approx -kT[n_l \ln(f_l/f_0) + n_c \ln(f_c/f_0) - {}^3/_2 \ln n_l] \tag{3}$$

where f_i represents the conformational freedom of the amino acid side chains in a region i of the protein and ${}^3/_2 \ln n_l$ is the

* To whom correspondence should be addressed.
‡ Current address: C/Alonso Saavedra 16 1A, 28033 Madrid, Spain. E-mail: Fbergasa@Telefonica.Net. Fax: 34 91 213 19 08.

10.1021/jp031023d CCC: $27.50
Published on Web 05/18/2004

classical Jacobson−Stockmeyer term.[55] The $n \ln(f/f_0)$ terms represent in a simple manner the internal structure and physical extension of the amino acid side chains that should become relevant when formation of a protein loop confines the amino acids within a sufficiently small volume. The conformational freedom of the amino acid side chains f in the free ends of the protein, f_0, is taken to be the same as that of the amino acids in the random coil. There might be some significant sequence dependent effects upon f. For example, sequences with high glycine content will have a lower value for $(f_l/f_0)^n$, as glycines do not have a side chain thereby increasing the free volume (i.e., the conformational freedom) available to the amino acids in the loop (i.e., for a segment rich in glycines $\ln(f_l/f_0) \rightarrow 0$). For such sequences with similar contact regions, eq 3 would show a Jacobson−Stockmeyer dependence with loop length, in accordance with recent experimental findings.[56] ΔG_{loop} can be shown to have two minima[8] as a function of loop length n_l: a deeper one at $65 < n_l < 85$ amino acids, called the optimal loop length, n_{op}, where $f_l \approx f_0$, and a shallower one at $n_l \approx 3-4$ amino acids. The shallow minimum represents the shortest length over which the protein chain can reverse its direction. Because a few amino acids are required to form a stable contact, we define a minimal loop to be generated by a protein contact between segments of ∼5 amino acids linked by a turn of 3−4 amino acids. Thus, the minimal loop size is $n_{\min} \approx n_l + 10 = 13-14$ amino acids. A heuristic estimate for the optimal distance can be derived from Gaussian chain statistics as follows: The conformational freedom f can be approximated to be proportional to the available volume v_{loop} in the loop defined by the contact, such that $f \propto v_{\text{loop}}$. To find an expression for v_{loop}, the probability distribution of the distance R_{nm} between any two monomers n and m is assumed to adopt the classical Gaussian form:[57]

$$P(R_{nm}) = [3/(2\pi b^2 n_{\text{eff}})]^{3/2} \exp[-3|\mathbf{R}_{nm}|^2/(2b^2 n_{\text{eff}})] \quad (4)$$

where $\mathbf{R}_{nm} = \mathbf{r}_n - \mathbf{r}_m$, n_{eff} is the effective number of monomers between n and m (i.e., for a loop of M residues, $n_{\text{eff}} = |n - m|[1 - |n - m|/(M - 1)]$) , and b is the average bond length. In the absence of topological constraints $R_{g,0} \sim bn^{1/2}6^{-1/2}$. By using the distribution in eq 3, it can be shown that the radius of gyration of a protein loop of length n, $R_{g,\text{loop}}$, becomes:

$$R_{g,\text{loop}} \approx R_{g,0}(1/2)^{1/2} \quad (5)$$

Now, we can take $v_{\text{loop}} \sim {R_{g,\text{loop}}}^3$, and write the side chain configurational entropic term as

$$\Delta G_{\text{loop,f}} = -kTn \ln(f_n/f_0) \propto n \ln({R_{g,\text{loop}}}^3/v_0), \qquad {R_{g,\text{loop}}}^3 < v_0 \quad (6)$$

Here $v_0 \approx {l_{aa}}^3$ is the minimal volume that allows for loop formation without chain interpenetration with l_{aa} being the characteristic amino acid side chain length, and $b = 3.5$ Å being equivalent to peptide bond length. Figure 1 shows the values of $R_{g,\text{loop}}$ as a function of loop length. Inspection of Figure 1 shows that for reasonable values of $l_{aa} \approx 8-9$ Å for the large side chains in the unfolded chain, the radius of gyration is $R_{g,\text{loop}} \approx l_{aa}$ and thus, $\Delta G_{\text{loop,f}} \approx 0$, for loop lengths of $n \approx 65-80$ amino acids. Because the Gaussian chain approach is only an approximation to the statistics of real polymers, a value for n_{op} of $65 \leq n_{\text{op}} \leq 85$ amino acids will be assumed in subsequent calculations, consistent with previous SCM estimates. A more refined analysis could specify l_{aa} for each amino acid type, but the present simplified form suffices for the goals of this paper.

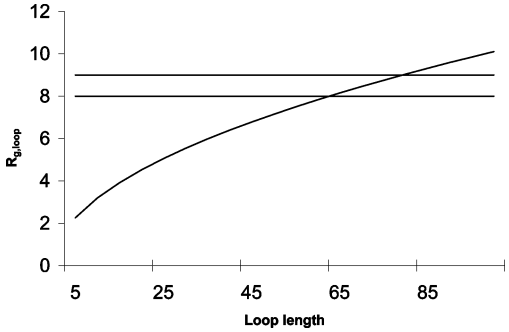

Figure 1. Radius of gyration in Å of a Gaussian chain loop $R_{g,\text{loop}}$ as a function of loop length. The horizontal lines are the values of 8 and 9 Å for $R_{g,\text{loop}}$.

On the basis of the formalism developed above, in the SCM, for proteins sufficiently long, the folding pathway is likely initiated by the formation of a contact between segments located at the optimal distance n_{op} of ∼65−85 amino acids.[8] Formation of this initial contact, referred to as the primary contact, leads to a multistate folding pathway that includes an intermediate state with many of the properties of a molten globule,[58] and therefore referred to as the molten globule-like intermediate state (MGLIS).[8] Formation of the initial contact is followed by the folding of the residues located outside the primary loop, referred to as the tails of the protein. This is followed by a hydrophobic collapse of the protein core in which the native topology of the protein is established.[9] The collapse is followed by an optimization subphase governed by the activation barriers generated by the need to fully eliminate any water not excluded in the hydrophobic collapse from the interior of the protein core in order to fully establish the interactions that stabilize the native structure.

The SCM is in the same spirit as recent theoretical efforts to develop models to describe the intermediates along the folding pathway in the context of the so-called "funnel" view of the folding process.[16−18] The SCM, however, shares much of the "old" view of the folding process in which the protein descends toward the free energy minimum through a few intermediate steps.[18] Although the SCM does not preclude the possibility that a protein might fold through several parallel pathways,[8] it suggests that there is a strong preference for those pathways that minimize the conformational entropy loss upon 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 as 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.[18]

2.1. Estimating the Relative Populations of Early Intermediates along Parallel Pathways. For many protein examples it is likely that more than one primary contact stable enough to nucleate early folding will exist. Then, a relevant question is to determine the relative populations of the primary contacts, as a first approximation to the populations of molecules fully folding through each of the parallel pathways. The primary contacts are defined by the condition:

$$\Delta G_{\text{cont}} \approx \Delta G_{\text{loop}} + \Delta G_{\text{hyd}} < 0 \quad (7)$$

and, from eq 3, up to a constant, ΔG_{loop} for a primary contact becomes:

$$\Delta G_{loop} = -kT[c \ln(f_c/f_0) - {}^3/_2 \ln n] \quad (8)$$

The difference in ΔG_{loop} between two distinct primary contacts l and k can be written as

$$\Delta\Delta G_{loop} = -kT[(c_l - c_k) \ln(f_c/f_0) - {}^3/_2 \ln(n_l/n_k)] \quad (9)$$

For most primary contacts $c_l \approx c_k$ and $n_l \approx n_k$ so $\Delta\Delta G_{loop} \approx 0$ and the difference in stabilization energy is $\Delta\Delta G_{cont} \approx \Delta\Delta G_{hyd}$, where ΔG_{hyd} is a measure of the hydrophobic interactions between amino acid side chains. Moreover, hydrophobic interactions are relatively weak as compared to the covalent interactions defining chain connectivity.[52] It is experimentally and theoretically well-established that (a) the overall protein stabilization free energy is low[52] and (b) it is much lower than the simple sum of the hydrophobic interactions providing most of the attractive interactions.[52] Then, it is expected that the absolute value of the configurational entropic term $|-kTc \ln(f_c/f_0)|$ is only marginally smaller than $|\Delta G_{hyd}|$ and that ΔG_{cont} must be small. The loop closure entropic term ΔG_{loop} is also expected to generate an activation barrier to primary contact formation of the form:[64,65]

$$E_{loop} = \Delta G_{loop} \quad (10)$$

The difference in magnitude of the activation barriers to formation of two distinct primary contacts $\Delta\Delta E_{loop}$ is then $\Delta\Delta E_{loop} = \Delta\Delta G_{loop} \approx 0$. Due to the absence of the large loop side chain entropy term $n \ln(f_l/f_0)$ in ΔG_{loop} for primary contacts, the activation barriers to formation of primary contacts are significantly smaller than the activation barriers to contact formation in the protein tails. This difference suggests that primary contact formation takes place on a time scale that probably allows for a quasiequilibrium distribution to arise, prior to the full formation of the MGLIS. Because of the weak interactions, the absence of discriminating activation barriers, and the difference in time scales between the formation of the primary contact and the subsequent folding of the tails, the relative populations N_k of any primary contact k within a set of possible primary contacts can be estimated to follow a Bolztmann-like distribution:

$$N_k \approx \exp[-\Delta G_{hyd}(\text{contact k})/(kT)]/\boldsymbol{Z} \quad (11)$$

where ΔG_{hyd}(contact k) is the free energy of stabilization due to the establishment of hydrophobic interactions in contact k, and Z is the partition function $Z = \sum_k \exp[-\Delta G_{hyd}(\text{contact k})/(kT)]$.

3. Folding Pathway of Staphylococcal Nuclease

Staphylococcal nuclease is a 149 amino acid long protein without disulfide bridges or prosthetic groups. Its structure[45] consists of an "OB" fold formed by a α-helix labeled α1 and a five-strand greek-key β-barrel, accompanied by two α-helices, labeled α2 and α3. It has been extensively studied over the years and has played an important role in the folding field. In this section, results are presented for the folding pathway of staphylococcal nuclease and a comparison is made with existing experimental and theoretical data. The SCM predictions are compared with experimental and theoretical efforts intended to elucidate the folding pathway of staphylococcal nuclease.[21−51]

3.1. Calculational Method. In this paper, we follow closely the method introduced previously to determine the sequence of folding events along the SCM folding pathway of apomyoglobin, cytochrome *c*, barnase and ribonuclease A,[8,60] hen lysozyme and α-lactalbumin,[9] apoleghemoglobin,[10] and β-lactoglobulin.[11] The primary contact is determined by the minimum value of ΔG_{hyd} for segments of five amino acids located 65 to 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[61] were assigned to each residue. To summarize the procedure, the hydrophobicity P_k of each residue is added over a contact window of five amino acid segments centered at residue i, resulting in a contact formation propensity P_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 to 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 to 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.[61] The relative populations of the early intermediates defined by the primary contacts are estimated to follow a Boltzmann-like distribution $\sim\exp[P_{ij}/0.45]/Z$, where Z is the partition function. The error in the populations is estimated to be ∼1−2% on the basis of the errors inherent in the hydrophobicity measurements in the Fauchère−Pliska scale alone.[61] The model, however, involves several other aproximations as explained in section 2, and the actual errors are likely to be substantially higher, probably closer to ∼10%.

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,[62] could be detected in the cooperative collapse sequence; it is also equivalent to the minimal loop length allowed by the model in the unfolded chain.[8]

The hydrophobicities P_k of 15 consecutive residues centered at residue j are summed, resulting in a hydrophobicity value H_j. The H_j values 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, with E^j for each protein segment taken 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).

The calculational method, employing only partition coefficients, is consistent with the underlying assumption in the SCM that secondary structure propensities do not usually play a critical role in determining the early sequence of contact-forming events along the folding pathway.[11] Secondary structure propensities could, however, be included if deemed necessary, for example adding a term reflecting the secondary structure free energy of formation to the contact formation propensity P_{ij}. This additional term could be derived from experimental data in order to be fully consistent with the chosen hydrophibicity coefficients.[54]

3.2. Primary Contacts of Staphylococcal Nuclease. Values for P_i are shown in Figure 2a. The best predicted primary contact in staphylococcal nuclease is established between (a) residues 34−38 and 111−115, with $P_{ij} = 10.6$ and an expected

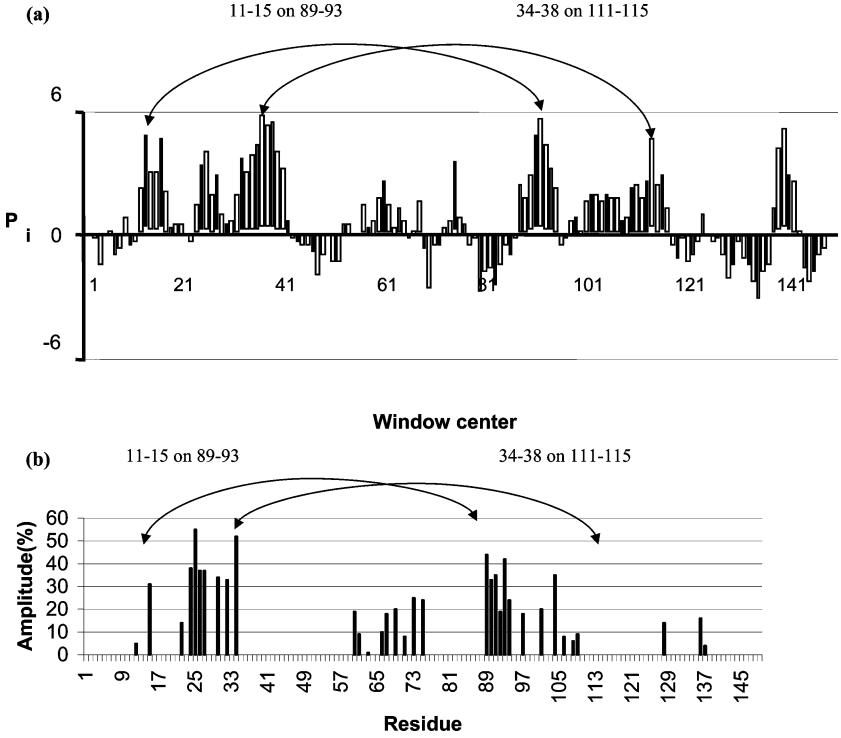

Figure 2. Early contact formation phase of staphylococcal nuclease: (a) the contact formation propensity P_i calculated for windows of five amino acids centered at amino acid *i*, and (b) amplitude of protected populations of experimental hydrogen exchange probes after 5 ms of refolding, no probes are available in segment 111−115 and just one in each of segments 34−38 and 11−15 (data from ref 29). The arrows in parts a and b indicate the two best primary contacts

population of ∼50%, and (b) residues 11−15 and 89−93, with $P_{ij} = 10.5$ and an expected population of ∼40%. The next best predicted primary contact is (c) established between segments 23−27 and 89−93 with $P_{ij} = 9.8$ and a expected population of ∼10%, about ∼1.5*kT* lower than the dominant pair of primary contacts. These P_{ij} values imply that the folding pathways of ∼90% of molecules will be initiated by either of the two energetically equivalent best primary contacts, leading to two early contact formation phase intermediates I(a) and I(b) for primary contacts (a) and (b), respectively, and the additional ∼10% will fold mostly through the formation of the [23−27, 89−93] primary contact leading to the early phase intermediate I(c). In the following sections the assumption will be made that the three-folding pathways coexist and the comparison with experimental data will mainly be done on this basis. There might be constraints downstream along the folding process preventing one of the coexisting folding pathways to proceed to the native structure.

A hydrogen exchange experiment probing the folding pathway of a P117G mutant form of staphylococcal nuclease based on the resonances of only 39 residues is available.[29] It is important to bear in mind that this comparison is only indicative, as the SCM does not require that a primary contact has nativelike secondary structure.[8,11] It is, however, likely that early definition of the contact might lead in many cases to early formation of nativelike secondary structure, especially when the secondary structure is defined locally (i.e., α-helices), rather than globally (β-sheets); the SCM predictions have reasonably reproduced proton exchange results for a number of proteins.[8−11] Also, it is likely that primary contact formation takes place on shorter time scales than those accessible to proton exchange techniques.[63] The hydrogen exchange experiment identified a fast phase that takes place within ∼5 ms. Figure 2b shows the experimentally determined populations of protected exchange probes at 5 ms.[29] Inspection of Figure 2b shows that the experimental results are in general agreement with the primary contact predictions, with all the highly populated probes located between the N-terminus and residue 40, and between the C-terminus and residue 89. All probes between 40 and 88 show low populations in this early folding stage. Only a hydrophobic residue, I15, was probed by hydrogen exchange within the 11−15 segment included in primary contact (a).[29] It has a high fast phase amplitude but shows low protection after 100 ms of refolding. I15 however, has a low intrinsic exchange rate, which may lead to incomplete labeling.[29] Residue T13 is not nativelike at this stage. All residues within segment 89−93, included in primary contact (a), have been probed by hydrogen exchange. All of them show high fast phase amplitudes. Protection factors within segment 89−93 are not too high after 100 ms, but L89, A90, and I92 also have low intrinsic exchange rates that may lead to incomplete labeling, thus artificially decreasing their experimental protection factors.[29]

Within segment 34−38 corresponding to primary contact (b), only F34 was probed. It shows the second highest fast phase amplitude of all the residues probed and its protection factor after 100 ms is very high.[29] No residues within segment 111−115 were probed, so no conclusions can be reached regarding their involvement in the earliest detectable intermediates.

A second proton exchange experiment on the H124L mutant including 60 probes has become recently available with results broadly comparable to those of the earlier experiment.[50] This experiment confirmed that there is significant folding before the experimental dead time (10 ms). Initial exchange amplitudes (rather than protection amplitudes) were measured for the 60 probes and the lowest of these were found for residues in β strands 2 and 3 including both segments 23−27 and 34−38 included in primary contacts (a) and (c), indicating that these regions are mostly protected in folding events that take place before the shortest time scale probed in the experiment. When protection factors were measured by varying the labeling pulse

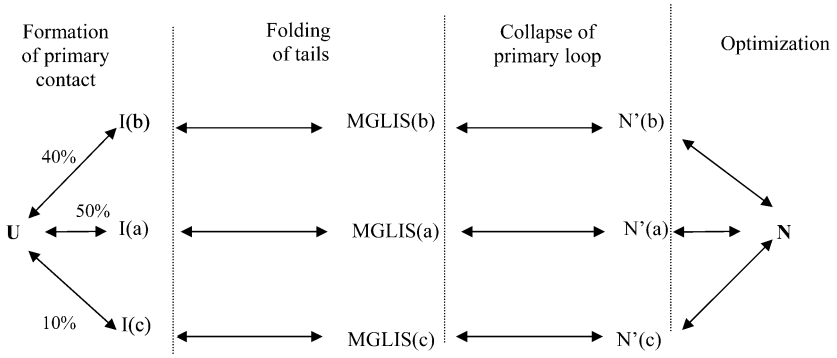

Figure 3. The three parallel folding pathways postulated by the SCM to be available for the folding of staphylococcal nuclease, and their relative populations.

at a fixed refolding time of 16 ms, it was observed that the highest protection factors corresponded to strands 2 and 3 and residue K134 close to the C-terminus and included in the C-terminal tail defined by the three predicted primary contacts. Protection factors in regions 1−38 and 88−143, including the three predicted primary contacts, were higher in general than those for the rest of the protein predicted to be included in the open fluctuating primary loop in all three early intermediates.[50]

To summarize, no definitive conclusions can be derived from comparison between the SCM predictions and the proton exchange experiment, although they are consistent. Also, due caution must be exerted when comparing hydrogen exchange results with SCM predictions as the SCM does not require that a primary contact has nativelike secondary structure, as it assumes that hydrophobic interactions are sufficient to determine the location of the contact. Better resolution experiments on the very early stages of the folding pathway, as already available for other proteins such as apomyoglobin,[63] could help clarify to what extent the SCM predictions are correct.

The segments included in primary contacts (a) and (c) are reasonably close in the native structure. Primary contact (b), however, is not a contact at all in the structure although the segments that define it are not too distant. Experimental evidence exists that, in the denatured state, residues 13−39 form a stable non-native structure with β strand-like character.[51] The main stabilization energy for this structure is provided by residues 36−39, all but one included in primary contact (a). These residues form a different set of tertiary interactions in the native state. This observation is consistent with the SCM results that predict that the N-terminal region up to residue 39 should become compact early along the folding pathway thus becoming one of the regions most protected to denaturation. Also, the fact that the region displays non-native interactions involving at least segment 34−38 included in primary contact (a) is consistent with the SCM prediction that the structure of the earliest folding intermediates, defined alone by hydrophobic interactions, need not be fully nativelike.

3.3. MGLIS along Pathways a, b, and c. The main events along the folding pathways a, b, and c are shown in Figure 3. Formation of primary contact (a) implies that the folded region primarily consists of the domain between the N-terminal region of the protein chain and segment 34−38 and the C-terminal region from segment 111−115. These regions should be topologically nativelike early along the folding pathway. Formation of primary contact (b) implies that the folded region primarily consists of the domain between the regions comprised between the N-terminus of the protein chain and segment 51−55 and the C-terminus and segment 116−120. These regions should be topologically nativelike early along the folding pathway. Formation of primary contact (c) implies that the folded region primarily consists of the domain between the N-terminal region of the protein chain and segment 23−27 and the C-terminal region from segment 89−93. These regions should be topologically nativelike early along the folding pathway. The combined populations of MGLIS (b), MGLIS (a), and MGLIS(c) should make up the majority of folding molecules. Because the N-terminal region up to segment 11−15 and the C-terminal region from segment 111−115 are included in all MGLISs, it is to be expected that they will show on average the highest protection factors in a refolding experiment. Also, because segment 40−88 is not included in any of the MGLISs, it should show the lowest degree of early folding. As explained in section 3.2. Figure 2b shows the observed experimental protection[29] of the 39 hydrogen exhange probes after 5 ms of refolding. Inspection of Figure 2b shows that the experimental results are consistent with the SCM predictions that the regions between the N-terminus and residue 39, and the C-terminus and residue 88 should fold earlier than the rest of the protein.[29]

It was suggested by Anfinsen that the folding of staphylococcal nuclease involves an intermediate state with folded N- and C-terminal regions and a disordered region between.[44] The SCM predicts that such an intermediate is nucleated by a long-range early contact (i.e., the primary contact) between the N- and C-terminal regions, while in Anfinsen's scheme, the two terminal regions are assumed to fold independently at first. Several truncated forms of staphylococcal nuclease have been shown to be compact although disordered.[25,27,30,32,33,35−37] The truncated protein Δ131Δ, which lacks residues 4−12 and 141−149, and Δ137−149, which lacks the 13 C-terminal residues, retain biological activity[25,32] and Δ131Δ has been shown to have essentially nativelike topology even in extreme denaturing conditions.[36,37,47] The existence of nativelike topology in Δ131Δ in the absence of the full secondary structure is consistent with the SCM postulate that the hydrophobic effect and the entropy loss associated with the increased steric hindrance due to loop-closure alone suffice to determine a protein's broad topology.[8] Truncated staphylococcal nuclease lacking the 16 C-terminal amino acids has been shown to be compact but lacks most of the native secondary structure.[27] The results of small-angle X-ray scattering experiments (SAXS) performed on the truncated molecules are consistent with a bipartite structure, in which there would be a compact domain and a second disorganized one, although this is not the only shape consistent with the results.[27] This experiment also suggests that the N- and C-terminal regions of the truncated protein are among the most compact, as the NMR resonances of the four histidine residues (H8, H46, H121, H124) included in the terminal regions

show very low dispersions.[27] Within the SCM, this observation suggests that in such a truncated form, the full native structure cannot be successfully reached because of missing interactions due to the excision of the C-terminal region, between the compact regions of the MGLIS and the primary loop. The existence of folding cooperativity requirements between the compact region of the MGLIS and the primary loop to reach the full native structure could explain why the MGLIS of a few proteins investigated within the SCM do not display a fully native compact region in their MGLIS.[10]

There is evidence for a hydrophobic collapse after early nucleation in staphylococcal nuclease.[43] Fluorescence energy transfer experiments on a C64→L64 mutant protein show that 60% of the native signal is gained when only ∼20−40% of the native secondary structure is in place.[43] This result is consistent with the SCM hypothesis that the hydrophobic collapse of the primary loop establishes the native topology of the protein core,[9] but not necessarily its native secondary structure.[10,11] The three MGLISs defined by the primary contacts include, if fully folded, up to ∼50% of the secondary structure of the native protein (considering equal populations for MGLISs (a), (b) and (c)). However, this is an upper bound, as it is likely that interactions between the compact region of the MGLIS and the folded primary loop play a role in defining the full native structure of the protein.

Numerous studies have shown that the folding of staphylococcal nuclease involves several intermediate steps, although a consensus picture of the folding pathway of staphylococcal nuclease has yet to emerge. In general, most studies have detected at least three folding phases.[29,34] A comprehensive picture of the folding pathway of staphylococcal nuclease was recently proposed on the basis of experimental evidence obtained employing a proline-free mutant.[39] The authors proposed that there were two basically unfolded states U_1 and U_2, leading to two distinct intermediates I_1 and I_2 that converge into a single intermediate state M, prior to the final formation of the native state.

The SCM prediction is partially consistent with this scheme. The two unfolded states U_1 and U_2 would probably correspond to (a) the formation of the two dominant primary contacts accounting for ∼90% of the early intermediates, with the protein mostly unfolded at this stage and little or no secondary structure formed at all, or (b) alternatively, it is possible that U_1 and U_2 represent in fact the three predicted SCM early intermediates, given that I(b) and I(c) are relatively similar, and the experiments upon which the proposed folding scheme is based do not have single amino acid resolution. Assuming alternative a, the two intermediate states I_1 and I_2 could be the two MGLISs defined by the two dominant primary contacts. Finally, state M would correspond to the protein after collapse of the primary loop. Because there are two different MGLIS states, two slightly different M states would likely exist with each one defined by the collapse of the corresponding primary loop. The two states would, however, be very similar, as the whole protein would have collapsed into a compact structure with native topology and only minor differences in the pattern of hydration likely to remain. These two states, N′(a) and N′ (b), and eventually N′-(c), are likely to be essentially indistinguishable. Also, the experimental data for refolding are also consistent with a fully parallel scheme in which every I state leads directly to the native state.[39] State M is introduced to fit additional data coming from unfolding experiments. Experimental data exist regarding the solvent-exposed area of each of the intermediate states.[39,46] Intermediate states I_1 and I_2 seem to represent states with $\sim^2/_3$ as much solvent-excluded area as the native state. In contrast, the transition I → M seems to involve a smaller change in solvent exposure.[39] This value is somewhat higher than the values expected for the MGLIS in the SCM, closer to $\sim^1/_2$ as the open primary loops represent $\sim^1/_2$ of the total protein length. It is possible that water accessibility is already somewhat lower for the amino acids in the open primary loop than for those in the random coil. It is also possible that the I states correspond to the already collapsed states of staphylococcal nuclease rather than to the MGLISs, while the U states could correspond to the MGLISs, given that very little secondary structure forms in the tails upon nucleation by the primary contact. Only more detailed experimental evidence about the nature of the intermediate states will help clarify this issue.

3.4. Kinetics of the Cooperative Collapse Phase: Activation Barriers. A stopped-flow CD experiment by Su et al.[34] on staphylococcal nuclease distinguished three folding phases: one faster than ∼20 ms, an intermediate phase involving at least three distinct folding events between ∼20 ms and ∼500 s, and one slower than ∼500 s. These findings are consistent with the phases observed in the proton exchange experiment.[29] The faster phase involved the formation of ∼20% of the native secondary structure, while the intermediate phase involved the formation of up to ∼70% of the native structure; the remaining 10% folded very slowly afterward. The slow phase is probably related to late proline isomerization in a fraction of the almost fully folded molecules.[34] Activation barriers were observed to control three folding events included in the intermediate phase,[34] rather than thermodynamic stability. Also, a linear sequential folding model with three partially unfolded intermediate states U, of the form $U_0 \leftrightarrow U_1 \leftrightarrow U_2 \leftrightarrow N$, was suggested. The three activation barriers in the intermediate phase were estimated to be of similar magnitude, $E_a \approx 17-19$ kcal mol^{-1}, on the basis of a simple Arrhenius transition state theory analysis, with the observed rate taken to be represented by the expression $\kappa = (kT/h)\exp(-\Delta G_{TS})$, where h is Plank's constant. On the basis of this analysis, two of the barriers were postulated to be dominantly entropic with $E_{a,S} \approx 11-12$ kcal mol^{-1}, where $E_a = E_{a,S} + E_{a,H}$, while the other barrier was mostly enthalpic with $E_{a,S} \approx 5$ kcal mol^{-1}. The two postulated entropic activation barriers were observed to correspond to relatively fast processes with rate constants of 0.15 and 1.15 s^{-1}, while the enthalpic barrier had an associated rate constant of 0.032 s^{-1}. In this section an alternative explanation for these results consistent with the three pathways suggested by the results obtained in the previous sections will be proposed: the three activation barriers will be argued to most likely correspond to the activation barriers in the hydrophobic collapse leading to the native state from the three MGLIS states.

Information about the size of the configurational entropic activation barrier for collapse of the primary loop can be obtained according to the SCM, from observations of the collapse of two-state folding proteins, as the hydrophobic collapse of the primary loop and the two-state collapse of small proteins are physically equivalent processes in the SCM.[64] Application of simple transition state theory within the SCM to the collapse of proteins with two-state kinetics has been shown to be sufficient to reproduce the regular relationship observed between the folding rate and the contact order of the native topology.[64,65] In the simplest kinetic model compatible with the SCM, the folding rate κ_{coll} of the collapsing region can be expressed as:

$$\kappa_{coll} \approx g_0 \exp[E^{loop}_{ij}/(NkT) - \Delta G_{TS}/(NkT)] \qquad (12)$$

where g_0 is a characteristic configurational transition frequency,

ΔG_{TS} is the free energy of stabilization of the transition state, N is the number of native contacts, and E^{loop}_{ij} is the overall configurational entropic activation barrier for formation of a contact upon collapse.[64,65] This formulation differs from the one employed in the analysis by Su et al. of their experimental results.[34] Instead of assuming a simple preexponential factor of the form (kT/h), a more complex form with $g = g_0 \exp[E^{loop}_{ij}/(NkT)]$ is employed as warranted by earlier theoretical results,[64,65] and existing experimental data.[66] The average SCM entropic barrier to contact formation is of the form $E^{loop}_{ij} = \langle n \rangle \ln[f(\langle n \rangle)/f_0]$,[64] where $\langle n \rangle$ is the average distance along the sequence between amino acids forming contacts $\langle n \rangle = (C.O.)\text{-}L$, where C.O. is the contact order of the native structure and L is the length of the collapsing region. Direct information about the size of the average entropic activation barrier $E^{loop}_{ij}/(NkT)$ can then be obtained by determining the slope of plots of ln κ_{coll} vs T^{-1} at fixed $\Delta G_{TS}/(kT)$.[66] Experimental evidence exists for the dependence of the rate of collapse with respect to the temperature for CspB and protein L, and a direct application of the formalism described above suggests that the size of the average activation barrier to collapse is $E^{loop}_{ij} N^{-1} \equiv E_a \approx 22$ kcal mol^{-1} [66] for $\langle n \rangle \approx 11$ amino acids, with $\Delta G_{TS} = 0.0135$ kcal mol^{-1} K^{-1}. The low observed transition state energy is expected within the SCM, as it is assumed that the transition state is held together by hydrophobic interactions that barely compensate the large configurational entropic loss due to compaction.[9,64] Then, it is reasonable to assume that, to a good approximation we can write eq 12 as

$$\kappa_{coll} \sim g_0 \exp[E^{loop}_{ij}/(NkT)] \qquad (13)$$

Within this framework, the three barriers experimentally observed in the experiment by Su et al.[34] would be mostly entropic, corresponding to the average configurational difussion barrier for collapse, and are of comparable size, $E_a \approx E^{loop}_{ij} \approx$ 17−19 kcal mol^{-1}. For staphylococcal nuclease, in first approximation we can take $\langle n \rangle \approx 20$ for the whole protein (with C.O. $\approx$ 13.7);[67] more precise calculations should consider only the contact order of those regions of the protein involved in the cooperative collapse. Because the collapse of the primary loops (a), (b), and (c) seems to involve significant interactions with the rest of the protein chain, it has been assumed here that consideration of the contact order of the whole native structure provides for a good approximation to the contact order of the collapsing region. The relative size of the activation barriers for staphylococcal nuclease and CsbP can be estimated as follows: the conformational freedom f can be taken as proportional to the available volume v for the amino acid side chain in a given configuration.[11] Following the discussion in section 2, this volume can be taken to scale as $v \approx n^{3/2}$, where n is the loop length. Then, E_a(Sn)/E_a(CsbP) can be written as

$$E_a(\text{Sn})/E_a(\text{CspB}) \approx \{\langle n(\text{Sn}) \rangle \ln[\langle n(\text{Sn}) \rangle / n_{op}]\}/\{\langle n(\text{CspB}) \rangle \ln[\langle n(\text{CspB}) \rangle / n_{op}]\} \qquad (14)$$

which yields a value of E_a(nucl.)/E_a(CspB) $\approx$ 1.25, for $n_{op} =$ 75 amino acids, suggesting that the two entropic activation barriers should be comparable to a reasonable first approximation. In fact, the value of E_a(CspB) $\approx$ 22 kcal mol^{-1} is comparable to those observed for the entropic activation barriers in the intermediate folding stages of staphylococcal nuclease if eqs 12 and 13 are applied to the analysis of the experimental data, suggesting that the three similar barriers observed in the intermediate folding stages of staphylococcal nuclease might correspond to the hydrophobic collapses of the three primary loops defined by the three distinct primary contacts (a), (b), and (c). Better experimental resolution will be necessary to establish the validity of these conclusions.

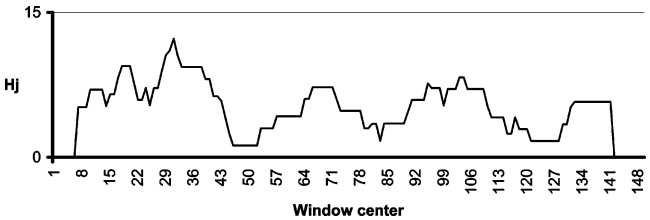

Figure 4. The relative hydrophobicity H_j versus the center j of the 15 amino acid segments for staphylococcal nuclease.

A least activation pathway (LAP) folding model was proposed on the basis of the experimental findings described above.[34] In such a model, activation barriers along the folding pathway would kinetically control protein folding. In the SCM, the situation is more complicated as inspection of eq 12 suggests that, while activation barriers are important in determining the folding rates, thermodynamic stability also plays a significant role through ΔG_{TS}. Also, the location of the primary contacts in the earlier folding stages is thermodynamically controlled in the SCM.

3.5. Optimization Subphase along Pathways (a), (b), and (c). Values of H_j for staphylococcal nuclease are shown in Figure 4. The regions around residues 50, 80−81, and 26 inside primary loop (a) should attain their native structure before the rest of the primary loop. Similarly, the regions around residues 50, 84, and 100 inside primary loops (b) and (c) should attain their native structure before the rest of the primary loop. Although the optimization subphase has been generally assumed within the SCM to be controlled by the activation barriers due to the need to fully exclude any remaining water still attached to the protein's buried surface, other activation barriers might also be relevant.[68,69]

4. Conclusions

In this paper the SCM was applied to reveal the folding pathway of staphylococcal nuclease and it was found that the model predicts two dominant independent folding pathways initiated by the formation of distinct primary contacts, plus probably an additional pathway nucleated by a weaker primary contact. These results were shown to be consistent with existing laboratory data. The activation barriers observed to control the folding rates in the experimental intermediate folding phase were postulated to reflect the entropic activation barriers to cooperative collapse of the primary loop predicted by the model. The significance of these results was also discussed in connection with the control of the folding pathway.

Acknowledgment. F.B.C. would like to thank Prof. Juan J. Saenz for useful discussions on polymer statistics. H.R. acknowledges partial support from the National Science Foundation.

References and Notes

(1) Levinthal, C. *J. Chim. Phys.* **1968**, *65*, 44.
(2) Alm, E.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11305.
(3) Galzitskaya, O. V.; Finkelstein, A. V. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11299.
(4) Muñoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.
(5) Shoemaker, B. A.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 657.
(6) Shoemaker, B. A.; Wang, J.; Wolynes, P. G. *J. Mol. Biol.* **1999**, *287*, 675.

(7) Portman, J. J.; Takada, S.; Wolynes, P. G. *J. Chem. Phys.* **2001**, *114*, 5082.
(8) Bergasa-Caceres, F.; Ronneberg, T. A.; Rabitz, H. A. *J. Phys. Chem. B* **1999**, *103*, 9749.
(9) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2001**, *105*, 2874.
(10) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2002**, *106*, 4818.
(11) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2003**, *107*, 3606.
(12) Radford, S. E.; Dobson, C. M.; Evans, P. A. *Nature* **1992**, *358*, 302.
(13) Chiti, F.; Taddei, N.; White, P. M.; Bucciantini, M.; Magherini, F.; Stefani, M.; Dobson, C. M. *Nat. Struct. Biol.* **1999**, *6*, 1005.
(14) Nishimura, C.; Prytulla, S.; Dyson, H. J.; Wright, P. E. *Nat. Struct. Biol.* **2000**, *7*, 679.
(15) Karplus M.; Weaver, D. L. *Nature* **1976**, *260*, 404.
(16) Sali, A.; Shakhnovich, E.; Karplus, M. *Nature* **1994**, *369*, 248.
(17) Bryngelson, J. D.; Onuchic J. N.; Socci, N. D.; Wolynes, P. G. *Proteins*: *Struct. Funct. Genet.* **1995**, *21*, 167.
(18) Lazaridis, T.; Karplus, M. *Science* **1999**, *278*, 1928.
(19) Wolynes, P. G.; Onuchic, J. N.; Thirumalai, D. *Science* **1995**, *267*, 1619.
(20) Harrison, S. C.; Durbin, R. *Proc. Natl. Acad. Sci. U.S.A.* **1985**, *82*, 4028.
(21) Schechter, A. N.; Chen, R. F.; Anfinsen, C. B. *Science* **1970**, *167*, 886.
(22) Epstein, H. F.; Schechter, A. N.; Chen, R. F.; Anfinsen, C. B. *J. Mol. Biol.* **1971**, *60*, 499.
(23) Davis, A.; Parr, G. R.; Taniuchi, H. *Biochim. Biophys. Acta* **1979**, *578*, 505.
(24) Fox, R. O.; Evans, P. A.; Dobson, C. M. *Nature* **1986**, *320*, 192.
(25) Shortle, D.; Meeker, A. K. *Biochemistry* **1989**, *28*, 936.
(26) Sugawara, T.; Kuwajima, K.; Sugai, S. *Biochemistry* **1991**, *30*, 2698.
(27) Flanagan, J. M.; Kataoka, M.; Shortle, D.; Engelman, D. M. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 748.
(28) Nakano, T.; Antonino, L. C.; Fox, R. O.; Fink, A. L. *Biochemistry* **1993**, *32*, 2534.
(29) Jacobs, M. D.; Fox, R. O. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 449.
(30) Alexandrescu, A. T.; Shortle, D. *J. Mol. Biol.* **1994**, *242*, 527.
(31) Kalnin, N. N.; Kuwajima, K. *Proteins*: *Struct. Funct. Genet.* **1995**, *23*, 163.
(32) Alexandrescu, A. T.; Abeygunawardana, C.; Shortle, D. *Biochemistry* **1994**, *33*, 1063.
(33) Wang, Y.; Shortle, D. *Biochemistry* **1995**, *34*, 15895.
(34) Su, Z.-D.; Arooz, M. T.; Chen, H. M.; Gross, C. J.; Tsong, T. Y. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2539.
(35) Alexandrescu, A. T.; Janhke, W.; Wiltscheck, R.; Blommers, M. J. J. *J. Mol. Biol.* **1996**, *260*, 570.
(36) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 158.
(37) Gillespie, J. R.; Shortle, D. *J. Mol. Biol.* **1997**, *268*, 170.
(38) Frye, K. J.; Royer, C. A. *Prot. Sci.* **1997**, *6*, 789.
(39) Walkenhorst, W. F.; Green, S. M.; Roder, H. *Biochemistry* **1997**, *36*, 5795.
(40) Uversky, V. N.; Karnoup, A. S.; Segel, D. J.; Seshadri, S.; Doniach, S.; Fink, A. L. *J. Mol. Biol.* **1998**, *278*, 879.
(41) Panick, G.; Malessa, R.; Winter, R.; Rapp, G.; Frye, K. J.; Royer, C. A. *J. Mol. Biol.* **1998**, *275*, 389.
(42) Wrabl, J.; Shortle, D. *Nat. Struct. Biol.* **1999**, *6*, 876.
(43) Nishimura, C.; Riley, R.; Eastman, P.; Fink, A. L. *J. Mol. Biol.* **2000**, *299*, 1133.
(44) Anfinsen, C. B. *Biochem. J.* **1972**, *128*, 737.
(45) Hynes, T. R.; Fox, R. O. *Proteins:Struct. Func. Genet.* **1991**, *10*, 92.
(46) Wrabl, J.; Shortle, D. *Nat. Struct. Biol.* **1999**, *6*, 876.
(47) Shortle D.; Ackerman, M. S. *Science* **2001**, *293*, 487.
(48) Woenckhaus, J.; Köhling, R.; Thiyagarajan, P.; Littrell, K. C.; Seifert, S.; Royer, C. A.; Winter, R. *Biophys. J.* **2001**, *80*, 1518.
(49) Sinclair, J. F.; Shortle, D. *Prot. Sci.* **1999**, *8*, 991.
(50) Walkenhorst, W. F.; Edwards, J. A.; Markley, J. L.; Roder, H. *Prot. Sci.* **2002**, *11*, 82.
(51) Wang, Y.; Shortle, D. *Prot. Sci.* **1996**, *5*, 1898.
(52) Dill, K. A. *Biochemistry* **1990**, *29* (31), 7133.
(53) Kuntz I. D.; Kauzmann, W. *Adv. Protein Chem.* **1978**, *28*, 239.
(54) O'Neil, K. T.; Degrado, W. F. *Science* **1990**, *250*, 646.
(55) Jacobson, H.; Stockmeyer, W. H. *J. Chem. Phys.* **1950**, *18*, 1600.
(56) Lapidus, L. J.; Eaton, W. A.; Hofrichter, J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 7220.
(57) Doi, M.; Edwards, S. F. *The theory of polymer dynamics*; Oxford Science Publications: New York,1986; pp 24−32, .
(58) Kuwajima, K. *Proteins*: *Struct. Func. Genet.* **1989**, *6*, 87.
(59) Loll, P. J.; Lattman, E. E. *Proteins*: *Struct. Funct. Genet.* **1989**, *5*, 183.
(60) Bergasa-Caceres, F. Ph.D. Thesis, Princeton University, 1996.
(61) Fauchère, J. L.; Pliska, V. *Eur. J. Med. Chem.* **1983**, *18*, 369.
(62) Leszcynski, J.; Rose, G. D. *Science* **1986**, *234*, 849.
(63) Ballew, R. M.; Sabelko, J.; Gruebele, M. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 5759.
(64) Bergasa-Caceres, F.; Rabitz, H. A. *Chem. Phys. Lett.* **2003**, *376*, 612.
(65) Bergasa-Caceres, F.; Rabitz, H. A. *J. Phys. Chem. B* **2003**, *107*, 12874.
(66) Scalley, M. L.; Baker, D. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10636.
(67) Contact order values for staphylococcal nuclease were obtained from the WorldWide Web at http://depts.washington.edu/bakerpg/contact_order/.
(68) Waldburger, C. D.; Jonsson, T.; Sauer, R. T. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 2629.
(69) Wu, Y.; Matthews, C. R. *J. Mol. Biol.* **2002**, *322*, 7.