Exploring structures in protein folding funnels with free energy functionals: the denatured ensemble.

ARTICLE *in* JOURNAL OF MOLECULAR BIOLOGY · MAY 1999

Impact Factor: 4.33 · DOI: 10.1006/jmbi.1999.2612 · Source: PubMed

2 AUTHORS, INCLUDING:

Peter Wolynes
Rice University
359 PUBLICATIONS **28,978** CITATIONS

SEE PROFILE

Article No. jmbi.1999.2612 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (1999) 287, 657–674

JMB

Exploring Structures in Protein Folding Funnels with Free Energy Functionals: The Denatured Ensemble

Benjamin A. Shoemaker and Peter G. Wolynes*

School of Chemical Sciences
University of Illinois
Urbana-Champaign, Urbana
IL 61801, USA

We discuss the formulation of free energy functionals that describe the formation of structure in partially folded proteins. These free energy functionals take into account the inhomogeneous nature of contact energies, chain entropy and cooperative contributions reflecting the many body character of some folding forces like hydrophobicity, but do not directly account for non-native contacts because they assume the validity of the minimal frustration principle. We show how the free energy functionals can be used to interpret experiments on partially folded proteins that probe the fractional occupancy of specific local structures. In particular, we study the hydrogen protection factors in lysozyme studied in transient experiments by Gladwin and Evans and by Nash and Jonas using equilibrium pressure denaturation and the NMR order parameters measured by Dobson and Kim for the homologous protein α-lactalbumin.

Keywords: denatured state ensemble; hydrogen exchange protection factor; NMR order parameters; protein structure

**Corresponding author*

Introduction

Even a completely folded protein exists in a myriad of different conformational substates (Frauenfelder *et al.*, 1991). Nevertheless, a single molecular structure like that produced by X-ray diffraction or NMR studies is a reasonable starting point for thinking about a folded protein's function (Perutz, 1992). On the other hand, the diversity of states in an unfolded or partially folded protein is so great that a more frankly statistical description is required even to begin. The energy landscape theory of protein folding explicitly recognizes this fact by using a quasi-thermodynamic ensemble description of the states involved in folding (Wolynes *et al.*, 1995; Abkevich *et al.*, 1996; Chan & Dill, 1997). This theory makes use of the language and techniques provided by the statistical mechanics of disordered and inhomogeneous systems (Bryngelson *et al.*, 1995; Onuchic *et al.*, 1997). Despite the statistical nature of partially folded protein conformations, describing their structure is both inherently fascinating and essential for a complete understanding of folding kinetics.

Recently, we described the structural features in the statistical ensemble of partially folded proteins by using correlation function descriptions and free energy functionals like those used for other inhomogeneous statistical systems such as liquids and magnets (Shoemaker *et al.*, 1997). Here, we explore in greater depth this approach and show how it can be used to interpret experiments on residual structure in the early folding intermediates in lysozyme that have been revealed by hydrogen exchange protection (Miranker *et al.*, 1991; Radford *et al.*, 1992; Gladwin & Evans, 1996) and on the residual structure lactalbumin molten globules as studied by NMR (Schulman *et al.*, 1997). In the accompanying paper (Shoemaker *et al.*, 1999) we use the same free energy functionals to examine the transition ensemble for protein folding kinetics and use this to interpret protein engineering experiments on the reaction kinetics of CI2 and λ-repressor (Itzhaki *et al.*, 1995; Burton *et al.*, 1997).

Determining a complete free energy functional for a folding protein would be a more exacting task than simply predicting a protein's structure from sequence. To avoid this difficulty, the free energy functionals that we investigate and use here are motivated by the idea that the energy landscape of a rapidly folding protein resembles a funnel (Leopold *et al.*, 1992; Bryngelson *et al.*, 1995); that is, the interactions present in the native structure are considerably more favorable than those in alternative structures. In other words, we assume the validity of the minimal frustration prin-

E-mail address of the corresponding author:
wolynes@scs.uiuc.edu

0022-2836/99/130657–18 $30.00/0

ciple (Bryngelson & Wolynes, 1987). This idea allows us to use the native structure as determined by X-ray or NMR as a reference point. Various measures of the similarity of variant protein structures to this native structure can be used as collective reaction coordinates for the folding process because of the funneled nature of the landscape (Bryngelson *et al*., 1995; Onuchic *et al*., 1995, 1996; Socci *et al*., 1996). Specific traps which have been observed in many cases (including the systems we study here, lysozyme and lactalbumin; Kiefhaber, 1995) cannot be completely described by the free energy functionals developed using this idea (Bryngelson & Wolynes, 1989), so the technique is actually most applicable to fast folding proteins. The collective coordinates that we use to describe the folding free energy profile involve the fraction of time any specific pair of residues that interact in the native structure can be found in their near-native relative locations. This is a reasonable set of collective coordinates to use, since tertiary contacts are a dominant source of the stabilization free energy for most proteins.

The free energy functional is constructed by expressing both the mean energy of an ensemble of structures specified by the fractional occupancy of each specific contact and also the entropy of that ensemble using the same generalized coordinates, the contact probabilities. The energetic part of the free energy functional depends then on residue-specific interaction parameters. In our work these energy parameters are taken from statistical potentials (Miyazawa & Jernigan, 1985, 1996) or database potentials (Goldstein *et al*., 1992) obtained using optimization schemes for protein structure prediction. The entropy term involves the entropy loss on forming specific contacts in an already partially structured polymer. We earlier used an expression for this entropy motivated by the Flory-Jacobson-Stockmayer theory of rubber vulcanization (Jacobson & Stockmayer, 1950; Flory, 1956). The same approximations when used in a homogeneous meanfield theory give a reasonable qualitative interpretation of the origin of the free energy barriers in folding (Plotkin *et al*., 1997). Here, we explore not only our earlier free energy functional but one that also includes a specific form of cooperativity of contact formation. This cooperativity mimics the effect of non-additive forces such as hydrophobicity, which depends on buried surface area. The cooperativity also describes the manner in which the polymeric entropy loss for forming highly spatially inhomogeneous collections of contacts can be reduced by bunching the contacts together. A strongly inhomogeneous set of contacts is envisioned to arise in theories of protein folding based on the capillarity model in which contiguous regions of a protein are thought to be fully folded or unfolded and to be separated by an interface (Bryngelson & Wolynes, 1990; Finkelstein & Badretdinov, 1997; Wolynes, 1997). The capillarity model is the other extreme from the homogeneous mean field theory picture used by us earlier. The present free energy functionals thus can interpolate between the two limits in a flexible realistic way. It also becomes possible to test the two approximations by comparison with experiment.

Since the free energy functionals which we use depend on a reduced description of the protein structure, comparing with experiment is crucial for understanding the range of validity of the approximations involved and of the semi-empirical energy parameters used. For this comparison, the study of equilibrium denatured states of the protein is particularly valuable, since no additional assumptions about the relation of dynamics to equilibrium thermodynamics are required in making such a comparison. While such an assumption is needed for interpreting rate data, as we have done earlier and do in the companion paper to this one (Shoemaker *et al*., 1999), the current study focuses on purely equilibrium experiments probing denatured lysozyme and lactalbumin ensembles.

After a discussion of the methods, we discuss the qualitative behavior of the structural correlations when we change the energetic heterogeneity of native contacts and cooperativity parameters. We then provide a tour of the folding funnels of lysozyme and lactalbumin. Here our attention is focused on the comparison with known NMR and protection experiments. This is followed by a discussion of directions for further improvement of the free energy functional approach. We also compare the approach with other computational routes to the structure of denatured states of proteins based on molecular dynamics simulation (Boczko & Brooks, 1995, 1997; Daggett *et al*., 1996; Schiffer & van Gunsteren, 1996).

The Free Energy Functional Approach

Summary of methods

In the methods section we describe the equations of the free energy functional approach. These free energy functional equations are written in terms of contact probabilities defined by amino acid residue pairs which come within a cutoff distance of each other in the native state. The contact probabilities vary continuously from zero in the unfolded state to one in the native state. We briefly motivate the physics behind each of the terms in the functional. The functional accounts for energetic forces dependent on residue identity and entropic forces dependent on chain collapse. In addition, the free energy functional accounts explicitly for cooperative forces corresponding to hydrophobic collapse and side-chain ordering that help bring together blocks of contacts in concert. These various cooperativity parameters for α-helix formation and contiguous core formation can be chosen rationally using experiments on peptides and statistical mechanical theories of how collapse affects secondary structure (Cantor & Schimmel, 1971; Luthey-Schulten *et al*., 1995; Finkelstein & Badretdinov, 1997). Once a free

energy functional and its parameters are chosen, structure formation in partially denatured ensembles is computed by minimizing the functional with constraints on the total amount of structure present. This leads to a series of self-consistent equations like those used to find titration curves for complex mixtures. These are solved through self-consistent iteration. Finally, we indicate how experimental observables, such as protection factors, are computed and compared with experiment.

Free energy

In order to describe mathematically the folding or unfolding of a protein we must develop a usable yet realistic free energy functional. In folding (unfolding) this functional models the competition between a decreasing (increasing) configurational space in which the residues can move and an increasing (decreasing) number of favorable interactions between the residues. If we want to study specific protein systems, the functional should include a combination of generic properties of a poly-amino acid chain and residue-specific properties. Homogeneous generic properties describe the general collapse of a polymer chain of length N, the backbone's tendency to form secondary structure, etc. Residue specific heterogeneous properties describe those interactions dependent on a specific amino acid sequence, such as the energy released on forming a hydrophobic contact or salt bridge.

The free energy functional can be written in terms of the set of Q_{ij}, the contact probabilities. Each Q_{ij} varies continuously from zero to one as the residues i and j in an ensemble of structures move from being more often away from each other in an extended state to being as close as in a folded state. In an ensemble of structures they quantitatively represent the fraction of time a given contact is made. The use of such density-like coordinates is familiar in the theory of hydrodynamics (Zwanzig, 1972), liquid structure (Morita & Hiroike, 1961; Munakata, 1975), and electron transfer reactions (Calef & Wolynes, 1983a,b,c). We will define as contacting pairs, all residue pairs i,j whose C^{β} atoms fall within a contact radius of 6.5 Å in the crystal structure. From this we define also a global reaction coordinate, $Q = \Sigma_{ij} Q_{ij} / \Sigma_{\text{native}} Q_{ij} = Q^*$, in which $Q = 0$ in the extended state when all $Q_{ij} = 0$ and $Q = 1$ in the folded state when all $Q_{ij} = 1$. Occasionally we give equivalent information with μ, the number of contacts formed, which is defined as $\mu = \Sigma_{ij} Q_{ij}$.

In our calculations the physico-chemical properties of interacting amino acids enter the functional through a database derived contact potential, $E = \Sigma_{ij} \varepsilon_{ij} Q_{ij}$. This represents a potential of mean force averaging over solvent configurations and side-chain configurations. It contains both entropic and enthalpic components. Other effects of backbone rigidity, etc. are neglected. Potentials based on both the information theoretic approach (Miyazawa & Jernigan, 1985, 1996) and on optimization schemes using a single contact radius (Goldstein *et al*., 1992) have been used by us. The single contact diameter potentials were calculated using the published procedure with the addition of a residue dependent potential for $i,i+4$ contacts within an α-helix. The resulting values are presented in Table 1. The results using either set of parameters are very similar, thus we focus on results with the optimized potential. The dependence of the energy in the functional on the entire set of Q_{ij} values rather than just the total Q reflects the heterogeneity of the contact potential.

For the entropy functional, we must account for the different regimes of entropy loss through the course of folding from an extended to a folded state. When the first few pairs of residues come within the contact radius of each other, the chain undergoes a large reduction in its configurational space. On the other hand, once some native structure has formed, developing additional contacts confines the chain to a much smaller extent. There are various ways to approximate the entropy functional in terms of contact pair probability assuming a protein can be modeled as an effective Gaussian polymer chain. The simplest entropy functional comes from the work of Jacobson and Stockmayer (1950) $S_0^{ij} = k_B \log[\Delta V / |i-j|^{3/2}]$ where $\Delta V = (3/2\pi)^{3/2}\ \Delta\tau / l_0^3$. $\Delta\ \tau$ is the volume of the interaction range and l_0 is the persistence length of the chain. This functional depends only on the sequence distance, $|i-j|$, of a loop formed in a random coil chain. The Jacobson-Stockmayer functional is the first term in the virial expansion of the entropy functional using contact pairs:

$$S(\{Q_{ij}\}) = \sum_{ij} S_0^{ij} Q_{ij} + \sum_{ijkl} S_0^{ijkl} Q_{ij} Q_{kl} + \cdots \qquad (1)$$

The virial expansion used by Bohr *et al*. (1994) and earlier by Chan and Dill (1990), leads to a slowly convergent expansion. This slowly convergent expansion must be resummed. The long-range terms are best treated using Flory's theory of rubber elasticity, while short range in sequence terms can be modeled as an explicit additional cooperativity (see below). We use functionals with both resummations present and adjust appropriately the coefficient of the effective cooperativity term.

Flory showed the entropy loss of forming cross-links in rubber in the mean field limit $S_0^{ij} = k_B \log[(\Delta V/(N/\mu))^{3/2}]$ where μ is the number of contacts made and N is the number of residues in the protein (Flory, 1956). As in our previous work, we use an interpolation of the Jacobson-Stockmayer and Flory entropy functionals:

$$S_0^{ij}(\mu) = k_B \log[\Delta V(|i-j|^{-3/2} + (N/\mu)^{-3/2})] \qquad (2)$$

For an extended configuration, contact formation is sensitive to the loop length $|i-j|$ as in the Jacobson-Stockmayer functional, which makes the

Table 1. Two-body single contact optimized potential interactions $-\varepsilon_{ij}$ for the 20 amino acids

Name	Ala	Arg	Asn	Asp	Cys	Gin	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	0.35	−0.24	−0.2	−0.28	−0.26	0.01	−0.22	−0.15	0.15	1.15	0.57	−0.18	0.07	0.22	−0.24	−0.13	−0.05	0.24	0.36	0.72
Arg		0.22	−0.4	−0.04	−0.57	0.23	0.14	−0.24	−0.26	−0.21	−0.22	−0.59	−0.91	0.16	−0.37	−0.23	−0.13	−0.2	−0.09	−0.11
Asn			0.08	0.28	0.27	−0.05	−0.24	0.1	0.08	−0.63	−0.32	−0.06	−0.18	−0.47	−0.29	0.35	−0.04	−0.15	−0.12	−0.08
Asp				−0.43	−0.32	−0.39	−0.29	−0.04	−0.02	−0.32	−0.54	−0.09	−0.55	−0.88	−0.4	0.45	0.03	−0.11	−0.42	−0.38
Cys					5.98	0.24	−0.22	0.22	1.87	1.53	0.25	−0.52	1.62	1.45	−0.13	−0.42	−0.19	2.64	1.46	0.41
Gln						−0.24	−0.21	−0.31	−0.93	−0.24	0.09	−0.06	−0.13	−0.43	−0.42	−0.19	−0.08	−0.71	−0.44	0.04
Glu							−0.46	−0.32	−0.5	−0.15	−0.27	0.19	−0.57	−0.29	−0.34	−0.26	−0.03	0.06	−0.14	−0.3
Gly								0.64	0.44	−0.17	−0.03	−0.29	−0.2	−0.12	0.02	−0.01	0.39	0.49	0.47	−0.05
His									1.64	0.29	0.48	−0.27	1.4	0.86	−0.53	−0.04	−0.27	−0.81	0.18	0.65
Ile										2.09	1.24	−0.21	0.97	0.81	−0.6	−0.28	−0.32	0.91	1.56	1.47
Leu											1.06	−0.34	0.9	1.68	−0.13	−0.28	−0.41	1.08	0.7	1.33
Lys												−0.56	−0.71	−0.26	−0.48	−0.2	0.07	0.04	−0.03	−0.11
Met													2.3	0.91	0.31	−0.29	−0.59	2.75	0.44	1.06
Phe														1.16	0.24	−0.38	−0.17	1.5	1.31	0.97
Pro															0.05	−0.37	−0.37	0.26	0.34	0.21
Ser																0.21	−0.06	−0.39	−0.29	−0.28
Thr																	0.16	−0.94	−0.15	0.06
Trp																		0.05	1.03	1.22
Tyr																			0.93	0.27
Val																				1.65

An optimized database potential with a single contact radius of 6.5 Å is used to generate the interactions between pairs of amino acid residues using the algorithm by Goldstein *et al.* (1992).

dominant contribution to the interpolated functional in this regime. Once the chain structures itself, the $|i-j|$ dependence for the entropy loss of contact formation saturates to (N/μ) as in Flory's mean field limit. At high Q, the atomistic limit associated with the core-halo model introduced by us in the homogeneous mean field theory becomes appropriate (Plotkin *et al.*, 1997). In the high Q limit, entire regions of contacts contiguous in sequence form or melt out instead of individual contacts, thus giving smaller, more uniform entropy changes per contact regardless of the loop length, $|i-j|$. These effects are partly accounted for by entropic cooperativity.

Combining the energy and entropy functionals, the free energy functional becomes:

$$F(\{Q_{ij}(\mu)\}) = \sum_{i,j} \varepsilon_{ij} Q_{ij}(\mu) - T\Bigg(\sum_{i,j} S_0^{ij}(\mu) Q_{ij}(\mu) + \sum_{\mu'=1}^{\mu} \sum_{i,j} (\partial S_0^{ij}(\mu')/\partial\mu)\delta Q_{ij}(\mu') + N\log(\nu)\Bigg) + \sum_{i,j} T\Big(Q_{ij}(\mu)\log[Q_{ij}(\mu)] + (1-Q_{ij}(\mu))\log[1-Q_{ij}(\mu)]\Big) \quad (3)$$

where $\delta Q_{ij}(\mu') = Q_{ij}(\mu') - Q_{ij}(\mu'-1)$. The last term accounts for the entropy associated with the possibility of forming contacts in multiple configurations in a partially ordered protein, i.e. the combinatorial entropy of mixing.

Cooperativity

Contact formation is directly cooperative for two reasons. First, it is entropically favorable to bunch contacts together. Second, hydrophobic forces depend on buried surface area and therefore have an explicitly many body cooperative component not accounted for by pair potentials. Averaging over side-chain orientations may generate such many body forces too. Incorporating cooperativity into the free energy functional is straightforward. The addition of explicit cooperativity to the functional also allows us to examine various conceptual limits of protein folding, corresponding to analytical models such as the capillarity theory. Nonadditivity can be introduced in the form of a term depending on the local density around a contact:

$$F_t(Q) = \alpha_t \sum_k (Q_{i,j}Q_{i,k} + Q_{i,j}Q_{k,j}) \quad (4)$$

This mimics the hydrophobic many body effect. The collapse of the chain encourages helix formation. This is an essential element of the theory of secondary structure formation in molten globules (Luthey-Schulten *et al.*, 1995; Yee *et al.*, 1994). This can be accounted for by an α-helical-local density interaction free energy:

$$F_{h-\rho}(Q) = \alpha_{h-\rho} T \sum_i^{hx} Q_{i-4,i} \sum_k Q_{i,k} \quad (5)$$

These forms of local cooperativity make it easier for a contact to form when the contacts near it have already started to form. Once a contact has formed, other contacts near in sequence can form more easily because they impose relatively small additional constraints on the configurational space of the protein chain.

In contrast to such cooperative free energies which act to stabilize the native state, some cooperativity is neutral to the folded state and is manifested as surface energies between folded and unfolded regions. One such surface energy is the interface energy between α-helices and coils:

$$F_h(Q) = \alpha_h \sum_i^{hx} \left(Q_{i-4,i} - \frac{1}{2}\right)\left(Q_{i,i+4} - \frac{1}{2}\right) \quad (6)$$

This free energy is directly analogous to the helical initiation free energy calculated from σ in helix-coil theory (Cantor & Schimmel, 1971). The magnitude of F_{hx} is estimated according to σ from helix-coil models as discussed below (Luthey-Schulten *et al.*, 1995).

For the free energy functional without explicit cooperativity, i.e. $\alpha_h = \alpha_c = \alpha_{h-\rho} = 0$, even when these are highly heterogeneous contact energies, the contact probabilities in the denatured state ensemble tend to remain relatively uniform as a slice of the protein is traversed. The entropic part of the non-additive free energies plays a role similar to the Flory resummation, but in this case the effects are local to specific regions of the protein while the Flory term treats all contacts equally and accounts for the long-range part of the contact-contact cooperativity.

Like the helix-coil theory, theories based on the capillarity limit explicitly view the distribution of contacts in the denatured state ensemble as clusters of probability in specific contiguous spatial regions of the protein (Wolynes, 1997). This means that there is a large surface tension between native and disordered structures. Outside a structured cluster little remnant order exists. A cooperative interaction explicitly reflecting the tendency toward capillarity behavior is:

$$F_c(Q) = \alpha_c \sum_{i,j} \sum_k \sum_l \Bigg(\left(Q_{i,j} - \frac{1}{2}\right)\left(Q_{l,k} - \frac{1}{2}\right) Q_{i,k}^{Nat} + \left(Q_{i,j} - \frac{1}{2}\right)\left(Q_{k,l} - \frac{1}{2}\right) Q_{k,j}^{Nat} \Bigg) \quad (7)$$

where $Q_{i,j}^{Nat}$ equals one if residues i and j form a contact in the native state and zero if they do not. This implies that $Q_{i,j}$ is favored by the formation of contacts near in space, but not necessarily near in

sequence. F_c and F_t provide similar cooperative stabilizations.

Some cooperativity already exists in the original functional, since contacts very near to i,j, such as $i+1,j+1$ or $i-2,j-1$, were already taken to form automatically when bringing i and j within the defined contact radius would also force these neighbors to be together because of the covalent chain connectivity. This cooperativity was expressed by coarsegraining the contact matrix into blocks of $i \pm 1,j \pm 1$.

Taking account of these three explicit forms of cooperativity, the free energy functional in its complete form is written as:

$$\begin{aligned} F(\{Q_{ij}(\mu)\}) = &\sum_{i,j} \varepsilon_{ij} Q_{ij}(\mu) - T\Bigg(\sum_{i,j} S_0^{ij}(\mu) Q_{ij}(\mu) \\ &+ \sum_{\mu'=1}^{\mu} \sum_{i,j} (\partial S_0^{ij}(\mu')/\partial\mu)\delta Q_{ij}(\mu') + N\log(\nu)\Bigg) \\ &- \alpha_h \sum_{i}^{hx} \left(Q_{i-4,i}(\mu) - \frac{1}{2}\right)\left(Q_{i,i+4}(\mu) - \frac{1}{2}\right) \\ &- \alpha_{h-\rho} \sum_{i}^{hx} Q_{i-4,i}(\mu) \sum_{k} Q_{i,k}(\mu) \\ &- \alpha_c \sum_{i,j} \sum_{k} \sum_{l} \Bigg(\left(Q_{i,j}(\mu) - \frac{1}{2}\right) \\ &\times \left(Q_{l,k}(\mu) - \frac{1}{2}\right) Q_{i,k}^{Nat} + \left(Q_{i,j}(\mu) - \frac{1}{2}\right) \\ &\times \left(Q_{k,l}(\mu) - \frac{1}{2}\right) Q_{k,j}^{Nat}\Bigg) \\ &+ \sum_{i,j} T(Q_{ij}(\mu)\log[Q_{ij}(\mu)] \\ &+ (1 - Q_{ij}(\mu))\log[1 - Q_{ij}(\mu)]) \end{aligned} \tag{8}$$

By varying the parameters, α_h, α_c, and $\alpha_{h-\rho}$, we can effectively turn on any combination of the various types of explicit cooperativity. The different effects of non-additivity can be explored. Here, we limit our study to $\alpha_t = 0$, since its effects are so similar to the α_c term. We discuss how results change as we vary these parameters for pedagogic reasons, but when comparing to laboratory experiment we fix the cooperativity parameters to physically reasonable values. These are chosen to fit the free energies of structural ensembles that can be independently calculated or measured.

Variational formulation

To calculate $Q_{i,j}$ from the free energy functional at a given value of the reaction coordinate, Q, we minimize the free energy with respect to Q_{ij} but include a Lagrange multiplier from the constraint $\Sigma_{ij} Q_{ij}/\Sigma_{\text{native}} Q_{ij} = Q^*$. This constraint allows us to follow the structural progress down the folding funnel using a reaction coordinate Q. This gives a self-consistent equation:

$$\begin{aligned} Q_{ij} = 1/\Bigg(1 + \exp\Bigg[&\varepsilon_{ij}/T - \lambda/T - S_0^{ij}(\mu) \\ &+ \alpha_{h-\rho} 2 \sum_{i}^{hx} \left(Q_{i-4,i} + \sum_{k} Q_{ik}\right) \\ &+ \alpha_h \frac{1}{T} \sum_{i}^{hx} (Q_{i,i-4} + Q_{i+4,i+8}) \\ &+ \alpha_c \frac{1}{T} \sum_{k} \sum_{l} (Q_{l,k} Q_{i,k}^{Nat} + Q_{k,l} Q_{k,j}^{Nat})\Bigg]\Bigg) \end{aligned} \tag{9}$$

where λ is a Lagrange multiplier enforcing the constraint as $(\Sigma_{ij} Q_{ij}/\Sigma_{\text{native}} Q_{ij} - Q^*)\lambda = 0$. When any of the explicit cooperativity terms are present, the nonlinear equation for Q_{ij} must be solved iteratively. Otherwise it may be solved without iteration.

Calculating thermodynamic input parameters

For the numerical values of the parameters in our free energy functional we choose values to give proper thermodynamics. Table 2 contains a brief description of the parameters in our model and their values used in our calculations. The constraint that the folding and unfolding minima are at the same free energy at the folding temperature, T_f, fixes the energy scale: the total amount of energy gained by the formation of all the native

Table 2. Parameters used in $Q_{i,j}$ calculations for α-lactalbumin and hen lysozyme

Parameter	Description	α-Lac value	Lys value
ν	Conformations/residue,	4	4
T_f	Folding temperature ($k_B T$)	1.2	1.0
b_o/μ	Hydrophobic collapse ($k_B T$)	−2.1	−2.1
r_o	Contact radius (Å)	6.5	6.5
l_0	Kuhn length, (Å)	5	5
ξ	Contact matrix cell size	4	4
α_h	Magnitude of helical cooperativity	0.58	0.58
α_c	Magnitude of capillarity cooperativity	0.03/0.08	0.03/0.08
$\alpha_{h-\rho}$	Magnitude of interaction cooperativity	0.6	0.4

All parameters are fixed to realistic estimates according to experiment and polymer physics theory.

contacts must be equal to the total amount of entropy lost by these contacts during the folding process.

As with other statistical potentials on comparing with experiment, one must use a parameter to control the temperature, T_f, and a parameter to control the hydrophobic collapse, b_o. These parameters fix the total energetic gain of the protein to the total entropic loss from the extended state to the native state. We assume each residue can take one of ν conformations, where $\nu = 4$, giving a total entropy of $N \log(\nu)$ in the extended state. In addition, the energetic difference between the molten globule and folded states must match the entropy in the globule state, $N \log(\nu/e)$. This gives the following equations: $N \log(\nu) = -(\varepsilon_f + b_0)/T_f$ and $N \log(\nu/e) = -(\varepsilon_f - \tilde{\varepsilon}_{mg})/T_f$. Here $\tilde{\varepsilon}_{mg}$ is computed by averaging the results of threading the sequence through alternate protein structures. This way of setting the average hydrophobicity parameter assumes that proteins are near a triple point where non-specific collapse is almost as favorable as folding. Other choices are possible.

The capture volume in the entropy functional, the parameter $\Delta V = 2$ is set by the interaction volume corresponding to our assumed 6.5 Å cutoff ($\Delta\tau^{1/3}$) and a reasonable persistence length of amino acid polymers (l_0) of 5 Å. Although we explore the effects of pure tertiary cooperativity, all the results we present use $\alpha_t = 0$.

The magnitude of the helical cooperativity was earlier set by varying the parameter α_h to match contact probabilities to clear Overhauser enhancements measurements made by the Wüthrich group (Neri *et al.*, 1992; Wüthrich, 1994; Shoemaker *et al.*, 1997). Where explicit free energies are calculated instead of structural averages alone, it is reasonable to set the helical cooperativity and helical boost energy from known thermodynamic σ and S values, respectively.

The two body statistical potential described earlier is optimized with a helical propensity for $i, i+4$ contacts within helical domains. This gives a more physical basis to the relative residue dependent interactions inside the helix. Extending this treatment, the s and σ parameters from helix-coil theory are used to estimate the magnitude of these propensities and of the helical cooperativity. From the parameter s the free energy difference of extending the helix by one unit relative to the coil state is obtained. We isolated an individual helix (α-helix B) from lysozyme and calculated the free energy change of the fully formed helix relative to the unformed coil state. This free energy difference relates in the simple zipper model to s and σ with the relation $F = -\log \sigma - n_h \log s$ where n_h is the number of helical units. Measured s values calculated by Baldwin (Chakrabartty *et al.*, 1994) are averaged for the amino acid identities of the helical contacts. For the formation of the entire helix an average measured value of $s = 0.24$ is obtained, while from the calculated free energy difference of the helix formation an average value of $s_{\text{th}} = 0.2$ is obtained. Within the approximation of the simple zipper model, we find that our statistical potential including helical propensities gives a correct free energy change for helix formation when applied to a helix from lysozyme.

Helical propensities by themselves do not seem to account for helical stability in the transition state (Muñoz *et al.*, 1996). Helical cooperativity is included in the free energy functional in two different forms to better account for helix formation. Again from the zipper model and peptide measurements, a value for the parameter σ is obtained for the energetic cost to helical end formation. This is analogous to a helix-coil surface tension in the free energy functional written as $F_h = \Sigma_i^{hx}(Q_{i-4,i} - 1/2)(Q_{i,i+4} - 1/2)$ which acts to destabilize the helical ends as they increase in probability. In order to properly adapt this cooperativity to the surface tension, σ, additional terms are added to the sum of the form $(Q_{i-4,i}^{\text{hx}} - 1/2)(\{Q^{coil} = 0\} - 1/2)$ in which i is an end residue interacting with a coil region as defined by the crystal structure. This gives an increasing destabilization free energy as the helix forms which saturates to a fixed value along the progression coordinate once both ends of the helix have fully formed. The magnitude of this helical cooperativity is first estimated using the simple poly-alanine measurement, $\sigma = 8 \times 10^{-4}$ (Cantor & Schimmel, 1971). However, similar results are obtained with residue dependent σ values from Serrano and coworkers (Muñoz & Serrano, 1994). In order to account for the collapsed state of the structure in which the helical cooperativity is being manifested, Luthey-Schulten *et al.* (1995) show that the appropriate magnitude of helical surface tension is reduced to $\sigma = 10^{-1}$ in the confined conformational space. Fixing the surface free energy change between helix and coil states for the fully formed isolated helix of lysozyme, $F = -\log \sigma$, to the helical cooperativity free energy, $\alpha_h = 0.58$ is obtained.

Another form of cooperativity involving helical interactions arises from the interaction of contacts within the helix and the neighboring contacts which surround the helix. In the free energy functional a term is introduced of the form $F_{h-\rho} = \alpha_{h-\rho} \Sigma_i^{hx} Q_{i-4,i} \Sigma_k Q_{i,k}$. The magnitude of this free energy can be estimated by looking again at the helix-coil theory (Luthey-Schulten *et al.*, 1995). The authors show that a free energy of the same form as this helical-local density interaction contributes roughly 1 $k_B T$ of stabilization energy per helical residue. Corresponding to this stabilization, the helical-local density free energy cooperativity parameter is set to $\alpha_{h-\rho} = 0.6/0.4$ in α-lactalbumin/lysozyme, respectively.

The magnitude of the capillarity cooperativity is determined by studying the entropy of compact configurations with protruding loops. The entropic loss of random coil loops constrained to the surface of a completely formed structure was calculated by Plotkin *et al.* (1997) and by Finkelstein & Badretdinov (1997). The capillarity cooperativity is

expected to be of the order of this surface entropy. To fit the parameters we arbitrarily choose structures with dividing surfaces cutting through the protein in which $Q_{i,j} = 1$ for all contacts on one side of the surface and $Q_{i,j} = 0$ for all contacts on the other side of the surface. α_c is adjusted so the total capillarity cooperativity calculated for each surface matches the total surface loop entropy loss for the same surface as calculated by the polymer physics arguments. This overestimates the entropy contribution, since some of the cooperativity is already in the Flory term. For this reason we also present $\alpha_c = 0$ results. The part of α_c coming from hydrophobic cooperativity can be estimated in a similar way. We have not explicitly done this, but expect the value to be comparable to the pure entropic component, thus compensating for overcounting.

Calculating experimental quantities

In the denatured state ensemble one of the best experimental probes is the hydrogen exchange protection factor determined from NMR measurements (Kim *et al.*, 1993; Wüthrich, 1994; Woodward, 1994; Houry *et al.*, 1998). Large protection factors come from residues whose amide hydrogen atom is sufficiently buried in the core of the protein so that it cannot undergo exchange with the polar solvent. There is a great deal of complex chemical mechanism in determining protection factors so they are difficult to precisely relate to the reduced contact description. However it is natural to assume that only when all contacts are made will complete protection be achieved.

Thus, we compare measured protection factors to the quantity $Q_i^{\text{prod}} = \Pi_j Q_{i,j}$. Q^{prod} is large only if all of the neighboring contacts of a residue have a high probability of being formed. Like Q_i^{prod}, the protection factor varies over several orders of magnitude and is large only for residues protected from solvent on all sides. When even one of a residue's contacts is not well-formed, its Q_i^{prod} value will diminish significantly, modelling a residue's ability to undergo H exchange with only a small amount of surface area exposed to solvent.

Another useful quantity we can calculate is $Q_i = \Sigma_j Q_{i,j}$. This can be compared directly to results of molecular dynamics when they are suitably averaged. By using it to color three-dimensional protein representations, Q_i provides a useful visual aid to discuss the structural effects of changing various thermodynamic conditions and free energy parameters. If all contacts in the native protein had the same energy, Q_i would also be related to site directed mutation changes of the stability of the denatured ensembles. Q_i can be thought of as the fraction of time a residue is properly positioned as in the native state.

In addition to monitoring Q_i and Q_i^{prod} on a residue by residue basis we can average either of these quantities over a particular secondary structure element. These averages are often better determined by experiment. The study of Q_i or its block averages, as Q*, the depth in the funnel is varied, is analogous to obtaining a titration curve with the slope of the funnel as the tunable experimental parameter.

Results

To get a better appreciation of how the functional models the folding free energy landscape we first examine the effects of varying the cooperativity parameters and energetic heterogeneity before comparing to specific experimental results at the physical values of these parameters. We use the 129 residue protein hen lysozyme to display this analysis since it is one of the proteins studied in the denatured state using NMR techniques.

Figure 1(a) shows the three-dimensional crystal structure of lysozyme. Residues are colored by Q_i in the denatured state $Q^* = 0.3$ with no explicit cooperativity turned on. Blue indicates the most ordered and red the least native-like. With only the Flory cooperativity the contact probability is fairly diffuse throughout the protein with little in the way of well-defined ''hot spots''. This represents the inhomogeneous mean field limit, since cooperativity comes only from the Flory part of the entropy and from the coarsegraining.

In Figure 1(b) we see the effect of having explicit helical cooperativity present, $\alpha_h = 3$, at $Q^* = 0.3$. An unrealistically high value of the α_h parameter is used to clearly demonstrate its effect. The Q_i values show clearly that the denatured state has become structured primarily by residues outside of the helices. In this limit the helical cooperativity's tendency to avoid helix-coil interfaces becomes evident as a result of its overwhelming contribution to the free energy.

Another type of explicit cooperativity affecting helix formation in our free energy functional model is shown in Figure 1(c). Here only the helical-local density interaction cooperativity is used at its realistic magnitude, $\alpha_{h-\rho} = 0.4$, at $Q^* = 0.3$. In contrast to the scenario in Figure 1(b), Figure 1(c) illustrates the denatured state has become structured primarily through secondary structure formation by residues in the helices. The helices are frayed at their ends. Due to the constraint $\Sigma_{ij} Q_{ij} / \Sigma_{\text{native}} Q_{ij} = Q^*$, the increased stability of helical hydrogen bonds sometimes forces contacts in the rest of the protein with marginal stability to lose contact probability. Only contacts with relatively large stabilities persist outside of the helices.

Figure 2 shows the progression of denatured state contact structure of hen lysozyme with increasing capillarity cooperativity from no cooperativity ($\alpha_c = 0$) to high cooperativity ($\alpha_c = 0.1$) at $Q^* = 0.47$. Beyond this point additional cooperativity does not change the Q_{ij} values. This progression illustrates the development from the mean field limit to the capillarity limit. As expected, when capillarity cooperativity is increased, the diffuse

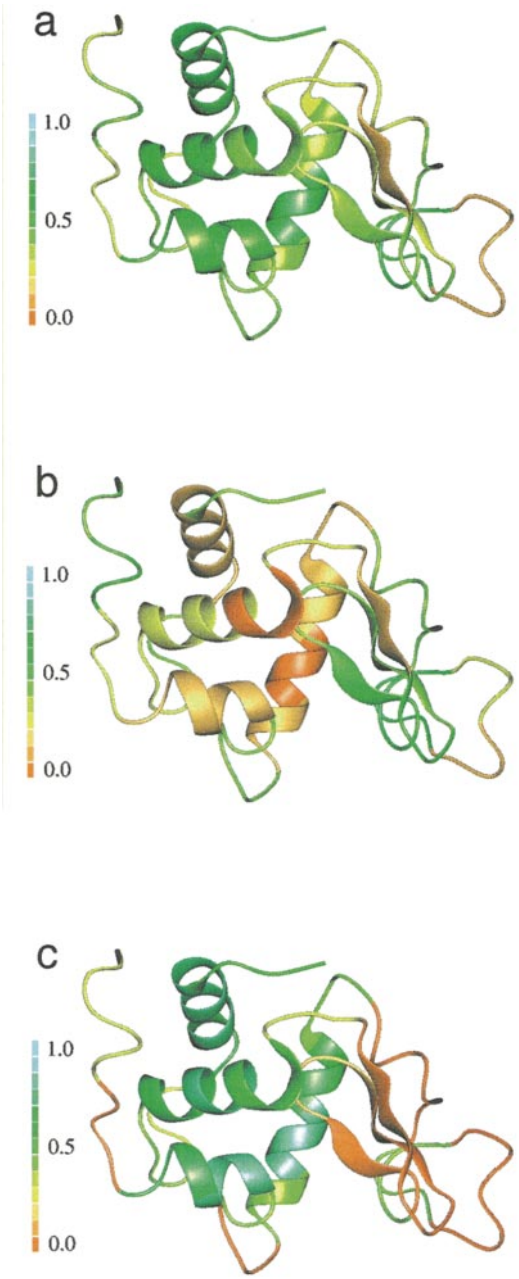

Figure 1. The effect of helical cooperativity in the denatured state. Ribbon representation of the hen lysozyme crystal structure colored by Q_i in the denatured state ($Q^* = 0.3$) with (a) no cooperativity, (b) with helical cooperativity, $\alpha_h = 3$, and with (c) helical-local density interaction cooperativity, $\alpha_{h-\rho} = 0.4$. Red indicates $Q_i = 0$ and blue indicates $Q_i = 1$.

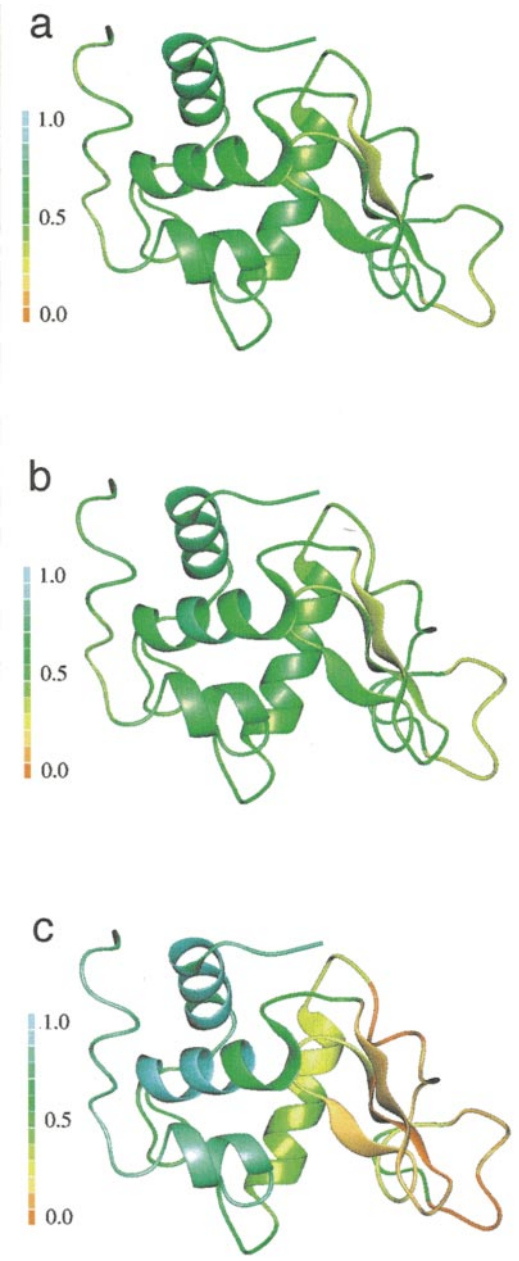

Figure 2. The effects of increasing capillarity cooperativity in the denatured state. Ribbon representation of the lysozyme crystal structure colored by Q_i in the denatured state ($Q^* = 0.47$) with capillarity cooperativity at (a) $\alpha_c = 0$, (b) $\alpha_c = 0.05$, and (c) $\alpha_c = 0.1$.

contact probability throughout the protein becomes more localized to specific spatial regions. In the limit of high capillarity cooperativity a distinct boundary between ordered and unordered regions emerges across the surface of the protein. This high capillarity cooperativity scenario shown in Figure 2(c) illustrates the theory described by Wolynes (1997). The ordered regions, or nuclei, that form are pockets of residues in the core, determined by their large number of neighbors.

Notice that such a sharp structural interface does not really form until a very large value of α_c is reached. The contact probabilities are fairly diffuse at the intermediate values of capillarity cooperativity, which our analysis suggests are physically reasonable. For actual proteins some structure persists in regions outside regions identifiable specifically with the sharp nuclei found in the capillarity limit. The contacts that form in the capillarity limit

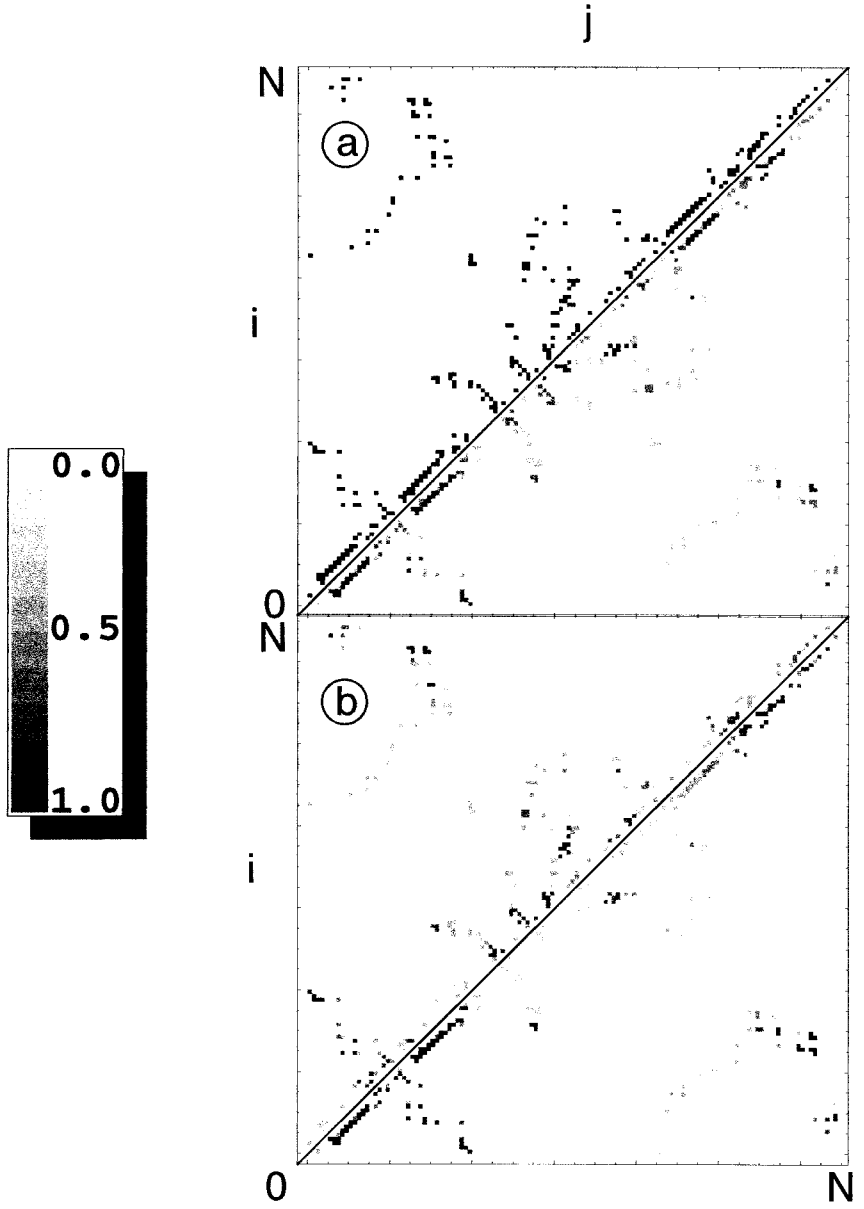

Figure 3. The effects of cooperativity in the denatured state. Contact maps for lysozyme using a gray scale for the calculated $Q_{i,j}$ values in the (a) native state (top panel), in the denatured state ($Q^* = 0.47$) with no explicit cooperativity bottom panel), and (b) with a large amount of helical cooperativity, $\alpha_h = 3$, $\alpha_c = 0$, $\alpha_{h-\rho} = 0$, (top panel), and with high capillarity cooperativity, $\alpha_h = 0$,$\alpha_c = 0.1$, $\alpha_{h-\rho} = 0$, (bottom panel).

arise primarily from their having a large number of neighboring contacts interacting with each other, not from any unusually stable individual contacts.

For another visual representation Figure 3 shows the two dimensional contact maps for some of the conditions in Figures 1 and 2. In the top panel of Figure 3(a) the native contact map is shown as a reference state. The contact map for $Q^* = 0.47$ with no explicit cooperativity is in the bottom panel. As in the three-dimensional representation, probability is fairly diffuse throughout the map. The top panel in Figure 3(b) shows the result of turning on a large amount of helical cooperativity ($\alpha_h = 3$) at $Q^* = 0.47$. The intensity along the diagonal decreases, resulting from the penalty of creating helix-coil interfaces. The bottom panel contains the contact map for the high capillarity cooperativity limit ($\alpha_c = 0.1$) at $Q^* = 0.47$. Here we see contact probability becomes confined to a few hot spots on the map, which lie in the core of the protein.

The three-dimensional structure and two-dimensional contact map representations illustrate the locations of contacts with highest probability and give a good idea of the degree of localization of these hot spots. We can display the localization of contact probability also by plotting the distribution of Q_{ij} values at different points throughout the progress coordinate, Q^*, as in Figure 4(a) with low but physically reasonable cooperativity ($\alpha_c = 0.03$,$\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.4$) for hen lysozyme. In this case the distributions are unimodal throughout the reaction coordinate with small tails at the middle Q^* values. In kinetic studies the character of the Q^* distribution distinguishes the ''specific nucleus'' scenario (where this description should be distinctly bimodal at $Q_{i,j} = 0$ or $Q_{i,j} = 1$) from the more general nucleation-condensation descrip-

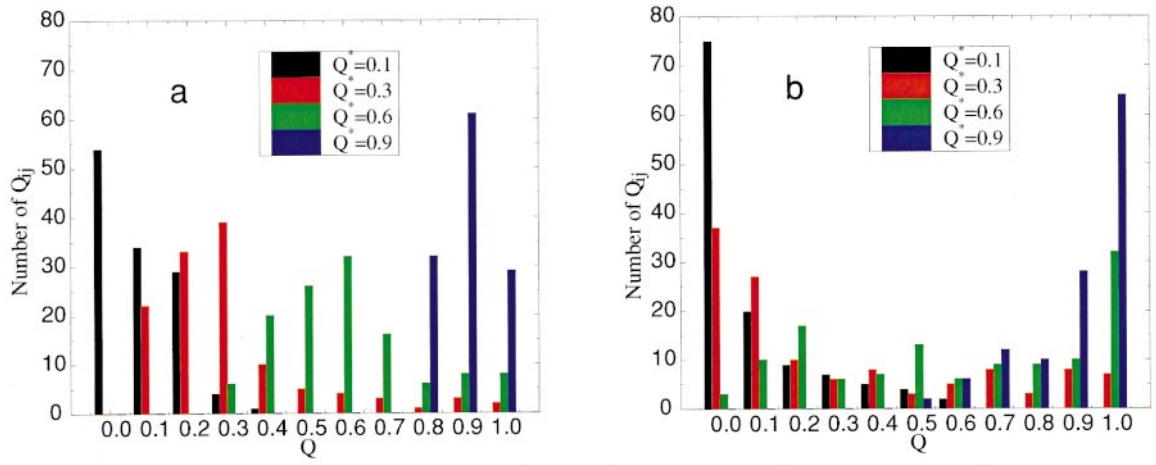

Figure 4. The localization of contact probability in the denatured states of lysozyme. A progression of distributions of Q_{ij} values from $Q^* = 0.1$ to $Q^* = 0.9$ with (a) physically reasonable low capillarity cooperativity $\alpha_c = 0.03, \alpha_h = 0.58, \alpha_{h-\rho} = 0.4$ and (b) with the large limit of capillarity cooperativity $\alpha_c = 0.15$.

tion (Onuchic *et al.*, 1996). At each value of Q^* the contact probabilities are closely centered around the corresponding value of Q^* indicating the relative homogeneity of $Q_{i,j}$ values.

Figure 4(b) shows the same progression in the high capillarity cooperativity limit ($\alpha_c = 0.15$). Here the distribution tends to be bimodal throughout Q^*. Notice especially at the middle Q^* values the large variance of $Q_{i,j}$ in Figure 4(b), which is in sharp contrast to the same Q^* values in Figure 4(a). We see then that strongly many body forces may lead to the specific nucleus scenario.

These statistical Q_i distributions are consistent with the visual representations of the effects of cooperativity on structure formation in lysozyme. In the denatured state localized hot spots occur at even moderate levels of cooperativity. As the protein becomes more native-like though, well-formed structure occurs throughout the protein unless a significant amount of cooperativity forces the structure to be confined primarily to a small region. This corresponds to the limit of a protein with a very large surface tension between native and denatured phases.

We now turn to the specific experiments on denatured protein. Human α-lactalbumin is an α/β protein that has been the subject of considerable experimental work (Miranker *et al.*, 1991; Radford *et al.*, 1992; Schulman *et al.*, 1995; Wu & Kim, 1998). We focus first on the two-dimensional NMR studies of individual residues in denatured α-lactalbumin states at varying denaturant strengths and temperatures performed by the Dobson group (Schulman *et al.*, 1997). This is an extremely interesting study because it characterizes the shape of the overall folding funnel by varying thermodynamic conditions, rather than only giving information on a single ensemble of intermediates at one point in the funnel. Changing denaturant conditions corresponds to modifying the stability gap or slope of the folding funnel. Note, however, that in all comparisons made with kinetic experiments, caution must be exercised to avoid interpreting the results in a purely equilibrium manner.

To compare to the NMR experiment Q_i values calculated from the free energy functional are averaged over both the α and β domains of α-lactalbumin (Figure 5). Beginning at the bottom of the funnel ($Q^* = 1$) using the free energy functional, each calculated Q_i probability is near unity for most residues with fractional probabilities mainly being present in the β-sheet domain. Traversing up to the middle and upper regions of the folding funnel ($Q^* = 0.4$), calculated Q_i probabilities become very low across the β domain, while significant structure persists in many of the helical residues. Not until the denatured ensembles of the uppermost part of the folding funnel are reached ($Q^* < 0.1$) do the last of the helices become unstructured. The behavior found throughout the progress coordinate from the NMR experiment agrees with the free energy functional method shown in Figure 5 for the two domains.

The Oxford/MIT group also monitors the formation of structure in denatured states by calculating a titration curve from the fraction of visible resonances for each of the helices in α-lactalbumin as denaturant strength and temperature are changed. In Figure 6(a) and (b) we calculated an analogous theoretical titration curve for each of the helices in α-lactalbumin by averaging the Q_i values for each helix at increasing values of Q^* in both the low and high surface tension regimes.

At the highest denaturant strength and temperature, 8 M guanidine hydrochloride, 50 °C, Dobson observes resonances from only the D and 3_{10}-helices. Using conditions less biased towards the unfolded state, 10 M urea, 50 °C, half of the cross-peaks in the A and B helices are also observed. At the lowest denaturant strengths and temperature, 0-6 M urea, 20 °C, resonances are observed in the C and short 3_{10}-helices(α/β domain interface). In general, resonances are observed only in helical

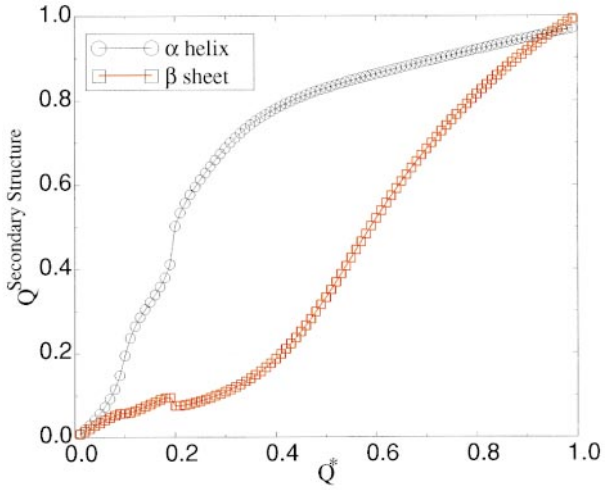

Figure 5. α/β Domain titration curves for α-lactalbumin. The average contact probability for all helical residues and for all sheet residues are plotted in black (α-helix) and red (β-sheet) continuous lines at ($\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.6$).

regions and not in β-sheet or coil regions in nonnative environments.

We find that the residues with the highest Q_i values are all located in the helices. Specifically, helix C forms first between $Q^* = 0.1$ and $Q^* = 0.15$ (see below) in the high capillarity regime. This helix is followed closely by the formation of helices A and B at $Q^* = 0.2$. Among the helices, helix D forms last but still within the early denatured states at around $Q^* = 0.3$.

Of particular interest in comparing the low and high surface tension regimes in Figsure 6(a) and (b) is the change in the cooperative nature of individual helix formation. With low surface tension each helix formation event is weakly cooperative. In this case a complete helix is not a compact object *vis-a-vis* its interactions with the rest of the protein. In contrast for the high surface tension regime the helices of lysozyme taken one by one form with very high cooperativity. This increase in the cooperative nature of helix formation illustrates well the general effect of cooperativity on secondary structure formation.

Figure 7 shows the same progression down the folding funnel by plotting Q_i values along the sequence at various values of Q^*. The main point here is the overall importance of helix formation over the sheet and domain interface regions as is found in experiment.

Another way to probe the landscape of the folding funnel using the free energy functional is to vary the temperature. Figure 8 shows the effect of increasing temperature on the helix titration plots for α-lactalbumin. When the temperature is increased to $3/2T_f$ the degree of cooperativity of the curves changes only slightly, with the relative ordering of the helices preserved. We conclude that the general nature of structure formation in the unfolded states remains qualitatively unchanged throughout a relatively wide range of temperature. Not until the temperature of $3T_f$ in Figure 8(c) do we see a substantially less cooperative picture of folding. Even at this high temperature, however, the relative ordering has changed very little.

One feature to note in the theoretical titration plot for α-lactalbumin (Figure 6(b)) is the emergence of multiple solutions in certain regions of Q^*. These solutions arise from the competition between different helices when there is a large amount of explicit helical cooperativity. As the magnitude of helical cooperativity is diminished the multiple solutions disappear, even in the presence of capillarity cooperativity and energetic heterogeneity. The existence of multiple solutions means that the pro-

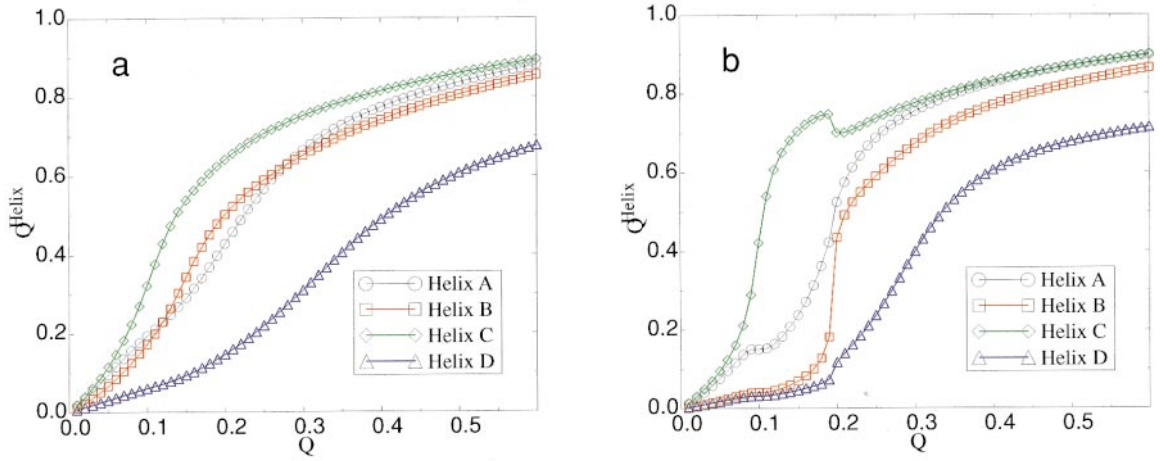

Figure 6. Theoretical titration curves for the α-helices in α-lactalbumin. (a) Calculated Q_i values in the low capillarity cooperativity regime ($\alpha_c = 0.03$, $\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.6$) of the free energy functional theory are averaged for each α-helix throughout the reaction coordinate, Q, giving a theoretical profile of the folding funnel. This is analogous to the NMR resonances averaged for each α-helix at various temperatures and denaturant strengths from experiment. (b) The same curves are calculated in the high capillarity cooperativity regime ($\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.6$).

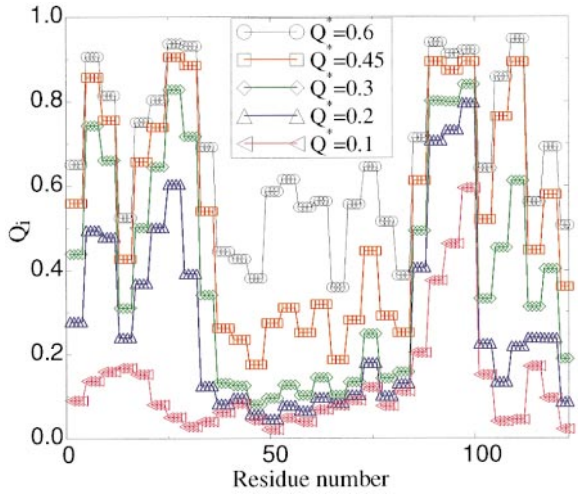

Figure 7. Residue structure formation in a progression down the funnel for α-lactalbumin. Q_i is plotted *versus* sequence number at representative values of the folding process in the high capillarity cooperativity regime ($\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-p} = 0.6$).

tein can form structure in the early stages of folding in more than one way. Folding in this case is described by a bifurcated funnel in which qualitatively distinct multiple routes to the native state are possible. In our calculations each set of Q_{ij}'s is calculated at equilibrium for a given value of μ, the number of contacts formed. As a result, for certain parameter values one set of Q_{ij} values conforms to a different local minimum than does the Q_{ij} values at a different but similar set of parameters, giving jagged characteristics to the otherwise smooth titration plots in Figure 6(b). In an experiment, both populations would be sampled giving smoother monotonic titration curves. The analogous effect for folding kinetics is a piecewise linear stability plot as observed by Fersht's group in barnase (Johnson & Fersht, 1995). For both α-lactalbumin and hen lysozyme, the multiple solutions do not affect the main conclusions. Helices form first in both proteins in the same order and at approximately the same values of Q^* regardless of the slightly different multiple solutions.

A protein homologous to α-lactalbumin is the 129 residue α/β protein hen lysozyme with four

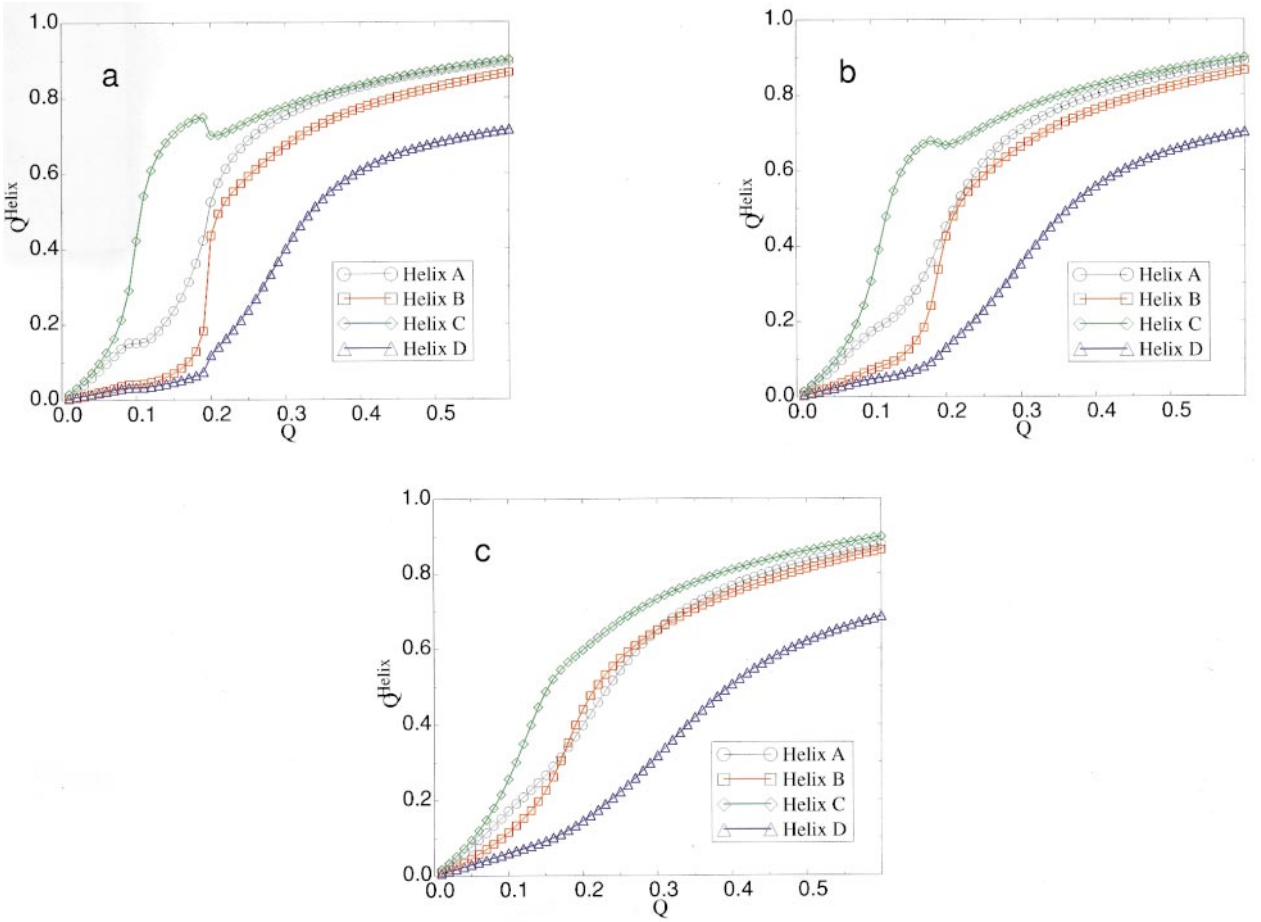

Figure 8. Variation of the theoretical titration curves with temperature for the α-helices in α-lactalbumin. Calculated Q_i values in the high capillarity cooperativity regime ($\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-p} = 0.6$) of the free energy functional theory are averaged for each α-helix throughout the reaction coordinate, Q, giving a theoretical profile of the folding funnel at (a) T_f, (b) $3/2T_f$, and (c) $3T_f$.

main helices. This has also been studied with various experimental probes (Miranker *et al.*, 1991; Radford *et al.*, 1992; Gladwin & Evans, 1996; Nash & Jonas, 1997). Both proteins have very similar topologies, but do show different patterns of residual structure in the denatured state. Here we see the role of sequence heterogeneity. Gladwin & Evans (1996) use a variant of the hydrogen exchange labelling technique to characterize early folding intermediates. By constraining our expression for Q_{ij} to a value of Q^* in the denatured state, we can calculate and compare the quantity Q_i to the experimental quantity I_D, the dead time exchange inhibition factor.

Figure 9(a) shows a bar graph of intensities of I_D (shaded) and Q_i (unshaded) for the residues in hen lysozyme. Figure 9(b) shows the same graph of I_D intensities (shaded) compared in this case to Q_i^{prod} intensities (unshaded). Residues with no I_D intensity are not necessarily zero, but did not have data collected for them. The four main α-helices, the β-sheet and the β-domain interface are labeled. Within each of the labeled regions there is very good qualitative agreement between theory and experiment in Figure 9(a) and (b).

Gladwin & Evans (1996) finds at least two large I_D values in each of the helices. The largest I_D values are found in the center of the C-helix and throughout the A-helix. The other significant region of protected structure is at the α/β domain interface, including residues Trp63, Cys64, Asn65, and Ile78. There are no appreciable I_D values in the β sheet region or the nearby 3_{10}-helix.

We find very similar regions of ''hot spots'' from both Q_i values at $Q^* = 0.4$ and Q_i^{prod} values at $Q^* = 0.57$ as does Evans in the early intermediates of folding. All of the large Q_i and Q_i^{prod} values in our analysis of hen lysozyme come from the helices and the domain interface region in agreement with the findings of the Evans group. We find large $Q_{\{i\}}$ and Q_i^{prod} values from residues in the middle of the C-helix, from the middle of the B-helix, and from residues throughout the A-helix. We also find high signals in the domain interface region from residues Trp62, Gly67, Arg68, Arg73, and Asn77 according to Q_i^{prod} intensities and from residues Leu56, Trp62-Cys64, and Arg73-Cys76 according to Q_i intensities. In accordance with experiment we find little structure in the β-sheet and nearby 3_{10}-helix.

Comparison of Figure 9(a) with (b) shows the same qualitative results in each. The differences lie in the more varied character of the Q_i^{prod} intensities in Figure 9(b), which seems to give better agreement overall to the experimental I_D intensities. Q_i^{prod} is generally smaller than the protection values. Thus the requirement of all contacts being needed for protection is probably too strong. We note the pattern of protection factors is quite reminiscent of earlier equilibrium values on lysozyme (Miranker *et al.*, 1991).

Figure 10 shows the three-dimensional representation of lysozyme at $Q^* = 0.6$ colored by Q_i^{prod}. At lower values of Q^* the Q_i^{prod} are all relatively weak. This representation in Figure 10 shows that helix D contains a residue completely ''protected'', having all of its contacts formed, along with a few residues in helix A and the interfacial region.

In Figure 11 we plot the theoretical titration curves for hen lysozyme, similar to the earlier plots for α-lactalbumin. There is not enough experimental data to characterize the continuous formation of structure in the folding funnel as we did earlier for α- lactalbumin. In this case we monitor the struc-

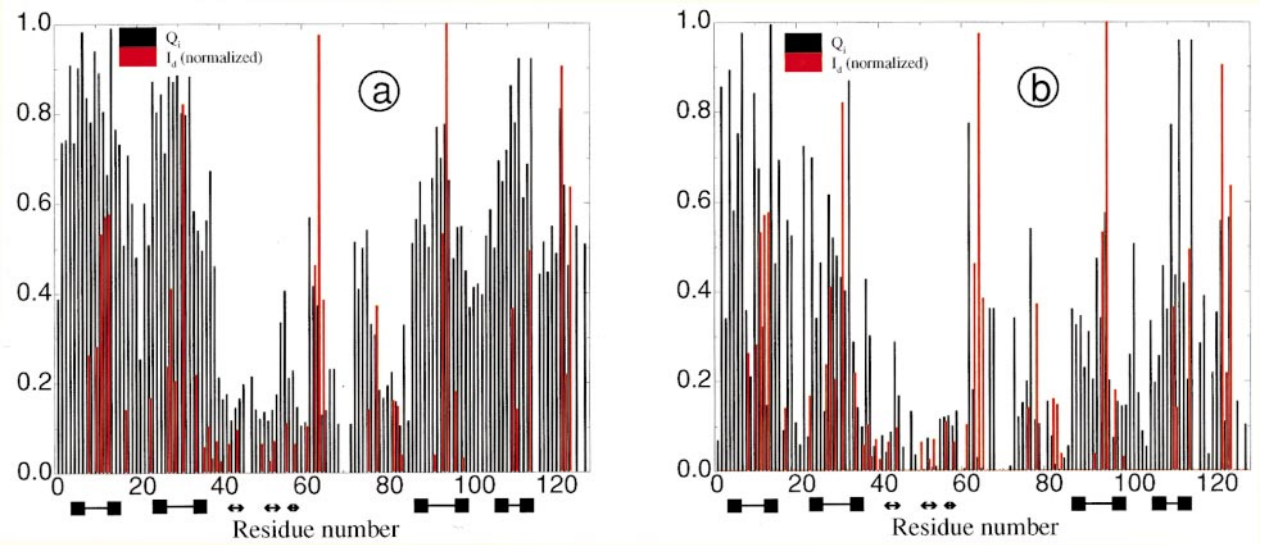

Figure 9. Dead time inhibition factors versus calculated Q_i and Q_i^{prod} for hen lysozyme. I_D values are shown in red for the selected residues for which they were measured. Calculated (a) Q_i values (a) and (b) Q_i^{prod} values are shown in black for all residues in the physically motivated high capillarity cooperativity regime ($\alpha_c = 0.08, \alpha_h = 0.58$, $\alpha_{h-p} = 0.4$). Notice that qualitative agreement is achieved for the α-helical, β-sheet domain interface, and β-sheet regions.

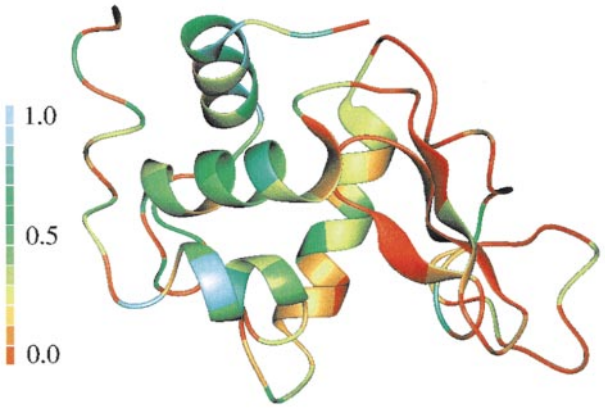

Figure 10. Hen lysozyme residues shown with the fraction of completely formed native contacts in a molten globule state. The ribbon representation is colored by Q_i^{prod} at $Q^* = 0.6$ with $\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.4$. Blue indicates $Q_i = 1$ and red indicates $Q_i = 0$.

ture in the β-domain interface as well as each of the four α-helices. The very early structure formation occurs in α-helix *A* and the β-domain interface in the low capillarity cooperativity case (Figure 11(a)). If $Q^* = 0.3$ then the other three α-helices all have considerable structure. In the high capillarity cooperativity case (Figure 11(b)) α-helix A also comes in first, but helices B and D come in ahead of C and the domain interface. Notice that in Figure 11(b) the additional cooperativity makes the helices form much more as individual units, but the domain interface remains unaffected forming in a gradual manner (disregarding the difference in free energy due to multiple solutions). To give a complete picture of the folding progress along the entire sequence Figure 12 shows Q_i *versus* residue number, *i*, at various values of Q^* in the high capillarity cooperativity regime. Residues form primarily in helical regions at the lowest Q^* values and also in the domain interface region at moderate Q^* values in the denatured states.

In concert with the Q_i values, the Q_i^{prod} values indicate that helix A structure forms first. The remaining helices and the domain interface region show the dominant structure of the denatured states and form between $Q^* = 0.1$ and $Q^* = 0.4$, although the specific ordering varies between low and high capillarity regimes. Both Q_i and Q_i^{prod} contain very little probability in the β-sheet region until late in the folding process. In most cases helices are frayed towards their ends.

Using static rather than transient measurements, Nash & Jonas (1997) measured hydrogen exchange protection factors in cold denatured, high pressure states of hen lysozyme. Their results are qualitatively similar to those obtained from the inhibition dead times measured by Gladwin & Evans (1996) in early folding intermediates with some additional native-like structure filling in the α-helices and β-domain interface. In particular α-helix D (residues 109-115) and the β-domain interface have much larger protection factors than Evans finds in the transients.

We can describe this experiment as the detection of intermediates at a larger value of the reaction coordinate, *Q*. Figure 11 shows that initially α-helix D has very little contact probability. When the structures are probed at $Q^* = 0.3$, however, α-helix D forms. This agrees with the picture found by Nash & Jonas (1997). In addition, the Q_i^{prod} values from the free energy functional indicate a small region in helix D which becomes protected in denatured states with more native-like structure.

We see that free energy functionals give models of the denatured state structures that are in semi-quantitative agreement with experiments on hen lysozyme and α-lactalbumin as measured by protection factors and NMR. The qualitative picture is

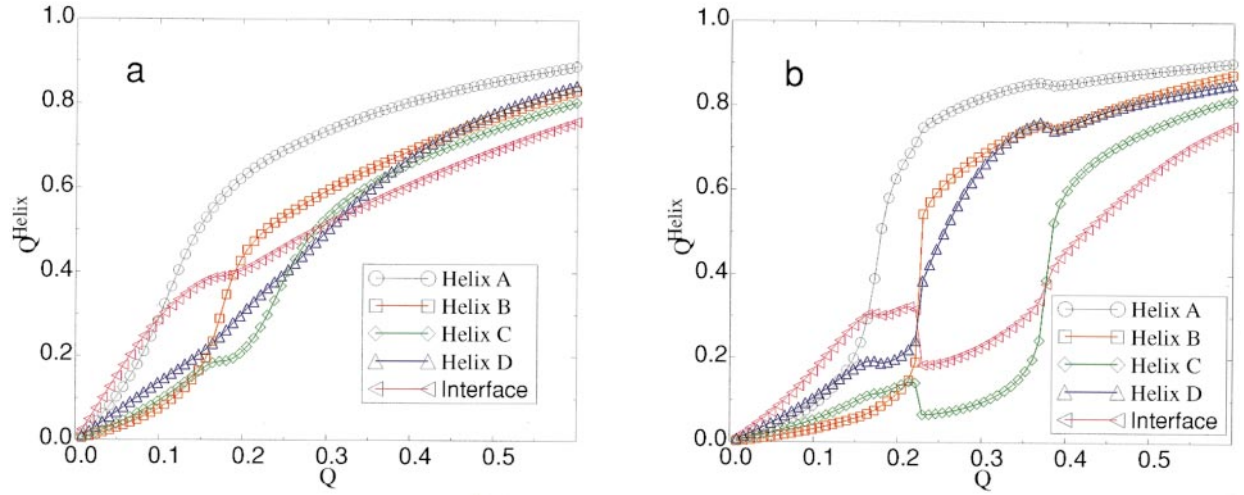

Figure 11. Theoretical titration curves for hen lysozyme. Q_i values are averaged for each of the helices and for the domain interface region at various values of Q^* for the (a) low and (b) high capillarity cases.

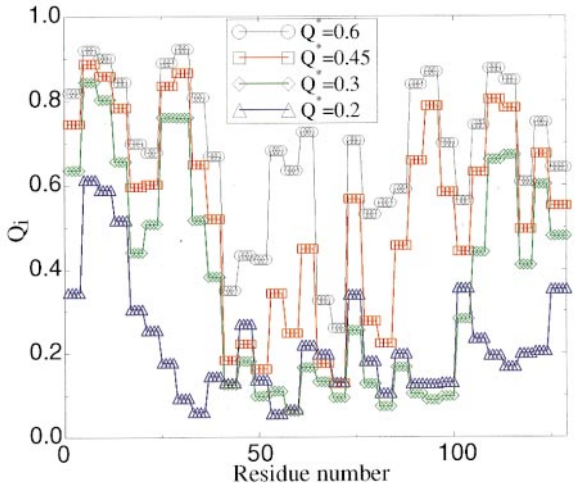

Figure 12. Residue structure formation in a progression down the funnel for hen lysozyme. Q_i is plotted *versus* sequence number at representative values of the folding process in the physically motivated high capillarity cooperativity regime ($\alpha_c = 0.08$, $\alpha_h = 0.58$, $\alpha_{h-\rho} = 0.4$).

that in α-lactalbumin the protected structure seems to be mainly in frayed helices while in hen lysozyme there is protected structure in both the β-domain interface and the helices.

Conclusions

The free energy functional technique provides a versatile way to study the denatured state of protein systems of arbitrary length using modest computational resources. Many quantities can be computed, which can be compared to many types of experiments. In addition the approach yields a detailed view of properties of systems currently too difficult to be measured experimentally.

Free energy functionals also allow many physical effects to be dissected. An example of this is our study of explicit non-additive force effects on the structure in the denatured state. We can contrast the limit of high capillarity cooperativity to the limit of no cooperativity in order to understand the effects that nonadditive forces might have on early structure formation. In the case of α-lactalbumin in the high capillarity cooperativity limit, we see contact probability completely confined to the center of the native protein structure with little structure in the solvent-exposed regions. This is in sharp contrast to the limit of vanishing explicit cooperativity in which there is diffuse contact probability throughout the protein.

The primary alternative to free energy functional based approaches to understanding structural correlations in the denatured ensemble is the use of molecular dynamics or Monte Carlo sampling methods in concert with complete all atom models of the protein (Boczko & Brooks, 1995, 1996; Daggett *et al.*, 1996). Such methods enjoy the advantage that they begin from the familiar starting point of a molecular mechanics Hamiltonian and, in principle, need not utilize any additional statistical mechanical approximations unlike the free energy functional approach. Presently, the main advantage of the free energy functional approach would seem to be its speed which allows a survey of many results. At the moment the computer time limitations of the direct simulation approach are severe, if approaching manageability. Significant conformational motions of a protein are sufficiently slow that all atom molecular dynamics calculations must make do with limited sampling or use complex sampling schemes whose accuracy is hard to assess. For this reason, denatured structures are sampled often by quickly unfolding the protein at highly elevated temperatures (Boczko & Brooks, 1995; Daggett *et al.*, 1996; Lazaridis & Karplus, 1997). Since the properties of both the protein and water are highly temperature dependent, such schemes are viewed by some with skepticism. Also it is important to point out that the averaged statistical quantities which are sought often involve a fine balance of entropy and energy. Small errors in the all atom potential can translate to large errors in the effective interactions between large segments of the protein. Thus the experimentally accessible quantities that average over such units may actually be modeled less accurately than hoped. For this reason we cannot be sure that such schemes are truly any less approximate than the free energy functional approach discussed here which uses coarse grained empirical potentials. Nevertheless, it is heartening that there is reasonably good agreement between the results of free energy functional studies and atomistic molecular dynamics in the one case where a direct comparison has been made that is, for the molecule CI2, discussed by us earlier and in the companion paper (Shoemaker *et al.*, 1999). Beyond this the direct access to free energy quantities possible in the free energy functional approach makes it considerably easier to use for understanding thermodynamics and the extra thermodynamic relations for protein folding kinetics that can be studied *via* protein engineering (Fersht *et al.*, 1992; Itzhaki *et al.*, 1995; Silow & Oliveberg, 1997; Ladurner *et al.*, 1998). In the long run, however, free energy functional and atomistic simulation approaches should not be viewed as being in competition, rather as in the theory of liquids and polymers, the two approaches can be fruitfully combined. It is much easier to determine the parameterization of a free energy functional if reliable simulation data on smaller segments of the entire protein are available. For these smaller regions, sampling problems can be more easily controled. In addition, the residue specific parameters in free energy functionals can be directly parameterized against laboratory experiments of the type discussed in this article. Using widespread site-directed mutagenesis and protein engineering it should be possible to obtain rather reliable effective potentials at the residue level. For

these reasons we believe the free energy functional methods will continue to grow as an important tool in interpreting the nature of the energy landscape of folding proteins.

Acknowledgments

We thank Jin Wang, Chris Dobson, Jiri Jonas, David Nash, John Portman and Shoji Takada for helpful discussions. This work was supported by a grant from the National Institutes of Health, No. PHS 5 R01 GM44557-07.

References

Abkevich, V. I., Gutin, A. M. & Shakhnovich, E. I. (1996). Improved design of stable and fast-folding model proteins. *Fold. Design,* **1**, 221-230.

Boczko, E. M. & Brooks, C. L. (1995). First principles calculation of the folding free energy for a three helix bundle protein. *Science,* **269**, 393-396.

Boczko, E. M. & Brooks, C. L. (1997). Exploring the folding free energy surface of a three-helix bundle protein. *Proc. Natl Acad. Sci. USA,* **94**, 10161-10166.

Bohr, H., Wang, J. & Wolynes, P. G. (1994). Growth of domains in distance geometry through protein folding. In *Protein Structure by Distance Analysis* (Bohr, H. & Brunak, S., eds), pp. 98-109, IOS Press, Amsterdam.

Bryngelson, J. D. & Wolynes, P. G. (1987). Spin glasses and the statistical mechanics of protein folding. *Proc. Natl Acad. Sci. USA,* **84**, 7524-7528.

Bryngelson, J. D. & Wolynes, P. G. (1989). Intermediates and barrier crossing in a random energy model (with applications to protein folding. *J. Phys. Chem.* **93**, 6902-6915.

Bryngelson, J. D. & Wolynes, P. G. (1990). A simple statistical field theory of heteropolymer collapse with applications to protein folding. *Biopolymers,* **30**, 177-188.

Bryngelson, J. D., Onuchic, J., Socci, N. D. & Wolynes, P. G. (1995). Funnels, pathways and the energy landscape of protein folding. *Proteins: Struct. Funct. Genet.* **21**, 167-195.

Burton, R. E., Huang, G. S., Daugherty, M. A., Calderone, T. L. & Oas, T. G. (1997). The energy landscape of a fast-folding protein mapped by ala-¿gly substitutions. *Nature Struct. Biol.* **4**, 305-310.

Calef, D. F. & Wolynes, P. G. (1983a). Smoluchowski-vlasov theory of charge solvation dynamics. *J. Chem. Phys.* **78**, 4145-4153.

Calef, D. F. & Wolynes, P. G. (1983b). Classical solvent dynamics and electron transfer. I. Continuum theory. *J. Phys. Chem.* **87**, 3387-3400.

Calef, D. F. & Wolynes, P. G. (1983c). Classical solvent dynamics and electron transfer. II. Molecular theory. *J. Chem. Phys.* **78**, 470-482.

Cantor, C. R. & Schimmel, P. R. (1971). *Biophysical Chemistry. Part III: The Behavior of Biological Macromolecules*, W. H. Freeman, New York.

Chakrabartty, A., Kortemme, T. & Baldwin, R. L. (1994). Helix propensities of the amino acids measured in alanine-based peptides without helix-stabilizing side-chain interactions. *Protein Sci.* **3**, 843-852.

Chan, H. S. & Dill, K. A. (1990). The effects of internal constraints on the configurations of chain molecules. *J. Chem. Phys.* **92**, 3118-3135.

Chan, H. S. & Dill, K. A. (1997). From levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10-19.

Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430-440.

Fersht, A. R., Matouschek, A. & Serrano, L. (1992). I. Theory of protein engineering analysis of stability and pathway of protein folding. *J. Mol. Biol.* **224**, 771-782.

Finkelstein, A. V. & Badretdinov, A. Y. (1997). Rate of protein folding near the point of thermodynamic equilibrium between the coil and the most stable chain fold. *Fold. Design,* **2**, 115-121.

Flory, P. J. (1956). Theory of elastic mechanisms in fibrous proteins. *J. Am. Chem. Soc.* **78**, 5222-5235.

Frauenfelder, H., Sligar, S. G. & Wolynes, P. G. (1991). The energy landscapes and motions of proteins. *Science,* **254**, 1598-1603.

Gladwin, S. T. & Evans, P. A. (1996). Structure of very early protein folding intermediates: new insights through a variant of hydrogen exchange labelling. *Fold. Design,* **1**, 407-417.

Goldstein, R. A., Luthey-Schulten, Z. A. & Wolynes, P. G. (1992). Optimal protein folding codes from spin glass theory. *Proc. Natl Acad. Sci. USA,* **89**, 4918-4922.

Houry, W. A., Sauder, J. M., Roder, H. & Scheraga, H. A. (1998). Definition of amide protection factors for early kinetic intermediates in protein folding. *Proc. Natl Acad. Sci. USA,* **95**, 4299-4302.

Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor-2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.

Jacobson, H. & Stockmayer, W. H. (1950). Intramolecular reaction in polycondensations. I. The theory of linear systems. *J. Chem. Phys.* **18**, 1600-1606.

Johnson, C. M. & Fersht, A. R. (1995). Protein stability as a function of denaturant concentration: the thermal stability of barnase in the presence of urea. *Biochemistry,* **34**, 6795-6804.

Kiefhaber, T. (1995). Kinetic traps in lysozyme folding. *Proc. Natl Acad. Sci. USA,* **92**, 9029-9033.

Kim, K. S., Fuchs, J. A. & Woodward, C. K. (1993). Hydrogen exchange identifies native-state motional domains important in protein folding. *Biochemistry,* **32**, 9600-9608.

Ladurner, A. G., Itzhaki, L. S., Daggett, V. & Fersht, A. R. (1998). Synergy between simulation and experiment in describing the energy landscape of protein folding. *Proc. Natl Acad. Sci. USA,* **95**, 8473-8478.

Lazaridis, T. & Karplus, M. (1997). ''New view'' of protein folding reconciled with the old through multiple unfolding simulations. *Science,* **278**, 1928-1930.

Leopold, P. E., Montal, M. & Onuchic, J. N. (1992). Protein folding funnels - a kinetic appoach to the sequence structure relationship. *Proc. Natl Acad. Sci. USA*, **89**, 8721-8725.

Luthey-Schulten, Z. A., Ramirez, B. E. & Wolynes, P. G. (1995). Helix-coil, liquid crystal, and spin glass transitions of a collapsed heteropolymer. *J. Phys. Chem.* **99**, 2177-2185.

Miranker, A., Radford, S. E., Karplus, M. & Dobson, C. M. (1991). Demonstration by NMR of folding domains in lysozyme. *Nature,* **349**, 633-636.

Miyazawa, S. & Jernigan, R. L. (1985). Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules,* **218**, 534-552.

Miyazawa, S. & Jernigan, R. L. (1996). Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* **256**, 623-644.

Morita, T. & Hiroike, K. (1961). A new approach to the theory of classical fluids. III. *Prog. Theor. Phys.* **25**, 537-578.

Munakata, T. (1975). Velocity correlation and relative diffusion in simple liquids. *Prog. Theor. Phys.* **54**, 1635-1647.

Muñoz, V. & Serrano, L. (1994). Elucidating the folding problem of helical peptides using empirical parameters. *Nat. Struct. Biol.* **1**, 399-409.

Muñoz, V., Cronet, P., López-Hernández, E. & Serrano, L. (1996). Analysis of the effect of local interactions on protein stability. *Fold. Design,* **1**, 167-178.

Nash, D. & Jonas, J. (1997). Structure of pressure-assisted cold denatured lysozyme and comparison with lysozyme folding intermediates. *Biochemistry,* **36**, 14375-14383.

Neri, D., Billeter, M., Wider, G. & Wüthrich, K. (1992). NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science,* **257**, 1559-1563.

Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z. A. & Socci, N. D. (1995). Towards an outline of the topography of a realistic protein folding funnel. *Proc. Natl Acad. Sci. USA,* **92**, 3626-3630.

Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z. A. & Wolynes, P. G. (1996). Protein folding funnels: the nature of the transition state ensemble. *Fold. Design,* **1**, 441-450.

Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. (1997). Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 539-594.

Perutz, M. (1992). *Protein Structure. New Approaches To Disease And Therapy*, W. H. Freeman, New York.

Plotkin, S. S., Wang, J. & Wolynes, P. G. (1997). Statistical mechanics of a correlated energy landscape model for protein folding funnels. *J. Chem. Phys.* **106**, 2932-2948.

Radford, S. E., Dobson, C. M. & Evans, P. A. (1992). The folding pathway of hen lysozyme involves partially structured intermediates and multiple pathways. *Nature,* **358**, 302-307.

Schiffer, C. A. & van Gunsteren, W. F. (1996). Structural stability of disulfide mutants of basic pancreatic trypsin inhibitor: a molecular dynamics study. *J. Chem. Phys.* **26**, 66-71.

Schulman, B. A., Redfield, C., Peng, Z. Y. & Dobson, C. M. (1995). Different subdomains are most protected from hydrogen exchange in the molten globule and native states of human α-lactalbumin. *J. Mol. Biol.* **253**, 651-657.

Schulman, B. A., Kim, P. S., Dobson, C. M. & Redfield, C. (1997). A residue-specific nmr view of the non-cooperative unfolding of a molten globule. *Nature Struct. Biol.* **4**, 630-634.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1997). Structural correlations in protein folding funnels. *Proc. Natl Acad. Sci. USA,* **94**, 777-782.

Shoemaker, B. A., Wang, J. & Wolynes, P. G. (1999). Exploring structures in protein folding funnels with free energy functionals: The transition state ensemble. *J. Mol. Biol.* **287**, 675-694.

Silow, M. & Oliveberg, M. (1997). Transient aggregates in protein folding are easily mistaken for folding intermediates. *Proc. Natl Acad. Sci. USA,* **94**, 6084-6086.

Socci, N. D., Onuchic, J. N. & Wolynes, P. G. (1996). Diffusive dynamics of the reaction coordinate for protein folding funnels. *J. Chem. Phys.* **104**, 5860-5868.

Wüthrich, K. (1994). NMR assignments as a basis for structural characterization of denatured states of globular proteins. *Curr. Opin. Struct. Biol.* **4**, 93-99.

Wolynes, P. G. (1997). Folding funnels and energy landscapes of larger proteins within the capillarity approximation. *Proc. Natl Acad. Sci. USA,* **94**, 6170-6175.

Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. (1995). Navigating the folding routes. *Science,* **267**, 1619-1620.

Woodward, C. K. (1994). Hydrogen exchange rates and protein folding. *Curr. Opin. Struct. Biol.* **4**, 112-116.

Wu, L. C. & Kim, P. S. (1998). A specific hydrophobic core in the α-lactalbumin molten globule. *J. Mol. Biol.* **280**, 175-182.

Yee, D. P., Chan, H. S., Havel, T. F. & Dill, K. A. (1994). Does compactness induce secondary structure in proteins? A study of poly-alanine chains computed by distance geometry. *J. Mol. Biol.* **241**, 557-573.

Zwanzig, R. (1972). Collective modes in classical liquids. In *Proceedings of the Sixth IUPAP Conference on Statistical Mechanics* (Rice, S. A., Freed, K. F. & Light, J. C., eds), University of Chicago, Chicago.

Edited by A. R. Fersht

(Received 6 October 1998; received in revised form 5 February 1999; accepted 5 February 1999)