FEBS Letters 576 (2004) 468–476 FEBS 28891

A flexible approach for understanding protein stability

D.R. Livesay[a,1], S. Dallakyan[b,1], G.G. Wood[b], D.J. Jacobs[b,*]

[a]Department of Chemistry, California State Polytechnic University, Pomona, 3801 W Temple Ave, Pomona, CA 91768, USA
[b]Department of Physics and Astronomy, California State University, Northridge, 18111 Nordhoff Street, Northridge, CA 91330-8268, USA

Received 19 August 2004; accepted 20 September 2004

Available online 6 October 2004

Edited by Robert B. Russell

Abstract A distance constraint model (DCM) is presented that identifies flexible regions within protein structure consistent with specified thermodynamic condition. The DCM is based on a rigorous free energy decomposition scheme representing structure as fluctuating constraint topologies. Entropy non-additivity is problematic for naive decompositions, limiting the success of heat capacity predictions. The DCM resolves non-additivity by summing over independent entropic components determined by an efficient network-rigidity algorithm. A minimal 3-parameter DCM is demonstrated to accurately reproduce experimental heat capacity curves. Free energy landscapes and quantitative stability-flexibility relationships are obtained in terms of global flexibility. Several connections to experiment are made.
© 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.

Keywords: Protein stability; Conformational flexibility; Heat capacity; Free energy decomposition; Transition state; Network rigidity

1. Introduction

In recent years, insight about protein flexibility [1], unfolding pathways [2], nucleation processes [3] and folding cores [4] has been obtained using FIRST (Floppy Inclusion and Rigid Substructure Topology). By modeling protein structure as a mechanical framework of distance constraints that represent microscopic interactions (i.e., covalent bonds, hydrogen bonds, etc.), FIRST provides quantitative mechanical stability measures based on network rigidity calculations [5]. However, nature must delicately balance protein conformational flexibility with thermodynamic stability. For example, a functioning enzyme must be flexible enough to mediate a reaction pathway, rigid enough to support specificity [6–9], and do both in a thermodynamically stable state. Building upon FIRST, a computational methodology that directly relates protein stability to conformational flexibility is presented. The approach, called the Distance Constraint Model (DCM), restores the efficacy of free energy decompositions [10–12] by rigorously accounting for non-additivity of component entropies using network rigidity [13,14]. Working directly with free energies, the DCM is more than 10^{10} times faster than standard molecular dynamics simulations, but not without precedence. The DCM resembles COREX [15] in that both approaches connect free energy decomposition directly to native structure while considering ensembles of fluctuating native-like and disordered topologies. Specifically, only native contacts are considered, but they are allowed to break to account for disordered topologies. This simplification makes the calculation tractable in practical computing times. A key advantage of the DCM, compared to all prior methods, is that network rigidity [1], an inherently long-range mechanical interaction [16], is explicitly calculated to model enthalpy–entropy compensation in a way that resolves the long-standing problem of non-additivity of component entropies.

The DCM has recently been applied to polypeptides undergoing normal [13] and inverted [14] α-helix to coil transitions, where exact transfer matrix methods are employed. In this letter, the DCM is applied to proteins using a mean-field treatment [17] akin to Landau theory. This approach provides a flexible modeling paradigm that can be custom-tailored to predict a variety of phenomena (i.e., stability, binding, folding kinetics, etc.). We demonstrate the utility of the approach by employing a minimal 3-parameter DCM. Despite its simplicity, measured heat capacity curves are reproduced across a heterogeneous protein dataset. Temperature-dependent flexibility measures and free energy as a function of a global flexibility order parameter are calculated. Using a global flexibility order parameter as an unfolding reaction coordinate, the transition state is identified. The location of the mechanical and thermodynamic transitions along the unfolding pathway allows inferences regarding transition state compactness to be made.

2. The distance constraint model

Covalent bonding is modeled as quenched distance constraints. Non-covalent interactions are modeled as fluctuating constraints. A component enthalpy and entropy (H_i, S_i) is assigned to each constraint. Entropy assignment is used as a measure for the strength of a constraint, where weaker constraints correspond to greater entropy. A framework is defined by a specification of constraints with a particular topological arrangement. For framework $\mathscr{F}$, let $n_i(\mathscr{F}) = (1,0)$ when the ith constraint is/is not present. The free energy of a framework is $G(\mathscr{F}) = H(\mathscr{F}) - TS_c(\mathscr{F})$, where enthalpy, $H(\mathscr{F})$, and conformational entropy, $S_c(\mathscr{F})$, are determined by:

$$H(\mathscr{F}) = \sum_i H_i n_i(\mathscr{F}) \text{ and } S_c(\mathscr{F}) = \sum_i S_i I_i(\mathscr{F}) \quad (1)$$

Except for long-range electrostatic contributions, the conformational entropy term $S_c(\mathscr{F})$ in principle accounts for solvent

* Corresponding author.
E-mail address: donald.jacobs@csun.edu (D.J. Jacobs).

[1] These authors contributed equally to this work.

0014-5793/$22.00 © 2004 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2004.09.057

effects implicitly through component enthalpy and entropy contributions [18]. However, hydration is not explicitly modeled in the implementation presented here, although this has been done successfully for polypeptides undergoing cold and heat denaturation [14]. For each framework, Eq. (1) expresses total enthalpy as an additive property, but total entropy is calculated as a sum over independent constraints reflecting non-additivity in entropy [10,11]. Naively summing all entropy components generally results in a grave overestimate. Graph-rigidity algorithms [1,5,13,14,16] are employed to identify the independent constraints. Variable I_i gives the number of independent distance constraints within the ith constraint. Note that a constraint may contain one or more elementary distance constraint(s). Unlike a normal mode vector space, a complete set of independent constraints generally does not constitute an orthogonal decomposition. Therefore, multiple answers for total entropy result for different independent constraint sets, but all give upper bounds for the true entropy. A preferential set of independent constraints, given by $\{I_i\}$ in Eq. (1), is calculated by selecting stronger constraints to be independent before weaker constraints, thereby obtaining the lowest upper bound. This approximation appears adequate for practical applications [13,14,17].

As fluctuating constraints break and form, distinct frameworks appear. Over an ensemble of accessible frameworks, Eq. (1) expresses an enthalpy–entropy compensation mechanism. Within a collection of energetically favorable constraints, rigid substructures having low energy and entropy will form in regions of high constraint density. Regions of low constraint density will be flexible, having higher energy and entropy. Placing a weak constraint in a rigid region lowers energy but not entropy. The loss in conformational entropy is limited to the cost of forming a rigid substructure. Compensation arises when a strong constraint breaks permitting a redundant (weaker) constraint to become independent, yielding a net increase in energy and entropy. This compensation mechanism yields a natural nucleation process because a group of constraints acting together is harder to melt compared to what would be expected if their individual free energies were simply added. Thus, non-additivity in entropy leads directly to *molecular cooperativity*. Rigid substructures generally form at low temperature, spontaneously breaking apart when temperature increases to a point where the entropic penalty is too great to maintain. The verdict specifying if local regions are more stable as rigid or flexible strongly depends on constraint topology and temperature. Therefore, a partition function is defined over an ensemble of topologically distinct frameworks, from which thermodynamic averaged macroscopic quantities are derived.

A minimal three free-parameter DCM is briefly described, with full description found elsewhere [17]. Given a known protein structure, native constraints are identified. Torsion interactions are modeled as fluctuating constraints representing conformations in native or disordered geometries. When present, the αth H-bond (salt bridges are modeled as special types of H-bonds) contributes energy E_α, otherwise formation of alternative H-bond to solvent contributes energy u. The energy E_α used in this work is a geometry dependent empirical potential [19] ranging from 0 to –8 kcal/mol. The entropy per distance constraint is set as $S_\alpha = 1.986R(1 + E_\alpha/8)$, where R is the ideal gas constant and E_α is measured in kcal/mol. Based on prior works on polypeptides [13,14], this linearity assumption is not necessary, but it offers a convenient way to express the general property that as the well depth becomes shallow the constraint will become weaker. Torsion constraints fluctuate between native-like or disordered, respectively, having (energy, entropy) components of (v, $R\delta_{\mathrm{nat}}$) and (0, 2.56R). Parameterization of covalent bonds is unnecessary as they do not fluctuate [13,15]. Per protein $\{u,v,\delta_{\mathrm{nat}}\}$ is determined by fitting to thermodynamic data, such as heat capacity. This three-parameter model is a derivative of a five-parameter model, where the values 1.986 and 2.56 were previously obtained by simultaneously fitting to 6 heat capacity curves of ubiquitin and histidine binding protein [17].

A free energy landscape is defined in a two-dimensional constraint space. As shown schematically in Fig. 1, the numbers of native-like torsion, N_{nt}, and H-bond, N_{hb}, constraints specify a macrostate of the protein. For each node (N_{hb}, N_{nt}), a Landau free energy function is expressed as:

$$G(N_{\mathrm{hb}}, N_{\mathrm{nt}}) = U(N_{\mathrm{hb}}) - uN_{\mathrm{hb}} + vN_{\mathrm{nt}} - T(S_{\mathrm{c}}(\delta_{\mathrm{nat}}) + S_{\mathrm{mix}}) \qquad (2)$$

where $U(N_{\mathrm{hb}})$ is the average total intramolecular H-bond energy, S_{mix} is the mixing entropy for the number of ways to have N_{hb} H-bonds, and N_{nt} native-torsion constraints. Mean-field probabilities for H-bonds to be present or not, or torsion constraints to be native-like or disordered, are expressed as independently distributed variables that are calculated self-consistently treating each constraint as a subsystem. For each node, mean-field probabilities are calculated and then used for Monte Carlo sampling to generate an ensemble of frameworks. Results from graph-rigidity calculations for each framework in a node are averaged, where Eq. (1) gives, $S_{\mathrm{c}} = \sum_\alpha S_\alpha\langle I_\alpha\rangle + \delta_{\mathrm{nat}}R\langle I_{\mathrm{nat}}\rangle + 2.56R\langle I_{\mathrm{dis}}\rangle$. Other terms in Eq. (2) are analytically calculated. In the minimal DCM presented here, only network rigidity provides interaction between constraints via the set

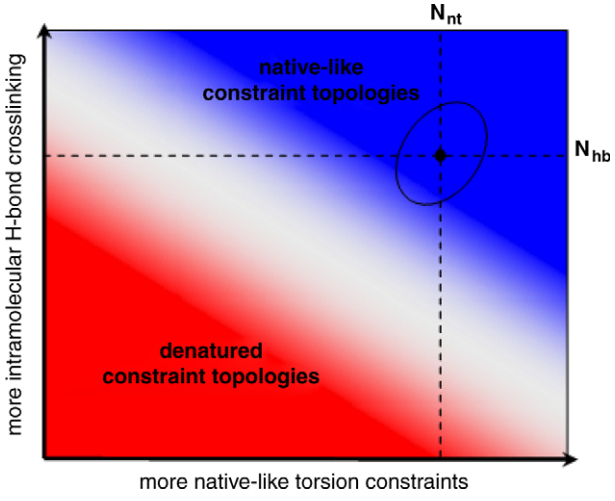

Fig. 1. Schematic DCM free energy landscape in constraint space. The average number of H-bonds, N_{hb}, and native-like torsion constraints, N_{nt}, defines a grid of nodes. Within each node, an ensemble average over frameworks is performed. The highlighted node (intersection of dashed lines) is part of an ensemble characterizing a functional protein (defined by the oval). Other frameworks are too flexible (white to red regions) or too rigid (blue regions) to function optimally. There will be a range of temperatures where the minimum in free energy is within the oval. On unfolding, the native protein must pass through the white (marginally mechanically stable) regions.

$\{I_i\}$. Although limited in prediction capability by its simplicity, the minimal DCM clearly demonstrates the physically important role of entropy non-additivity, as network rigidity is the only cooperative interaction presently modeled.

A flexibility order parameter, characterizing global flexibility, is defined as:

$$\theta \equiv \frac{\langle I_{\mathrm{dis}} \rangle}{N_{\mathrm{res}}} = \frac{(\text{average number of independent disordered torsion constraints})}{(\text{number of residues in protein})} \quad (3)$$

The flexibility order parameter effectively measures the number of biologically relevant degrees of freedom, given by $\langle I_{\mathrm{dis}} \rangle$. That is, disordered torsion constraints imply partial unfolding and greatest possible conformational flexibility. All nodes in constraint space (Fig. 1) with the same θ are grouped together to obtain a free energy function, $G(T,\theta)$, in terms of the flexibility order parameter. $G(T,\theta)$ serves the important role of directly relating protein thermodynamic stability to global flexibility. At $\theta = 0$, the protein is 100% rigid, too rigid to be functional. The native-like ensemble is typically centered at $\theta \approx 1 \pm 0.4$. When θ is large, the protein is globally disordered and regarded as unfolded. Generating atomic coordinates for conformations is unnecessary because graph-rigidity only requires connectivity information – *the crucial factor making the DCM computationally fast*. Local variations in rigidity and flexibility are quantified as ensemble averaged measures. The DCM is easily adaptable to different free energy decompositions and constraint types to explicitly account for other important effects.

3. Bridging network rigidity and thermodynamics

We demonstrate remarkable generality of the minimal DCM by fitting to heat capacity curves for a structurally and functionally diverse protein dataset (see Table 1). Over great structural diversity, three free parameters reproduce essential features of heat capacity. Fig. 2A shows four typical best fits to differential scanning calorimetry data (remaining cases are provided in supplementary material). To our knowledge, no other all-atom model, nor free energy decomposition scheme, has reproduced entire heat capacity curves. Note that Hedwig and Hinz [20] have successfully used free energy additivity to obtain heat capacities of unfolded proteins, which is naturally explained by the small percentage of redundant constraints present in unfolded conformations. Table 1 lists best-fit parameters for our exemplar protein dataset.

After $\{u,v,\delta_{\mathrm{nat}}\}$ are determined, the Landau free energy, $G(T,\theta)$, is calculated, as shown in Fig. 2B for lysozyme at four temperatures as a typical case. The most important feature of the free energy landscape is that at the melting point, T_{m}, there are two minima, indicative of a first-order (two-state) phase transition. Enthalpy–entropy compensation is demonstrated over these temperatures as the competition between energetically favorable cross-linking interactions and increased degrees of freedom (dof) is born out. In most cases, the minima are generally not of equal depth at T_{m} (although they are very close in lysozyme). Because $G(T,\theta)$ provides more information than whether the protein is merely folded or not, multiple θ values map onto the native or unfolded two-state model. In simple terms, T_{m} does not correspond to two minima of precisely equal depth whenever the two local wells have different shape. The scale of $G(T,\theta)$ is ≈8–13 kcal/(mol residue), or ~1000 kcal/mol for lysozyme. Fig. 2C shows that the DCM is capturing small stability differences (0–3 kcal/mol) between large absolute values. Raising temperature above T_{m} increases the relative concentration of unfolded (vs. folded) protein, eventually eliminating all traces of native structure. Lowering the temperature below T_{m} has the opposite effect.

The three phenomenological parameters, which account for a wide array of protein diversity, are reminiscent of the Lifson–Roig model [21] for helix–coil transitions. In the DCM, however, substructure nucleation is an explicit outcome of network rigidity calculations that depend on the crosslinking constraint topology [13,14]. In the minimal DCM, cooperativity is only mediated through the H-bond network, which governs fluctuating rigid and flexible regions. Although too simple for complete parameter transferability, all best-fit parameters are physically reasonable (see Table 1). It is worth mentioning that arbitrarily rescaling the measured heat capacity curves by a factor of 2 results in not being able to fit to the heat capacity data, which implies a physically sound model. Moreover, the parameters are found to be somewhat transferrable across homologous proteins, meaning *blind* predicted heat capacity

Table 1
Proteins investigated in this work with parameter values

Protein	SCOP class	# res	pH	u	v	δ_{nat}	BH[a]	θ_{nat}[b]	θ_{TS}	θ_{RP}	θ_{den}	LSE[c]
Protein G [42]	$\alpha+\beta$	56	6.0	−1.88	−0.65	1.18	1.80	0.87	1.48	1.20	2.18	0.061
BPTI [43]	Small	58	4.3	−1.83	−0.98	0.83	2.64	0.82	1.39	1.39	2.10	0.085
CSP [44]	β	66	4.0	−2.18	−0.66	1.80	0.80	0.68	1.07	1.28	1.41	0.018
			8.0	−2.86	−0.97	1.85	1.07	0.50	0.86	1.05	1.13	0.022
Ubiquitin [45]	$\alpha+\beta$	76	2.0	−1.76	−0.44	1.60	1.14	1.38	1.66	1.75	2.15	0.000
			4.0	−2.01	−0.82	1.60	1.29	1.02	1.29	1.43	1.81	0.001
Fibronectin [46]	β	91	5.0	−2.45	−0.89	1.69	0.91	0.76	1.02	0.98	1.29	0.080
Lysozyme [47]	$\alpha+\beta$	130	3.0	−1.98	−0.59	1.18	2.71	1.15	1.65	1.04	2.24	0.004
RNase H [48]	α/β	146	5.5	−2.08	−1.04	1.35	2.82	0.67	1.05	1.10	1.54	0.030
GDH domain 2 [49]	α/β	150	6.5	−1.52	−0.58	1.16	2.72	0.98	1.24	1.32	1.64	0.004
HIV protease [50]	β	198	3.4	−1.52	−0.35	1.85	0.61	1.11	1.31	1.40	1.57	0.009
HBP (apo) [51]	α/β	238	8.3	−1.91	−0.64	1.42	6.15	0.99	1.30	1.14	1.80	0.016
HBP (bound) [51]	α/β	238	8.3	−2.23	−0.86	1.24	13.3	0.84	1.20	1.26	1.84	0.027

[a] BH = barrier height at T_{m} normalized by RT_{m}.
[b] θ_{nat}, θ_{TS}, θ_{RP} and θ_{den} = global flexibility order parameter values corresponding to the native structure, transition state, rigidity percolation threshold and denatured state at T_{m}.
[c] Least squares fitting error per residue normalized by experimental C_{P} peak height.

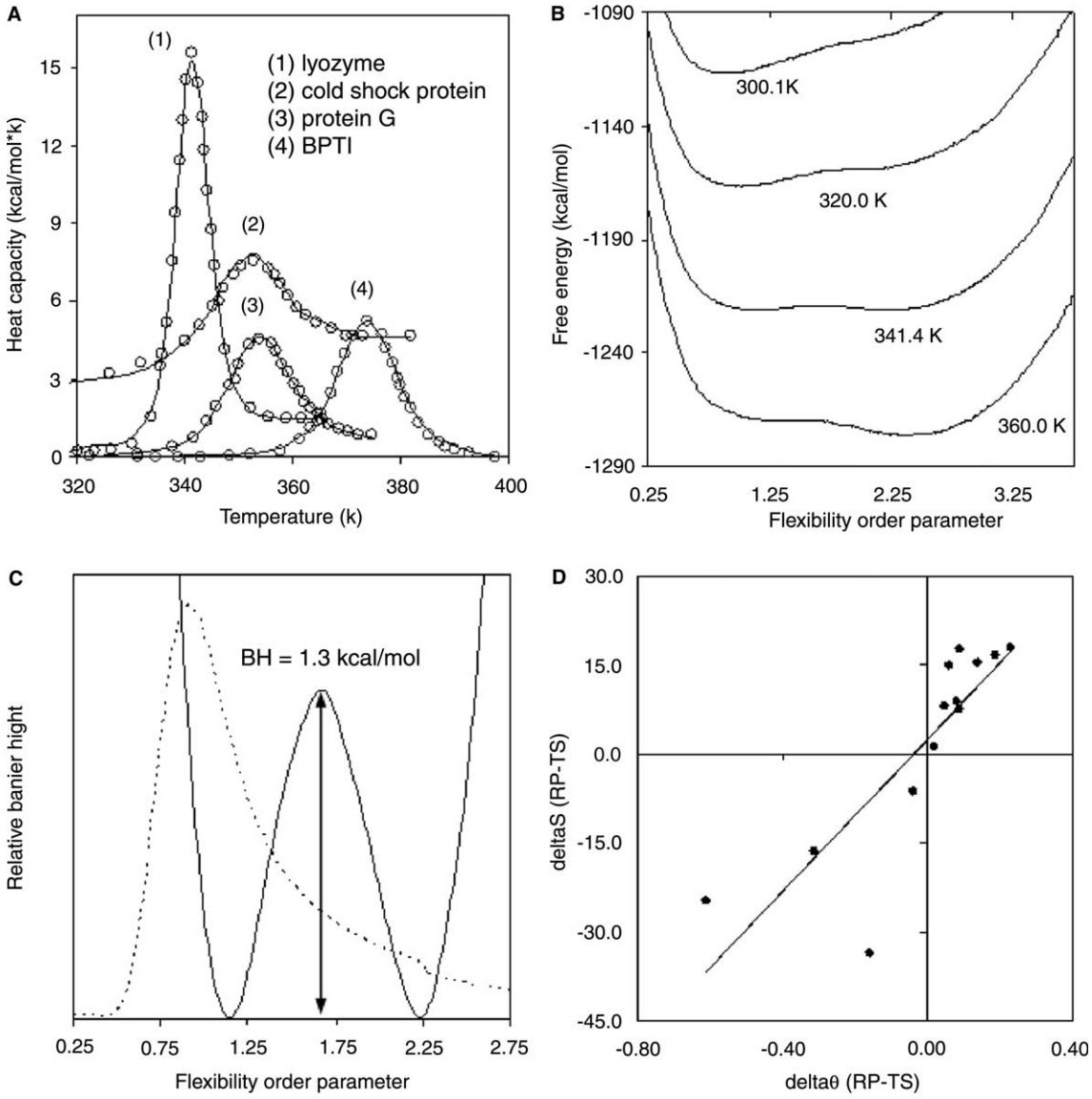

Fig. 2. (A) Typical best-fits to four heat capacity curves. (B) Landau free energies for lysozyme where $T_m = 341.4$ K. (C) Magnified look at $G(T,\theta)$ (solid line) for lysozyme at T_m highlighting the barrier separating the two phases. The rigid cluster size susceptibility for lysozyme (dashed line). The susceptibility, denoted as $rcsRM_2$, is defined as the 2nd moment of the size of rigid clusters with the biggest cluster size excluded (i.e., reduced). The peak in $rcsRM_2$ locates a percolation threshold, even in finite size systems. (D) $T_m\Delta S_{RP-TS}$ vs. $\Delta\theta_{RP-TS}$. The enthalpic and entropic (shown) portion of the free energy decomposition has positive correlation to the difference in the location of the mechanical and thermodynamic transitions. Correlation coefficient = 0.84.

curves have peaks typically shifted by ±30 °C with ±30% error in peak height, while same qualitative Landau free energy curves are obtained. However, transferring parameters between unrelated structures generally does not produce a transition within the temperature range between 0 and 100 °C. The generality of this minimal DCM, combined with the transferrable energy and entropy parameterization for the intramolecular H-bonds, strongly supports the idea that the hydrogen bond network is a dominant factor affecting heat capacity [22].

The three values of the flexibility order parameter $\{\theta_{nat}, \theta_{den}, \theta_{TS}\}$ characterize $G(T_m,\theta)$, where the first two, respectively, correspond to local minimum for the native and denatured states, and the third corresponds to the straddling local maximum. Note that the native and denatured states, $\{\theta_{nat}, \theta_{den}\}$,correspond to most probable sub-ensembles, which are weighted differently depending on temperature. Thus, change in constraint topology is driven by temperature, which changes the flexibility profile (Fig. 3). Furthermore, flexibility is quantified by rigidity susceptibility curves (see Fig. 2C) that characterize the mechanical transition in the protein by the amount of fluctuation in the rigid cluster size. The peak signifies the point, called the rigidity percolation threshold (θ_{RP}), where there is maximum rigid cluster size fluctuation. The rigidity threshold corresponds to when the protein transitions from predominantly one large rigid cluster to many smaller ones [16].

In previous investigations, Thorpe and co-workers [2,3] use a H-bond dilution to produce the rigidity transition and model folding kinetics. It was shown that the rigid cluster decomposition correlates well to folding pathways. Here, the

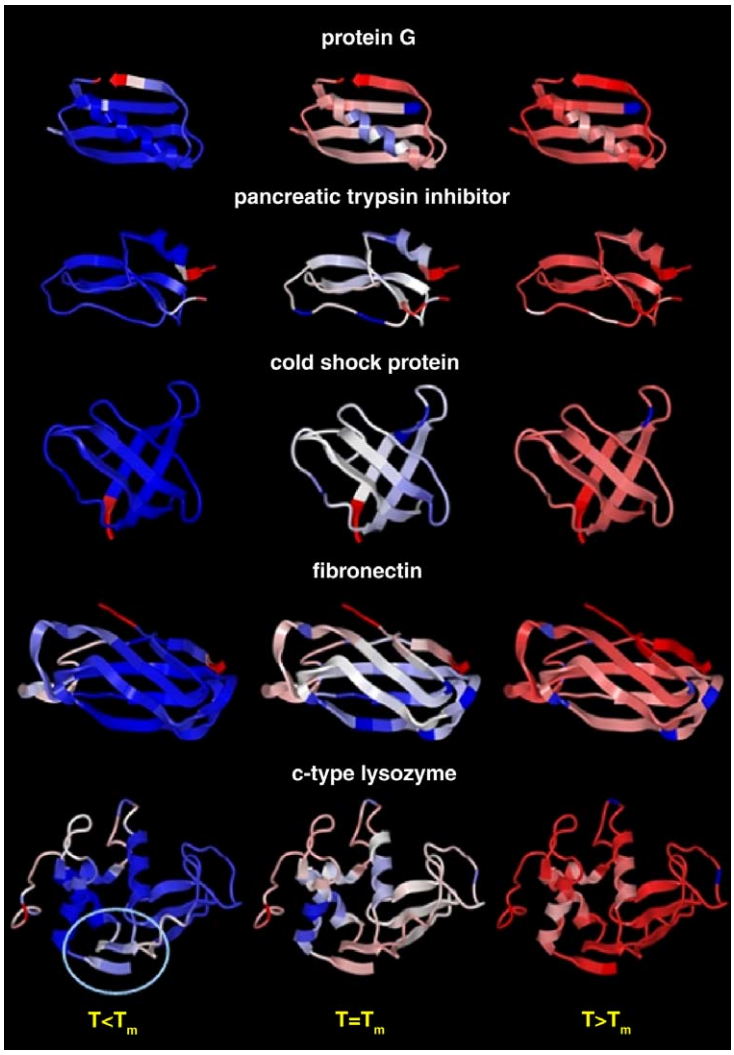

Fig. 3. Backbone flexibility at temperatures below, equal to, and above the melting point is shown. (Blue, red) colors indicate more (rigid, flexible) structures. The proteins are organized from smallest to largest (top to bottom). The temperatures of c-type lysozyme correspond to the highest three Landau free energies shown in Fig. 2B; the region highlighted by the oval corresponds to the hinge-bending motion in lysozyme.

rigidity transition appears as an equilibrium property. Both views compliment each other, and together they suggest that network rigidity provides a direct link between protein structure, stability, flexibility and kinetics. Although a single reaction coordinate, or progress variable, may be insufficient to describe all folding processes [23], we assume that the flexibility order parameter provides an operational reaction coordinate, and θ_{TS} defines the transition state. Most progress variables are equally good when the folding funnel is minimally frustrated [24]. Nevertheless, several other possibilities have been considered, such as number of H-bonds. Empirical results (not shown) suggest that θ is special because all proteins studied to date exhibit semi-universal behavior in all quantities calculated only when plotted against θ. It is encouraging that the predicted barrier height of known fast folders (e.g., ubiquitin, cold shock protein, and fibronectin [25]) is much lower than some of the other two-state proteins investigated. Surprisingly, HIV protease is predicted to have the smallest barrier. Realistically, HIV protease is unlikely to have such a low barrier. However, these results are in qualitative agreement with the observed shallow folding landscape of HIV protease [26]. Additionally, we have successfully fit heat capacity [27] for met-myoglobin, but did not find two-state behavior (unpublished). The connections between $G(T,\theta)$ and kinetics, and in particular comparing barrier heights with two-state kinetics will be explored in subsequent work after hydrophobic interactions are explicitly modeled. Here, emphasis is on the model independent result that the barrier in $G(T_m,\theta)$ at θ_{TS} is distinctly different from the rigidity transition at θ_{RP}. In all examples, except for lysozyme, θ_{TS} and θ_{RP} tend to parallel each other, both occurring between θ_{nat} and θ_{den}, but are seldom exactly the same. A comparison of the rigid cluster susceptibility and $G(T_m,\theta)$ is shown for lysozyme in Fig. 2C. Deviations between θ_{TS} and θ_{RP} are ascribed to entropy effects having important consequences on folding and unfolding kinetics depending on whether θ_{TS} occurs before or after θ_{RP} along the folding pathway. We suggest that when

$\theta_{RP} > \theta_{TS}$ the transition state is compact (as it is overall mechanically rigid), otherwise it is voluminous, being much more flexible.

3.1. Flexibilty predictions

Many quantitative characteristics of conformational flexibility are calculated (i.e., probability for a residue to be disordered, rigid cluster decomposition and rigid cluster size statistics). A key measure is the flexibility index, defined in prior work [1], except, the number of independent dof in FIRST is synonymous to the number of independent disordered torsion constraints in the DCM. Within a (flexible, rigid) region, local density of independent dof and redundant constraints is calculated as (ρ_{dof}, ρ_{rdc}). Isostatic rigid regions have $\rho_{dof} = \rho_{rdc} = 0$. The flexibility index is defined as the ensemble averaged difference ($\rho_{dof} - \rho_{rdc}$). In Fig. 3, the flexibility index for five proteins at temperatures below, at, and above T_m is shown. Not surprising, the backbone has limited flexibility below T_m and is very flexible at high temperature. Nevertheless, the denatured ensemble retains fluctuating rigid substructures, which indicates that a random coil is thermodynamically unstable.

The S^2 order parameters from NMR and X-ray structure temperature factors are frequently used to identify flexible regions within protein structures (for example, see [28,29], respectively). Table 2A compares DCM ubiquitin flexibility predictions to two popular theoretical flexibility models: the Gaussian network model (GNM) [30] and the athermal FIRST [1] network rigidity calculation. The DCM (see Fig. 4) and GNM perform nearly equally well; both methods do better than FIRST. While encouraging, this result should be taken

Table 2
Comparisons of theoretical model predictions and experimental results

	DCM		FIRST [1]		GNM [30]		Zhang [32]	
(A) Correlation coefficients between model and experimental results								
$1 - S^2$ [52]	0.87		0.70		0.89[a]		0.98	
B-factors [53]	0.85		0.64		0.72		n/a	
	Frag. 1	Frag. 2	Frag. 3	Frag. 4	Frag. 5	Frag. 6	Frag. 7	Frag. 8
(B) H/D exchange vs. average probability to rotate[b]								
Experiment [34]	30.1	59.9	45.2	35.7	100.0	43.2	37.7	66.1
DCM	32.3	15.0	21.8	13.2	80.5	24.7	34.4	100.0

[a] GNM translational mobility was used, not rotational mobility introduced in [31].
[b] DCM values are the average probability to rotate over the corresponding fragment. To facilitate comparisons, the raw data have been normalized such that the fastest exchange (experiment) and most flexible (model) regions equal 100.

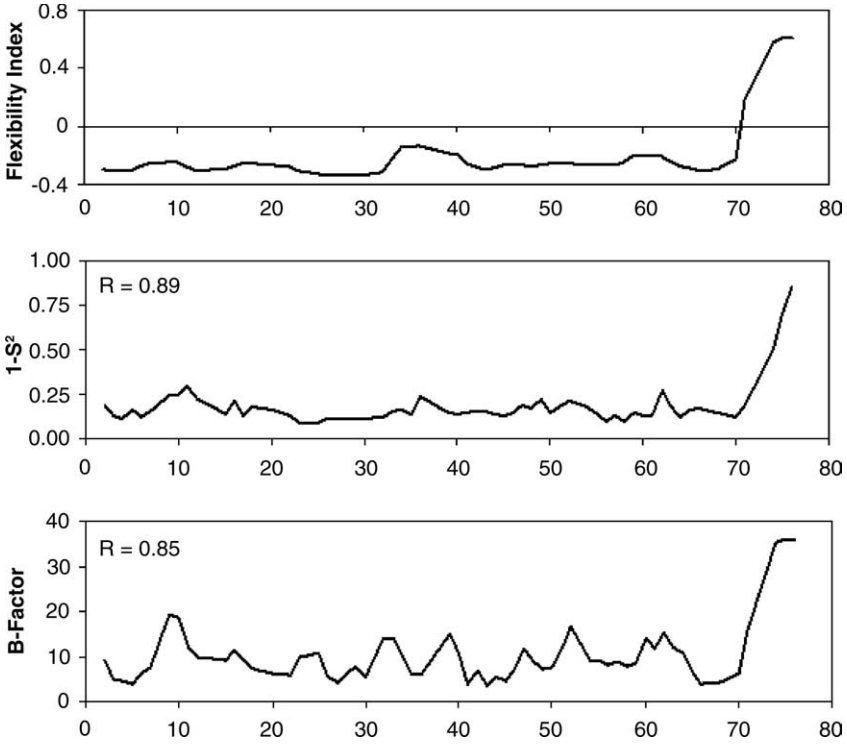

Fig. 4. DCM flexibility index for ubiquitin compared to NMR S^2-order parameters ($1 - S^2$) and X-ray crystallography temperature factors (B-factors). The S^2 data are incomplete; in all three cases, only positions with reported [52] S^2 values are plotted. Correlation coefficients quantify the similarity between the DCM flexibility index and respective experimental data.

with caution because S^2 order parameters are generally associated with fast (~ns) fluctuations [28]. The GNM, FIRST, and as a consequence, the DCM are explicitly designed to predict long timescale quantities. Haliloglu and Bahar [31] have demonstrated that S^2 values can be transposed into a reflection of rotational (vs. translational) mobilities, which should correlate better. On the other hand, it is possible to predict S^2 values accurately with relatively simple contact models. In the case of ubiquitin, Zhang and Bruscweiler [32] report a very high correlation coefficient using such an approach that reproduces the experimental data better than all three of long timescale models compared.

X-ray crystallography temperature factors (B-factors) are also frequently thought of as indicators of flexibility. However, problems arise because B-factors are better indicators of mobility, reflecting a continuum of timescales, crystal lattice packing, crystal quality, etc. Nevertheless, temperature factors are commonly used to qualitatively benchmark flexibility predictions. Fig. 4 also compares ubiquitin B-factors to the DCM flexibility index. The results correlate well, slightly better than GNM and FIRST. Comparison with all other proteins in our dataset (data not shown) indicates that the three methods perform approximately the same, with a slight advantage to the DCM. The calculated correlation coefficients between the DCM flexibility index and experimental B-factors range from no correlation ($R = 0.03$ for protein G) to ubiquitin as a best case ($R = 0.85$). *Note*: None of the three models are able to correlate to protein G B-factors or S^2 values [33]. Overall, these comparisons indicate that DCM flexibility predictions are in line with commonly used long timescale flexibility predictors.

Hydrogen–deuterium (H/D) exchange results are frequently interpreted in the context of longer timescale motions. Complicating these interpretations, however, are solvent accessibility issues. In a rigid substructure, completely exposed residues allow exchange, but buried residues do not facilitate exchange until the region opens up, which is related to flexibility. Ignoring solvent accessibility factors, a naive comparison (Table 2B) is made to H/D exchange within ubiquitin fragments [34]. In the experimental investigation, ubiquitin is incubated in heavy water, allowing exchange to occur, and then fragmented. Afterward, electrospray mass spectrometry (ESI-MS) is employed to quantify the percent deuterium in eight individual fragments. Averaging the DCM flexibility index of each fragment and comparing experimental data to this yields a low correlation ($R = 0.41$). However, the probability to rotate, which is a unique flexibility measure provided by DCM (not FIRST), gives a modest correlation ($R = 0.69$). Both DCM measures correctly predict the two fastest exchanging fragments as the most flexible. Not surprisingly, discrepancies with experimental data occur in the more rigid and less solvent exposed portions of the protein.

On face value, comparison of DCM flexibility measures (and FIRST and GNM results for that matter) to experimental results is disappointing. However, it can be argued that the majority of the discrepancies result because of improper comparison between quantities that are at best only semi-related. For example, experimental S^2 values, B-factors and H/D exchange data themselves generally do not correlate well. It is therefore prudent to incorporate better descriptors within the DCM to predict different types of physical measurements. Since the approach presented here is based on using native structure as input, any descriptor based on native-like contacts can be incorporated in future implementations of the DCM.

4. Relating protein flexibility and stability

The utility of quantified flexibility-stability measures is illustrated using human c-type lysozyme as the center of discussion. While it is known that lysozyme folds via a kinetic intermediate [35,36], two-state models are frequently employed for simplicity. For example, all experimental heat capacity measurements using DSC assume a two-state model. Our inability to obtain two-state behavior in myoglobin encourages us that more sophisticated implementations of the DCM will discriminate lysozyme intermediates as well. From Table 1, $\theta_{\mathrm{nat}} = 1.15$ for lysozyme at T_{m}, indicating that there are ≈11 independent dof for every 10 residues. Relative to all other proteins studied ($\langle\theta_{\mathrm{nat}}\rangle = 0.88$ at their respective T_{m}), lysozyme is generally more flexible protein in its native state at T_{m}. At functional temperatures, Fig. 3 reveals that the backbone is almost completely rigid. The ambient flexibility is mainly due to sidechain motions, specifically side-chains not participating in the hydrogen bond network. The predicted backbone rigidity has been demonstrated experimentally [37]. As temperature is lowered, more dof become quenched until the protein eventually becomes too rigid to function, as schematically depicted in the upper-right corner of Fig. 1. Upon heating, the number of independent dof that become available approximately doubles ($\theta_{\mathrm{den}} \approx 2.2$) with dramatic increase in backbone flexibility. The unfolding free energy barrier at T_{m}, located at $\theta_{\mathrm{TS}} = 1.65$, is found to be 1.3 kcal/mol. This free energy barrier is decomposed into enthalpy and entropy contributions, using the DCM calculations for $H(T,\theta)$ and $S(T,\theta)$. The enthalpic (H) and entropic (TS) contributions to the lysozyme barrier upon unfolding are found to be 68.0 and 66.7 kcal/mol, respectively, which quantifies the relative degree of enthalpy–entropy compensation.

Based on rigid cluster statistics above and below the rigidity transition, rigid substructures fluctuate at the transition state greatly when $\theta_{\mathrm{TS}} > \theta_{\mathrm{RP}}$. Conversely, when $\theta_{\mathrm{TS}} < \theta_{\mathrm{RP}}$, well formed rigid regions presumably foreshadow native structure. For lysozyme, finding $\theta_{\mathrm{RP}} = 1.04$ precedes $\theta_{\mathrm{TS}} = 1.65$ leads to the prediction that its transition state is voluminous, consisting of many small rigid clusters. Moreover, Fig. 3 identifies a stable folding nucleus within the α-helical domain, which is the same as identified by H/D exchange experiments [38]. Although no experimental Φ-value data characterizing the transition state of this particular lysozyme isoform exists, our predictions and the H/D exchange results on the human ortholog are consistent with experimental descriptions of the hen egg-white ortholog [39].

A particularly interesting feature of lysozyme is that $\theta_{\mathrm{RP}} < \theta_{\mathrm{nat}}$ at T_{m}, indicating that the native conformational ensemble retains some rigid substructure fluctuations. At functional temperatures, θ_{nat} decreases to 1.09. Across all proteins investigated, rigid cluster susceptibility, in particular θ_{RP}, is virtually independent of temperature. Therefore, lysozyme functions optimally slightly above the rigidity threshold. The lysozyme backbone flexibility profile shown in Fig. 3 at functional temperatures sheds some light into the nature of these native fold rigid cluster fluctuations. The highlighted

hinge region is marginally rigid, but the hinge does fluctuate between being flexible and rigid (see Fig. 2 and supplementary material plots describing lysozyme conditional backbone flexibility profiles and 3D renderings of typical rigid cluster decompositions). The identified pocket of flexibility in the hinge region (residues 84–90) is consistent with experimentally observed lysozyme hinge-bending motions [40]. At functional temperatures the backbone of lysozyme is overall rigid, but possesses a hinge-motion that is "sticky" due to rigidity fluctuations. Future work will explore the possibility of clustering the statistical rigid substructure decompositions to deconvolute the rigid cluster susceptibility into unfolding/folding and functional (i.e., hinge-bending) fluctuations.

DCM predictions for lysozyme are now compared to results from other proteins. Unlike lysozyme, the transition state of cold shock protein, CSP (from *T. maritima*), is predicted to be native-like because $\theta_{RP} - \theta_{TS} > 0$. This prediction is consistent with experiment, where Perl et al. [41] have reported that the transition state of *T. maritima* CSP (as well as the one from *B. subtilis*) is "unusually native-like." The difference $\theta_{RP} - \theta_{TS}$ appears to provide a simple indicator for finding native-like characteristics in the transition state. Over our protein dataset, $\Delta S \equiv S(T_m, \theta_{RP}) - S(T_m, \theta_{TS})$ is plotted against the difference $\theta_{RP} - \theta_{TS}$ in Fig. 2D. A correlation ($R = 0.84$) is observed. The (positive, negative) changes in entropy required to reach the transition state from the rigidity threshold are consistent with (voluminous, compact) structure. Another hinge-bending protein investigated here is histidine binding protein (HBP). In apo-HBP, $\theta_{RP} = 1.14$ which is smaller than $\theta_{TS} = 1.30$ but greater than $\theta_{nat} = 0.99$ – indicating that rigid cluster fluctuations occur in the transition state. Upon substrate binding halo-HBP has a more rigid backbone, with $\theta_{RP} = 1.26$ increasing and now greater than $\theta_{TS} = 1.20$ – indicating a reduction of rigid cluster fluctuations in the transition state. Not surprising, the total number of dof in native halo-HBP ($\theta_{nat} = 0.84$) is reduced in relation to the native apo-HBP.

5. Conclusions

A bridge between network rigidity and protein thermodynamics has been made using a distance constraint model (DCM) based on a free energy decomposition scheme. These concepts have been demonstrated using a minimal 3-parameter DCM that remarkably reproduces experimental heat capacity curves across a diverse protein dataset. A Landau free energy function is calculated to directly relate protein stability to global flexibility. First-order (two-state) folding transitions are predicted for all proteins in the dataset, indicating that folded and unfolded conformations co-exist. The rigidity percolation transition tends to parallel the thermodynamic transition state, but does not generally coincide. Differences in these transition locations are linked to the enthalpy–entropy decomposition of the barriers and transition state characteristics.

This letter firmly establishes the generality of the DCM; implemented minimally several predictions are found to be in qualitative agreement with experiment. This novel approach allows detailed questions about flexibility/rigidity to be probed as a function of thermodynamic condition (i.e., pH, ionic strength, and temperature). Moreover, working directly with free energies makes DCM calculations fast. For example, given an a priori determined set of parameters, quantified stability and flexibility relationships for all 13 proteins in the dataset considered here can be obtained overnight on a modern desktop computer. In future work, we will augment to the DCM appropriate descriptors for different types of experimental measurements that probe conformational flexibility, and expand the parameterization to account for sequence and structural (i.e., solvent exposed vs. buried) characteristics. Ultimately, a robust set of transferrable parameters is sought, eliminating the need for fitting to thermodynamic data.

Acknowledgements: We thank Shankar Subramaniam (University of California, San Diego) for critiquing an early version of our manuscript, James Wrabl (University of Texas Southwestern Medical Center) for stimulating discussions concerning flexibility measures and the referees for their insightful comments. This work is partially supported by National Institute of Health Grants S06 GM48680-0952 (D.J.J.) and S06 GM53933-07 (D.R.L.); Research Corporation grant CC5141 (D.J.J.); and California State University Program for Education and Research in Biotechnology (CSUPERB) grant (to D.J.J. and D.R.L.). The graph-rigidity algorithm is claimed in U.S. Patent Number 6,014,449, which has been assigned to the Board of Trustees Michigan State University. Used with permission.

References

[1] Jacobs, D.J., Rader, A.J., Kuhn, L.A. and Thorpe, M.F. (2001) Proteins 44, 150–165.
[2] Hespenheide, B.M., Rader, A.J., Thorpe, M.F. and Kuhn, L.A. (2002) J. Mol. Graph. Model. 21, 195–207.
[3] Rader, A.J., Hespenheide, B.M., Kuhn, L.A. and Thorpe, M.F. (2002) Proc. Natl. Acad. Sci. USA 99, 3540–3545.
[4] Rader, A.J., Anderson, G., Isin, B., Khorana, H.G., Bahar, I. and Klein-Seetharaman, J. (2004) Proc. Natl. Acad. Sci. USA 101, 7246–7251.
[5] Jacobs, D.J. and Thorpe, M.F. (1998) US Patent # 6014449.
[6] Fields, P.A. (2001) Comp. Biochem. Physiol. A Mol. Integr. Physiol. 129, 417–431.
[7] Frauenfelder, H. (1989) Nature 338, 623–624.
[8] Gerstein, M., Lesk, A.M. and Chothia, C. (1994) Biochemistry 33, 6739–6749.
[9] Janin, J. and Wodak, S.J. (1983) Prog. Biophys. Mol. Biol. 42, 21–78.
[10] Dill, K.A. (1997) J. Biol. Chem. 272, 701–704.
[11] Mark, A.E. and van Gunsteren, W.F. (1994) J. Mol. Biol. 240, 167–176.
[12] Gomez, J., Hilser, V.J., Xie, D. and Freire, E. (1995) Proteins 22, 404–412.
[13] Jacobs, D.J., Dallakyan, S., Wood, G.G. and Heckathorne, A. (2003) Phys. Rev. E. 68, 061109.
[14] Jacobs, D.J. and Wood, G.G. (2004) Biopolymers 75, 1–31.
[15] Hilser, V.J. and Freire, E. (1996) J. Mol. Biol. 262, 756–772.
[16] Jacobs, D.J. and Thorpe, M.F. (1995) Phys. Rev. Lett. 75, 4051–4054.
[17] Jacobs, D.J. and Dallakyan, S. (2004) Biophys. J. (in review).
[18] Lazaridis, T. and Karplus, M. (2003) Biophys. Chem. 100, 367–395.
[19] Dahiyat, B.I., Gordon, D.B. and Mayo, S.L. (1997) Protein Sci. 6, 1333–1337.
[20] Hedwig, G.R. and Hinz, H.J. (2003) Biophys. Chem. 100, 239–260.
[21] Lifson, S. and Roig, A. (1961) J. Chem. Phys. 34, 1963–1974.
[22] Cooper, A. (2000) Biophys. Chem. 85, 25–39.
[23] Socci, N.D., Onuchic, J.N. and Wolynes, P.G. (1998) Proteins 32, 136–158.
[24] Private communication with Jose Onuchic (UCSD).
[25] Plaxco, K.W., Simons, K.T. and Baker, D. (1998) J. Mol. Biol. 277, 985–994.
[26] Panchal, S.C. and Hosur, R.V. (2000) Biochem. Biophys. Res. Commun. 269, 387–392.
[27] Kelly, L. and Holladay, L.A. (1990) Biochemistry 29, 5062–5069.

[28] Sahu, S.C., Bhuyan, A.K., Majumdar, A. and Udgaonkar, J.B. (2000) Proteins 41, 460–474.
[29] Smith, D.K., Radivojac, P., Obradovic, Z., Dunker, A.K. and Zhu, G. (2003) Protein Sci. 12, 1060–1072.
[30] Bahar, I., Wallqvist, A., Covell, D.G. and Jernigan, R.L. (1998) Biochemistry 37, 1067–1075.
[31] Haliloglu, T. and Bahar, I. (1999) Proteins 37, 654–667.
[32] Zhang, F. and Bruschweiler, R. (2002) J. Am. Chem. Soc. 124, 12654–12655.
[33] Idiyatullin, D., Nesmelova, I., Daragan, V.A. and Mayo, K.H. (2003) Protein Sci. 12, 914–922.
[34] Akashi, S., Naito, Y. and Takio, K. (1999) Anal. Chem. 71, 4974–4980.
[35] Radford, S.E., Dobson, C.M. and Evans, P.A. (1992) Nature 358, 302–307.
[36] Bieri, O. and Kiefhaber, T. (2001) J. Mol. Biol. 310, 919–935.
[37] Howarth, O.W. and Lian, L.Y. (1984) Biochemistry 23, 3522–3526.
[38] Hooke, S.D., Radford, S.E. and Dobson, C.M. (1994) Biochemistry 33, 5867–5876.
[39] Motoshima, H., Ueda, T. and Imoto, T. (1996) J. Biochem. (Tokyo) 119, 1019–1023.
[40] Kidera, A., Inaka, K., Matsushima, M. and Go, N. (1992) J. Mol. Biol. 225, 477–486.
[41] Perl, D., Welker, C., Schindler, T., Schroder, K., Marahiel, M.A., Jaenicke, R. and Schmid, F.X. (1998) Nat. Struct. Biol. 5, 229–235.
[42] Honda, S., Kobayashi, N., Munekata, E. and Uedaira, H. (1999) Biochemistry 38, 1203–1213.
[43] Buczek, O., Krowarsch, D. and Otlewski, J. (2002) Protein Sci. 11, 924–932.
[44] Wassenberg, D., Welker, C. and Jaenicke, R. (1999) J. Mol. Biol. 289, 187–193.
[45] Wintrode, P.L., Makhatadze, G.I. and Privalov, P.L. (1994) Proteins 18, 246–253.
[46] Clarke, J., Hamill, S.J. and Johnson, C.M. (1997) J. Mol. Biol. 270, 771–778.
[47] Takano, K., Ogasahara, K., Kaneda, H., Yamagata, Y., Fujii, S., Kanaya, E., Kikuchi, M., Oobatake, M. and Yutani, K. (1995) J. Mol. Biol. 254, 62–76.
[48] Robic, S., Guzman-Casado, M., Sanchez-Ruiz, J.M. and Marqusee, S. (2003) Proc. Natl. Acad. Sci. USA 100, 11345–11349.
[49] Consalvi, V., Chiaraluce, R., Giangiacomo, L., Scandurra, R., Christova, P., Karshikoff, A., Knapp, S. and Ladenstein, R. (2000) Protein Eng. 13, 501–507.
[50] Todd, M.J., Semo, N. and Freire, E. (1998) J. Mol. Biol. 283, 475–488.
[51] Kreimer, D.I., Malak, H., Lakowicz, J.R., Trakhanov, S., Villar, E. and Shnyrov, V.L. (2000) Eur. J. Biochem. 267, 4242–4252.
[52] Tjandra, N., Feller, S.E., Pastor, R.W. and Bax, A. (1995) J. Am. Chem. Soc. 117, 12562–12566.
[53] Vijay-Kumar, S., Bugg, C.E. and Cook, W.J. (1987) J. Mol. Biol. 194, 531–544.