Predicting the Equilibrium Protein Folding Pathway: Structure-Based Analysis of Staphylococcal Nuclease

Vincent J. Hilser[1] **and Ernesto Freire**[1]*
[1]*Department of Biology and Biocalorimetry Center, Johns Hopkins University, Baltimore, Maryland*

ABSTRACT **The equilibrium folding pathway of staphylococcal nuclease (SNase) has been approximated using a statistical thermodynamic formalism that utilizes the high-resolution structure of the native state as a template to generate a large ensemble of partially folded states. Close to 400,000 different states ranging from the native to the completely unfolded states were included in the analysis. The probability of each state was estimated using an empirical structural parametrization of the folding energetics. It is shown that this formalism predicts accurately the stability of the protein, the cooperativity of the folding/unfolding transition observed by differential scanning calorimetry (DSC) or urea denaturation and the thermodynamic parameters for unfolding. More importantly, this formalism provides a quantitative account of the experimental hydrogen exchange protection factors measured under native conditions for SNase. These results suggest that the computer-generated distribution of states approximates well the ensemble of conformations existing in solution. Furthermore, this formalism represents the first model capable of quantitatively predicting within a unified framework the probability distribution of states seen under native conditions and its change upon unfolding. Proteins 27:171–183**

INTRODUCTION

The sequence of events that allow a polypeptide chain to assume a unique native conformation is known as the protein folding pathway, and elucidation of this pathway has been of prime importance to biologists since Anfinsen[1] demonstrated that the native state of a protein represents a minimum in the free-energy surface. The energy landscape that leads to the native state is believed to be rugged and characterized by two distinct features; peaks that represent high energy transition states and valleys that represent local energy minima and correspond to states with higher than average probabilities of becoming populated.[2,3] Equilibrium studies of protein intermediates provide information about the location of the minima in the free-energy surface, while kinetic studies on the rates of folding and unfolding provide information about the heights of the energy barriers in the folding pathway. Both approaches complement each other and provide useful information for mapping the energy surface and characterizing the folding pathway.

The most successful experimental approach for identifying protein regions that have a higher probability of being folded in intermediate states has been nuclear magnetic resonance (NMR), particularly the technique of hydrogen exchange performed under native conditions, which also provides a means of estimating apparent stability constants for individual residues.[4–13] Approaches that rely on the utilization of conditions that destabilize the native structure (i.e., extremes of pH, temperature, or denaturant) have also been useful, even though the intermediate conformations stabilized by these treatments are not necessarily equal to the ones most prevalently populated under native conditions (several reviews[14–19] have appeared on this subject). A quantitative model of the equilibrium folding pathway should be able to predict the intermediate conformations that are accessible under native, denaturing, and transition conditions. Additionally, an accurate model will reproduce both the thermodynamic parameters for unfolding and the cooperativity of the folding process. In short, it is evident that any quantitative formalism of the folding pathway must account for a broad range of experimental observables.

In recent years, we have been engaged in the development of a statistical thermodynamic formalism aimed at modeling the equilibrium distribution of conformational states of a protein.[19–23] This formalism utilizes the high-resolution crystallographic or NMR structure as a template to generate a large ensemble of partially folded states. The energetics and probability of each state in the ensemble are calculated using an empirical parametrization of the folding free energy and from the resulting probability distribution different properties are predicted.[20,21,24–26]

Contract grant sponsor: NIH, contract grant number RR04328; contract grant sponsor: National Science Foundation, contract grant number MCB-9118687.
*Correspondence to: Dr. Ernesto Freire, Department of Biology and Biocalorimetry Center, Johns Hopkins University, Baltimore, MD 21218. E-mail: bcc@biocal2.bio.jhu.edu.
Received 1 August 1996; accepted 8 August 1996.

Here we provide a quantitative account of the experimental hydrogen exchange protection factors measured under native conditions for staphyloccocal nuclease (SNase). It is demonstrated that the computer-generated ensemble of conformations also predicts accurately the stability of the protein, the cooperativity of the folding/unfolding transition observed by differential scanning calorimetry (DSC) or urea denaturation and the thermodynamic parameters for unfolding. The ability to quantitatively predict both the calorimetric and the NMR results validates the statistical thermodynamic formalism and suggests that the computer-generated ensemble reproduces the general features of the ensemble of SNase conformations in solution.

RESULTS

Modeling the Ensemble of Partially Folded States of Staphyloccocal Nuclease

The simulated ensemble of partially folded states was generated by using the crystallographic structure of SNase[27,28] as a template. In this algorithm (COREX) the entire protein is considered as being composed of different folding units. Partially folded states are generated by folding and unfolding these units in all possible combinations as described before.[21–23]

The division of the protein into a given number of folding units is called a partition. In order to maximize the number of distinct partially folded states, different partitions were included in the analysis. Each partition is defined by placing a block of windows over the entire sequence of the protein. The folding units are defined by the location of the windows irrespective of whether or not they coincide with specific secondary structure elements. By sliding the entire block of windows one residue at a time different partitions of the protein are obtained. For two consecutive partitions the first and last amino acids of each folding unit are shifted by one residue. This procedure is repeated until the entire sequence is exhausted. The algorithm is illustrated in Figure 1 for a window of size 10, which results in a total of 163,822 different states for SNase. In this paper, windows ranging in size from 9 to 15 residues were used. In all cases, the results were similar, which suggests that the conclusions presented here are not strongly dependent on window size within the range considered.

Each of the states generated by the COREX algorithm is characterized by having some regions folded and some other regions unfolded. There are two basic assumptions in this algorithm: (1) the folded regions in partially folded states are nativelike; and (2) the unfolded regions are assumed to be devoid of structure. The thermodynamic quantities (ΔH, ΔS, ΔC_p, and ΔG) for each state are calculated using the structural parametrization of the folding energetics developed earlier.[20,21,24–26,29] The probability of each

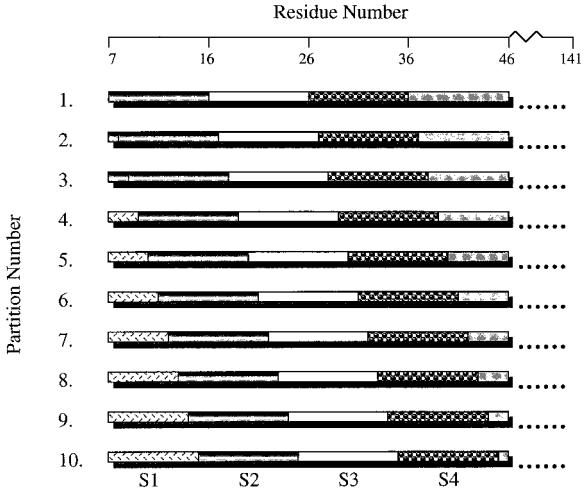

Fig. 1. Schematic representation of the COREX algorithm for generating partially folded states. The partitioning scheme for SNase is shown, using a window size of 10 residues, which results in 10 different partitions. In the first partition, the protein is divided into a group of 10-residue windows that are located along the sequence starting at the first residue (residue 7). This gives 13 windows of 10 residues and a final window of 5 residues: 14 total. The windows define the locations of folding units (labeled S1-S14). For cases in which the amino or carboxy terminal windows contain less than 3 residues, they are added to the adjoining window and counted as a single folding unit (see partitions 2 and 3). The partially folded conformations are generated with the computer by folding and unfolding the folding units in all possible combinations. This process uses the high-resolution structure as a template and results in $2^{14} - 2 = 16{,}382$ partially folded states. The second partition is then made by sliding the group of windows one residue in the sequence and repeating the procedure. After 10 partitions the window locations have made a complete cycle and the number of folding combinations involving 10 residues has been exhausted.

state is evaluated using the standard statistical thermodynamic equation:

$$P_i = \frac{exp\ (-\Delta G_i/RT)}{\Sigma\ exp\ (-\Delta G_i/RT)} \qquad (1)$$

The set of $\{P_i\}$ values defines the probability distribution function of states and is used in the calculation of all physical quantities. A way of testing how well the simulated distribution of states mimics or reproduces the most important features of the actual distribution of states in solution is by predicting a wide range of physical properties using the same set of elementary parameters. Two types of properties are considered in this paper: global properties like the overall stability of the protein or the cooperativity of the folding reaction, and properties of individual residues that also reflect the distribution of conformational states like hydrogen exchange protection factors.

Stability of SNase

The first test for the simulated ensemble is its ability to predict the stability of the native state and

the cooperativity of the folding/unfolding transition. This test is performed by predicting the excess heat capacity function, $<\Delta C_p>$, obtained experimentally by differential scanning calorimetry. The heat capacity function is a useful indicator of how well the calculations predict the temperature stability, cooperativity, and overall energetics of the transition. This is a purely thermodynamic prediction not influenced by nonthermodynamic factors. The excess heat capacity function, $<\Delta C_p>$, is defined as

$$<\Delta C_p> = \Sigma\,(\Delta H_i \cdot (\partial P_i/\partial T) + P_i \cdot \Delta C_{p,i}) \qquad (2)$$

where the native state is the reference state and the summation runs over all states in the ensemble.

Figure 2A shows the simulated temperature dependence of the excess heat capacity function for SNase. Analysis of the predicted and experimental overall thermodynamic parameters indicate a close correspondence. At the transition temperature of 66.8°C the predicted enthalpy change is 101 kcal/mol compared to the experimental value of 105 kcal/mol. The predicted and experimental heat capacity changes for unfolding are identical within error (ΔC_p = 2 kcal/K·mol).[22] A number of important features emerge from these data. First, the stability of the native state is predicted well. Second, the overall thermodynamic parameters (ΔH, ΔS, ΔC_p) are predicted essentially within experimental error. The entropy change required to reproduce the experimental T_m of SNase at neutral pH was 0.995 of the one predicted by the parametrization. Third, the simulated data are described accurately by a two-state model even though 40,938 states are included in the calculation. Likewise, the ratio of the van't Hoff to calorimetric enthalpies ($\Delta H_{vH}/\Delta H = 0.96$) is close to unity, indicating that the transition is essentially two-state from a practical standpoint.[30,31] This result is remarkable when one considers the extraordinary number of intermediate states that are accessible to the system, and underscores the importance of cooperative interactions that essentially reduce to zero the population of intermediate conformations during the unfolding transition. For SNase and other proteins, only under very low pH conditions, mild concentrations of denaturants, or extreme salt concentrations, do partially folded conformations become significantly populated.[15,17–19]

Prediction of the urea or GuHCl denaturation curves requires an additional parameter, the so-called *m* value, which accounts for the dependence of ΔG on denaturant concentration. In general, for any state, *i*, in the partition function the Gibbs energy in the presence of denaturant will be given by the equation

$$\Delta G_i = \Delta G_i^{\circ} - m_i \cdot C_d \qquad (3)$$

where ΔG_i° is the Gibbs energy in the absence of denaturant, C_d the concentration of denaturant, and m_i the *m* value for state *i*. The denaturation curve as a function of denaturant concentration can be defined in terms of the average degree of unfolding ($<F_u>$) at any denaturant concentration:

$$<F_u> = \Sigma F_{u,i} \cdot P_i \qquad (4)$$

where $F_{u,i}$ is the fractional degree of unfolding of state *i* (i.e., the number of unfolded residues in state *i* divided by the total number of residues) and P_i the probability of state *i* given by Equation (1).

Calculation of $<F_u>$ requires the assignment of an *m* value to each state in the partition function. The *m* values are not thermodynamic quantities, and currently their magnitude cannot be predicted from structure; however, Myers and colleagues[32] have recently observed that the *m* values and total ΔASA changes are linearly related for a large number of proteins. In our calculations, the *m* value of each state was assumed to be directly proportional to the change in accessible surface area for that state (i.e., $m_i = (\Delta ASA_i/\Delta ASA_u) \cdot m_u$), where m_u is the *m* value for global unfolding of the native state. We empirically obtained m_u by identifying the value that reproduces the experimental denaturant concentration at which the transition is half completed. In our calculations, the midpoint and the shape of the denaturation transition are not defined by m_u, as would be the case if a two-state model were assumed.

Figure 2B shows the predicted urea denaturation curve for SNase. The midpoint of the calculated curve was centered at 2.7 M urea, which is the experimental midpoint observed by Loh and colleagues.[9] To obtain the curve in Figure 2B, a microscopic m_u value of 2539 cal/mol.M was required. Analysis of the predicted curve using the linear extrapolation model yields an apparent *m* value of 2193 cal/mol.M and a ΔG° of 5.9 kcal/mol. The experimental ΔG° published by Loh and colleagues[9] is 6.1 kcal/mol. The close correspondence between the microscopic and apparent *m* values indicates that the transition does not deviate significantly from the two-state model. In fact, the solid line in Figure 2B is the best fitted curve using the two-state linear extrapolation model. It is clear again that the predicted denaturant unfolding transition can be modeled by a two-state mechanism even though 40,938 states were included in the calculation of the equilibrium ensemble. The COREX algorithm correctly predicts the temperature as well as denaturant observed cooperativity in the unfolding transition reaction.

A notable observation in Figure 2B is that even in the absence of denaturant $<F_u>$ is not zero. The

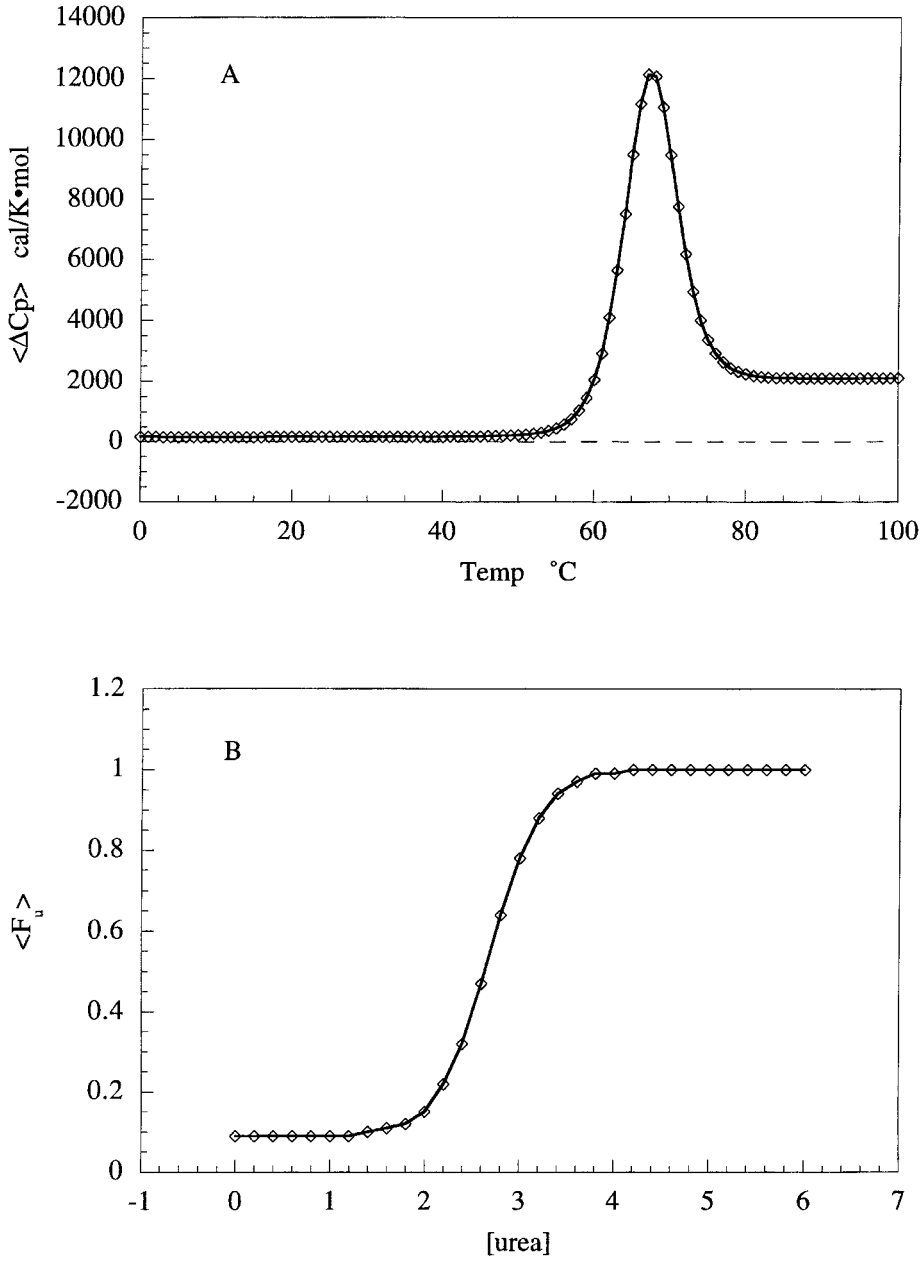

Fig. 2. **A:** The excess heat capacity function for SNase generated according to Equation (2) (open diamonds). The best-fit thermodynamic parameters are ΔH = 101 kcal/mol, T_m = 66.8°C and ΔC_p = 2.0 cal/K·mol. **B:** The urea denaturation profile generated according to Equation (4). The best-fit thermodynamic parameters using the linear extrapolation model are $\Delta G°$ = 6.1 kcal/mol, m = 2193 cal/mol.M and $C_{1/2}$ = 2.7 M. The partitioning scheme involved a window size of 12 residues, resulting in 40,938 states in both calculations. The solid lines through the points represents the curve generated from the best-fit (minimum in χ^2) parameters to a two-state model.

reason for this behavior is that the loop region between the third β strand (β3) and the first α helix (α1) is predicted to be disordered even under native conditions (see Figure 4 and discussion below).

Hydrogen Exchange Protection Factors

The COREX algorithm correctly predicts the cooperativity of the folding/unfolding transition, that is, the fact that the population of intermediate conformations is negligible. The accuracy of the predicted equilibrium ensemble of conformations can be tested by NMR detected hydrogen exchange. This technique has become one of the most valuable tools for the analysis and quantitation of intermediate conformations.[4–13,33–37] For a subset of the amide protons, exchange with solvent protons can only occur through

the following equilibrium:

$$\mathrm{NH(closed)} \overset{K_{op}}{\leftrightarrow} \mathrm{NH(open)} \overset{k_{int}}{\rightarrow} \mathrm{exchanged}$$

where open and closed refer to all states in which a specific proton is exchange-competent or -incompetent, respectively. In the limiting case where the open-to-closed reaction is fast compared to the intrinsic exchange rate (k_{int}) (i.e., in the so-called EX2 or bimolecular exchange regime), the measured exchange rate (k_{ex}) is represented by[10]

$$k_{ex} = K_{op} \cdot k_{int}. \tag{5}$$

Since k_{int} is known from exchange studies of peptides,[38] the equilibrium (K_{op}) for each residue can be evaluated directly from the measured exchange rate. Of particular interest among hydrogen exchange experiments has been the results obtained under conditions that strongly favor the folded state. Under this mechanism, protons that are buried from the solvent become exchange-competent as a result of being exposed to the solvent by partial or complete unfolding of the protein. Under such conditions, the protection factors ($PF = 1/K_{op}$) observed for amide protons reflect the energetics of those unfolding reactions.[6,8–10,12,13,39] These energetics can be compared to the one predicted by the COREX algorithm.

Residue Stability Constants

The prediction or analysis of residue level quantities like hydrogen exchange protection factors requires the introduction of a different statistical descriptor: the apparent residue stability constant.[23] This quantity is equal to the ratio of the summed probabilities of all states in which a particular residue is folded to the summed probabilities of all states in which that residue is unfolded:

$$\kappa_{f,j} = \frac{\Sigma P_{f,j}}{\Sigma P_{u,j}} \tag{6}$$

The $\kappa_{f,j}$ values can be used to define apparent folding free energies per residue as $\Delta G_j = -R \cdot T \cdot \ln \kappa_{f,j}$. The residue stability constants are the quantities that can be compared to a subset of the experimental protection factors. In analogy to the stability constant for a particular residue [Equation (6)], the protection factor can be defined as the ratio of the summed probabilities of all states in which a particular residue is closed (exchange incompetent) to the summed probabilities of all states in which that residue is open (exchange-competent). Since not all residues that are in the folded state are also exchange incompetent (closed), Equation (6) needs to be modified in order to describe the protection factors:

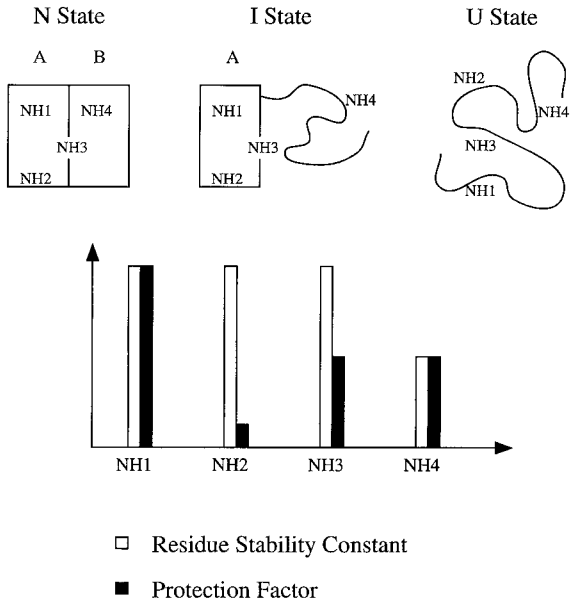

Fig. 3. Schematic representation of a protein with four exchangeable amide protons. Three (NH1, NH2, and NH3) are located in region A, which is unfolded only in the completely unfolded state (U state), and one (NH4) in region B, which is unfolded in the I and U states. Also shown are the relative stability constants and protection factors (see text for details).

$$PF_j = \frac{\Sigma P_{f,j} - \Sigma P_{f,xc,j}}{\Sigma P_{u,j} + \Sigma P_{f,xc,j}} \tag{7}$$

where $\Sigma P_{f,xc,j}$ is the sum of the probabilities of all states in which residue j is folded, yet exchange-competent. Only for those residues in which $\Sigma P_{f,xc,j} = 0$ or close to zero will the protection factors be equal to the stability constants. In other words, while the stability constant is a purely thermodynamic quantity, the protection factor has both thermodynamic and nonthermodynamic components. For example, a residue can be folded but exposed to the solvent or become exposed to the solvent as a result of the unfolding of a different part of the protein. This situation is illustrated in Figure 3 for two regions of a protein that contain four exchangeable amide protons (NH), each site representing a different situation that may be encountered in partially folded states. According to the figure, NH1 is buried in the most stable region (region A), while NH2 is located at the solvent-exposed surface of region A. NH3 represents an interesting case because it is part of region A, but it is buried at the interface between region A and the less stable region B. In other words, it is part of the complementary surface to region B. Lastly, NH4 is buried in region B.

Figure 3 also shows stability constants and protection factors that would be observed for this hypo-

thetical protein, assuming that the system had three accessible states: fully folded, fully unfolded, and the intermediate state in which region A is folded and region B is unfolded. It is clear that only for the cases in which the exchangeable groups are deeply buried in the specific region (NH1 and NH4), will the protection factors resemble the stability constants. For these situations, the nonthermodynamic contributions to the protection factors are minimal, and the equilibrium between the open (exchange-competent) and closed (exchange-incompetent) is identical to the folding/unfolding equilibrium.

For residues that are solvent-exposed or are complementary to regions that display lower stability, the protection factors and the stability constants can differ significantly. Residues that are exposed to the solvent in the native state are expected to show little or no protection even if they are located in regions of high stability. Such cases can, in principle, be identified from the high-resolution structure of the native state.[40] Residues located in complementary regions, on the other hand, are not readily apparent. In the example in Figure 3, NH3 would display a stability constant consistent with it belonging to the more stable region. However, as NH3 is complementary to a less stable region, it will become solvent-exposed, and hence exchange-competent, when the least stable region unfolds. As a consequence, exchange at NH3 can occur in both the intermediate and the completely unfolded state, resulting in a protection factor that significantly underestimates the stability of region A. A quantitative prediction of the protection factor for NH3 requires an accurate prediction of the conformational equilibrium.

Finally, the prediction of hydrogen exchange protection factors requires knowledge of the limiting exchange rates that can be measured under a given set of experimental conditions. This constraint sets a limit to the magnitude of the protection factors that can be determined for a given amino acid in the sequence. This is a purely experimental constraint, the magnitude of which depends on the actual experimental setup.[5] For the calculations in this paper, the expected exchange rates for each amide were estimated by using the intrinsic exchange rates calculated according to the method of Bai and colleagues.[38] For SNase, the cutoff criteria for limiting detection was set as the experimental dead time reported by Loh and colleagues.[9]

Pattern of Stability for SNase

Figure 4 shows the stability constants for residues 7–141, calculated according to Equation (6). For illustration purposes, the experimental protection factors obtained by Loh and colleagues[9] are also shown. Residues 1–6 and 141–149 are not resolved in the crystal structure and are presumed disordered.[27] Likewise, no hydrogen exchange protection

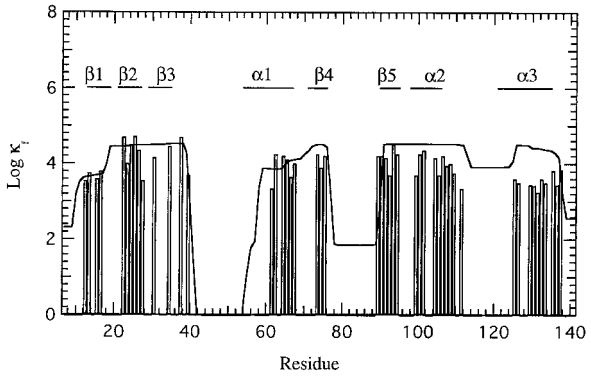

Fig. 4. The residue stability constants for SNAse calculated according to Equation (4). For comparison the experimental hydrogen exchange protection factors (Loh et al.[9]) for those residues that exhibit protection are shown (bars). Also indicated are the corresponding secondary structural elements.

was observed for these residues, supporting the notion that they are devoid of structure. As discussed before, the stability constants are defined for all residues whereas the protection factors are only defined for a subset of the residues in the protein. It is clear, however, that for those residues that exhibit significant protection, the magnitude of the stability constants and the protection factors are similar, indicating that the predicted probability distribution of states accurately predicts the apparent stability free energies for those residues. The identification of those residues that are expected to exhibit protection and those that are not are discussed in the next two sections.

Inspection of the residue stability constants reveals that the sequence of SNase can be divided into three stretches of residues that show high stability (residues 10–42, 58–76, and 90–138) and two stretches that show low stability (residues 42–57 and 77–89). The regions of the protein showing high stability, although being distal in sequence, are contiguous in the three-dimensional structure. This is shown in Figure 5 where the structure of SNase is displayed according to the rank order of the stability constants. Three levels of stability are shown that demonstrate that the residues with the highest stability constants represent those residues that comprise the bulk of the hydrophobic core of the protein. For SNase, this region corresponds mostly to the residues comprising the β barrel and most of the helical structures. As noted previously,[23] the highest resolution for stability constants and protection factors is expected under conditions at which the native state is maximally stable. Due to the relatively low overall stability of SNase at the conditions of the comparison (i.e., 37°C, $\Delta G \sim 6$ kcal/mol), little resolvable difference between the stability constants of the three most stable regions is observed as well as predicted. Additional calculations performed to simu-

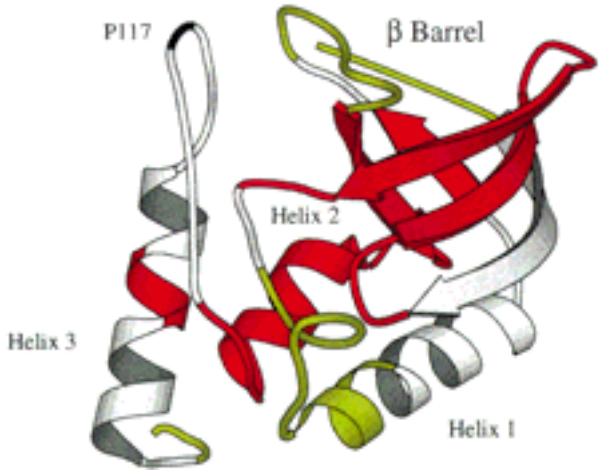

Fig. 5. Ribbon structure of SNase depicting the different levels of stability constants in the protein. For simplicity three levels are shown, which constitute high (red), medium (gray), and low (yellow) stability.

late conditions of maximal stability indicate higher stability constants for the β barrel, especially for the third, fourth, and fifth β strands, than for the helical domain. The notable exception involves residues in the second α helix, which are predicted to have a stability comparable to that of the β barrel.[23]

Shown also in Figure 4 are the corresponding elements of secondary structure. As noted by Loh and colleagues,[9] the regions of high and low protection correspond to the presence and absence of α/β secondary structure, respectively. In particular, the same regions of irregular secondary structure, which show little or no protection from exchange (residues 40–60, and 76–88), correspond to those regions that show decreased stability constants, as described above. The one notable exception, which will be discussed below, concerns the loop region between the second and third α helices (residues 112–122). Hence the trend in the predicted stability constants mimics the overall pattern of the protection factors as shown before.[23]

Comparison of Calculated and Experimental Protection Factors

As illustrated in Figure 3, not all residues in a protein exhibit protection even if they belong to highly stable regions. Figure 6 presents an expanded view of the experimental hydrogen exchange protection factors [9] and the protection factors calculated according to Equation (7). A less detailed analysis was presented before.[23] The calculation of protection factors requires evaluation of the sum of the probabilities of all states in which a particular residue is folded yet exchange-competent (i.e., $\Ɗ P_{f,xc,j}$). All folded residues whose amide nitrogens were more than 5% exposed in partially folded states were included in $\Ɗ P_{f,xc,j}$. Identified in Figure 6 are (1) All proline residues, which have no amide proton and hence no protection factor (labeled P); (2) residues that are solvent exposed (>5%) in the native structure (labeled S) and are expected not to show protection irrespective of whether they are folded or unfolded; (3) residues that are predicted to have protection factors beyond the experimental limit of detection; (4) residues for which calculated and experimental protection factors agree (labeled M); and (5) residues for which calculated and experimental values do not agree (labeled X).

Under the experimental conditions of Loh and colleagues,[9] only 49 residues exhibited protection factors that could be resolved over the dead time of the experiment (~20 minutes). Of these 49 residues, 44 are correctly predicted to show protection. For these 44 residues, the difference between the predicted residue free energies [$\Delta G_{calc,j} = -RT \cdot \ln PF(\kappa_{f,j})$] and those determined from the protection factors ($\Delta G_{exp,j} = -RT \cdot \ln PF_j$) is shown in Figure 7. The average $\Delta\Delta G_{calc\text{-}exp}$ of 300 + 630 cal/mol is indicative of the good agreement between the calculated and experimental values. This figure is equivalent to an average difference of 6.2% with a standard deviation of 12%. In addition, 56 residues (excluding prolines) are both predicted and experimentally determined to have no protection (Fig. 6). As the experimental $\Delta G_{exp,j}$ value is not known for these residues, a numerical comparison could not be performed. Of the 135 residues in SNase, 6 are prolines and therefore unable to exchange. Hence of a total of 129 residues with the potential to show protection, the protection factors of 100 or 78% are predicted reasonably well. This leaves 24 residues for which protection is predicted but not observed experimentally and five residues for which protection is observed but not predicted (labeled X in Figure 6).

There is a qualitative difference between a prediction that overestimates and one that underestimates the protection factors. In the first case either the stability prediction is incorrect or exchange occurs through additional mechanisms. In the second case the stability prediction is simply incorrect for that residue. The fact that only five of all misses are overestimates suggests that for at least a subset of mispredicting the observed exchange is not related to the intrinsic stability of the region.

Examination of the residues whose protection factors are overpredicted reveals that 3 are isolated and are surrounded by regions that are predicted reasonably well (residues 40, 95, and 134). There is one case of four consecutive misses (residues 68–71) and another corresponding to 3 misses (residues 18, 19, and 21); both groups represent turn regions (Fig. 6). The stability constants predict that the regions are stable, yet the amides are not protected even though they are buried from solvent in the crystal structure. It must be noted that residues 68–71 exhibit high-temperature factors in the crystallo-

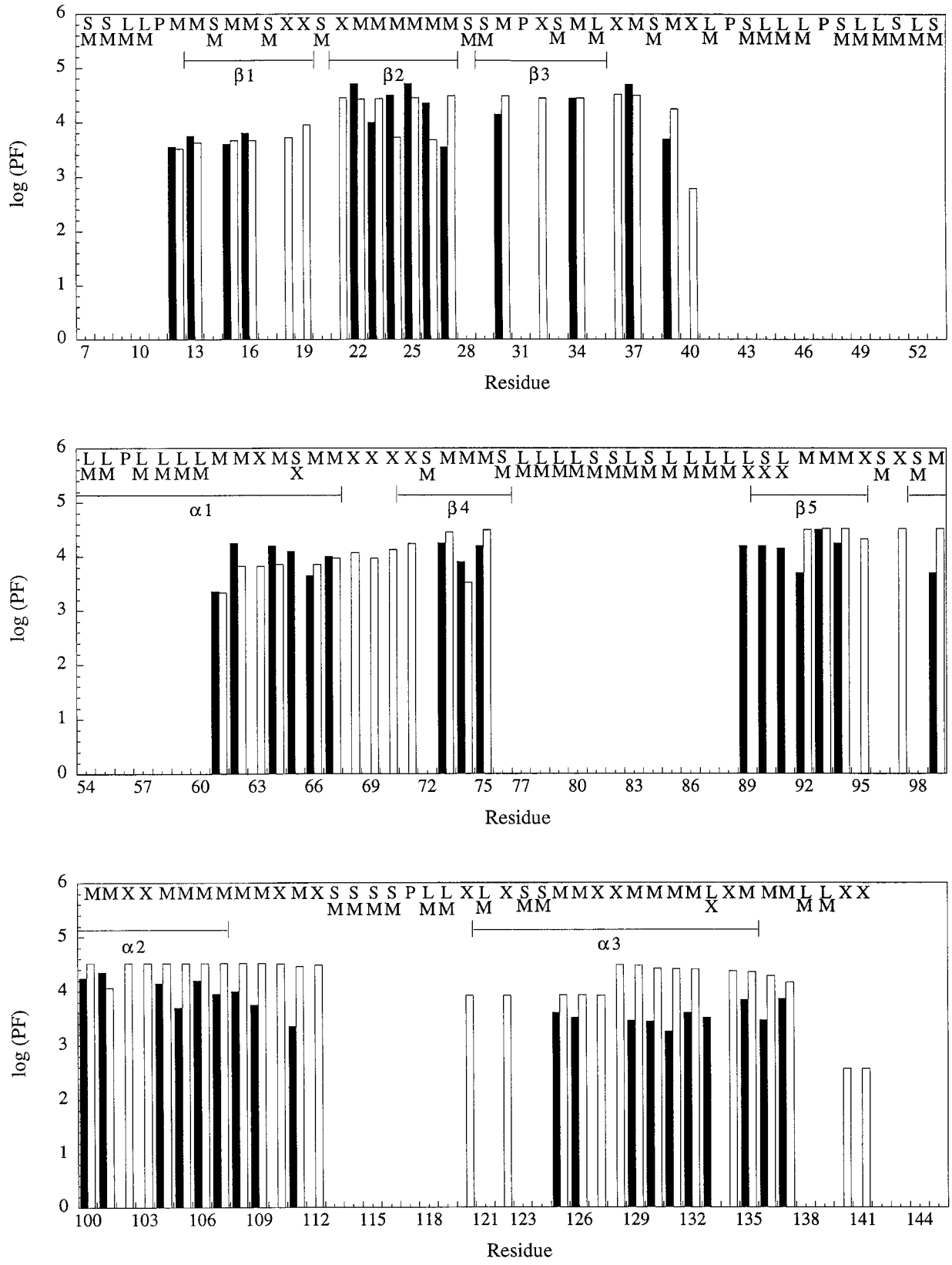

Fig. 6. Comparison of predicted (open bars) and experimental protection factors (Loh et al.[9]) (shaded bars). The predicted values were calculated by using Equation (5). Residues that exhibit measurable protection are labeled M if the predicted and experimental values match and X if they do not. Residues that do not exhibit measurable protection are labeled P if they are prolines, S if their amide groups are solvent-exposed in the native state, and L if the calculated protection is beyond the limit of detection. See text for details.

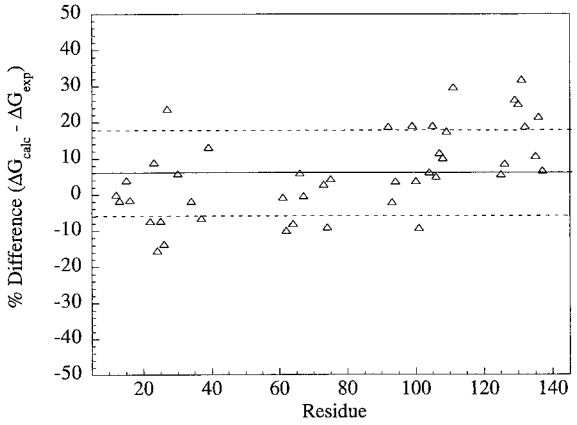

Fig. 7. Percent difference between predicted and experimental apparent free energies per residue ($\Delta\Delta G = \Delta G_{calc,j} - \Delta G_{exp,j}$) for those residues showing protection (open triangles). The predicted and experimental values are given by the equations $\Delta G_{calc,j} = -RT \cdot \ln PF(\kappa_{f,j})$ and $\Delta G_{exp,j} = -RT \cdot \ln PF_j$. The average $\Delta\Delta G$ is 300 ± 630 cal/mol.

graphic structure and it is likely that they are not resolved in the current analysis. The amide protons of residues 20 and 21 face the loop region following the third β strand (residues 40–50)—a region of low stability. Upon unfolding of the loop region, the amide protons of residues 20 and 21 face solvent, although only residue 20 shows a measurable change in solvent accessible surface. Residues 18 and 19, on the other hand, are involved in a bifurcate hydrogen bond with the backbone carboxyl of residue 22, while the amide of residue 22 hydrogen bonds with the backbone carboxyl of residue 19. It is possible that exchange at residue 21 is faciliatated by the same mechanism as residue 20 (i.e., it is complementary to a region of lower stability), while exchange at both 18 and 19 are the result of solvent penetration, which affects the entire bifurcate network. If ex change were indeed due to the intrinsic instability of the loop, it would be expected that residue 22 also would show no protection, and this is not the case.

The region of least agreement between predicted and experimental protection factors is the loop region between the second and third helices (residues 110–122) in which four residues are not predicted correctly. The difference in the predicted and observed behavior perhaps originates from the proline at position 117 (Fig. 5). The current calculation was performed on a crystal structure in which the K116—P117 peptide bond is in a *cis* configuration. This configuration is associated with a higher stability, and more compact structure, while the *trans* configuration is somewhat more expanded.[9,41] Since the equilibrium *cis-trans* ratio is ~10:1 in unligated nuclease[42], both forms may contribute substantially to the protection factor pattern. Loh and colleagues[9] concluded that the dynamic of *cis-trans* isomerization results in a more "open" form of the protein, thereby resulting in decreased interactions near the N terminus of helix 3.

Finally, a comparison of the original assignments for SNase (Figure 4 in Wang et al.[43]) and the spectra obtained by Loh and colleagues[9] suggests that the protection factors for some residues with overlapping (residues 32, 63, 103, 140, and 141) or unidentified (residues 36, 97, 102, and 127) peaks could not be evaluated. If these residues are not included in the statistics, then the total of correctly predicted residues amounts to 83%.

There are five residues that show protection even though protection is not predicted. Two of these residues (65 and 90) represent cases where the amide groups are greater than 5% exposed to solvent in the native state and therefore were considered to be exchange-competent. The remaining three residues (89, 91, and 133) represent cases in which the calculated protection factors were below the limit of experimental protection.

For residues 65 and 133 the numerical agreement between the calculated stability constant and the measured protection factor is good (0.3 kcal/mol and 1.2 kcal/mol), suggesting that the source of the error for these residues involves uncertainties in the cutoff criteria for solvent exposure. For residues 89–91, however, this is not likely to be the case. The stability constants for these residues are significantly less than the observed protection factors (1.8–3.5 kcal/mol). Figure 6 shows that these three residues represent the first of a series of stable residues that follow a relatively unstable loop region (residues 77–88). We have no explanation for this result.

The Urea Dependence of the Residue Stability Constants

The urea concentration dependence of the natural logarithm of the apparent residue stability constants is shown in Figure 8. In general, as the urea concentration increases and the stability of the protein diminishes, the magnitude of the stability constants decreases. At low urea concentration, the rate of decrease is not the same for all residues; several groups of residues with similar stability constants and similar *m* values can be recognized, as shown in Figure 8. At increasing urea concentrations the stability constants progressively merge into a single curve characterized by the parameters corresponding to the global unfolding of the protein. This is the same type of behavior observed experimentally for the denaturant dependence of the protection factors,[10] which have been used to define cooperative folding units or partially unfolded forms (PUF's).[10]

As shown in Figure 8, the β barrel, particularly strands 2, 3, and 5 as well as α helix 2 define the group of residues with the highest stability constants and *m* values. These residues define the most stable core of SNase. The unfolding of these residues only occurs by complete unfolding of the protein. α

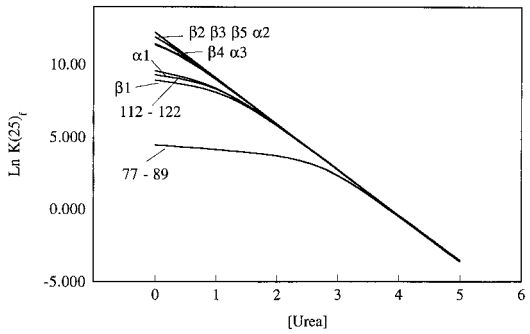

Fig. 8. The urea concentration dependence of the natural logarithm of the apparent residue stability constants. For clarity a single line is shown for groups of residues that exhibit similar behavior. The calculations in this figure corresponds to 25°C, a temperature at which the protein is more stable and predicted to show better resolution in the protection factors of different residues.

helix 1, the first β strand and the loop region defined by residues 112–122 come next, followed by the loop region defined by residues 77–89. The loop between β strand 3 and α helix 1 (residues 42–57) is unstable at all urea concentrations and is not shown in Figure 8. These conclusions are similar to those summarized in Figure 5.

A notable feature of the calculated curves, also noticed in the experimental denaturant dependence of the protection factors,[10] is the low *m* value observed at low denaturant concentrations for most residues. This effect underscores the statistical nature of the folding equilibrium, since, intrinsically, the Gibbs energy is a linear function of the denaturant concentration. For any given residue the observed or effective *m* values ($m_{\mathrm{eff},j}$) are statistical quantities equal to the difference between the average *m* value for the conformations in which the residue is not folded and the average *m* value for the conformations in which that residue is folded (see Hilser and Freire[23]): $m_{\mathrm{eff},\,j} = -(\partial \ln \kappa_{\mathrm{f},\,j}/\partial[\mathrm{D}] = \langle m \rangle_{\mathrm{nf},\,j} - \langle m \rangle_{\mathrm{f},\,j}$). At low denaturant concentrations the two averages are about equal and cancel each other to a large extent. This formalism explains why it is possible for some residues to exhibit high stability constants or protection factors and simultaneously *m* values close to zero.

Nature of the Conformational Ensemble of Staphyloccocal Nuclease Under Native Conditions

The agreement between predicted and experimental hydrogen exchange protection factors suggests that the probability distribution of partially folded states generated with the computer mimics the general features of the ensemble of conformations of SNase under native conditions and can be used to examine general aspects of the equilibrium folding pathway.

Under conditions that stabilize the native state, the total population of partially folded states is extremely small; nevertheless, the vast majority of partially folded states that become populated are nativelike. The uniqueness of the NMR hydrogen exchange technique is that it allows examination of this very small pool of intermediate conformations without interference from the native state. The quantitative agreement between predicted and experimental protection factors strongly suggests that the population of nonnative partially folded states must be extremely small. It must be recalled that all the calculations have been made under the assumption that the folded regions in partially folded states are nativelike. Therefore, if nonnative states were making a significant contribution, quantitative agreement would not have been observed.

A comprehensive view of the equilibrium folding pathway can be obtained if the entire ensemble of partially folded states is subdivided according to the degree of folding. This strategy allows identification of the most prominent features of the folding pathway. At the early stages of folding, some regions of the protein might exhibit a higher relative stability than at later stages and occupy a higher order in the stability rank. This is illustrated in Figure 9 where the apparent stability constants per residue for the subensemble of partially folded states having a degree of folding lower than 10% are shown. At the early stages of folding, a very large number of conformations are observed, none of which acquires a high probability. The overall pattern of stability constants for this subensemble differs significantly from that corresponding to the entire ensemble. In particular, the regions that exhibit the highest stability correspond to the three α helices in SNase and more specifically to the central portions of those helices. For low degrees of folding, most of the amino acid residues are still exposed to the solvent, and the peaks observed in Figure 9 essentially report the higher helical propensities of those regions of the molecule. Remarkably, at these low degrees of folding the turn regions, especially the turn between the fourth and fifth β strands, occupy a higher order in the stability rank than at later stages. A relatively high early stability for critical turn regions might be important in the definition of the overall fold of the protein. At low degrees of folding, the residue stabilities are determined more by intrinsic or local interactions than by long-range cooperative interactions. For this reason, the β barrel structure of SNase, which is stabilized by significant tertiary interactions involving a large number of residues, shows a negligible probability at these early stages. This structure becomes the most probable element when the degree of folding reaches about 40%, that is, the necessary number of residues required to define it, and its formation is associated with the hydrophobic collapse of the protein. Once this level is reached,

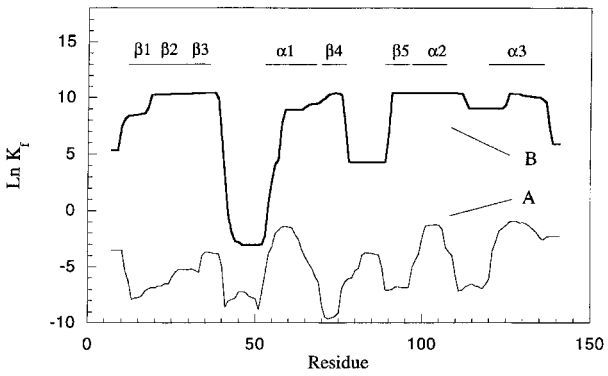

Fig. 9. Natural logarithm of the residue stability constants for the subensemble of SNase conformations in which the degree of folding is less than 10% (curve A). For comparison, the curve for the entire ensemble is also shown (curve B). For the calculation of the stability constants for this subensemble, a window size of 6 residues was used in the generation of partially folded states.

helices and other structures become stabilized by tertiary interactions. All these events occur in a highly fluctuating environment characterized by the existence of a large number of states.

Conclusions

The approach presented here emphasizes the statistical nature of the folding pathway rather than the traditional deterministic picture in which a single protein molecule progressively gains structure until reaching the native state. Examination of the ensemble of partially folded states suggests the existence of a large number of folding paths, many of which are highly divergent from the common view in which the folding process is conceptualized as a gradual, more or less linear, accretion of structure. Under equilibrium conditions, average properties reflect the most probable distribution in the ensemble of conformations. In the time domain, on the other hand, the observed folding kinetics reflects the "most traveled" folding path, which is also an average property contributed by the individual path of all the molecules in the ensemble.

The picture of the equilibrium folding pathway that emerges from these simulations is consistent with the hydrogen exchange and the stability data and can be summarized as follows: All the states with a degree of folding <10% can be characterized as premolten globule or precollapsed states, in the sense that the hydrophobic collapse of the molecule has not occurred yet. The hydrophobic collapse of SNase is associated with the formation of the β barrel, which requires the participation of a minimum of ~40% of the residues. The collapsed state is not formed in a progressive, cumulative fashion but rather in an abrupt, cooperative fashion. Importantly, most of the residues that define the collapsed state are not those that exhibit the highest structural propensities when exposed to solvent. The hydrophobic collapse reshuffles the stability rank of individual structural elements. In essence, the hydrophobic collapse can be thought of as the transfer from a polar solvent to a hydrophobic solvent of those residues that become buried. Thus, for example, helical structures that have a relatively high stability compared to other structural elements in precollapsed states may be overcome by those other structures in the stability rank in postcollapse states. An important consequence of this observation is the breakdown in the linearity or gradual progression of the folding process: For a given degree of folding, the most probable structures are not necessarily formed from the most probable structures having lower degrees of folding. While the calculations reported here refer to the equilibrium ensemble of conformations, it is noteworthy that the pattern of hydrogen protection in kinetic experiments[33] show striking similarities to the equilibrium pattern under native conditions. In particular, a significant portion of the β barrel (notably β2, β3, β4) and also a portion of helix 2 show early protection during refolding. In this respect, Woodward[44] has advanced the idea that "the folding pathway would correspond approximately to the reverse order of native-state exchange rates."

METHODS

Calculation of Gibbs Energies

The free energy ($\Delta G \equiv \Delta H - T \cdot \Delta S$) of each conformational state was calculated using the empirical parametrization of the free energy developed before.[20,21,24–26] This parametrization has been derived from the analysis of protein data. Briefly, this procedure involves calculation of the relative heat capacity (ΔC_p), enthalpy (ΔH), and entropy (ΔS) of each state at the desired temperature.

The heat capacity change is a weak function of temperature and has been parametrized in terms of changes in solvent accessible surface areas (ΔASA), since it originates mainly from changes in hydration[45,46]:

$$\Delta C_p = \Delta C_{p,ap} + \Delta C_{p,pol} \quad (6a)$$

$$\Delta C_p = a_C(T) \cdot \Delta ASA_{ap} + b_C(T) \cdot \Delta ASA_{pol} \quad (6b)$$

where the coefficients $a_C(T) = 0.45 + 2.63 / 10^{-4} \cdot (T - 25) - 4.2 / 10^{-5} \cdot (T - 25)^2$ and $b_C(T) = -0.26 + 2.85 / 10^{-4} \cdot (T - 25) + 4.31 / 10^{-5} \cdot (T - 25)^2$. In the equation above, ΔASA changes are in «^{2} and the heat capacity in cal/K·mol. In general, for low temperature calculations ($T < 80$°C) the temperature-independent coefficients are sufficient.[46]

The bulk of the enthalpy change also scales in terms of ΔASA changes and at the reference tempera-

ture of 60°C it can be written as

$$\Delta H_{gen}(60) = a_H(60) \cdot \Delta ASA_{ap} + b_H(60) \cdot \Delta ASA_{pol} \quad (7)$$

where $a_H(60) = -8.44$ and $b_H(60) = 31.4$.[20,21]

In the calculation of the entropy change two primary contributions are included, one due to changes in solvation and the other due to changes in conformational degrees of freedom ($\Delta S = \Delta S_{solv} + \Delta S_{conf}$). The entropy of solvation can be written in terms of the heat capacity if the temperatures at which the apolar and polar hydration entropies are zero ($T^*_{S,ap}$ and $T^*_{S,pol}$) are used as reference temperatures:

$$\Delta S_{solv} = \Delta S_{solv,ap} + \Delta S_{solv,pol} \quad (8a)$$

$$\Delta S_{solv} = \Delta C_{p,ap} \cdot \ln(T/T^*_{S,ap}) + \Delta C_{p,pol} \ln(T/T^*_{S,pol}) \quad (8b)$$

$T^*_{S,ap} = 385.15$ K has been known for some time,[29,45] and recently, it has been found that $T^*_{S,pol} \approx 335.15$ K.[25]

Conformational entropies are evaluated by explicitly considering the following three contributions for each amino acid:

1. $\Delta S_{bu \rightarrow ex}$, the entropy change associated with the transfer of a side chain that is buried in the interior of the protein to its surface
2. $\Delta S_{ex \rightarrow u}$, the entropy change gained by a surface-exposed side chain when the peptide backbone unfolds
3. ΔS_{bb}, the entropy change gained by the backbone itself upon unfolding.

The magnitude of these terms for each amino acid has been estimated by computational analysis of the probability of different conformers as a function of the dihedral and torsional angles.[25,47] For each conformational state, ΔASA_{ap} and ΔASA_{pol} are calculated using the Lee and Richards algorithm using software developed in this laboratory. These ΔASA values are then used to calculate ΔH, ΔC_p, and ΔS_{solv} values. In addition, for each residue in each conformational state, the state of the side chain (buried, exposed in a folded region, exposed in an unfolded region) and the backbone (folded or unfolded) are determined in order to evaluate conformational entropies.

ACKNOWLEDGMENTS

We thank Benjamin Townsend for developing the new accessible surface calculations library.

This work was supported by grants from the National Institutes of Health (RR04328) and the National Science Foundation (MCB-9118687).

REFERENCES

1. Anfinsen, C.B. Principles that govern the folding of protein chains. Science 181:223–230, 1973.
2. Bryngelson, J.D., Onuchic, J.N., Socci, N.D., Wolynes, P.G. Funnels, pathways and the energy landscape of protein folding: A synthesis. Proteins 21:167–195, 1995.
3. Wolynes, P.G., Onuchic, J.N., Thirumalia, D. Navigating the folding routes. Science 267:1619–1620, 1995.
4. Wagner, G., Wuthrich, K. Correlation between the amide proton exchange rates and the denaturation temperatures in globular proteins related to the bovine pancreatic trypsin inhibitor. J. Mol. Biol. 130:31–37, 1979.
5. Radford, S.E., Buck, M., Topping, K.D., Dobson, C.M., Evans, P.A. Hydrogen Exchange in Native and Denatured States of Hen Egg-White Lysozyme. Proteins 14:237–248, 1992.
6. Mayo, S.L., Baldwin, R.L. Guanidine chloride induction of partial unfolding in amide proton exchange in RNase A. Science 262:873–876, 1993.
7. Kim, K.-S., Fuchs, J.A., Woodward, C. Hydrogen exchange identifies native-state motional domains important in protein folding. Biochemistry 32:9600–9608, 1993.
8. Kim, K.-S., Woodward, C.K. Protein internal flexibility and global stability: Effect of urea on hydrogen exchange rates of bovine pancreatic trypsin inhibitor. Biochemistry 32: 9609–9613, 1993.
9. Loh, S.N., Prehoda, K.E., Wang, J., Markley, J.L. Hydrogen exchange in unligated and ligated staphylococcal nuclease. Biochemistry 32:11022–11028, 1993.
10. Bai, Y., Sosnick, T.R., Mayne, L., Englander, S.W. Protein folding intermediates: Native-state hydrogen exchange. Science 269:192–197, 1995.
11. Morozova, L.A., Haynie, D.T., Arico-Muendel, C., Van Dael, H., Dobson, C.M. Structural basis of the stability of a lysozyme molten globule. Nature Struct. Biol. 2:871–875, 1995.
12. Orban, J., Alexander, P., Bryan, P., Khare, D. Assessment of stability differences in the protein G B1 and B2 domains from hydrogen-deuterium exchange: Comparison with calorimetric data. Biochemistry 34:15291–15300, 1995.
13. Swint-Kruse, L., Robertson, A.D. Temperature and pH dependence of hydrogen exchange and global stability for ovomucoid third domain. Biochemistry 35:171–180, 1996.
14. Ptitsyn, O.B. Protein folding: Hypotheses and experiments. J. Protein Chem. 6:272–293, 1987.
15. Kuwajima, K. The molten globule state as a clue for understanding the folding and cooperativity of globular-protein structure. Proteins 6:87–103, 1989.
16. Kim, P.S., Baldwin, R.L. Intermediates in the folding reactions of small proteins. Annu. Rev. Biochem. 59:631–660, 1990.
17. Haynie, D.T., Freire, E. Structural energetics of the molten globule state. Proteins 16:115–140, 1993.
18. Fink, A.L. Compact intermediate states in protein folding. Annu. Rev. Biophys. Biomol. Struct. 24:495–522, 1995.
19. Freire, E. Thermodynamics of partly folded intermediates in protein folding. Ann. Rev. Biophys. Biomolec. Struct. 24:141–165, 1995.
20. Xie, D., Freire, E. Molecular basis of cooperativity in protein folding. V. Thermodynamic and structural conditions for the stabilization of compact denatured states. Proteins 19:291–301, 1994.
21. Xie, D., Freire, E. Structure based prediction of protein folding intermediates. J. Mol. Biol. 242:62–80, 1994.
22. Xie, D., Fox, R., Freire, E. Thermodynamic characterization of a staphylococcal nuclease folding intermediate. Protein Sci. 3:2175–2184, 1994.
23. Hilser, V.J., Freire, E. Structure based calculation of the equilibrium folding pathway of proteins. Correlation with

hydrogen exchange protection factors. J. Mol. Biol. 262:756–772, 1996.
24. Gómez, J., Freire, E. Thermodynamic mapping of the inhibitor binding site of the aspartic protease endothiapepsin. J. Mol. Biol. 252:337–350, 1995.
25. D'Aquino, J.A., Gómez, J., Hilser, V.J., Lee, K.H., Amzel, L.M., Freire, E. The magnitude of the backbone conformational change in protein folding. Proteins 25:143–156, 1996.
26. Hilser, V.J., Gómez, J., Freire, E. The enthalpy change in protein folding and binding: Refinement of parameters for structure based calculations. Proteins In press: 1996.
27. Loll, P.J., Lattman, E.E. The crystal structure of the ternary complex of staphylococcal nuclease, Ca2+, and the inhibitor pdTp, refined at 1.65 «. Proteins 5:183–201, 1989.
28. Hynes, T.R., Fox, R.O. The crystal structure of staphylococcal nuclease refined at 1.7 angstroms resolution. Proteins 10:92–99, 1991.
29. Baldwin, R.L. Temperature dependence of the hydrophobic interaction in protein folding. Proc. Natl. Acad. Sci. USA 83:8069–8072, 1986.
30. Lumry, R., Biltonen, R., Brandts, J.F. Validity of the "two-state" hypothesis for conformational transitions of proteins. Biopolymers 4:917–944, 1966.
31. Freire, E., Biltonen, R.L. Statistical mechanical deconvolution of thermal transitions in macromolecules. I. Theory and application to homogeneous systems. Biopolymers 17:463–479, 1978.
32. Myers, J.K., Pace, C.N., Scholtz, J.M. Denaturation *m* values and heat capacity changes: Relation to changes in accessible surface areas of protein unfolding. Protein Sci. 4:2138–2148, 1995.
33. Jacobs, M.D., Fox, R.O. Staphylococcal nuclease folding intermediate characterized by hydrogen exchange and NMR spectroscopy. Proc. Natl. Acad. Sci. USA 91:449–453, 1994.
34. Jennings, P.A., Wright, P.E. Formation of a molten globule intermediate early in the kinetic folding pathway of apomyoglobin. Science 262:892–896, 1993.
35. Roder, H., Elove, G.A., Englander, S.W. Structural characterization of folding intermediate in cytochrome C by H-exchange labelling and proto-NMR. Nature 335:700–704, 1988.
36. Schulman, B.A., Redfield, C., Peng, Z., Dobson, C.M., Kim, P.S. Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α lactalbumin. J. Mol. Biol. 253:651–657, 1995.
37. Udgaonkar, J.B., Baldwin, R.L. NMR evidence for an early framework intermediate on the folding pathway of ribonuclease A. Nature 335:694–699, 1988.
38. Bai, Y., Milne, J.S., Mayne, L., Englander, S.W. Primary structure effects on peptide group hydrogen exchange. Proteins 17:75–86, 1993.
39. Qian, H., Mayo, S.L., Morton, A. Protein hydrogen exchange in denaturant: Quantitative analysis by two-process model. Biochemistry 33:8167–8171, 1994.
40. Pedersen, T.G., Sigurskjold, B.W., Andersen, K.V., Kjaer, M., Poulsen, F.M., Dobson, C.M., Redfield, C. A nuclear magnetic resonance study of the hydrogen-exchange behavior of lysozyme in crystals and solution. J. Mol. Biol. 218:413–426, 1991.
41. Royer, C.A., Hinck, A.P., Loh, S.N., Prehoda, K.E., Peng, X., Jonas, J., Markley, J.L. Effects of amino acid substitutions on the pressure denaturation of staphylococcal nuclease as monitored by fluorescence and nuclear magnetic resonance spectroscopy. Biochemistry 32:5222–5232, 1993.
42. Evans, P.A., Dobson, C.M., Kautz, R.A., Hatfull, G., Fox, R.O. Proline isomerization in staphylococcal nuclease characterized by NMR and site-directed mutagenesis. Nature 329:266–268, 1987.
43. Wang, J., Hinck, A.P., Loh, S.N., LeMaster, D.M., Markley, J.L. Solution studies of staphylococcal nuclease H124L. 2. 1H, 13C, 15N chemical shifts assignments for the unligated enzyme and analysis of chemical shift changes that accompany formation of the nuclease-thymidine 3′,5′-bisphosphate-calcium ternary complex. Biochemistry 31: 921–936, 1992.
44. Woodward, C. Is the slow-exchange core the protein folding core? TIBS 18:359–360, 1993.
45. Murphy, K.P., Freire, E. Thermodynamics of structural stability and cooperative folding behavior in proteins. Adv. Protein Chem. 43:313–361, 1992.
46. Gómez, J., Hilser, V.J., Xie, D., Freire, E. The heat capacity of proteins. Proteins 22:404–412, 1995.
47. Lee, K.H., Xie, D., Freire, E., Amzel, L.M. Estimation of changes in side chain configurational entropy in binding and folding: General methods and application to helix formation. Proteins 20:68–84, 1994.