Protein Science (1996), 5:991-1000. Cambridge University Press. Printed in the USA.

REVIEW

Protein folding for realists: A timeless phenomenon

DAVID SHORTLE, YI WANG, JOEL R. GILLESPIE, AND JAMES O. WRABL
Department of Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205

(RECEIVED February 14, 1996; ACCEPTED March 29, 1996)

Abstract

Future research on protein folding must confront two serious dilemmas. (1) It may never be possible to observe at high resolution the very important structures that form in the first few milliseconds of the refolding reaction. (2) The energy functions used to predict structure from sequence will always be approximations of the true energy function. One strategy to resolve both dilemmas is to view protein folding from a different perspective, one that no longer emphasizes time and unique trajectories through conformation space. Instead, free energy replaces time as the reaction coordinate, and ensembles of equilibrium states of partially folded proteins are analyzed in place of trajectories of one protein chain through conformation space, either in vitro or in silico. Initial characterization of the folding of staphylococcal nuclease within this alternative conceptual framework has led to an equilibrium folding pathway with several surprising features. In addition to the finding of two bundles of four hydrophobic segments containing both native and non-native interactions, a gradient in relative stability of different substructures has been identified, with the most stable interactions located toward the amino terminus and the least stable toward the carboxy terminus. Hydrophobic bundles with up-down topology and stability gradients may be two examples of numerous tactics used by proteins to facilitate rapid folding and minimize aggregation. As NMR methods for structural analysis of partially folded proteins are refined, higher resolution descriptions of the structure and dynamics of the polypeptide chain outside the native state may provide many insights into the processes and energetics underlying the self-assembly of folded structure.

Keywords: denatured state; folding intermediates; hydrophobic bundles; protein folding; stability gradients; structure prediction

Part of the affinity biochemists feel toward proteins comes from their fascinating three-dimensional structures, with their odd mix of symmetry and asymmetry. Perhaps an even greater attraction comes from the excitement of working with a class of molecules that plays such a central role in all life processes. At the heart of our thinking about each protein is the conviction that specific biological functions emerge directly from the details of its unique and highly individualistic three-dimensional structure. By a process known as folding, the long and rather uninteresting organic polymer that emerges from a ribosome spontaneously assumes its biologically active native structure. In many respects, protein folding represents, at the molecular level, the step where life begins, where chemistry makes the jump to biology.

We suspect that many whose research touches on the subject of protein folding occasionally have day dreams in which they imagine, by some omnipotent power, that they can actually watch as their favorite polypeptide chain folds, moving through a series of increasingly structured forms to arrive at the biologically active conformation. In an attempt to realize this dream, some protein scientists pursue a deeper understanding of folding by experiments that track the changes in structure spectroscopically as proteins refold as a function of time. Others, with a more theoretical bent, are using computer representations of the protein chain in attempts to model what takes place during folding.

From the large number of research papers published over the past 10 years, one can infer that there is no shortage of experiments that can be done to study protein folding. Yet not all experiments that can be done hold out reasonable prospects of increasing our understanding of this complex phenomenon. From the perspective of the research field today, based on what we currently know about folding and about the experimental and theoretical methods we have at our disposal, is it realistic to anticipate that one day we will be able to follow proteins as they fold, either real chains in solution or virtual chains in silico? This review begins with the premise that the answer to both questions is "no." The justification for such a pessimistic position is outlined, and a proposal is made to modify the definition of

Reprint requests to: David Shortle, Department of Biological Chemistry, The Johns Hopkins University School of Medicine, Baltimore, Maryland 21205.

the phenomenon of protein folding in a way that allows more realistic goals to be set for both experimental and theoretical approaches to the problem.

Trajectories through time

On consideration of Levinthal's paradox with its estimate of the astronomically large number of wrong conformations (Levinthal, 1968), it is easy to be convinced that proteins face an impossible task of finding the correct native structure in a time shorter than the history of the universe. Yet the indisputable reality is that proteins fold in milliseconds to seconds. Obviously, they must employ one or more strategies for consistently beating what appear to us to be enormous odds stacked against them. How do they do it?

The most straightforward experimental approach to resolve this paradox is to study proteins as they refold from the denatured state by monitoring how structure increases as a function of time and attempting to identify the strategies that permit fast folding. Obtaining a kinetic description of the sequence of structure-forming events has been the principal objective of many experimental studies of protein folding for the past 30 years (Matthews, 1993). Although no detailed model of the folding pathway of any protein has yet emerged, work over this period has lead to improvements in spectroscopic methods, and data have been collected on a surprisingly large number of proteins.

For small, single-domain proteins, a common pattern of events is observed. Within a few milliseconds or less, the polypeptide chain appears to form a compact core of hydrophobic residues. A multiplicity of kinetic phases can usually be identified in the time range from 10 ms to several hundred milliseconds, and, within a second or less, the majority of molecules have reached the native state. In the time interval of milliseconds to seconds that defines the time scale of folding, fluorescence spectroscopy can track changes in the solvent accessibility of tryptophan or tyrosine residues; CD and vibrational spectroscopy can monitor the average content of secondary structure; and pulsed hydrogen exchange labeling can report on the kinetics of exposure/protection for one-fourth to one-half of backbone amide protons. However, the mainstays of structural analysis of folded proteins—X-ray crystallography and high-resolution NMR spectroscopy—have fundamental limitations that make them practically useless for tracking changes in solution structure on this time scale.

Experimental studies of protein folding must confront a major dilemma: following the buildup of structure in atomic detail over time intervals shorter than hours is simply not feasible at present. Furthermore, there is no realistic prospect on the horizon for a revolutionary new method that will fundamentally change this situation. How then are we to proceed to study folding and extend our understanding of the physical chemistry underlying the folding process? If time is the fundamental variable upon which structure formation depends, then it appears we may never succeed in our attempts to watch how protein chains self-assemble into the native state. The relevant time scales are simply too short.

In the face of such a dilemma, one especially pertinent question should be asked by all experimentalists: "Must time be considered the one truly fundamental variable that governs the appearance of protein structure?" "Should kinetic studies of folding serve as the Gold Standard that defines what is physiological, 'real' folding, and what is some sort of artifact?"

It must be remembered that, in the study of simple chemical reactions, kinetic investigations are treated as one of several strategies for probing reaction mechanisms. Most of the important alternatives are based on characterization of reactions at equilibrium, which requires a shift in conceptual framework that removes time as the principal variable and replaces it with free energy. Although thermodynamic studies provide significantly less information than can in theory be acquired by kinetics (equilibrium constants can be derived from kinetic constants but not vice versa), much of the lost information involves transition states.

Thus, the size of the penalty to be paid for shifting from a kinetic to an equilibrium framework depends on the importance of transition states to a deeper understanding of the reaction. For the sake of argument, the position is taken in this review that transition states are less central to the process of protein folding than the stable and meta-stable intermediates formed during folding; the more fundamental question is, "Where is the polypeptide chain going?" not, "How does it get there?" In other words, enumeration and structural characterization of these partially folded conformations is more likely to provide profound mechanistic insights than characterization of the intervening transition states. And the most convincing argument for redefining protein folding as a timeless phenomenon is that it leads to workable strategies for studying some of these intermediate states in great detail.

The removal of time as the fundamental variable for protein folding would require a major change in our perspective of this phenomenon. Instead of a single molecule moving through conformation space, the description of a folding protein would involve equilibrium ensembles of conformations of molecules. Instead of a unique trajectory through time, the description of a folding protein would consist of a series of ensembles advancing toward the native state as the free energy of the polypeptide chain is lowered in a series of equilibrium steps. In other words, the rigor and clarity that unique three-dimensional structures bring to the issues of protein chemistry would have to be abandoned, to be replaced by the vague and fuzzy features that accompany statistical descriptions of ensembles of conformations.

It is interesting to note that theoretical approaches to predicting structure from sequence must confront, at some point in the not-so-distant future, a dilemma that is somewhat analogous to that facing experimental approaches. As explained more fully below, when folding is modeled as the unique trajectory of a single molecule, errors in the calculated energy of the final conformation lead to uncertainties in the "best" structure that the prediction can provide. A solution to this dilemma proposed by Finkelstein et al. (1995) is in broad respects analogous to the solution presented above for the dilemma faced by experimentalists: biochemists must shift their thinking about protein folding away from unique trajectories through time and toward a view consisting of ensembles of structures with different free energies.

Equilibrium predictions of structure

Strategies for predicting the structures of proteins employ energy functions whose value depends on the quality of fit between the amino acid sequence and a candidate structure. Although for empirical strategies, such as the secondary structure meth-

ods of Chou and Fasman (1974), this function may not be defined in terms on energy, it can be viewed as analogous to the explicit energy function used by strategies with a physical chemistry basis. Predicting a structure from sequence proceeds by an iterative process of generating a series of candidate structures and evaluating the energy of each in turn. The predicted structure is the single conformation declared the best candidate on the basis of its having the lowest calculated energy.

Last year several ab initio methods met with significant success in predicting secondary structure and elements of supersecondary structure for peptides and/or small proteins (Avbelj & Moult, 1995; Srinivasan & Rose, 1995; Sun et al., 1995). The details of tertiary structure were usually, but not always, rather poorly predicted. These methods, which are based on current ideas about the physical chemistry involved in folding, treat hydrophobic interactions in an approximate way and omit one or more of the other three basic energy terms—van der Waals interactions, electrostatics, and hydrogen bonds. Considering these obvious simplifications in the energy function, the initial successes of these methods have generated considerable optimism that perhaps predictive strategies are at last on the right track.

Common sense suggests that a more realistic representation of the energy function, though perhaps more computationally costly, will always translate into more accurate predictions. But just how good does the energy function need to be to get the right answer? What is the relationship between errors in the calculated energy and errors in the structure of the predicted conformation?

In a recent article, Finkelstein et al. (1995) present a plausible argument for the following assertion: the number of conformations of a polypeptide chain with a given energy increases exponentially as the energy increases. The significance of this statement for protein structure prediction comes from the fact that the calculated energy will always be an approximation of the "true" energy, because the energy function used in computer programs will never exactly correspond to the function used by nature. Thus, one expects that the native conformation (which we assume to be the structure of lowest free energy) will not be the structure with the lowest calculated energy, an expectation amply supported by many instances of predictions over the years. The calculated energy of the true or correct structure will be higher than the calculated energy of the structure with the lowest energy E_{lowest} by an amount ΔE, the error in the calculated energy. The important point is that, as the value of ΔE increases, the number of different wrong conformations with energies the same as or lower than that calculated for the correct native conformation will increase exponentially.

This is not good news when one's objective is to predict the one correct structure. However, as Finkelstein et al. (1995) point out, there is a way to extract the maximal amount of reliable information provided by a predictive method when the search strategy converges to give a large number of promising candidate conformations that vary in energy from E_{lowest} to $E_{lowest} + \Delta E$, with ΔE representing the error in the value of E_{lowest}. By viewing the output of structure prediction not as a single conformation, but rather as an ensemble of conformations, the structural features common to many or all conformations can be identified. These common features will be found in segments of the polypeptide chain where the energy function can make a specific prediction. For segments of the chain that form multiple structures with no common features, the energy function must be viewed as indeterminate; these parts of the protein cannot be predicted with confidence. As summarized by Finkelstein et al., " . . . one has to compute the equilibrium state of the molecule and pay attention only to its most probable structural features."

From this vantage point, it makes sense to view the process of protein structure prediction in the same framework as equilibrium folding experiments—the virtual polypeptide chain does not follow a single trajectory through conformation space to end in a unique structure. Rather, it proceeds through multiple trajectories that, when stopped at any given stage, can be viewed as comprising an equilibrium ensemble.

One graphical way of representing the output of such an equilibrium prediction would be to align and superimpose a set of five or more of the lowest energy conformations. As illustrated by a hypothetical example in Figure 1, visual inspection of these sets of structures provides a qualitative assessment of the degree to which each part of the structure is determined by the energy function. Alternatively, the individual structures within the ensemble of lowest energy conformations can be converted to distance maps, which can then be added together to obtain an average map (J.R. Gillespie & D. Shortle, in prep.). Just as in NMR spectroscopy with coaddition of a series of free-induction decays, those features that survive the averaging process become the signal, those that do not are considered to be noise.

Fig. 1. A hypothetical example of an equilibrium prediction. A set of seven predicted structures (thin lines) optimally superimposed on the least-squares model of these structures (bold line). Chain segments with good superpositioning correspond to substructures that can be predicted by the energy function, whereas chain segments with poor superpositioning indicate failure of the energy function to make a unique prediction. (Figure taken from Mosimann et al., 1995.)

One attractive feature of equilibrium predictions is that the uncertainty in the predicted ensemble of structures could be rigorously quantified by calculating the RMSD of one or more atom types. In addition, each energy function could be evaluated by two separate sets of criteria: (1) the characteristics of those chain segments for which it can and cannot make a reliable prediction, and (2) the accuracy of the structure obtained for these predicted segments when the native conformation is known. In one sense, an improvement in an energy function should yield a tighter envelope of superimposed structures. More importantly, there should be a better correspondence between the prediction and the real high-resolution structure. By breaking down the evaluation of energy functions into multiple, independent steps, it may be possible to systematically optimize their performance for structure prediction.

Equilibrium folding experiments

By relinquishing the idea of protein folding as the unique trajectory of a single molecule and redefining it as a time-independent and trajectory-independent ensemble phenomenon, a strategy for predicting protein structure can be devised, one that lends itself to rigorous evaluation of its merits. In a somewhat analogous manner of redefining protein folding as an ensemble phenomenon, an equilibrium-based strategy for designing and interpreting folding experiments can be developed. This experimental framework interfaces well with theoretical approaches to equilibrium predictions and could lead to a unified framework that encompasses both theory and experiment. In this section, the logic behind equilibrium folding experiments is outlined.

If one looks at the behavior of proteins in solution, a folding equilibrium can be demonstrated that, at low resolution, is often well described as a simple interconversion of two states: the folded or native state N, and a much less structured state, perhaps best referred to as the "denatured state," D. It was pointed out by Lumry et al. (1966) that these two states are actually distributions (or macrostates) of many different conformations (or microstates). As shown diagrammatically in Figure 2A, a key feature of the equilibrium denaturation reaction is the presence of a significant free energy barrier between these two distributions, leading to very low levels of conformations intermediate in structure between N and D states.

If the reaction coordinate is the number of chain-chain interactions or contacts, the native state must be represented as a very narrow distribution. Evidence from theoretical models plus experimental data on the denatured states of several proteins indicate that the D state is a much broader distribution, one whose position and breadth depend on the conditions under which the equilibrium is established (Dill & Shortle, 1991). In other words, a number of chain-chain interactions persist in the D state; how many and which ones depends on the perturbation that caused the breakdown of the N state. By definition, a random coil state is one lacking side-chain-side-chain interactions. Under conditions mild enough to permit significant population of the N state, the random coil state will be minimally populated because of its high free energy.

An experimental strategy for characterizing the process of spontaneous self-assembly of the folded state can be based on the empirical observation that the residual structure in the denatured state depends on solution conditions and on the amino acid sequence (Dill & Shortle, 1991). Instead of attempting to monitor the buildup of structure as a function of time, as kinetic experiments do, the buildup of structure can be determined as a function of changes in solution conditions (Shortle, 1993). In a series of equilibrium steps, conditions can be made progressively more favorable for folding, as shown diagrammatically in Figure 2B for the series 5-1. For example, urea concentration, glycerol concentration, pH, etc., can be treated as variables that effectively alter "the free energy distance" between the polypeptide chain, frustrated in its efforts to complete the folding process, and the native state it is attempting to reach. By quantitating the structural changes that occur between one equilibrium step and the next, an equilibrium folding pathway can be constructed that reflects the relative strengths and interdependencies of different substructural elements.

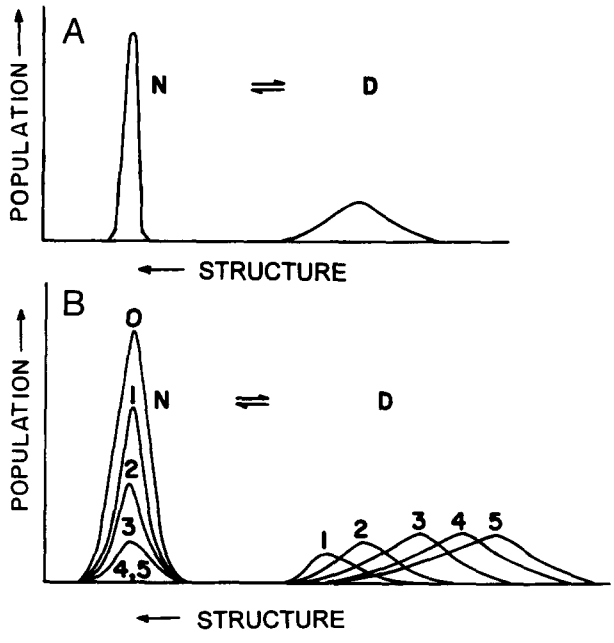

Fig. 2. An equilibrium view of reversible denaturation. **A:** Diagram of the populations of protein conformations (y-axis) versus a parameter that describes structure (x-axis). N represents the native state and D the denatured state. **B:** Schematic diagram of the shifts in the populations of N and D and in the residual structure of the D state as a function of changing conditions, for example, urea concentration from 0 M to 5 M. In experiments to characterize the equilibrium folding pathway, conditions must be used that do not allow significant population of the N state.

The single most important feature of this strategy is that time has been removed as the central variable; time is no longer an obstacle to detailed structural analysis. But the price to be paid for this shift in emphasis from structure as a function of time to structure as a function of conditions of solution is considerable; at least two significant problems arise. Firstly, whereas time has a universal meaning and can be measured precisely, the connection between a variable such as urea concentration or pH and "the free energy distance" is only approximate. To the extent that the mechanism is known by which such variables change the free energies of protein structures, however, this connection can be made semi-quantitative. For example, urea (glycerol) promotes structure breakdown (formation) by making the exposure of buried surface area more favorable (unfavorable).

Secondly, the relevance of an equilibrium pathway to folding in cells or refolding in vitro will be highly uncertain. Because the system is at equilibrium, there are no longer prospects for identifying and characterizing partially folded intermediates on the basis of the free energy barriers between them. Similarly, the issue of whether an equilibrium intermediate is kinetically competent (i.e., on pathway) may be difficult if not impossible to resolve. Although all information on transition states will be lost, the phenomenology of the folding process has been greatly simplified and direct access to the structure of stable, partially folded states has been gained.

Perhaps the most persuasive argument in favor of examining the persistent structure in the denatured state under a variety of conditions is that such experiments can be done with currently available methods, the most important of which is high-resolution NMR spectroscopy. New pulse sequences employing ^{15}N- and ^{13}C-labeled proteins developed for folded proteins can often be applied with good effect to proteins that are not folded (Shortle, 1996a). And new NMR experiments to resolve some of the highly problematic features unique to denatured proteins are being developed. With the opening of this experimental window on the behavior of polypeptide chains outside of the native state, there is a need for a conceptual framework in which to plan and interpret the NMR experiments. The concept of an equilibrium pathway described above provides the first such framework.

The zone of high energy intermediates between the denatured and native states will undoubtedly impose limits on analysis of the most compact unfolded states, because these high energy forms can only be minimally populated. However, these limits are not absolute. A promising new method allows some of the unstable forms on the native side of the intermediate zone to be detected and their structure partially characterized (Bai et al., 1995). By quantitating the kinetics of deuterium exchange of amide protons in the native state and how it is affected by changes in denaturant concentration, inferences can be drawn about the residues that are exposed to solvent in partially *un*folded states, states that have lost some structure, but not enough to reach the transition state for cooperative unfolding.

Thus, it may be feasible to establish connections across the energy barrier separating native and denatured states. With such connections, a more-or-less continuous tracking of equilibrium structure from the denatured to the native state may be possible, yielding a series of snapshots of the ensemble of stable, highly populated conformations under varying equilibrium conditions. At one extreme would be a true random coil state; at the other, the native structure as seen by high-resolution methods. For states in between, one expects to see structural units, such as helices, beta hairpins, and perhaps turns form transiently, then more stably, and then merge into still larger structural units as conditions are made progressively more favorable.

On the basis of a variety of arguments (Srinivasan & Rose, 1995), there is reason to anticipate that equilibrium folding pathways will be hierarchical, in which case the increase in structure can be displayed diagrammatically in a hierarchical cluster diagram (an example of which is shown in Fig. 3B). The free energy distance becomes the independent variable along the *x*-axis and the state of association between chain segments the dependent variable along the *y*-axis. Through a succession of coalescence events involving substructural elements (turns, helices, and extended strands either singly or in combination) that combine to form larger structural components, the native state is reached as the final step.

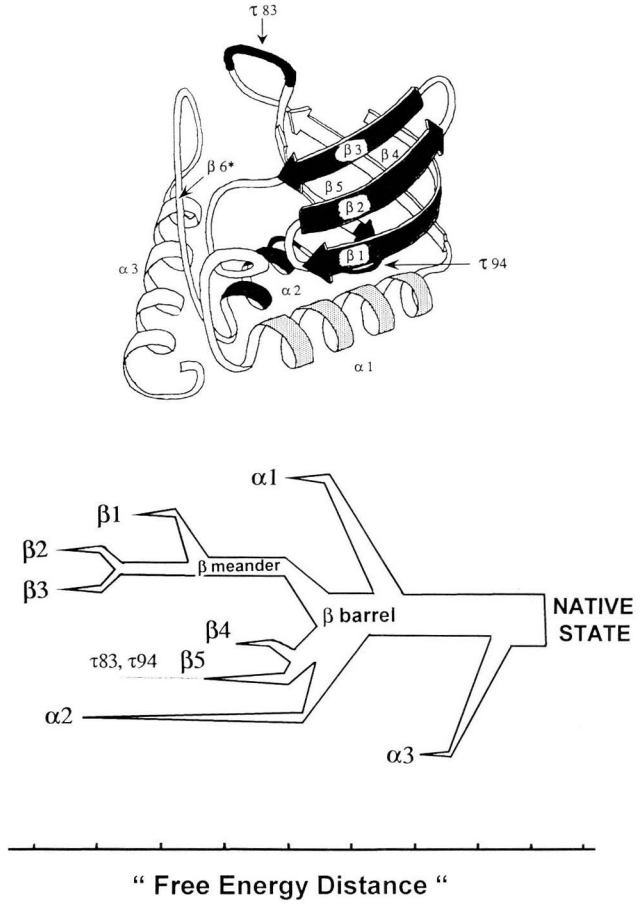

Fig. 3. Top: Ribbon diagram showing the secondary structure of native staphylococcal nuclease. Dark chains segments form a similar structure in the denatured state with modest to high stability. Stippled chain segments are only weakly stable in the denatured state. Beta strands (β1-β5) and alpha helices (α1-α3) are numbered sequentially from the amino terminus. Segment β6* is extended in the native state, but does not form hydrogen bonds with other beta strands. Bottom: Hierarchical cluster diagram showing a tentative model of the equilibrium folding pathway for staphylococcal nuclease (Wang & Shortle, 1995). The *x*-axis is a measure of the difference in free energy between the partially folded polypeptide chain and the native state, i.e., the free energy distance. The *y*-axis corresponds to the protein sequence from amino (top) to carboxy (bottom) terminus, except for α1, which is placed out of position for readability. Chain segments correspond to those in the figure above. An approximate gradient of substructure stability extends from the upper left to the bottom right.

An example: Equilibrium folding pathway of staphylococcal nuclease

Over the past several years, our laboratory has been working on the equilibrium folding pathway of staphylococcal nuclease, a small, 149-residue enzyme that hydrolyzes both DNA and RNA nonspecifically. The overall structure of this $\alpha+\beta$ protein, shown in Figure 3A, consists of a five-stranded Greek key antiparallel beta barrel, three alpha helices, and a several tight turns, loops, and connecting segments. The beta barrel forms the ma-

jor hydrophobic core, with the three helices interacting tangentially with this core and aligned approximately parallel to each other. Because the nuclease structure contains the small OB domain motif – the five-strand barrel plus one peripheral helix or loop – it may have evolved from a precursor with a smaller fold through elongation of one helix ($\alpha 1$) plus the addition of ~50 residues that form helices $\alpha 2$, $\alpha 3$, and a long extended segment running between them (Alexandrescu et al., 1995).

On the basis of a series of experiments using NMR and CD to monitor the structure of denatured nuclease as a function of urea, glycerol, pH, and amino acid sequence, a tentative equilibrium folding pathway has been proposed (Wang & Shortle, 1995). As shown in Figure 3B, the two most stable substructures are a two-strand beta hairpin $\beta 2$-$\beta 3$ and a short alpha helix $\alpha 2$. Both are significantly populated in 6 M urea. As conditions are made more favorable for structure formation, strand $\beta 1$ adds to $\beta 2$-$\beta 3$ to from a three-strand meander $\beta 1$-$\beta 2$-$\beta 3$. Type I and I′ beta turns involving residues 83–86 and 94–97, respectively, appear to be highly populated at this point, even though there is little evidence for formation of strand $\beta 5$, the segment bracketed by these two turns. As described in the next section, $\alpha 2$ and the extended segment $\beta 6$* exhibit a set of non-native interactions with segments $\beta 4$ and $\beta 5$, interactions that are undone upon coalescence of the five beta strands to form the barrel structure with its large hydrophobic core. Helices $\alpha 1$ and $\alpha 2$ become more stable after formation of the beta barrel, and the final event appears to be the docking of the longest helix $\alpha 3$ with the rest of the protein to form the native state.

Most of the data on which this pathway was built were obtained from a fragment of nuclease lacking five structural residues at the amino terminus and one structural residue at the carboxy terminus. This fragment, referred to as $\Delta 131\Delta$, is more than 99.5% denatured under optimal conditions, yet upon the addition of ligands that bind tightly at the active site, it refolds to give a native state virtually indistinguishable from wild-type nuclease. NMR analysis of $^{15}N/^{13}C$-labeled samples under nondenaturing conditions has yielded assignments for most of the backbone and side chain resonances (Alexandrescu et al., 1994). A variety of residue-specific NMR parameters (secondary chemical shifts, coupling constants, and medium-range NOEs) indicate at least four secondary structural elements are highly populated in buffer: two tight beta turns (t83–86 and t94–97), helix $\alpha 2$, and the beta meander $\beta 1$-$\beta 2$-$\beta 3$.

The initial evidence for persistence of the beta meander was indirect: *all* of the amide protons H_N for the residues that form this supersecondary structure were missing from the ^1H-^{15}N NMR spectrum. Considering the solution conditions under which the data were collected, the only reasonable explanation for loss of proton peaks from a contiguous set of residues was intermediate exchange broadening – the presence of a dynamic set of conformations with large differences in chemical shifts that interconvert within a few milliseconds (Alexandrescu et al., 1994). In 6 M urea, however, NMR analysis revealed no "missing peaks," presumably because the structures responsible for intermediate exchange broadening had been disrupted (Wang et al., 1995). With assignments of the ^1H, ^{15}N, and ^{13}C resonances of virtually all residues in 6 M urea, the changes in structure of $\Delta 131\Delta$ could be tracked between 0 M and 6 M urea in 1 M intervals by NMR methods and by CD.

The CD data shown in Figure 4B indicate that the structure being formed as the urea concentration is lowered is predominantly beta sheet-like in character. From the pattern of simultaneous disappearance of sets of amide protons as a function of urea (Fig. 4C,D), it appears that strands $\beta 2$ and $\beta 3$ form a dynamic structure below 6 M urea and that strand $\beta 1$ incorporates into this structure below 3 M urea. Although direct NMR analysis of these structures is limited by severe line broadening of most NMR resonances for residues in the $\beta 1$-$\beta 2$-$\beta 3$ segment, direct detection of the carbonyl ^{13}C resonances provides supporting evidence that the structure being formed has considerable beta strand character (Wang & Shortle, 1995).

Helix $\alpha 2$ is also partially populated in 6 M urea. On the basis of secondary chemical shifts of several backbone resonances, the fractional helical content of this chain segment can be estimated as 10–15%, down from the estimated 30% in 0 M urea. The high intrinsic stability of the $\alpha 2$ helix, the $\beta 2$-$\beta 3$ hairpin, and the $\beta 1$-$\beta 2$-$\beta 3$ meander were predicted by Moult and Unger (1991) on the basis of the unusually large amount of nonpolar surface area buried upon forming these secondary structures from a random coil conformation. Preliminary characterization of a second, distinctly different denatured state – WT nuclease at pH 3.0 and low salt – has revealed a state very similar in average residual structure to $\Delta 131\Delta$ at low concentrations of urea (Wang & Shortle, 1995), a finding that suggests that many features of the denatured state of staphylococcal nuclease are independent of the mode of denaturation.

The changes in the structure of $\Delta 131\Delta$ that accompany addition of glycerol [which effectively strengthens hydrophobic interactions (Gekko & Timasheff, 1981)] have also been monitored by CD and NMR spectroscopy (Wang et al., 1995). As shown in Figure 5B, the structure induced by 30% glycerol is predominantly beta-sheet like, whereas continued addition of glycerol to a concentration of 70% appears to induce significant amounts of helical structure. Although the changes in the intensity of the amide proton peaks shown in Figure 5C, D, E, and F cannot be interpreted quantitatively because of viscosity-induced broadening effects, they are consistent with rapid incorporation of $\beta 4$ and $\beta 5$ into a dynamic structure (which produces severe line broadening by intermediate exchange), a somewhat slower incorporation of $\alpha 1$ and $\alpha 2$ into a dynamic structure, and only a small amount of structural change for residues in the carboxy-terminal helix $\alpha 3$. Changes in the fluorescence of the unique tryptophan at the end of $\alpha 3$ as a function of glycerol concentration strongly support the conclusion that docking of this alpha helix with a nearly completed folded structure involving the rest of the polypeptide chain is the final event on the equilibrium pathway.

Hydrophobic bundles and stability gradients

This initial characterization of the folding pathway of nuclease indicates that equilibrium intermediates do not always form a simple hierarchy of exact subsets of native structure that converges symmetrically on the native state. Two examples have been found of specific interactions between chain segments that are not seen in the native state, interactions that may play a role as scaffolding that transiently stabilizes supersecondary structures in preparation for subsequent coalescence events. In addition, a gradient in the relative stabilities of substructures has been identified, with the most stable located toward the amino terminus and the least stable toward the carboxy terminus.

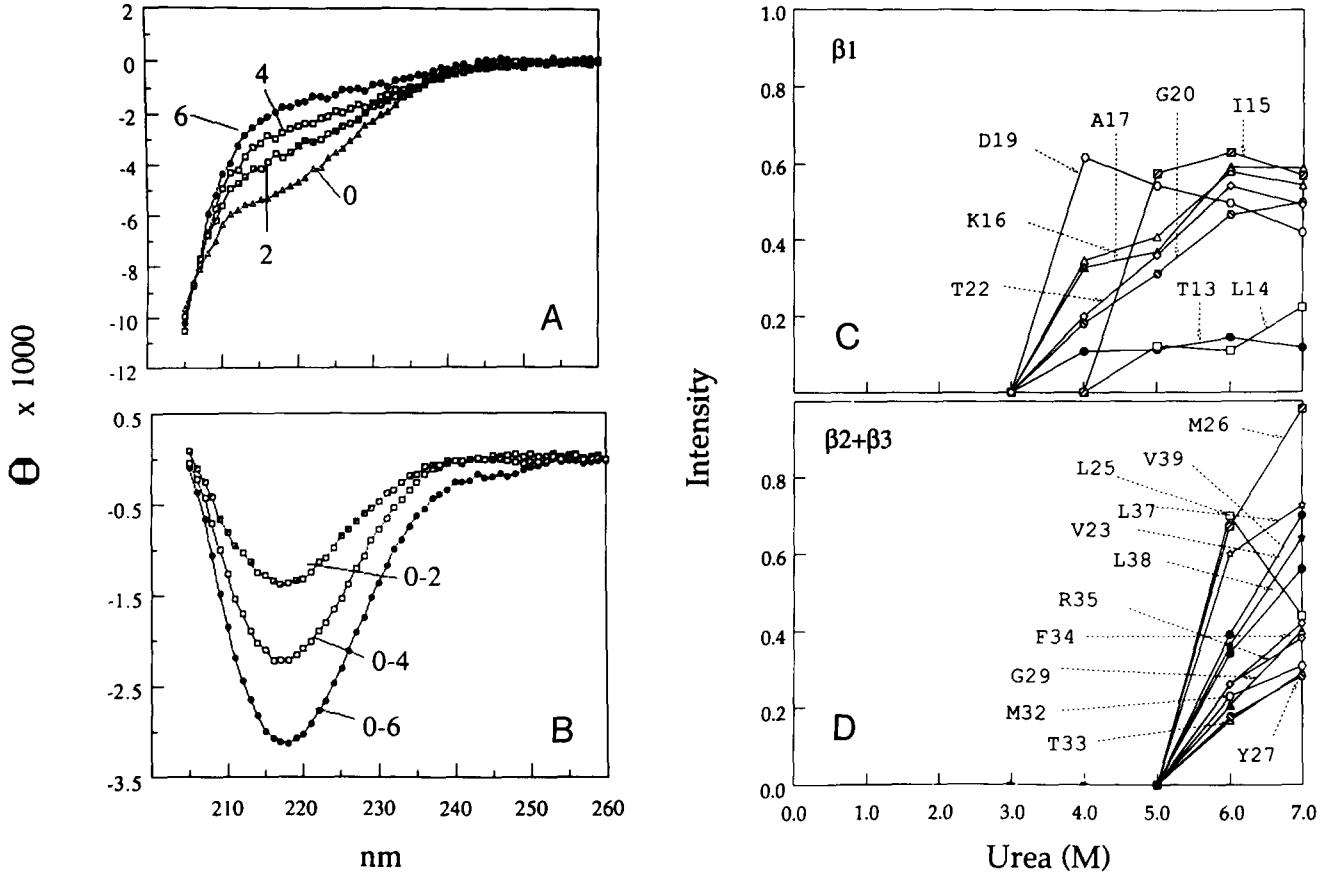

Fig. 4. Structural changes produced by urea. **A:** Far-UV CD spectrum of Δ131Δ as a function of urea concentration. **B:** Difference spectra labeled with the urea concentrations in M. Intensities of H_N resonances (peaks heights from the HSQC spectrum) of Δ131Δ as a function of urea concentration for (**C**) residues between T13 and T22, which includes the β1 segment; and (**D**) residues between V23 and V39, which includes the β2-β3 segment plus the four large hydrophobic residues. Taken from Wang and Shortle (1995).

The data in Figure 4D demonstrate that, in addition to the residues from the β2-β3 hairpin, several residues nearby in sequence are included in the dynamic structure formed at high urea concentrations. In particular, the H_N of four large hydrophobic residues, L36, L37, L38, V39, disappear in concert with those of the beta hairpin. The ^{13}C carbonyl chemical shifts of V39 and L37 (or L36) are suggestive of beta strand character, and single glycine substitutions for each of these four residues profoundly lower the stability of this structure, leading to the appearance of the "missing" H_N resonances at 0 M or very low urea concentrations (Y. Wang & D. Shortle, in prep.). Because preliminary measurements of the hydrogen exchange kinetics of the β1-β2-β3 segment in the denaturated state show little if any protection, hydrogen bond pairing between beta strands (as shown in Fig. 6B) must be either transient or absent altogether. Thus, it may be more appropriate to describe this structure as a dynamic bundle of four hydrophobic segments with up-down topology rather than as a beta meander.

A second bundle of four hydrophobic segments appears to form in a contiguous stretch of polypeptide chain that includes β4, β5, α2, and β6*, an extended segment between 110 and 116. In an effort to characterize the long-range interactions between residues in Δ131Δ, paramagnetic relaxation effects on the amide protons produced by PROXYL spin labels (Kosen, 1989) attached at sites of unique cysteine mutations are being quantitated (J.R. Gillespie & D. Shortle, in prep.). Enhancements of both T_1 and T_2 can be translated into approximate interresidue distances in the range of 10–20 Å and used as long-range distance restraints for structural calculations in the same way that NOEs serve as short-range restraints. Although this study is not yet complete, it has so far provided unambiguous evidence of close interactions among these four hydrophobic chain segments, with the predominant orientations shown in Figure 6D. As with the bundle involving β1-β2-β3 plus residues 36–39, some of these interactions are not seen in the native state.

For both of these dynamic structures, a contiguous stretch of protein chain containing four hydrophobic segments appears to have folded back on itself in a simple up-down or hairpin topology. In both cases, some of these chain-chain interactions are retained in the native state, whereas others are lost through changes in the angle between two of the hydrophobic segments. From this overall pattern, it is tempting to conclude that these two hydrophobic bundles are formed by a local hydrophobic zippering process (Dill et al., 1993) that establishes part of the native topology and stabilizes, in an approximate form, two native supersecondary structures (β1-β2-β3 for the first bundle and β4-β5 for the second bundle). In subsequent folding steps, the non-native interactions are replaced by native interactions

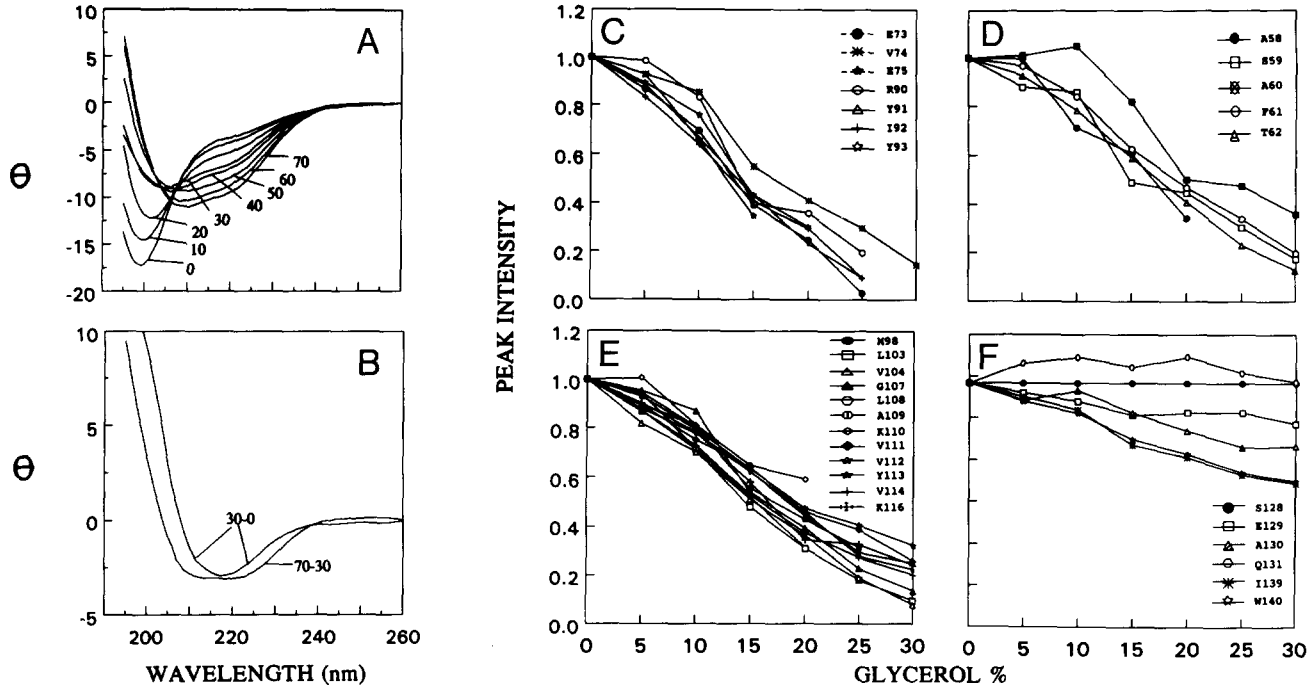

Fig. 5. Structural changes produced by glycerol. **A:** Far-UV CD spectrum of Δ131Δ as a function of glycerol concentration. **B:** Difference spectrum labeled with the glycerol concentration in percent by volume. Intensities of H_N resonances (peak heights from the HSQC spectrum) of Δ131Δ as a function of glycerol concentration for **(C)** residues from the β4 and β5 segments; **(D)** residues from the α segment; **(E)** residues from a segment that includes α2 plus β6*; **(F)** residues from the α3 segment plus carboxyl terminal loop. Taken from Wang et al. (1995).

through changes in angles between hydrophobic segments driven by the formation of long-range tertiary contacts.

Although local non-native interactions in partially folded states must be paid for indirectly as a reduced stability of the native state (Shortle, 1996b), there may be several general advantages to employing them in the folding process. They could provide a simple strategy for facilitating the formation of supersecondary structures by fixing chain segments in favorable relationships prior to their incorporation into independently stable folding domains. An even more important function may be to sequester hydrophobic segments into locally favorable clusters to prevent their incorporation into inappropriate long-range intramolecular or intermolecular interactions, the types of interactions that chaperones are required to break down before folding can continue. The energetic cost of local non-native hydrophobic interactions may be more than offset by prevention of even more costly long-range non-native interactions.

A second unexpected feature of the nuclease folding pathway is the presence of a gradient of relative stabilities of substructures from the amino to the carboxy terminus. As can be seen in Figure 3B, the most stable structure involves the β2-β3 chain segment; the least stable secondary structure, the α3 segment; and structures of intermediate stability derive from β4-β5, α2 plus β6*. This gradient makes the hierarchical arrangement of the pathway inherently asymmetric, with some branches well underway or completed before others initiate. Two types of statistical data suggest that such a gradient may be a general feature of proteins: prediction of secondary structure is, on average, more successful near the amino terminus than near the carboxy terminus (Schultz, 1988) and structure near the amino terminus is, on average, more locally compact than near the carboxy terminus (Alexandrov, 1993).

Such a gradient could contribute to efficient folding in several ways, such as reducing the opportunities for incorrect hydrophobic interactions and facilitating rapid folding in vivo by structuring the amino terminus immediately as it emerges from the ribosome or from a membrane after translocation. Preliminary studies of the hydrophobic zippering of HP model chains configured on a square lattice (Dill et al., 1993) indicate that a modest gradient in H-H contact probability can increase the efficiency of finding the native state by the zippering process by more than tenfold for some HP sequences (J.O. Wrabl & D. Shortle, unpubl. results).

The pathway proposed in Figure 3B represents a tentative model based on a limited amount of data. Hopefully, future NMR experiments will provide additional data to refute or support the more speculative features of this model and also add new structural details. Although analyses of partially folded proteins by NMR spectroscopy are providing many rather formidable challenges (Shortle, 1996a), there are good reasons to expect that new pulse sequences employing ^{15}N- and ^{13}C-labeled proteins plus improvements in instrumentation will lead to steady progress in the years ahead on staphylococcal nuclease and other model systems with the requisite high solubility of the denatured state.

Connecting predictions and experiments

The equilibrium perspective of protein folding described in this review, with its emphasis on ensembles and free energy rather

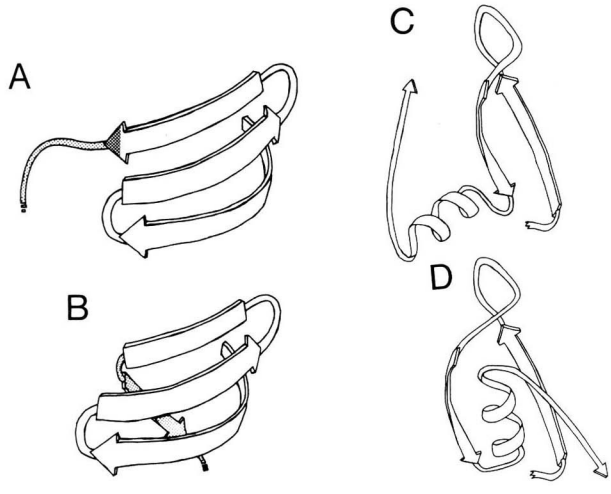

Fig. 6. Schematic diagrams showing the topological relationships between chain segments for two non-native structures observed in the denatured state of staphylococcal nuclease. NMR data suggest that these two four-strand bundles are *much more variable and heterogeneous in structure* than suggested by these diagrams, with little or no hydrogen bonding between adjacent beta strands. Arrangement of segments formed by the β1-β2-β3+ four hydrophobes group (**A**) in the native state and (**B**) in the denatured state. Taken from Wang and Shortle (1995). Spatial arrangement of segments formed by the β4-β5-α2-β6* group (**C**) in the native state and (**D**) in the denatured state.

than on time and unique trajectories, may serve as a unified framework that encompasses both predictions and experiments. Progress in theory could be accelerated by the availability of experimental tests for predictions of secondary and supersecondary structural features based on NMR data. As mentioned earlier, Moult and Unger (1991) correctly predicted several features of the denatured state of staphylococcal nuclease—the β1-β2-β3 meander and helix α2—from simple considerations of burial of hydrophobic surface. [Interestingly, their prediction that an unusual loop involving hydrophobic residues between A132 and W140 should also be stable in solution is not supported by NMR analysis (Maciejewski & Zehfus, 1995; Wang et al., 1995)].

Similarly, equilibrium experiments stand to benefit from what theory and simulations have to offer. At present, NMR characterization of partially folded states of proteins supplies only a limited number of structural restraints per residue, far fewer than needed to define all of the basic features of ensembles of dynamic structures. Thus, the optimal approach to interpreting the NMR data will probably be similar to that used in other situations in which the data is insufficient to uniquely determine a solution. Simulations will generate multiple, alternative models of what the structure and dynamics of the chain may be like; from these models, the observable NMR parameters can be calculated, and the model providing the best fit to the experimental data will be selected as the best description of the structure.

A common framework for experimental design and data interpretation could generate opportunities for synergistic interactions between theorists and experimentalists. By providing a common language for designing and interpreting experiments and simulations, the two groups could collaborate in more productive ways, leading to data that can more effectively support or refute ideas about how the polypeptide chain interacts with solvent and with itself.

Conclusions

For the forseeable future, it is unrealistic to expect that kinetic studies of protein refolding will provide high-resolution structural information about the very early, very important chain–chain interactions that establish the global topologies of protein folds. Similarly, it may be unrealistic to expect that efforts at protein structure prediction will succeed any time soon in routinely providing unique, high-resolution models that approach the accuracy of models obtained by X-ray crystallographic methods. As outlined in this review, one solution to both dilemmas is to redefine the phenomenon of protein folding in a way that reduces both problems to a more manageable form and then to combine theory and experiment into a joint enterprise. By removing time from the dominant position it currently holds, and describing protein folding in terms of free energy and ensembles of equilibrium structures, a more doable agenda for theory and experiment can be formulated. With the advantages to be gained by close interactions between calculation and measurement, such an equilibrium-based approach holds considerable promise for future progress toward a quantitative understanding of how amino acid sequence encodes protein structure.

Initial equilibrium folding experiments on staphylococcal nuclease have uncovered two unexpected features of folding intermediates: the presence of bundles of hydrophobic chain segments with simple topologies and a gradient in stabilities of local structures from the amino terminus to the carboxy terminus of the chain. Obviously, the generality of such features can only be established after the equilibrium intermediates formed by other proteins have been characterized. Nevertheless, these early results hint at the possibility that a wealth of new and unexpected patterns of chain–chain interactions may soon be discovered in partially folded proteins, patterns that, when properly understood, may provide the insights needed to solve the protein folding problem.

Acknowledgments

We thank John Moult for helpful discussions, Al Mildvan for his many persuasive examples where kinetics does not provide all of the answers, and Andrei Alexandrescu and Chitrananda Abeygunawardana for their contributions to the NMR spectroscopy. This work was supported by an NIH grant (GM34171).

References

Alexandrescu AT, Abeygunawardana C, Shortle D. 1994. Structure and dynamics of a denatured 131-residue fragment of staphylococcal nuclease: A heteronuclear NMR study. *Biochemistry 33*:1063–1072.

Alexandrescu AT, Gittis AG, Abeygunawardana C, Shortle D. 1995. NMR structure of a stable "OB-fold" subdomain isolated from staphylococcal nuclease. *J Mol Biol 250*:134–143.

Alexandrov N. 1993. Structural argument for N-terminal initiation of protein folding. *Protein Sci 2*:1989–1992.

Avbelj F, Moult J. 1995. Determination of the conformation of folding initiation sites in proteins by computer simulation. *Proteins Struct Funct Genet 23*:129–141.

Bai Y, Sosnick TR, Mayne L, Englander SW. 1995. Protein folding intermediates: Native-state hydrogen exchange. *Science 269*:192–197.

Chou PY, Fasman GD. 1974. Prediction of protein conformation. *Biochemistry 13*:222–245.

Dill KA, Fiebig KM, Chan HS. 1993. Cooperativity in protein-folding kinetics. *Proc Natl Acad Sci USA 90*:1942–1946.

Dill KA, Shortle D. 1991. Denatured states of proteins. *Annu Rev Biochem 60*:795–825.
Finkelstein AV, Gutin AM, Badretdinov AY. 1995. Perfect temperature for protein structure prediction and folding. *Proteins Struct Funct Genet 23*:142–150.
Gekko K, Timasheff SN. 1981. Mechanism of protein stabilization by glycerol: Preferential hydration of glycerol-water mixtures. *Biochemistry 20*:4667–4676.
Kosen PA. 1989. Spin labeling of proteins. *Methods Enzymol 177*:86–121.
Levinthal C. 1968. Are there pathways for protein folding? *J Chim Phys 65*:44–45.
Lumry R, Biltonen R, Brandts JF. 1966. Validity of the "two-state" hypothesis for conformational transitions of proteins. *Biopolymers 4*:917–944.
Maciejewski MW, Zehfus MH. 1995. Structure of a compact peptide from staphylococcal nuclease determined by circular dichroism and NMR spectroscopy. *Biochemistry 34*:5795–5800.
Matthews CR. 1993. Pathways of protein folding. *Annu Rev Biochem 62*:653–683.
Mosimann S, Meleshkdo R, James MNG. 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins Struct Funct Genet 23*:301–317.
Moult J, Unger R. 1991. Analysis of protein folding pathways. *Biochemistry 30*:3816–3824.
Schultz GE. 1988. A critical evaluation of methods for prediction of protein secondary structures. *Annu Rev Biophys Bioeng 12*:183–210.
Shortle D. 1993. Denatured states of proteins and their roles in folding and stability. *Curr Opin Struct Biol 3*:66–74.
Shortle D. 1996a. Structural analysis of non-native states of proteins by NMR methods. *Curr Opin Struct Biol 6*:24–30.
Shortle D. 1996b. The denatured state (the other half of the folding equation) and its role in protein stability. *FASEB J 10*:27–34.
Srinivasan R, Rose GD. 1995. LINUS: A hierarchical procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.
Sun S, Thomas PD, Dill KA. 1995. Simple protein folding algorithm using a binary code and secondary structure constraints. *Protein Eng 8*:769–778.
Wang Y, Alexandrescu AT, Shortle D. 1995. Initial studies of the equilibrium folding pathway of staphylococcal nuclease. *Philos Trans R Soc Lond [Biol] 348*:27–34.
Wang Y, Shortle D. 1995. The equilibrium folding pathway of staphylococcal nuclease: Identification of the most stable chain–chain interactions. *Biochemistry 34*:15895–15905.