A Statistical Thermodynamic Model of the Protein Ensemble

Vincent J. Hilser,*,† Bertrand García-Moreno E.,‡ Terrence G. Oas,§ Greg Kapp,§ and Steven T. Whitten†

Department of Biochemistry and Molecular Biology and Sealy Center for Structural Biology and Molecular Biophysics, University of Texas Medical Branch, Galveston, Texas 77555-1068, Department of Biophysics, The Johns Hopkins University, Baltimore, Maryland 21218, Departments of Biochemistry and Chemistry, Duke University, Durham, North Carolina 27710

Received April 11, 2005

Contents

1. Introduction

To understand the structural basis of biology, it is necessary to understand the transformations of macromolecules during functional cycles. These changes can involve large-scale conformational transformations, changes in the state of ligation or association, changes in dynamics, or alterations in covalent bond structure. X-ray crystallography and NMR spectroscopy have been invaluable for the description of the structures of proteins and their transformations. However, proteins are far from being the static structures used to represent them. Deeper understanding of key functional properties of proteins, such as allostery, catalysis, and the roles of proteins in signal and energy transduction, will require improved understanding of the conformational excursions around the state represented by the high-resolution structure.

Proteins, even under native conditions, exist as ensembles of related, interconverting, transient microstates that, as an average, describe the canonical high-resolution structures observed by crystallography or NMR spectroscopy. The view of the native state as a set of structural microstates raises the possibility that many of the physical and functional properties of proteins (e.g., stability, solubility, their ability to recognize, bind, and respond to the binding of ligands) are influenced significantly by the same structural fluctuations that give rise to the ensemble. The observed biological activity of proteins represents the energy-weighted contributions of the component microstates of the ensemble. The next step toward elucidation of the structural basis of biological organization will require understanding the structural character and energetics of the constituent microstates of protein ensembles.

Here we review the background, the physical basis, and the experimental validation of a structural thermodynamic model of the protein ensemble, known as COREX. Over the past decade, this simple model has been shown to reproduce a surprising number of apparently disparate biophysical and functional properties of proteins. It has afforded novel interpretations of solution properties and suggested experiments to test previously unrecognized roles of local conformational fluctuations on functional aspects of proteins.

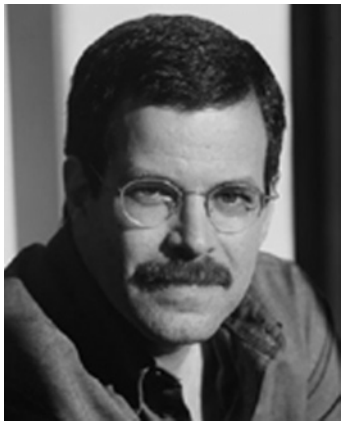
In light of the apparent simplicity of the model, it is important to explore the implications of its success at explaining a range of physical observations. More important than what experimental tests have revealed about the accuracy of COREX is what these tests suggest about the behavior of proteins in general and about the robustness of the observed behavior. The underpinnings of the COREX algorithm are rather unique; they were adopted for their simplicity. Here these underpinnings are contrasted with alternative approaches for modeling conformational fluctuations in proteins. The self-consistent view of proteins that is emerging from studies with the COREX model suggests that without detailed understanding of the structural and energetic character of the ensemble, it will not be possible to further dissect the relationship between structure and function of proteins.

* Corresponding author. Fax: 409-747-6816. Tel: 409-747-6813. E-mail: vjhilser@utmb.edu.
† University of Texas Medical Branch.
‡ The Johns Hopkins University.
§ Duke University.

10.1021/cr040423+ CCC: $59.00
Published on Web 03/23/2006

Vincent J. Hilser has earned a B.S. degree in chemistry from St. John's University (1987), an M.S. degree in biotechnology from Manhattan College (1991), and a Ph.D. in biochemistry from The Johns Hopkins University (1995). From 1995 through 1997, he did postdoctoral research in the lab of Ernesto Freire (John Hopkins University). In 1997, he accepted an assistant professor's position at the University of Texas Medical Branch, where currently he is Director of the Sealy Center for Structural Biology and Molecular Biophysics. Research in his lab is directed toward elucidating the physical and energetic basis, as well as the functional consequences, of conformational heterogeneity in proteins and applying this information to the development of novel fold classification schemes and protein design strategies.

Bertrand García-Moreno was born and raised in Mexico City. He earned an A.B. degree in biochemistry from Bowdoin College and a Ph.D. in chemistry under the direction of Prof. Frank Gurd at Indiana University in Bloomington. He did postdoctoral research with Prof. Gary Ackers at Johns Hopkins University and at Washington University School of Medicine in St. Louis. He returned to Johns Hopkins in 1992 to join the Department of Biophysics. Research in his laboratory is currently focused on experimental and computation studies of structure-energy relationships and electrostatic effects in proteins.



Terry Oas received a Ph.D. from the University of Oregon in 1986. Currently, he is an Associate Professor of Biochemistry and Chemistry at Duke University. His lab is primarily interested in the mechanisms of protein folding.

Greg Kapp finished his undergraduate degree in biology and chemistry at the University of Richmond in 1997. He continued his education at the Duke University Medical Center where he pursued his Ph.D. with Terry Oas in the Department of Biochemistry. Greg's Ph.D. project focused on experimental and theoretical investigation of the kinetics of protein folding. Greg's experimental work involved investigation of individual sites in the model protein monomeric λ repressor and how sequence changes in these positions affect the kinetics of protein folding and unfolding. In his theoretical work, Greg concentrated on implementing microscopic reversibility and modeling of substrate unfolding rates in the diffusion–collision model for protein folding. Greg received his Ph.D. in 2003 and is now a postdoctoral fellow in Tanja Kortemme's laboratory in the Department of Biopharmaceutical Sciences at the University of California, San Francisco. Greg is currently computationally and experimentally redesigning protein and protein–protein interactions involved in cellular signaling events.

2. The Protein Ensemble—Framing the Problem

2.1. The Energy Landscape of Proteins

A powerful strategy for elucidating the physical and structural basis of function of biological macromolecules consists of correlating the observed changes in energy for a particular biological process with the structural changes observed using high-resolution X-ray crystallography or NMR spectroscopy. Thermodynamic information is an essential element in the dissection of the structural basis of biological function. Specifically, the Gibbs free energy (ΔG) is useful to describe quantitatively the probability of each state and of transitions between them, whereas the enthalpy (ΔH) and entropy (ΔS) functions are useful because they inform on the noncovalent forces stabilizing the different structural states.

The native state of each protein is usually considered as a single species represented by the high-resolution structure. Hydrogen exchange and NMR relaxation measurements convey a much different picture of proteins, requiring that the native state be viewed as an ensemble of interconverting conformational states.[1−11] This is depicted schematically in Figure 1, which shows a hypothetical energy landscape of a protein with the high-resolution structure represented as the lowest energy state.[12] In this figure, the landscape that is represented is rugged, with various local minima, presumably representing states with non-native structure occupying basins that could influence the kinetics of processes such as folding, ligand binding, or interactions with other proteins. It is of considerable interest to know what the different conforma-

Steven Whitten received a B.S. degree (Physics) in 1994 from the University of Nebraska at Omaha and a Ph.D. from the Department of Biophysics at The Johns Hopkins University in 1999. Under the guidance of Prof. Bertrand García-Moreno, his Ph.D. thesis centered on experimental dissection of the residue-specific contributions to pH-dependent stability in staphylococcal nuclease. Following a postdoctoral fellowship in 2000 in the labs of Profs. Michael Blaber and Timothy Logan (Florida State University), he has since worked as a research scientist in the lab of Prof. Vincent Hilser at the University of Texas Medical Branch. There, his primary interest has been in developing computational models of the solution-dependent behavior of protein structure with a particular focus on pH, temperature, salt, and osmolyte effects.

tional microstates accessible to proteins are and how these alternative conformations affect their solution and functional properties. In other words, what and where are the minima in the energy landscape? This is equivalent to knowing the structures and the energies of the microstates that are populated under native conditions. It is also of interest to elucidate how environmental variables such as pH, temperature, salts, mutation, and additives such as denaturants, osmolytes, and ligands affect the structural character and distributions of microstates in an ensemble. The COREX model can be used to address these questions.

2.2. The Ergodic Hypothesis

2.2.1. Temporal vs Instantaneous Representations of Ensembles

Investigations of the conformational heterogeneity in proteins based on high-resolution structures can be grouped into two approaches: (i) temporal representations that consider the evolution of a single molecule over a trajectory in time and (ii) instantaneous or ensemble representations that consider the distribution of microstates at any given instant (see Figure 2). The ergodic hypothesis states that these two approaches can provide equivalent information. In practice, accurate modeling of protein fluctuations is difficult to achieve; protein molecules contain thousands of atoms that can be arranged in an almost infinite number of ways. Therefore, explicit consideration of all conformational permutations is computationally intractable. The inherent inability of any method to enumerate all possible microstates of a system impacts temporal and ensemble representations of proteins in different ways.

The most common method for exploring the temporal representation of the protein ensemble involves use of molecular dynamics (MD) simulation.[13−17] This atomistic approach is particularly useful when information about changes in state, such as the identification of plausible transition states, is of interest. One drawback of standard

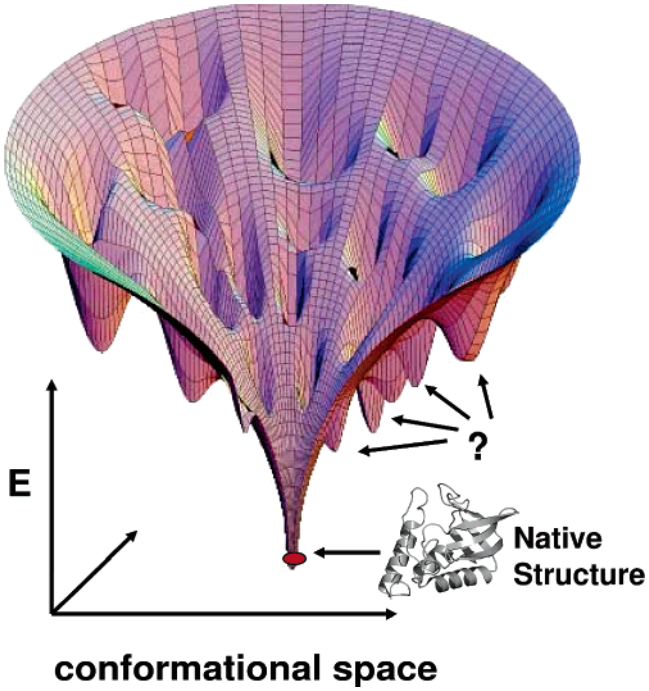


Figure 1. The hypothetical energy landscape of a protein.

MD methods is that despite the substantial computer power available today, simulations for proteins of even modest size (200 amino acids) are usually limited to the tens of nanoseconds time regime. The conformational excursions that are sampled in a standard MD simulation are generally the high frequency fluctuations, leaving a large section of conformational space unexplored. Methods such as replica exchange MD have been devised recently for improved conformational sampling.[18−20] Although information about the evolution of the system as a function of time is lost in these types of simulations, a larger conformational space is searched by means of Monte Carlo guided cycles of heating and annealing.[21] Notwithstanding these improvements, replica exchange methods are not yet applicable to large proteins, and it is challenging to derive from these calculations realistic probabilities for the species that are sampled.

All-atom simulations are not yet useful to routinely access the fluctuations and conformational states that are sampled in the relatively slow time scales (microseconds and slower) relevant for most biological equilibrium thermodynamic processes. Coarse-grained models aim to cover this gap. Models such as Cα Go, in which attractive forces are assigned to native contacts and repulsive ones to non-native contacts, represent a useful alternative.[22] Despite the minimalist nature of these models, they have been very successful in elucidating molecular details of the energetics and kinetics of folding.[23−26] An important implication of the success of these approaches is that they suggest that many seemingly complex phenomena are actually very robust features of the ensemble (i.e., they are not particularly sensitive to the precision of the energy function). This is an important result that was taken into consideration in the development of the COREX model described below.

There are other coarse-grained methods useful for the study of dynamic fluctuations about the native state that bypass the computational inefficiency of atomic approaches for the study of large macromolecules. For instance, in the Gaussian network model, simplified force fields are used to describe vibrational dynamics.[27,28] In these models, the fluctuations

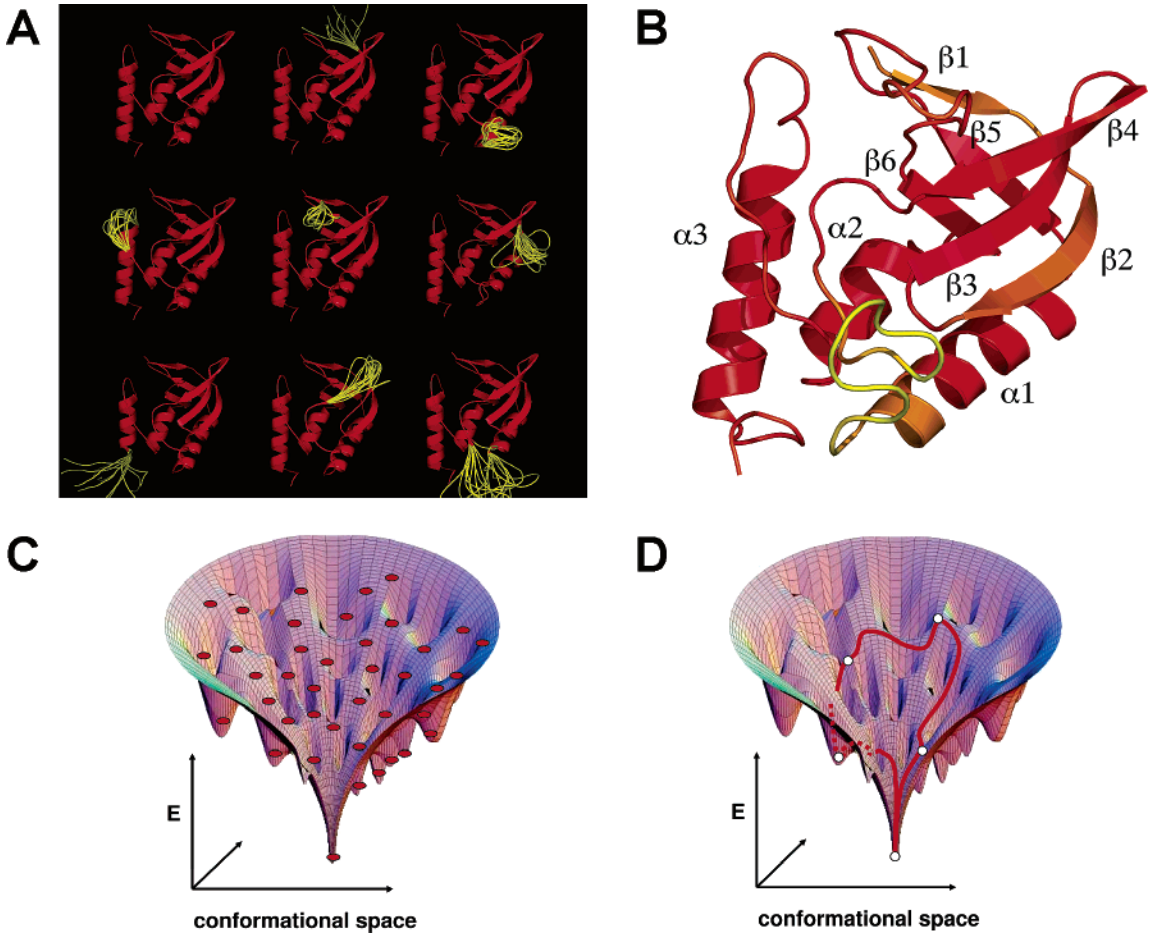


Figure 2. The ergodic hypothesis. A fundamental postulate of statistical mechanics is that the instantaneous probability distribution of an ensemble is equivalent to the time average of a single molecule. In practical terms, however, this requires sampling over long time trajectories. Panel A presents a hypothetical representation of the instantaneous ensemble of a generic protein showing regions that are non-native (yellow) and native (red) in each state. Panel B shows the temporal representation of a single molecule showing regions with different probabilities for conformational diversity (red = low; orange = medium; yellow = high). Regions colored orange are those that are either red or yellow in the different states shown in panel A. Panel C show the energy landscape representation of an ensemble of states (red points). In the instantaneous ensemble, the pathways between states are not considered. Panel D shows the energy landscape representation of a single molecule performing a search (red arrows) through conformational space.

of residues are assumed to be Gaussian-distributed about their mean positions and coupled by harmonic potentials. This approach is particularly appropriate for describing the collective motions of proteins, and it has been applied successfully to the study of various equilibrium properties of proteins governed by fluctuations, such as hydrogen exchange, crystallographic temperature factors, and NMR relaxation.[27,28]

One of the difficulties with all of the above methods for describing the protein fluctuations is that they do not account self-consistently for the effects of environmental variables such as pH, pressure, temperature, salt, mutations, and the chemical potential of salts, ligands, denaturants, and osmolytes. These models have not been designed to allow the calculation of Gibbs free energies of the different microstates in the ensemble (ΔG). Therefore, they cannot be used to calculate the probabilities of the states. In general, these methods were not designed to reproduce other thermodynamic variables of the system (e.g., ΔH, ΔS, and ΔC_p), which are the variables that allow the most direct comparisons with experimental observations. An ensemble representation of proteins that can be connected with equilibrium thermodynamic properties could offer many advantages, particularly with regard to interpreting experimental results or using experimental results to guide in the development of the model.

In temporal representations of protein fluctuations, the focus of the calculations is to determine pathways between alternative conformations. In contrast, the goal of an instantaneous representation of the ensemble (Figure 2) is to identify and characterize the most probable states (i.e., the minima in the landscape) in the ensemble. For this reason, in an instantaneous representation of the ensemble, the kinetic properties of the system are less relevant—only the states themselves matter (although see section 4.3). One practical benefit of the ensemble representation of proteins is that they avoid the computational burden of calculating energy barriers between minima; computational resources can be focused instead on the generation and thermodynamic characterization of a large number of conformational microstates.

The instantaneous representation of the ensemble described ahead is fundamentally a structural—thermodynamic method. Therefore it is well suited for calculation of thermodynamic observables, making it particularly amenable to testing directly against experimental comparisons. In this ensemble-based method, the problem of characterizing the energy landscape of a protein is reduced to determining the structure and energy of the low-energy states and their sensitivity to

various environmental variables. In principle, these are the states that will be populated in solution and the ones that will govern the solution and functional properties of proteins.

2.2.2. Energetics of the Ensemble

Once a particular microstate is identified in an ensemble representation of the energy landscape, it is possible to calculate the energetic contribution of that state to the overall properties of the ensemble. For each microstate, the statistical weight can be expressed as

$$K_i = e^{-\Delta G_i/(RT)} \quad (1)$$

where R is the gas constant, T is absolute temperature, and ΔG_i is the Gibbs free energy of state i. ΔG_i can be further divided into the component enthalpy (ΔH_i), entropy (ΔS_i), and heat capacity (ΔC_{p_i}) contributions. The assumption of a temperature-independent ΔC_{p_i} and use of a reference temperature (T_{ref}) leads to the familiar Gibbs−Helmholtz expression:

$$\Delta G_i(T) = \Delta H_i(T_{ref}) - T\Delta S_i(T_{ref}) + \Delta C_{p_i}[(T - T_{ref}) - T \ln(T/T_{ref})] \quad (2)$$

The importance of eq 1 is that the sum of the statistical weights of all N microstates in the ensemble corresponds to the partition function:

$$Q = \sum_{i=1}^{N_{states}} K_i \quad (3)$$

from which all important thermodynamic quantities, in particular the probability of each state, can be determined:

$$P_i = \frac{K_i}{Q} \quad (4)$$

Equations 1−4 show that the rigorous, formal description of energies in the context of the ensemble representation is straightforward. The challenge of this approach is 2-fold. First, it is necessary to develop a sampling scheme that sufficiently represents the real protein ensemble. Second, it is necessary to identify an accurate energy function that enables correlation between energetic and structural information. The representation of the conformational landscape for a hypothetical ensemble in more quantitative terms in Figure 3 illustrates the scope of these challenges (i.e., in terms of the fraction of the residues in each state that maintain native-like geometry, see Figure 3).

The hypothetical ensemble representation shown in Figure 3 raises a number of questions and issues. For instance, what are the structures of the segments of protein with non-native conformations in the microstates that have a large fraction of residues in the native geometry (region III)? In other words, what is the nature of the conformational fluctuations around the native structure? States with a small fraction of the residues in the native geometry (region I) can be viewed as consisting of two groups: (i) extended conformations and (ii) compact non-native conformations. Are compact non-native structures required to model the ensemble or are these rare, high-energy states of no functional relevance whatsoever? For the states with intermediate degrees of native contacts (region II), what is the relative energy of these states? Does a model that adequately captures the relative

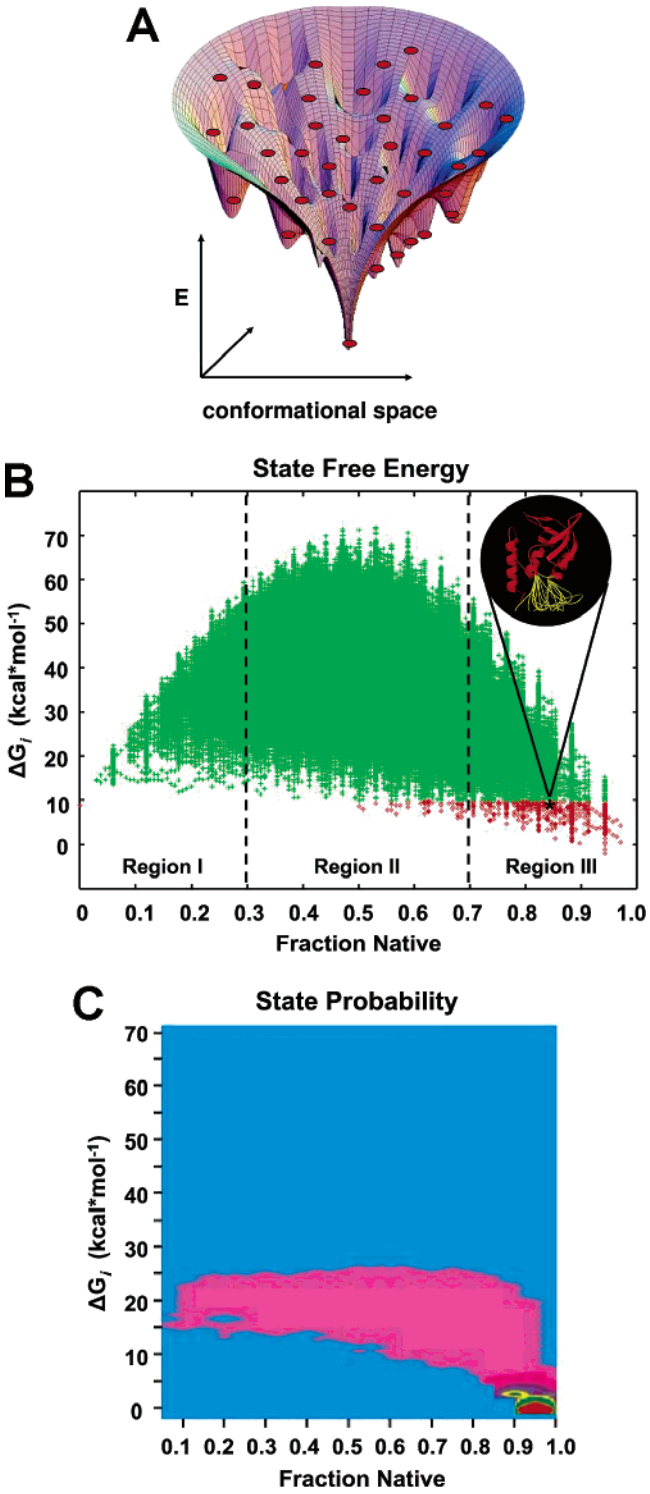


Figure 3. Representations of the ensemble for a hypothetical protein depicted schematically as a landscape (A), as fraction native contacts vs free energy (B), and as fraction native contacts vs probability (C). In Panels A and B, each of the points represents a state that is color-coded on the basis of its energy; green states are high-energy, and red states are low-energy, with the distinction being largely context-dependent. Also shown (B) is a ribbon representation of one state in the ensemble with ~80% native contacts (red corresponds to regions of native structure and yellow corresponds to regions of non-native structure). Panel C shows the probability distribution (determined from the free energies in panel B) indicating that for a specific set of conditions, only a subset of states attain a significant probability. To achieve high probability for other states in the ensemble, the system must be perturbed (e.g., by T, pH, denaturants, osmolytes, ligands, mutation, etc.) in a manner that preferentially stabilizes other states.

energies of native-like (region III) and denatured-like (region I) states also predict high energies for states with intermediate degrees of native structure (region II)?

2.2.3. Response of the Ensemble to Perturbations

Perhaps the most attractive feature of an ensemble representation of a protein is that it can provide the opportunity to quantitatively determine the statistical weight of each state under a variety of different environmental conditions. The partition function and the probabilities of the microstates are all dependent on environmental factors and conditions. For this reason, it is possible to directly access a variety of fundamental features of the protein ensemble and of protein function. The dependence of the probabilities of microstates on environmental conditions also enables direct experimental tests of an ensemble approach. For example, consider the case where the effects of pH and denaturant are included. The partition function becomes

$$Q(T,[\mathrm{den}],\mathrm{pH}) = \sum_{i=1}^{N_{\mathrm{states}}} K_i \prod_{j=1}^{m_{\mathrm{sites}}} (1 + K_{\mathrm{a},j,\mathrm{den}}[\mathrm{den}]) \prod_{j=1}^{m_{\mathrm{sites}}} (1 + K_{\mathrm{a},j,\mathrm{H}^+}[\mathrm{H}^+]) \quad (5)$$

In this equation, K_i refers to the intrinsic stability, and $K_{\mathrm{a},j,\mathrm{den}}$ and $K_{\mathrm{a},j,\mathrm{H}^+}$ are site-specific denaturant and H^+ binding constants. As eq 5 reveals, the probability of each microstate in the ensemble is modulated by changes in the various intensive parameters in the system, and the effects are relatively straightforward to consider in the context of an ensemble representation. In this way, a modeling strategy based on an instantaneous representation of the ensemble can provide an avenue for explicitly considering the dependence of the entire protein ensemble on temperature, pH, denaturant, ligands, mutation, etc.

3. Modeling the Protein Ensemble

3.1. Overview

It is widely acknowledged that the folds of proteins are over-determined; the primary sequence of a protein can be modified extensively without altering the fold. This observation has two important corollaries: (1) most of the microstates that are populated in the ensemble in solution will reflect the consequences of fluctuations about the mean state represented by the high-resolution structure; (2) compact states with a different fold are rarely populated, if at all. It follows from these corollaries that the high-resolution native structures of proteins can be used to enumerate the vast majority of the relevant microstates of an ensemble. It is of considerable interest to determine the stability of the different microstates in the ensemble and to characterize the extent to which native geometry is maintained.

Generating ensembles that represent structural deviations from the canonical structure has been the focus of significant recent efforts. MD methods have been combined with various types of experimental constraints such as hydrogen exchange protection factors and NMR-derived structural and dynamic parameters, to compute theoretical ensembles.[29,30] These methods have been shown to be self-consistent by demonstration that they can successfully recapitulate the parameters used to compute the ensemble. However, it is not yet known whether a structural ensemble trained from one type of experimental data can be used to predict other types of data (i.e., where results of the type that are predicted are not included in the training set). Notwithstanding these uncertainties, hybrid experimental−computational methods of this type will be of significant value in interpreting the origins of experimental difference.

Although it is clear that the conformational fluctuations around the canonical structure of a protein are native-like, the nature of these fluctuations is not known. For instance, are fluctuations well described by one or a few distinct conformational variants? A recent experimental result appears to illuminate the character of conformational fluctuations that exist under native conditions. In recent work by Wand and colleagues,[31] NMR of proteins encapsulated in reverse micelles was used to examine the protein ensemble under conditions that favor cold denaturation. A striking conclusion of those studies is that the protein ensemble appears to be dominated by states that have a *dual structural character*. This means that for a particular state, some regions retain a remarkable degree of native-like character, while other regions are best characterized as occupying *many* alternative conformational states. In other words, these results appear to point toward a model that resembles the local order-to-disorder (or local unfolding) transitions that have been described previously.[32]

The experimental results obtained from these studies raise an important issue that must be addressed if the thermodynamic consequences of these types of states are to be adequately modeled. If the energy is distributed among a large number of conformational states for a particular region (i.e., if the local unfolding model of fluctuations is indeed accurate), how can this behavior be modeled in a computationally tractable and thermodynamically rigorous way? One approach is to avoid describing the unfolded regions of proteins in structural terms. Instead, unfolded regions can be described using strictly thermodynamic terms. This is exactly the strategy employed by the COREX algorithm, which is referred to as a hybrid structural−thermodynamic approach. It is considered a hybrid approach because folded regions are treated in explicit structural terms but unfolded regions are treated only in thermodynamic terms. This hybrid character is described in detail in the following sections.

3.2. A Hybrid Structural–Energetic Approach: The COREX Algorithm

3.2.1. Sampling Conformational Space

Macromolecular equilibria are usually modeled as transitions between fixed macroscopic states. Protein folding, for example, is usually described as a two-state process,[33] wherein the protein fluctuates between two macroscopic states, the native state and the unfolded (or denatured) state. Although this is a reasonable approximation of protein unfolding within the transition region,[33] it ignores other equilibria that exist, for example, under strongly native conditions. Under these conditions, the unfolded state is highly unstable, but the protein undergoes conformational excursions about the high-resolution structure. As noted, these conformational excursions can be detected by hydrogen exchange and NMR relaxation data.[34−54] As the protein is destabilized and the unfolded state probability begins to compete with the probability of the fluctuations seen in hydrogen exchange, the protein behavior becomes more classically "two-state". Thus, an accurate model of the protein equilibrium must be able to account self-consistently for the

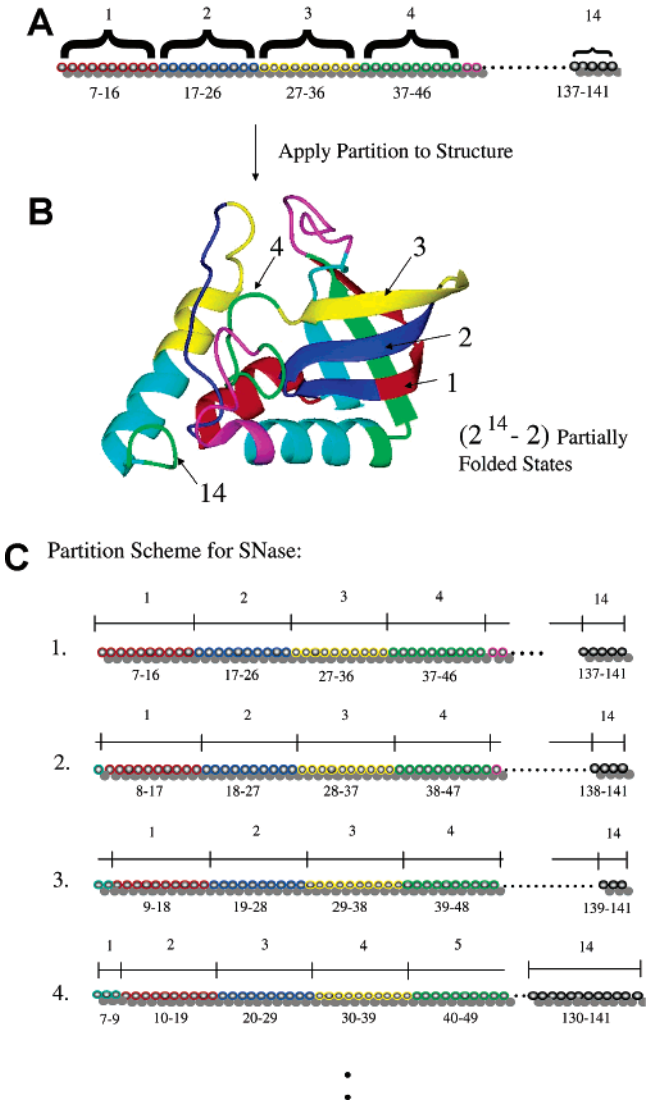


Figure 4. Limited enumeration of the protein ensemble: (A) The linear sequence of the protein is partitioned into folding units. (B) The folding units are applied to the three-dimensional structure, and all possible combinations of "unfolded" and "folded" states of each folding unit are created to define the ensemble. (C) End effects are accommodated by sliding the folding units along the linear sequence.

behavior in each regime, and it should capture the transition between these regimes.

The COREX algorithm was developed to model the native state heterogeneity and to connect it quantitatively to the classic folding/unfolding behavior observed for proteins.[34−35] The COREX approach employs a parametrized Gibbs free energy function (described ahead) to estimate the free energy of each structural microstate (eq 2) alongside a simple sampling scheme to generate structural information for a large number of microstates. To enumerate the microstates of the ensemble, a crystal structure is used as a template onto which a partitioning scheme is applied. Figure 4 shows the partitioning of staphylococcal nuclease (SNase) as an example. In this example, a folding unit window size of 10 residues is employed. To begin the partitioning, the first 10 residues are assigned to the first folding unit, the second 10 are assigned to the second folding unit, and so on. The partitioning is then overlaid onto the high-resolution structure and an ensemble of structures is generated by systematically assigning each folding unit as either fully folded (native) or unfolded. Note that this approach is consistent with the *dual structural character* of the microstates (some local regions being native and others unfolded) that was experimentally observed by Wand and colleagues.[31] For a system with N folding units, the partitioning strategy produces $2^N - 2$ partially native states, representing all possible combinations. To diminish the influence of the location of each partition, the partition boundaries are systematically varied by sliding the folding units one residue at a time in the sequence and repeating the procedure as described.

3.2.2. Modeling Fluctuations

The approach outlined in Figure 4 represents an efficient and systematic means of distinguishing the regions of proteins that are treated in the model as *native*-like regions from the regions that are treated as *non-native*-like. This approach produces an ensemble of states that display dual structural character. The crystal structure can be used to describe the native-like regions. The question then becomes, how should the non-native regions be treated? Should alternative conformations be considered explicitly for each region? If so, how many? To estimate the magnitude of this problem, we need only realize that if 10 residues are to be treated as non-native and each residue has 10 possible conformations, 10^{10} different conformations would have to be considered—an extremely large number to model explicitly. Because fluctuations in multiple regions of the molecule must also be considered, it becomes clear that exhaustive structural enumeration of the unfolded segments of proteins is not a tractable solution.

To avoid the computational intractability of exhaustive enumeration, as well as the approximation of considering only a minute fraction of the relevant states, the COREX approach treats the fluctuations in statistical thermodynamic rather than structural terms. This is a key aspect of this approach. With use of Boltzmann's equation,

$$S = R \ln \Omega \tag{6}$$

where S is the entropy, Ω is the number of conformations, and R is the gas constant, it is possible to estimate the energetic impact of all the conformational variants, provided an estimate of the number of possible conformational variants is known or can be calculated. In the context of this ensemble model, eq 6 corresponds to the conformational entropy (S_{conf}) of the protein. Although this approach does not provide explicit structural details of the alternative conformations, it does address the thermodynamic impact of the **entire** ensemble and is thus rigorous from a statistical thermodynamic standpoint.

3.2.3. Determination of the Energetics of the Ensemble

To determine the relative free energy of each of the structural microstates created by the partitioning scheme, a simple deconstruction of eq 2 is employed.[55] The approach is based on the observation that the enthalpy and the heat capacity of protein unfolding can be related to the difference in solvent-accessible surface between the crystal structure and a hypothetically fully extended unfolded state (see Figure 5).[56,57]

The agreement between the experimental and the calculated energetics for a wide range of proteins suggests that the thermodynamics of the partially folded states also calculated by this approach will provide a reasonable

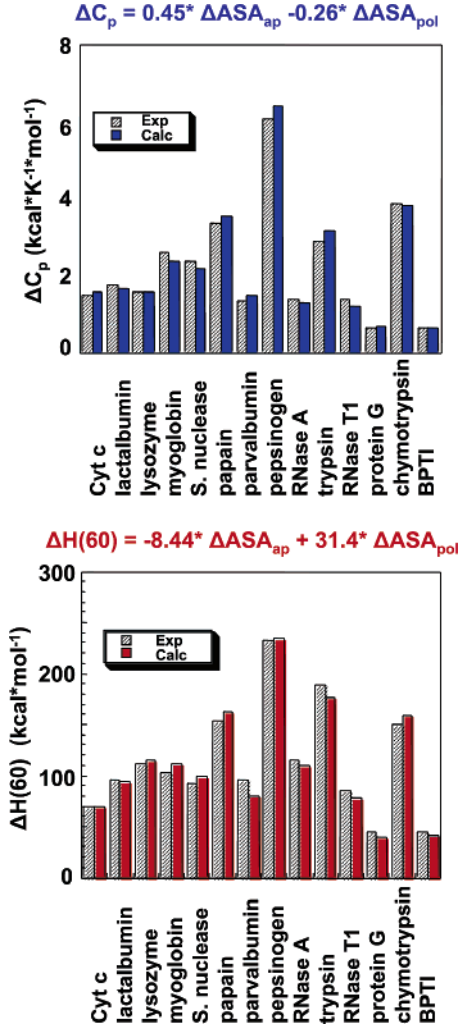


Figure 5. Parametrization of the heat capacity and enthalpy. Hatched bars show the calorimetrically obtained changes in heat capacity and enthalpy of unfolding for a database of proteins of various sizes. Colored bars represent those values for the heat capacity and enthalpy that are calculated using the parametric equations shown at the top of the graphs, which relate each quantity to the changes in solvent-accessible surface area upon unfolding.

approximation of the actual energetics of the ensemble. Most importantly, the energetics can be calculated and compared to experimental results under multiple environmental conditions, thus allowing direct tests of the validity of the ensemble-based model to study solution and functional properties of proteins.

The entropy difference between each state and a reference state can also be calculated for each microstate in the ensemble from parametrized energetics.[56−63] The entropy is divided into two components, the solvent entropy (ΔS_{solv}) and the conformational entropy (ΔS_{conf}):

$$\Delta S_{total} = \Delta S_{solv} + \Delta S_{conf} \quad (7)$$

The ΔS_{conf} represents the *number* of conformational variations that a particular energetic state can occupy. The important feature is that backbone and side chain conformational entropy values for each amino acid have been determined empirically,[59,60] thus providing a means of quantifying, in statistical thermodynamic terms, the conformational variability in the non-native segments of each state in the calculated ensemble shown in Figure 6. Although

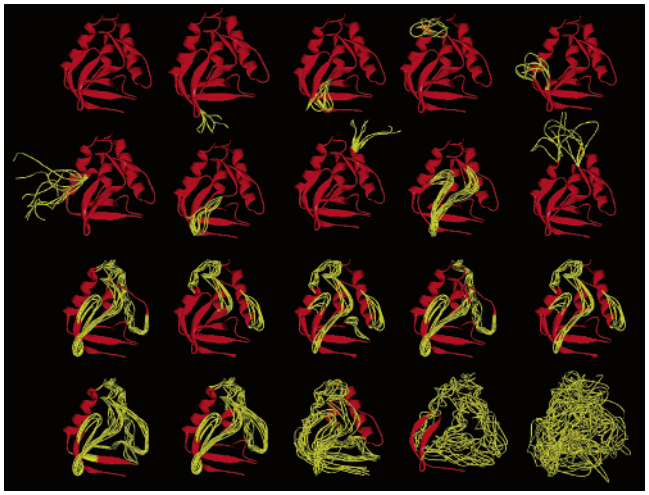


Figure 6. The COREX ensemble. Red regions in each state are portions that are treated as native-like. Yellow regions are treated as denatured-like. However, as indicated in Figure 2, the actual conformations accessible to the fluctuating regions (shown here for schematic purposes only) are not modeled explicitly. The energetic impact of those fluctuations on the free energy of each state is evaluated using eq 1.

clearly a coarse approximation of the conformational space available to a particular protein segment, such an approach provides an efficient and systematic alternative to exhaustive enumeration.

Similar to the case of ΔH and ΔC_p, the solvation entropy, ΔS_{solv}, is determined from changes in solvent-accessible surface area and is calculated from the apolar and polar heat capacity contributions shown in Figure 5:

$$\Delta S_{solv,tot}(T) = \Delta C_{p_{apol}} \ln(T/385) - \Delta C_{p_{pol}} \ln(T/335) \quad (8)$$

Although the surface-area-based energy function described here is coarse and provides little insight into an atomic level understanding of the origins of the energies of each state, it is nonetheless useful and accurate. The accuracy is a direct consequence of the fact that the energy function is parametrized from the thermodynamics of unfolding of real proteins measured calorimetrically (Figure 5). According to the experimental data, the changes in the extent of polar and apolar surface area is a valid metric of the expected thermodynamics of folding/unfolding, regardless of the mechanistic origins of the experimental values. In as much as the unfolding of subdomains or isolated regions of proteins is thermodynamically similar to the unfolding of an entire protein, the surface-area-based parametrization should provide very reasonable estimates of the energetics of each state.

3.2.4. Summary of COREX Capabilities

In applications of the COREX algorithm, more than 10^6 structural−thermodynamic microstates are routinely studied (Figure 4). Because the fluctuating regions in each microstate are treated as local folding/unfolding transitions, these 10^6 structural microstates account thermodynamically for more than 10^{100} conformations. Thus, the modeling strategy utilized by the COREX method represents an efficient and computationally tractable means of (i) accounting for states with a high degree of structural dissimilarity, (ii) determining the energies of each state based on parametrized energetics, and

(iii) estimating the impact of mutations and environmental perturbations on the energies of each state. Although the COREX approach has a number of fundamental limitations, which are discussed below, a significant benefit is that the approach generates and determines the energetics for an ensemble of states that spans the entire range of structure, from the completely folded native structure to the completely unfolded state. This provides a unique opportunity to unify quantitatively the description of unfolding with the description of fluctuations.

4. Use of COREX Ensemble to Understand Solution Properties of Proteins and Biological Function

The benefit of an equilibrium ensemble model is twofold. First, the effects of perturbations on each state can be calculated in a straightforward fashion, as described, for example, by eq 5. Second, observed equilibrium properties of the system, ⟨obs⟩, can be calculated as the probability-weighted contribution of the component states in the ensemble:

$$\langle \mathrm{obs} \rangle(T,[\mathrm{lig}],\mathrm{pH}) = \sum_{i=1}^{N_{\mathrm{states}}} \mathrm{obs}_i * P_i(T,[\mathrm{lig}],\mathrm{pH}) \qquad (9)$$

Equation 9 shows that the overall observed properties of an ensemble (e.g., the fluorescence, proton binding, CD, activity, etc.) are a consequence of the probability of each state and the sensitivity of each state to the environmental conditions. Thus, if the effects of these environmental perturbations are known for each state, the response of the ensemble can be calculated. To assess the validity of the COREX-derived ensemble, several types of perturbations have been studied.

4.1. The Response of the COREX Ensemble to Perturbations

Recently, we presented two examples of how COREX can be used to monitor the effects of an environmental perturbation on the protein ensemble.[31,64] The first study focused on the effects of pH on the stability of staphylococcal nuclease (SNase).[64] In that study, the effects of protons on the distribution of microstates in the ensemble were determined following a simple rule set: (i) ionizable groups in native regions were assigned the pK_a value calculated using standard continuum electrostatics models; (ii) ionizable groups in non-native regions or structurally adjacent to non-native regions were assigned the pK_a values of model compounds in water (i.e., unfolded state pK_a values). As a result of this rule set, each microstate in the ensemble had a unique titration behavior, and the overall titration behavior of the ensemble followed eq 9.

The performance of this simple model is noteworthy. As shown in Figure 7, the redistribution of the ensemble of SNase conformations in response to proton binding results in an apparent cooperative pH-induced unfolding at pH $\approx$ 3.8, which is in very good agreement with the experimentally observed pH midpoint of unfolding obtained by monitoring intrinsic fluorescence. In addition, the experimentally observed cooperativity of the pH-induced transition ($\partial \ln K/\partial \ln [\mathrm{H}^+] = \Delta\nu = 4.7$) is in excellent agreement with the value of 4.8 determined directly from the calculated ensemble. Furthermore, the COREX calculations were able to identify the ionizable groups that are responsible for the acid unfolding properties of the protein. This reasonable agreement between the COREX calculations and the experiment suggests that the assumption of "local unfolding" for the non-native regions inherent to the COREX model is a reasonable thermodynamic treatment. Similarly, the ability of this simple model to reproduce the acid unfolding behavior suggests that the energy function reproduces the Gibbs energies of the microstates nearly quantitatively.

In another recent study, the effect of temperature on the ubiquitin ensemble was determined and compared to the experimentally measured cold denaturation.[31] The calculations qualitatively reproduced the cooperative heat denaturation, the non-cooperative cold denaturation, and the location of the residual structure that persisted after the cold denaturation transition. These two studies directly validate the concept that the component microstates in the native state ensemble embodied in the COREX model have dual structural character. In other words, it is thermodynamically valid to invoke the concept of local unfolding to treat the non-native segments of proteins.

4.2. Cooperativity and Long-Range Communication in Proteins

The fact that COREX calculations can reproduce a disparate range of phenomena (at least qualitatively) suggests that the physical basis for the behavior can be established by interrogating the ensemble and identifying the dominant (i.e., most probable) states. The overall response of the system will depend on the nature of the perturbation and on the response of each state. More importantly, coarse representations of the ensemble are particularly well suited for identifying general phenomena. This is perhaps best demonstrated by considering how COREX was able to provide unique insight into several key aspects of cooperativity and communication in proteins.[65−67] A recent example of this involves the use of COREX to define the energetic connectivities between the different structural elements of dihydrofolate reductase (DHFR) from *Escherichia coli*.[67] Analysis of this protein has allowed us to describe two important aspects of intramolecular communication (see Figure 8). First, within the protein, pairwise couplings exist that define the magnitude and extent to which mutational effects propagate from the point of origin. Second, in addition to the pairwise energetic coupling between residues, functional connectivity exists; it is apparent in terms of energetic coupling between entire functional elements (i.e., binding sites) and the rest of the protein. Analysis of the energetic couplings provides access to the thermodynamic domain structure in DHFR, as well as the susceptibility of the different regions of the protein to both small-scale (e.g., point mutations) and large-scale perturbations (e.g., binding ligand). The results of that analysis point toward a view of allosterism and signal transduction wherein perturbations do not necessarily propagate through structure via a series of conformational distortions that extend from one active site to another. Instead, the observed behavior is a manifestation of the distribution of states in the ensemble and of how the distribution is affected by a perturbation such as mutations or ligand binding. This result and other similar ones[31,64−68] reveal that the coarse-level representation of the ensemble embodied in COREX provides unique and significant insight about the physical basis for a wide range of complex and functionally significant properties of proteins.

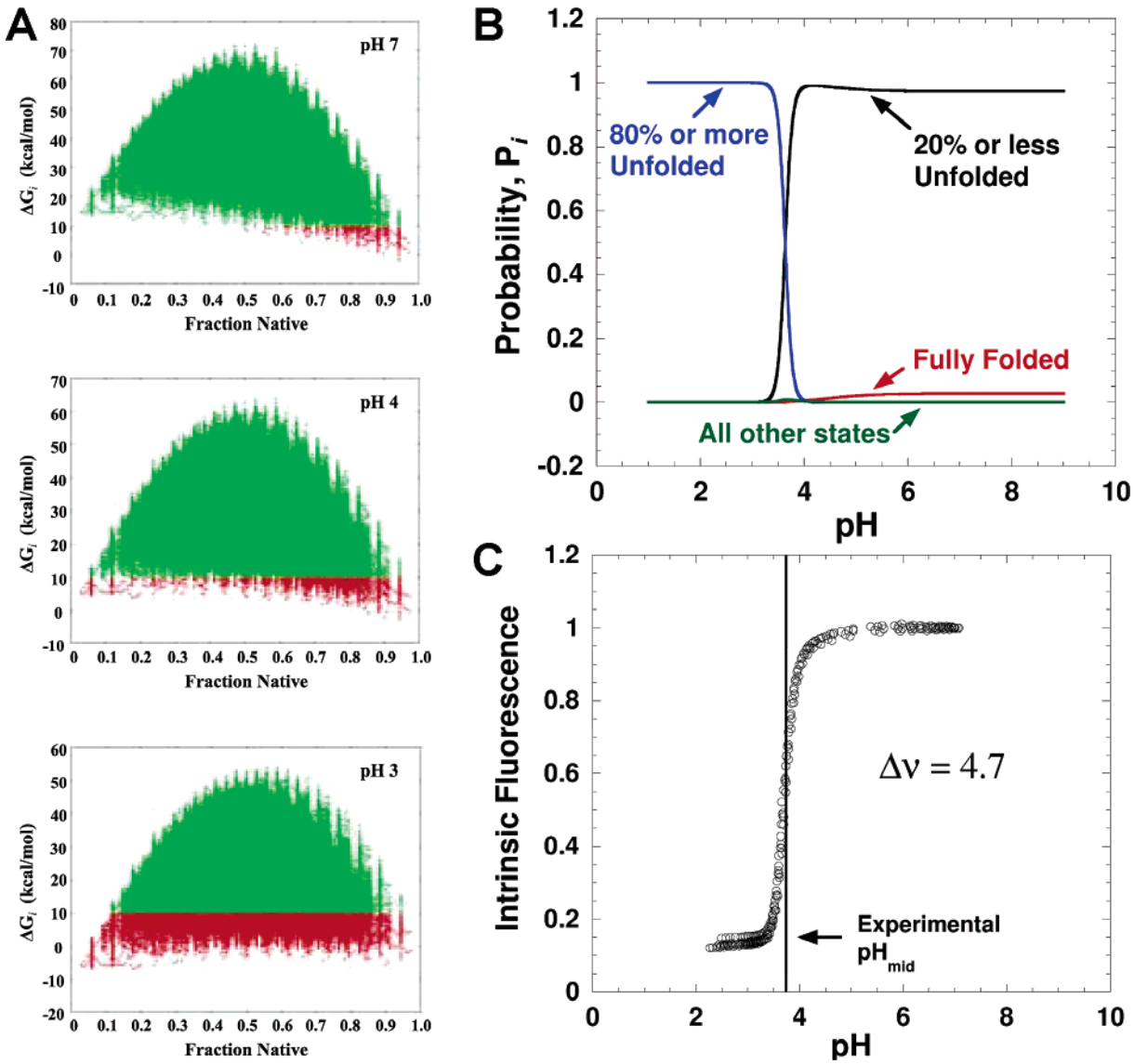


Figure 7. Actual calculated ensembles (ΔG vs fraction native) for staphylococcal nuclease (SNase) at (A) pH 7, pH 4, and pH 3, (B) probability of the different fraction native populations as a function of pH, and (C) experimental pH dependence of the fluorescence of SNase at 25 °C.

4.3. Kinetics of the Ensemble: Modeling Protein Folding Mechanisms

Although ensembles are used to represent the conformational properties of proteins at equilibrium, this does not preclude their use to describe the kinetics of interconversion among ensemble substates. The interface between statistical thermodynamic and kinetic models of protein folding has been elegantly presented by Zwanzig.[69] Zwanzig has shown that the classical two-state kinetic model of folding is compatible with an ensemble view of proteins. It should be noted that many proteins sample highly unfolded members of their ensemble many times per second or faster via an apparently two-state kinetic mechanism. The classic two-state rate equation is

$$\frac{dP_N}{dt} = k_f P_U - k_u P_N$$
$$P_N + P_U = 1 \qquad (10)$$

where P_N and P_U are the fractional populations of the native and unfolded subensembles and k_f and k_u represent the rate constants for folding and unfolding. Following the argument put forth by Zwanzig, the partitioning of the general protein ensemble into two (or more) subensembles is not based on a thermodynamic distinction. Rather, a particular conformation belongs to whichever subensemble contains other conformations with which it interconverts more rapidly than the interconversion between subensembles. Although Zwanzig demonstrated that such a scenario can satisfy eq 10 for an apparently two-state system, the analysis does not preclude more than two subensembles, which may or may not appear to be kinetically two-state. Subensembles corresponding to high-energy intermediates may be kinetically undetectable but may still interconvert more slowly than do the conformations within each subensemble. The kinetics of interconversion between subensembles (or equivalently, states) can be described with the following ordinary differential equation:

$$\frac{dP_a}{dt} = \sum_b k(b \rightarrow a) P_b - \sum_b k(a \rightarrow b) P_a \qquad (11)$$

which describes the rate of change of the population of any state a in terms of the sum of all fluxes into ($k(b \rightarrow a)P_b$) or

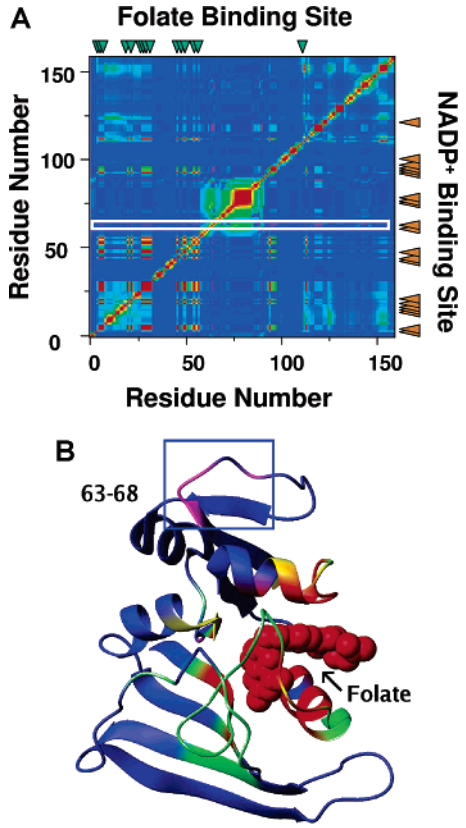


Figure 8. Analysis of cooperativity in the protein ensemble and its relationship to ligand binding. Panel A shows the calculated energetic connectivity of each residue in DHFR to the folate binding site (determined by the correlation between the probability that any specific residue is folded or unfolded and the probability that the folate binding site residues are also folded). Red corresponds to a large positive energetic connectivity, blue to the smallest, and purple to negative energetic connectivity. Panel B shows the high-resolution structure of DHFR with the van der Waals' surface of folate in its binding site. The structure is color coded according to the magnitude of the energetic connectivities in panel A. The negative connectivities between the folate site and residues 63−68 are highlighted. Note the absence of a propagation pathway from the folate site. Reprinted with permission from ref 67. Copyright 2000 National Academy of Sciences, U.S.A.

out of $-(k(\mathrm{a}\rightarrow\mathrm{b})P_\mathrm{a})$ state a, where k represents the rate constant for the transition and P represents the time-dependent instantaneous population. The rate constants must satisfy detailed balance so that at equilibrium the flux from state a to state b equals the flux from state b to a:

$$P_\mathrm{a}(\mathrm{eq})k(\mathrm{a}\rightarrow\mathrm{b}) = P_\mathrm{b}(\mathrm{eq})k(\mathrm{b}\rightarrow\mathrm{a}) \qquad (12)$$

A description of the interconversion between all states at equilibrium requires the calculation of the flux between each state. Some states may not be directly accessible from a given state, and the rate constants and fluxes between such states would be zero.

Several theories of protein folding seek to describe the folding mechanism as a series of kinetic steps involving the interconversion of discrete states along the folding pathway. These discrete states can be defined as subensembles of the complete COREX ensemble using the definition of a subensemble given above. The B domain of protein A (BdpA) is a 58 residue three-helix bundle protein whose COREX ensemble can be subdivided into five kinetically distinguishable subensembles (see below). According to the

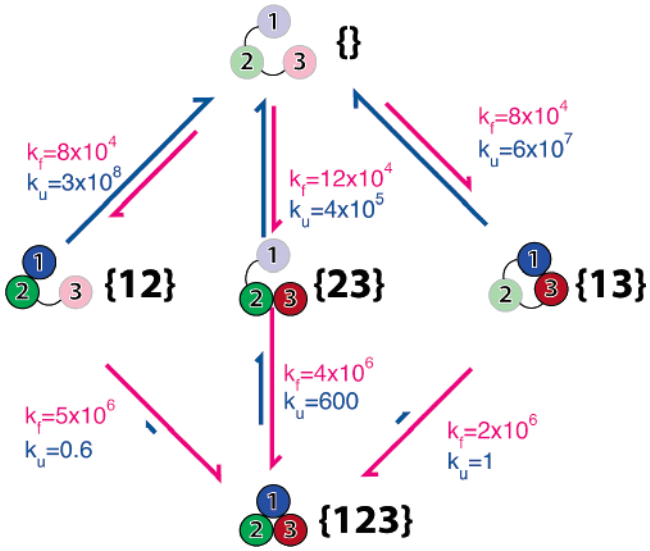


Figure 9. Calculated folding/unfolding mechanism of the B domain of protein A, based on the diffusion−collision model. The five substates are labeled by the numbers of the helices that form native interfaces in all members of the associated ensembles. The folding rate constants (magenta) were calculated using standard diffusion−collision theory. The reverse rate constants (cyan) are calculated with eq 12 and populations listed in Table 1. All rate constants are given in units of s^{-1}. The length of the arrows is proportional to $\ln(k)$.

diffusion−collision theory of Karplus and Weaver,[70] the interconversion between these subensembles involves the formation or dissociation of an interhelical interface, as depicted in the scheme of Figure 9. A similar theory has recently been proposed by Dill and co-workers.[71] Implicit in both theories is the assumption that conformations within a subensemble interconvert much more rapidly than do the subensembles. For example, the BdpA {12} subensemble with a native interface between helix 1 and 2 would include conformations that have frayed helix 1 or 2 but zipping up of the frayed helix would be very fast relative to either the dissociation of helix 1 and 2 (a transition to the {} ensemble) or the docking of helix 3 (a transition to {123}). Likewise, the {12} subensemble would include conformations with some residues from helix 3 that are helical but not docked with helix 1 or 2. The flickering formation and decay of such short helical segments would be much faster (10−50 ns) than the formation and docking of a helix of sufficient length (~60%) to cause a transition to the {123} ensemble. Note that to represent such helix 3 undocked conformations, the COREX conformational sampling algorithm had to be modified to allow native secondary structure without native tertiary interactions. In this case, these conformations can be represented by a structure in which helix 3 is translated away from helices 1 and 2 in the native structure.

When this logic is followed, it is possible to computationally inspect each member of a complete COREX ensemble and assign it to a kinetically defined subensemble. For BdpA, these subensembles are depicted in Figure 10. The fractional population of each subensemble i is

$$P_i = \frac{Q_i}{Q_\mathrm{total}} \qquad (13)$$

where Q_i is the partition function for subensemble i and Q_total is the partition function for the complete ensemble. Both partition functions are calculated using eq 3. With this

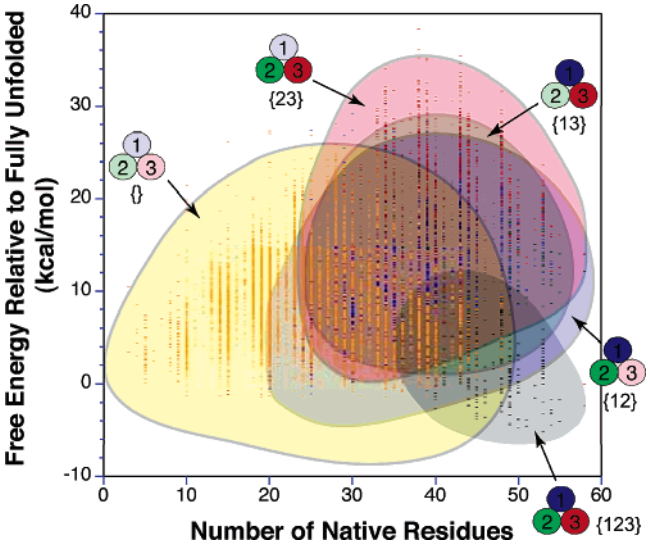


Figure 10. COREX BdpA subensembles constituting the five substates depicted in Figure 9. All free energies are relative to the fully denatured protein. Each conformation is categorized as described in the text.

Table 1. Equilibrium Populations of the Five Subensembles of BdpA Depicted in Figure 9

state	number of conformations	fractional population
{}	101224	4.3×10^{-4}
{12}	320	5.7×10^{-8}
{23}	276	1.8×10^{-4}
{13}	268	5.8×10^{-7}
{123}	312	0.9994

approach, the five subensembles of BdpA have the equilibrium populations given in Table 1.

The rate constants for the formation of each subensemble from its less folded precursors can be calculated using a variety of approaches, including that of Dill and co-workers,[71] that of Munoz and Eaton,[72] molecular dynamics,[73] or diffusion−collision theory.[70] The latter approach has been used to calculate the forward rate constants depicted in Figure 9. The reverse rate constants depicted are calculated using the equilibrium populations from Table 1 and eq 12.

The rate constants in Figure 9 combined with the equilibrium populations in Table 1 can be used to calculate the equilibrium fluxes between subensembles. The fluxes for BdpA are depicted in Figure 11. The flux along a sequential pathway is the inverse of the sum of the inverse fluxes (by analogy to the combination of capacitances in an electrical circuit). Thus, the flux along a path is dominated by the smallest value, as expected for a rate-limiting step. Even when two sequential steps have identical fluxes, the total flux through both steps is half the individual fluxes because half of the molecules in the intermediate between two steps return by the route by which they came. The aggregate flux through parallel pathways is the simple sum of the fluxes through each pathway. In the case of the fluxes depicted in Figure 11, the aggregate flux is 5.4%/ms. Note that this flux is very high and indicates that under native conditions a molecule of BdpA will sample the fully unfolded state every ~20 ms. This conclusion is supported by experimentally measured folding and unfolding rates.[74] With this number, it is possible to calculate the fractional flux through each parallel pathway, as listed in Table 2.

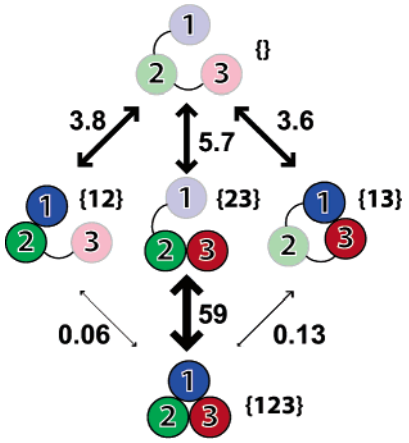


Figure 11. Calculated BdpA folding/unfolding fluxes at equilibrium based on the folding mechanism depicted in Figure 9. All fluxes are given in % ms^{-1}. Arrow thickness is proportional to the logarithm of the flux.

Table 2. Fluxes through the BdpA Folding Pathways

path	flux (%/ms)	fractional flux (%)
{} ↔ {12} ↔ {123}	0.06	1.1
{} ↔ {23} ↔ {123}	5.2	96.7
{} ↔ {13} ↔ {123}	0.12	2.2

It is worth noting that the middle pathway in Figures 9 and 11 is predominant not because the forward rate constant in the first step is significantly faster than that of the other two pathways but rather because the population of the {23} intermediate, as predicted by COREX is significantly higher than those of {12} or {13}. This higher population reduces the reverse rate constant and makes the forward reaction more probable, thereby dramatically increasing the flux through the {23} ↔ {123} step. The high flux of this step corresponds to a highly frequent sampling (59%/ms) of the {23} ensemble from the fully folded ensemble, indicating a highly labile interface between helix 1 and the rest of the protein. In any case, models developed as described here emphasize the importance of folding/unfolding reactions even under conditions that strongly favor the native state.

Thus, although the COREX algorithm was originally developed as a purely thermodynamic tool, it can also be used to describe the dynamics of a protein folding reaction at equilibrium. This provides a unique opportunity to connect the dynamic information with other seemingly disparate equilibrium processes, as noted above. These types of connections, in turn, provide unparalleled opportunities to elucidate the determinants of the observed thermodynamic and kinetic processes, and most importantly, they provide a more complete framework for experimentally challenging the predictions.

5. Concluding Remarks

The experimental tests of the COREX algorithm suggest that despite its inherent simplicity and approximations, the model reproduces many seemingly disparate solution properties of proteins rather well. We interpret this as evidence that the biophysical and functional behavior of proteins that we have rationalized on the basis of an ensemble-based model are governed by robust features of the ensemble. This is a significant result because it de-emphasizes the importance of one or a small number of specific interactions in

determining such seemingly complex behavior as allosterism. To the contrary, these results highlight the importance of the protein as a whole. What determines the properties of a protein, such as whether a perturbation at one site (i.e., binding) will affect another site, is the energetic hierarchy of states in the distribution (i.e., which states are most stable) and how each state in the distribution is affected by the perturbation.[64,67,75] In principle, the distribution of states that is sufficient to facilitate the selected behavior of each protein can be achieved through a variety of amino acid sequences. The fact that the COREX algorithm can capture this behavior indicates that the interactions themselves are not as important as the resultant hierarchy that the interactions help to preserve.[64,67] Although this result seems at first glance paradoxical, it is entirely consistent with the observations that homologous proteins with low sequence identity have identical functions. Indeed, it is noteworthy that such a robust "coding" of biological function, wherein a particular energetic hierarchy can be facilitated through a highly degenerate sequence space, is an ideal evolutionary strategy for maintaining function while at the same time allowing for selective improvement and diversification through random mutagenesis.

The ensemble-based model described in this review could be improved and extended in several ways. In section 4.3 above, for example, we have shown how the manner in which the microstates are counted could be modified to suit a specific application. Similarly, in some cases it might be necessary to account for contributions made by alternative folds to the properties of the ensemble. The parametrized energy function could also be fine-tuned, for example, by applying even more rigorous approaches to the manner in which electrostatic effects, described in section 4.1, are computed. A benefit of the relatively simple architecture of the COREX model is that these extensions and modifications of the ensemble-based approach are relatively straightforward and can be guided by direct comparisons with experimental data.[76]

6. Acknowledgment

This work was supported by NIH Grant GM63747, NSF Grant MCB9875689, and Welch Foundation Grant H-1461 to V.J.H. and NIH Grant GM061597 and NSF Grant MCB0212414 to B.G.M.E.

7. References

(1) Englander, S. W. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 213.
(2) Bai, Y.; Sosnick, T. R.; Mayne, L.; Englander, S. W. *Science* **1995**, *269*, 192.
(3) Swint-Kruse, L.; Robertson, A. D. *Biochemistry* **1996**, *35*, 171.
(4) Hvidt, A.; Nielsen, S. O. *Adv. Protein Chem.* **1966**, *21*, 287.
(5) Kim, K.-S.; Fuchs, J. A.; Woodward, C. K. *Biochemistry* **1993**, *32*, 9600.
(6) Chamberlain, A. K.; Handel. T.; Marqusee, S. *Nat. Struct. Biol.* **1996**, *3*, 782.
(7) Radford, S. E.; Buck, M.; Topping, K. D.; Dobson, C. M.; Evans, P. A. *Proteins* **1992**, *14*, 237.
(8) Fuentes, E. J.; Wand, A. J. *Biochemistry* **1998**, *37*, 3687.
(9) Fuentes, E. J.; Wand, A. J. *Biochemistry* **1998**, *37*, 9877.
(10) Itzhaki, L. S.; Neira, J. L.; Fersht, A. R. *J. Mol. Biol.* **1997**, *270*, 89.
(11) Feng, H.; Takei, J.; Lipsitz, R.; Tjandra, N.; Bai, Y. *Biochemistry* **2003**, *42*, 12461.
(12) Brooks, C. L., III; Onuchic, J. N.; Wales, D. J. *Science* **2001**, *293*, 612.
(13) Ren, J.; Kachel, K.; Kim, H.; Malenbaum, S. E.; Collier, J. R.; London, E. *Science* **1999**, *284*, 955.
(14) Hogle, J. M. *Annu. Rev. Microbiol.* **2002**, *56*, 677.
(15) Bullough, P. A.; Hughson, F. M.; Skehel, J. J.; Wiley, D. C. *Nature* **1994**, *371*, 37.
(16) Baker, D.; Agard, D. A. *Structure* **1994**, *2*, 907.
(17) Gamblin, S. L.; Haire, L. F.; Russell, R. J.; Stevens, D. J.; Xiao, B.; Ha, Y.; Vasisht, N.; Steinhauer, D. A.; Daniels, R. S.; Elliot, A.; Wiley, D. C.; Skehel, J. J. *Science* **2004**, *303*, 1838.
(18) Sugita, Y.; Okamoto, Y. *Chem. Phys. Lett.* **1999**, *314*, 141.
(19) Hansmann, U. *Chem. Phys. Lett.* **1997**, *281*, 140.
(20) Hukushima, K.; Nemoto, K. *J. Phys. Soc. Jpn.* **1996**, *65*, 1604.
(21) Paschek, D.; Garcia, A. E. *Phys. Rev. Lett.* **2004**, *93*, 238105.
(22) Cheung, M. S.; Chavez, L. L.; Onuchic, J. S. *Polymer* **2004**, *45*, 547.
(23) Cheung, M. S.; Garcia, A. E.; Onuchic, J. S. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 685.
(24) Yang, S.; Cho, S. S.; Levy, Y.; Cheung, M. S.; Levine, H.; Wolynes, P. G.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 13786.
(25) Levy, Y.; Wolynes, P. G.; Onuchic, J. N. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *101*, 511.
(26) Miyashita, O.; Onuchic, J. N.; Wolynes, P. G. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 12570.
(27) Atligan, A.; Durell, S. R.; Jernigan, R. L. Demirel, M. C.; Keskin, O.; Bahar, I. *Biophys. J.* **2001**, *80*, 505.
(28) Keskin, O.; Jernigan, R. L.; Bahar, I. *Biophys. J.* **2000**, *78*, 2093.
(29) Vendruscolo, M.; Paci, E.; Karplus, M.; Dobson, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14817.
(30) Vendruscolo, M.; Paci, E.; Dobson, C. M.; Karplus, M. *J. Am. Chem. Soc.* **2003**, *125*, 15686.
(31) Babu, C. R.; Hilser, V. J.; Wand, A. J. *Nat. Struct. Mol. Biol.* **2004**, *11*, 352.
(32) Miller, D. W.; Dill, K. A. *Protein Sci.* **1995**, *4*, 1860.
(33) Lumry, R.; Biltonen, R.; Brandts, J. F. *Biopolymers* **1966**, *4*, 917.
(34) Hilser, V. J.; Freire, E. *J. Mol. Biol.* **1996**, *262*, 756.
(35) Hilser, V. J.; Freire, E. *Proteins* **1997**, *27*, 171.
(36) Itzhaki, L. S.; Neira, J. L.; Fersht, A. R. *J. Mol. Biol.* **1997**, *270*, 89.
(37) Mayo, S. L.; Baldwin, R. L. *Science* **1993**, *262*, 873.
(38) Woodward, C. K.; Rosenberg, A. *J. Biol. Chem.* **1971**, *246*, 4114.
(39) Idiyatullin, D.; Nesmelova, I.; Daragan, V. A.; Mayo, K. H. *J. Mol. Biol.* **2003**, *325*, 149.
(40) Alexandrescu, A. T.; Rathgeb-Szabo, K.; Rumpel, K.; Jahnke, W.; Schulthess, T.; Kammerer, R. A. *Protein Sci.* **1998**, *7*, 389.
(41) Clore, G. M.; Driscoll, P. C.; Wingfield, P. T.; Gronenborn, A. M. *Biochemistry* **1990**, *29*, 7387.
(42) Crump, M. P.; Spyracopoulos, L.; Lavigne, P.; Kim, K. S.; Clark-Lewis, I.; Sykes, B. D. *Protein Sci.* **1999**, *8*, 2041.
(43) Farrow, N. A.; Muhandiram, R.; Singer, A. U.; Pascal, S. M.; Kay, C. M.; Gish, G.; Shoelson, S. E.; Pawson, T.; Forman-Kay, J. D.; Kay, L. E. *Biochemistry* **1994**, *33*, 5984.
(44) Feher, V. A.; Cavanagh, J. *Nature* **1999**, *400*, 289.
(45) Gagné, S. M.; Tsuda, S.; Spyracopoulos, L.; Kay, L. E.; Sykes, B. D. *J. Mol. Biol.* **1998**, *278*, 667.
(46) Kay, L. E.; Torchia, D. A.; Bax, A. *Biochemistry* **1989**, *28*, 8972.
(47) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559.
(48) Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4546.
(49) Mandel, A. M.; Akke, M.; Palmer, A. G. *J. Mol. Biol.* **1995**, *246*, 144.
(50) Wagner, G. *Q. Rev. Biophys.* **1983**, *16*, 1.
(51) Wagner, G. *Struct. Biol.* **1995**, *2*, 255.
(52) Yang, D. W.; Kay, L. E. *J. Mol. Biol.* **1996**, *263*, 369.
(53) Ye, J.; Mayer, K. L.; Stone, M. J. *J. Biomol. NMR* **1999**, *15*, 115.
(54) Zidek, L.; Novotny, M. V.; Stone, M. J. *Nat. Struct. Biol.* **1999**, *6*, 1118.
(55) Freire, E. *Methods Mol. Biol.* **2001**, *168*, 37.
(56) Hilser, V. J.; Gomez, J.; Freire, E. *Proteins: Struct., Funct., Genet.* **1996**, *26*, 123.
(57) Gomez, J.; Hilser, V. J.; Xie, D.; Freire, E. *Proteins: Struct., Funct., Genet.* **1995**, *22*, 404.
(58) Luque, I.; Mayorga, O. L.; Freire, E. *Biochemistry* **1996**, *35*, 13861.
(59) D'Aquino, J. A.; Gomez, J.; Hilser, V. J.; Lee, K. H.; Amzel, L. M.; Freire, E. *Proteins* **1996**, *25*, 143.
(60) Lee, K. H.; Xie, D.; Freire, E.; Amzel, L. M. *Proteins* **1994**, *20*, 68.
(61) Murphy, K. P.; Xie, D.; Thompson, K. S.; Amzel, L. M.; Freire, E. *Proteins* **1994**, *18*, 63.
(62) Murphy, K. P.; Freire, E. *Adv. Protein Chem.* **1992**, *43*, 313.
(63) Freire, E.; Murphy, K. P. *J. Mol. Biol.* **1991**, *222*, 687.
(64) Whitten, S. T.; García-Moreno E., B.; Hilser, V. J. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 4282.
(65) Hilser, V. J.; Dowdy, D.; Oas, T.; Freire, E. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 9903.
(66) Freire, E. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 10118.
(67) Pan, H.; Lee, J. C.; Hilser, V. J. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 12020.
(68) Wooll, J. O.; Wrabl, J. O.; Hilser, V. J. *J. Mol. Biol.* **2000**, *301*, 247.
(69) Zwanzig, R. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 148.

(70) Karplus, M.; Weaver, D. L. *Nature* **1976**, *260*, 404.
(71) Weikl, T. R.; Dill, K. A. *J. Mol. Biol.* **2003**, *329*, 585.
(72) Munoz, V.; Eaton, W. A. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311.
(73) Duan, Y.; Kollman, P. A. *Science* **1998**, *282*, 740.
(74) Myers, J. K.; Oas, T. G. *Nat. Struct. Biol.* **2001**, *8*, 552.
(75) Luque, I.; Leavitt, S. A.; Freire, E. *Annu. Rev. Biophys. Biomol. Struct.* **2002**, *31*, 235.
(76) Hilser, V. J. *Protein Struct., Stab. Folding. Methods Mol. Biol.* **2001**, 93.

CR040423+