Proc Indian Natn Sci Acad **82** No. 4 September 2016 pp. 1211-1228
© *Printed in India*. DOI: 10.16943/ptinsa/2016/48570

Review Article

Predictive Modeling of Protein Folding Thermodynamics, Mutational Effects and Free-Energy Landscapes

ATHI N NAGANATHAN*
Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai 600 036, India

(Received on 01 February 2016; Revised on 25 April 2016; Accepted on 25 April 2016)

Deciphering the folding mechanism of small single-domain proteins has a long and well-chartered history that has been and still is aided by numerous experimental and computational approaches. The computational tools at the disposal of the folding community range from all-atom molecular simulations to structure-based models. In this review, we highlight one such structure-based statistical mechanical model termed the Wako-Saitô-Munõz-Eaton (WSME) model. We have, over the past few years, made the model physically more realistic by systematically introducing mean-field terms for solvation and electrostatics apart from conventional packing interactions. The WSME model can simply be calibrated with equilibrium unfolding curves and various features such as heat capacity thermograms, free-energy surfaces or profiles and hence the folding mechanism, changes in stability upon point mutations or certain post-translational modifications, thermodynamic *vs.* dynamic effects and possible connections with function fallout of the model without additional calibration. The model requires only a small set of tunable thermodynamic parameters (~3-4) allowing for a tremendous scope in further improvement of its energy function. Most importantly, it can be employed as a rapid, physical and ensemble-based tool to directly characterize experimental equilibrium and kinetic rate and amplitude data (in real world units), that is not conventionally possible in other native-centric treatments. We believe that the WSME model is now poised to address numerous questions in the field of protein folding including pathway heterogeneity, structural-energetic relations, quantifying disorder and the effect of point mutations in disease.

Keywords: Statistical Mechanics; Ensemble; Function; Equilibrium Experiments; Intermediates; Kinetics; Landscape; Electrostatics; Solvation

Introduction

Understanding the intricate connection between the patterning of amino-acid residues in a protein sequence and the final structure, termed as the "folding problem", has been labeled as one of the 125 biggest unsolved problems in science (Editorial, 2005). It is unsolved because of two primary reasons: the incredibly large conformational space accessible to a protein chain (Levinthal, 1968) and the remarkable diversity of non-covalent interactions. The former disallows any computational protocol from sampling each and every one of the conformations possible, as a moderate protein domain of even 50 residues in length can theoretically sample more than 10^{25} conformations. Additionally, a folded protein is stabilized by a variety of non-covalent interactions, that include van der Waals, Coulombic, cation-π interactions, hydrogen bonds and importantly the enigmatic hydrophobic effect that includes both enthalpic and entropic contributions arising from not just the protein chain but also the solvent (Freire, 1995; Robertson and Murphy, 1997; Southall *et al.*, 2002; Baldwin, 2007). The magnitude of each of these terms is also dependent on the degree of burial in a structure, the immediate electronic environment, the intrinsic chemical nature of the possible 20 amino acids that constitute the protein chain and their relative effect on water structure. In effect, there is a large compensation between the various (free-) energetic terms (Bryngelson *et al.*, 1995; Akmal and Munõz, 2004; Naganathan *et al.*, 2006), resulting in a protein

**Author for Correspondence: E-mail: athi@iitm.ac.in*

stability of ~5-40 kJ/mol at 298 K, that is equivalent to the strength of just a few hydrogen bonds! The sequence-folding-structure-function is now additionally confounded by the observation of a large number of disordered proteins in several proteomes (Uversky *et al.*, 2000; Uversky, 2013).

Despite this complexity, tremendous strides have been made in understanding folding mechanisms - the order of formation of secondary structures (Englander *et al.*, 2007), pathway heterogeneity (single or multiple folding paths?) (Udgaonkar, 2008), the presence or absence of intermediates (Baldwin, 2008) and the magnitude of thermodynamic barriers (Sanchez-Ruiz, 2011) - from the perspective of experiments. They include ensemble and single-molecule measurements (Moffitt *et al.*, 2008; Schuler and Hofmann, 2013), stopped-flow to laser temperature-jump (T-jump) experiments for monitoring millisecond or (sub-) microsecond kinetics (Jones *et al.*, 1993), Nuclear Magnetic Resonance (NMR) in isolation (Sekhar and Kay, 2013) or in combination with hydrogen-exchange (HX) mass spectrometry (Hu *et al.*, 2013), multi-spectroscopic-probe techniques (Garcia-Mira *et al.*, 2002; Ma and Gruebele, 2005), and fluorescence life-time (Lakshmikanth *et al.*, 2001), multi-site Forster Resonance Energy Transfer (FRET) (Sinha and Udgaonkar, 2008), and infra-red (IR) based approaches (Vu *et al.*, 2004; Kubelka and Kubelka, 2014; Ma *et al.*, 2015).

The experiments are now routinely supplemented with additional evidence from molecular simulations that range from all-atom molecular dynamics simulations (MD) in explicit solvent (Vendruscolo, 2007; Best, 2012; Piana *et al.*, 2014) to advanced sampling protocols (Leone *et al.*, 2010; Doshi and Hamelberg, 2015; Perez *et al.*, 2016). The complexity observed in experiments and simulations is in fact predicted by theoretical treatments. Numerous groups have independently contributed to the understanding of the basic physics of the folding process through lattice-, off-lattice simulations and analytical models (Bryngelson *et al.*, 1995; Lumry *et al.*, 1966; Taketomi *et al.*, 1975; Freire and Biltonen, 1978; Wako and Saito, 1978; Ikegami, 1981; Finkelstein and Shakhnovich, 1989; Thirumalai, 1995; Abkevich *et al.*, 1995; Socci *et al.*, 1996; Hilser and Freire, 1996; Onuchic *et al.*, 1997; Dill and Chan, 1997; Munõz and Eaton, 1999; Mirny and Shakhnovich, 2001; Ghosh *et al.*, 2007; Naganathan *et al.*, 2007; Hyeon and Thirumalai, 2011; Chan *et al.*, 2011). The general consensus is that the amino acid sequence pattern not only defines the final three-dimensional structure, but also allows for a distribution of folding mechanisms, intermediate populations, stabilities and barrier heights.

Given the complexity and variety in available experimental protein folding data, and the time-intensive nature and limited sampling afforded in all-atom simulations, the focus has also been on alternate methods. Particularly, coarse-grained (CG) simulations that reduce the number of degrees of freedom associated with a protein chain have had tremendous successes in capturing the basic physics of the folding process (Hyeon and Thirumalai, 2011; Chan *et al.*, 2011; Brooks, 1998; Clementi *et al.*, 2003). A majority of CG models rely on Gõ-like energetics, i.e. they assume that only those interactions present in the native state of a protein (available from the PDB file) contribute the most to the folding mechanism (Taketomi *et al.*, 1975). This is in tune with the expectations from the energy landscape theory of protein folding that postulates that most non-native interactions (i.e. those that are not observed in the native PDB file) have been weeded out through millions of years of Natural Selection, thus effectively smoothening the folding landscape to result in minimal frustration (Bryngelson *et al.*, 1995). Strong evidence to this comes indirectly from work on designed proteins that show complex unfolding behavior compared to the natural proteins (Walters *et al.*, 2007; Sadqi *et al.*, 2009) and all-atom MD simulations that explicitly demonstrate the absence of productive non-native interactions during folding (Best *et al.*, 2013).

The CG simulations have numerous advantages over the conventional MD protocols notable among them being: (a) simulations converge faster, (b) allows for enhanced conformational sampling due to the smooth folding landscape, (c) a variety of energetic functions can be incorporated with minimal effort and (d) only a small subset of parameters is required (~10-15) compared to all-atom MD that needs more than 100 parameters. However, the intrinsically rapid and tunable nature of these models comes with a disadvantage: a careful calibration needs to be performed on the energy terms for a better and quantitative understanding of experimental outputs. In this review, we approach this problem through the

eyes of a simple yet physical statistical mechanical and coarse-grained structure-based model called the Wako-Saitô-Munõz-Eaton (WSME) model (Wako and Saito, 1978; Munõz and Eaton, 1999; Henry and Eaton, 2004). We have, over the last few years, supplemented the basic model with additional energetic terms that has made it highly predictive compared to the original versions. We discuss below the brief history of the model, the various energetic terms, the advantages afforded by this treatment with specific examples and directions for future improvements.

Wako-Saitô-Munõz-Eaton (WSME) Model

Defining the Ensemble and a Brief History of the Model

The basic idea behind the native-centric WSME model was developed independently by Wako and Saitô (Wako and Saito, 1978), and later by Mu Q }z and Eaton (Mu Q }z and Eaton, 1999). It accounts for the statistical nature of the folding process by constructing an ensemble with pre-defined rules that determines the nature of microstates that can be populated. Since the folded state is well defined, a binary variable *1* is allotted to conformations that sample native-like regions (folded-like conformations) of the Ramachandran map (Ramachandran *et al.*, 1963) while a binary variable of *0* is employed to represent all other possible non-native dihedral angles or the residue unfolded status. In effect, for a *N*-residue protein this translates to 2^N possible conformations or microstates. In their original work, Wako and Saitô (Wako and Saito, 1978) make an additional approximation that for any two residues to interact all the intervening residues should also be folded. This specific approximation enables the calculation of the total partition function *Z* over all the 2^N states,

$$Z = \sum_i w_i = \sum_i \mathrm{e}^{\left(-\Delta F_i / RT\right)}$$

where ΔF_i and w_i are intrinsic free energy and statistical weight of the state *i*, by a simple transfer-matrix formalism that involves multiplying N N-by-N matrices for a N-residue protein. Each matrix accounts for the interactions of that particular residue with all other residues following it. The free energy includes contributions from both the energetics (from specific interactions) and conformational entropy (see below).

While this work was lost to time, Munõz and Eaton developed a similar binary approach but limited their ensemble to specific collection of microstates defined solely by all possible single island of *1*s (single sequence approximation; SSA), all possible two non-interacting islands of *1*s separated by *0*s (double sequence approximation; DSA) and so on (Munõz and Eaton, 1999). This reduced the conformational space tremendously and the number of microstates for each of the approximations can be written employing the binomial coefficient $\left(\frac{N+1}{2m}\right)$ where *N* is the protein length and *m* is the sequence approximation. The partition function for this specific subset of conformations is then calculated algorithmically.

The original work of Wako and Saitô came to light in the publication of Bruscolini and Pelizzola who proposed an alternate transfer-matrix strategy to calculate the total partition function (Bruscolini and Pelizzola, 2002). At about the same time, Henry and Eaton developed an iterative method for the same (Henry and Eaton, 2004). Effectively, three different groups have proposed three different approaches for the calculation of the total partition function highlighting the interest in the folding community to exploit this simple binary approach to the folding problem. An important point to note here is that, in the WSME model, the ensemble is pre-defined while one generates the ensemble as a function of time in all other explicit chain representations. This allows for a rapid prediction of various experimental observables, providing a significant advantage over all other methods that are generally extremely time intensive. For example, it just takes less than a minute to calculate the temperature-dependent partition function for a, say, 50 residue protein. Various partial partition functions can also be generated by the derivative methodology proposed by Wako-Saitô (Wako and Saito, 1978) or, by employing symbolic operations (for example, in MATLAB).

Entropic Penalty and Energetics

An important parameter in the WSME model is the entropic penalty that refers to the cost of fixing a residue in the native conformation. In other words,

since the number of unfolded-like (*U*) conformations for a residue outnumber the folded-like conformations (*F*), i.e., Ω_F/Ω_U <<1, the model invokes an entropic penalty ΔS_{conf} defined as:

$$S_F - S_U = \Delta S_{conf} = R\ln\left(\Omega_F/\Omega_U\right)$$

where Ω represents the density of microstates and R is the universal gas constant. Size-scaling studies of two-state thermodynamic parameters reveal that the entropic penalty should be ~–16.5 J mol^{-1} K^{-1} per residue (Robertson and Murphy, 1997). Though this estimate is an average over different residue types in several proteins, it is in agreement with independent measures from statistical analysis of Ramachandran maps (Munõz and Serrano, 1994), microcalorimetry measurements (Daquino *et al.*, 1996) and all-atom molecular dynamics simulations (Baxa *et al.*, 2014). The magnitude of the entropic penalty estimated from the WSME model analysis of experimental data should therefore be within the range predicted by the various methods above. Many of the works that employ the WSME model do reveal similar numbers for the entropic penalty attesting to the physical reasonableness of the model (Naganathan, 2012).

Terms other than the entropic penalty can be grouped under the broad umbrella of 'stabilization energetics' (ΔG^{stab}). In this regard, it is important to note that the WSME model is structure-based or Gō-like and non-native interactions are not taken into consideration.The free energy of each microstate with a folded structure between residues *m* and *n* can be represented as

$$\Delta F = \sum \Delta G^{stab}_{m,n} - T\sum_{m}^{n}\Delta S_{conf}$$

WSME model has traditionally employed only van der Waals (vdW) interactions for its energetics (E_{vdW}). These interactions are obtained from the PDB file by constructing a spherical shell of a specific radius (usually 4-6 Å) around a particular atom, counting its interaction partners and grouping them into a residue-residue interaction matrix (contact-map). Each interaction is then assigned a vdW interaction energy, ξ, that can either be a single mean-field number or weighted according to the number of interactions. Additional considerations on the number of nearest neighbor residues to include or exclude can be exploited to eliminate spurious interactions that could arise purely from chain connectivity.

Towards a More Physically Reasonable Energetic Description in the WSME Model

Since the original work of Munõz and Eaton, the WSME model with microstates from up to DSA (i.e. SSA + DSA) has been extensively employed to characterize the folding of the villin headpiece domain. In many of the recent works of Eaton and coworkers, a modified WSME model is employed in which they additionally account for the statistical weights of microstates with interactions between islands of *1*s if they 'see' each other in the native structure. This model has been successful in reproducing several experimental variables and in predicting multiple folding pathways in Villin, in close agreement with all-atom MD simulations (Godoy-Ruiz *et al.*, 2008; Kubelka *et al.*, 2008; Henry *et al.*, 2013).

Despite these successes, the basic energetics of the model where only vdW interactions are considered is not physically accurate. For example, one of the fundamental features of protein thermodynamics is the observation of cold denaturation that is thought to arise primarily from the positive heat capacity change upon unfolding (Baldwin, 2007). While the origins of the positive heat capacity change is debatable (Cooper, 1976; Munõz and Sanchez-Ruiz, 2004; Cooper, 2010), the fact that it is universally observed in protein systems suggests that any model that attempts to reproduce unfolding thermodynamics should also capture this phenomenon. However, the basic WSME model does not reproduce cold denaturation.

Second, the surface of a protein is highly diverse with a specific pattern of charged residues. These charged residues aid in better protein solubility, ligand/co-factor binding and folding. Many protein active sites are inherently frustrated due to strong electrostatic repulsion between like-charged groups. This energetic frustration though undesired at the level of thermodynamic stability is however functionally critical. It has been shown through multiple independent approaches that there is a delicate balance between folding speed/stability and functional requirements (Shea *et al.*, 2000; Ferreiro *et al.*, 2007;

Levy *et al.*, 2007; Ferreiro *et al.*, 2011; Gosavi, 2013; Ferreiro *et al.*, 2014). For example, a spatially close cluster of negatively charged residues coordinate the binding to $Ca^{2+}/Mg^{2+}/Zn^{2+}$ ions, while a similar spatial cluster of positively charged residues determine the binding strength to the negatively charged DNA backbone. From the viewpoint of Gō-potentials (including the basic WSME model), such repulsive interactions would still be considered attractive and aiding in folding, while in reality their effect is the completely opposite.

Introducing Solvation and Electrostatic Effects into the WSME Model

The question then is how can these basic terms be introduced into the WSME model without compromising on the number of thermodynamic parameters. The ability to rapidly calculate the total partition function *Z*, allows for a simple approach to parameterize the model through quantitatively reproducing equilibrium experimental observables even on proteins as large as 200 residues. This is particularly possible when the data is available from differential scanning calorimetry measurements (DSC) that is also termed as the heat capacity profile (C_p). The advantage of a heat capacity profile over other experimental observations is multi-fold: a) it reports on the global unfolding thermodynamics and not local changes in structure, b) subtle structural changes in the folded or unfolded states can have strong effects in the pre- and post-transition DSC baselines, respectively, and that are generally invisible in other macroscopic experiments like CD or fluorescence, c) ΔC_p and hence the temperature of cold denaturation can be directly estimated by quantitatively analyzing the heat capacity profile, and d) importantly, no assumptions (in terms of the signals) need to be made when calculating the DSC profile from the WSME model (or any statistical model, for that matter) as it can be directly obtained from the energetic fluctuations or from the derivative of the partition function as follows

$$C_p = 2RT\left(\frac{d\ln Z}{dT}\right) + RT^2\left(\frac{d^2\ln Z}{dT^2}\right)$$

In this regard, the structurally similar proteins hen egg-white lysozyme (HEWL) and bovine lactalbumin (BLA) serve as critical systems to evaluate the performance of coarse-grained models. They possess near-identical structures in terms of the positions of the C_α-atoms (C_α RMSD ~1.5 Å), but exhibit widely different melting temperatures of 320 K and 351 K, respectively (Halskau *et al.*, 2008). The corresponding heat capacity profiles display dramatic differences in their broadness indicating that it does not arise from the temperature dependence of the unfolding enthalpy. A quantification of the sharpness of the heat capacity profile based on the variable barrier model (Munõz and Sanchez-Ruiz, 2004) revealed a large difference in thermodynamic barriers (~33 kJ/mol). It has been shown that this difference arises from the specific functionally critical distribution of charged residues on the protein surface (Halskau *et al.*, 2008).

The availability of the heat capacity profile of both proteins in absolute units enabled a direct parameterization of the functional form of the heat capacity term (ΔG_{solv}) and the magnitude of the effective dielectric constant (ε_{eff}) in the electrostatic interaction energy term (E_{elec}) of the WSME model.For the former, the solvation free energy ΔG_{stab} is assumed to scale with the number of intra-molecular interactions within that microstate (x_{cont}) with the proportionality constant being the heat capacity change upon forming a native contact (ΔC_p^{cont}). Thus, the functional form of solvation free energy, according to the fundamental dependencies of enthalpy and entropy on temperature, is

$$\Delta G_{solv} = x_{cont}^{m,n}\Delta C_p^{cont}\left[\left(T - T_{ref}\right) - T\ln\left(T/T_{ref}\right)\right]$$

where T_{ref} is the reference temperature and is set to the convergence temperature of 385 K (Robertson and Murphy, 1997). It is important to note that this approach in introducing solvation (Naganathan, 2012) is similar to two previous works that relied on estimating the accessible surface areas (ASAs) for various microstates from either the SSA-based approach (Garcia-Mira *et al.*, 2002) or the exact solution (Bruscolini and Naganathan, 2011). However, such methods are intrinsically cumbersome due to the additional approximations introduced in calculating ASAs and their reliance on the empirical parameters

of Freire and coworkers that relate the polar and apolar ASAs to the heat capacity change (Gomez *et al.*, 1995).

For electrostatics, a Debye-Hückel (DH) approximation is employed to account for interactions between charged residues. This automatically includes ionic-strength (I) and temperature (T) dependence in its energy function apart from an effective dielectric constant ε_{eff}.

$$E_{elec} = \sum_{m,n} K_{Coulomb} \frac{q_i q_j}{\varepsilon_{eff} r_{ij}} \exp\left(-r_{ij} \kappa\right)$$

where $K_{Coulomb}$ is the Coulomb constant (1389 kJ.Å/mol), q_i is the charge on the atom i, r_{ij} is the distance between charged atoms of residues$i$ and j and $1/\kappa$ is the Debye screening length. The energetics of the new model now includes a van der Waals interactions term (E_{vdW}), solvation free energy term (ΔG_{stab}) and an electrostatic term (E_{elec}).

$$\Delta G_{m,n}^{stab} = E_{vdW} + E_{elec} + \Delta G_{solv}$$

thus making it more realistic (Naganathan, 2012) compared to the older versions that relied solely on vdW interactions. The solvation term, in particular, introduces small temperature dependence in the stabilization free energy that is sufficient enough to capture cold denaturation (see below).

The heat capacity profiles of HEWL and apo-BLA were reproduced following an iterative and a well-constrained approach employing identical thermodynamic parameters. This resulted in an effective dielectric constant of 29 and a heat capacity change per contact of –0.23 J mol^{-1} K^{-1}. It is important to note the difference in their thermodynamic behaviors from the perspective of the model originates purely from the contact-map and hence from the distribution of charged residues. The model was also able to capture the large differences in thermodynamic barriers resulting in values of ~37 kJ/mol and ~13 kJ/mol at midpoint temperatures of HEWL and BLA, respectively (Fig. 1B). This magnitude is in accordance with independent estimates from the variable barrier model analysis of the heat capacity profiles (Halskau *et al.*, 2008). Moreover, the predicted structure of the intermediate states

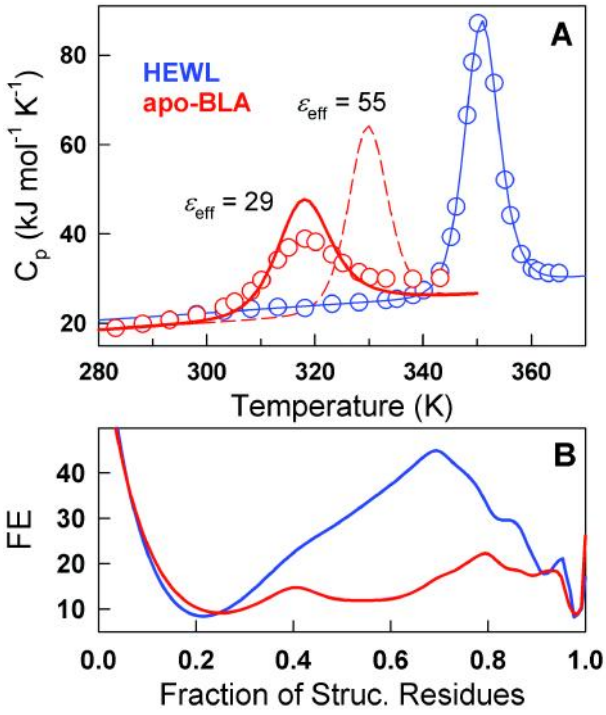

Fig. 1: Blue and red represent HEWL and apo-BLA, respectively. (A) Experimental DSC profiles (circles), fit to HEWL (blue line) and predictions of apo-BLA heat capacity profile for the different magnitudes of the effective dielectric constant ε_{eff} (red). (B) Free energy profiles in kJ mol^{-1} under iso-stability conditions as a function of the fraction of structured residues. Adapted with permission from(Naganathan, 2012). Copyright 2012, American Chemical Society

(Naganathan, 2012) is consistent with experimental observations (Halskau *et al.*, 2005).

The approach proposed above to characterize thermograms does not predict *a priori* the heat capacity change upon unfolding given a structure, but provides a simple and systematic avenue to quantify them in terms of fundamental parameters, given a particular experimental data. Therefore, the actual magnitude of is expected to vary from protein to protein, depending primarily on the degree of hydrophobic packing and experimental solvent conditions. The functional form, however, should be robust enough to be applicable to multiple proteins. In Fig. 2. we show the power of this approach in the characterization of two structurally similar alpha-helical proteins, LacR and CytR; the former is well folded in the absence of DNA (Felitsky and Record, 2003) and the latter is intrinsically disordered under

the same condition (Moody *et al.*, 2011). We performed an analysis similar to that of HEWL/BLA homologous pair; the unfolding curve of LacR was exactly reproduced using the WSME model with solvation and electrostatics to obtain the fundamental parameters together with the folded and unfolded baselines. Once this is fixed, the contact-map and the charge-distributions of LacR was substituted with that of the folded CytR (obtained in the presence of DNA). The conformational behavior of CytR was predicted to be disordered in good agreement with experimental observations (Fig. 2A) (Naganathan and Orozco, 2013). In parallel, the phenomenon of cold denaturation can be captured and the predicted T_c (cold denaturation midpoint) of ~253 K compares well with the ~245 K estimated from a two-state analysis of the LacR unfolding curve (Felitsky and Record, 2003). To our knowledge, this is the first time cold denaturation is captured from the perspective of an ensemble-based model. It is important to note that conventional coarse-grained treatments are not capable of the same (Naganathan, 2013).

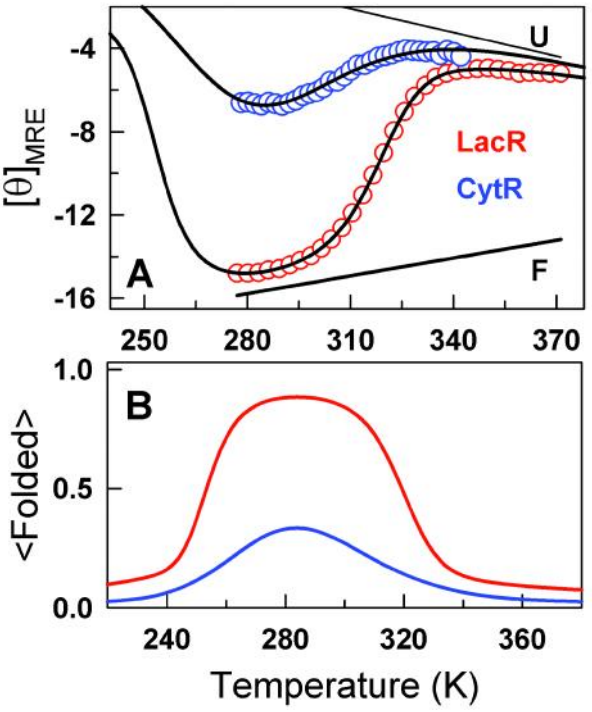

Fig. 2: Blue and red represent CytR and LacR, respectively. (A) far-UV CD monitored unfolding curves as function of temperature and in the absence of DNA reported in mean residue ellipticity units (MRE/1000). *F* and *U* stand for the folded and unfolded baselines, respectively, obtained by exactly fitting the unfolding curve of LacR to the WSME model. Adapted with permission from(Naganathan and Orozco, 2013). Copyright 2013, American Chemical Society. (B) Predicted folded state population with the roll-over signifying cold-denaturation

Role of Electrostatics in Protein Folding Thermodynamics

One interesting observation is the magnitude of the effective dielectric constant (ε_{eff}) that is predicted to be 29 in close association with experiments (Naganathan, 2012). This estimate is at odds with the dielectric constant of ~78.5 that is conventionally used to quantify the strength of two interacting charges in water. Protein charged residues, on the other hand, 'see' each other across the protein surface or chain that cannot be treated as a continuum of polarizable water molecules; in other words, apart from the charged/polar atoms, a significant fraction of the protein's surface is composed of weakly polarizable hydrogen and carbon atoms from methylene and methyl groups of side chains. Additionally, the motion of atoms linked to each other through the protein chain is restricted, thus limiting their response to local changes in the electric field. This dynamical effect is expected to further reduce the dielectric constant. These considerations suggest that the magnitude of the effective dielectric constant required to quantify charge-charge interactions on the protein surface should be lower than 78.5 and higher than ~4 which is generally employed for hydrophobic protein interior, as originally predicted by Warshel and co-workers (Vicatos *et al.*, 2009). Interestingly, Alexov and co-workers reported dielectric constants of 20-30 for charged residues on the protein surface employing a smooth Gaussian-based dielectric function (Li *et al.*, 2013). A correlation of 0.9 and above is also generally observed between the per-residue electrostatic interaction energy calculated by the Debye-Hückel approximation (with ε_{eff} = 29) and the more computationally intensive Tanford-Kirkwood (TK) algorithm (Tanford and Kirkwood, 1957; Ibarra-Molero *et al.*, 1999) for several proteins (see below), further validating the effective dielectric constant estimate of 29.

Point Mutations

How reliable is the ε_{eff} value of 29 in the context of folding thermodynamics? To explore this, we compiled a database of 138 single- and multiple-point mutations

involving charged residues (charge reversal, addition or deletion) from 16 different proteins and enzymes and whose thermal unfolding curves are also available (Naganathan, 2013). We reproduced the thermal denaturation midpoint (T_m) of each of the wild-type proteins exactly using the WSME model with electrostatics and solvation by modulating a single thermodynamic parameter, the strength of van der Waals interactions (ξ). Experimental mutations were simply introduced through PyMol (The PyMOL Molecular Graphics System) and the resulting structure was employed to predict the melting curves of the mutant proteins using identical parameters as the WT. The charged status of ionic residues was chosen according to the pH: D, E, K and R are charged at pH 7.0 while H is additionally charged at pH 5.0.

We obtained a reasonable correlation of 0.65 between experiments and prediction with a slope of 0.59 (against an expected value of 1) and a near zero-intercept of –0.45 K (Fig. 3A). The correlation increases to 0.71 and the slope to 0.73 upon eliminating just 5% of the mutants with maximum deviation from the expected 1:1 correlation line. Importantly, the overall success rate (fraction of true positives) is 81% and increases to 90% when considering multiple-point mutations.This suggests that the model captures the changes in stability induced by even point mutations of charged residues very well (Naganathan, 2013). However, the correlation is not very high simply because of the fact that point mutations in secondary structural elements also modulate the intrinsic secondary structure propensity, which is not taken into consideration in this mean-field approach.

Mesophilic-Thermophilic Protein Pairs

The corollary to the above observation is that multiple point mutations or changes in stabilities of homologues arising specific charge-charge interactions should be captured well by the model as large changes in sequence is expected to average out the conformational entropic effects. In this regard, it is well known that proteins from thermophiles exhibit a higher thermodynamic stability than their mesophilic cousins due to a larger network of charge-charge interactions (Kumar *et al.*, 2000). If this is true, one expects the model to perform very well in reproducing the differences in T_ms between mesophile-thermophile protein pairs. To check for this, we compiled a

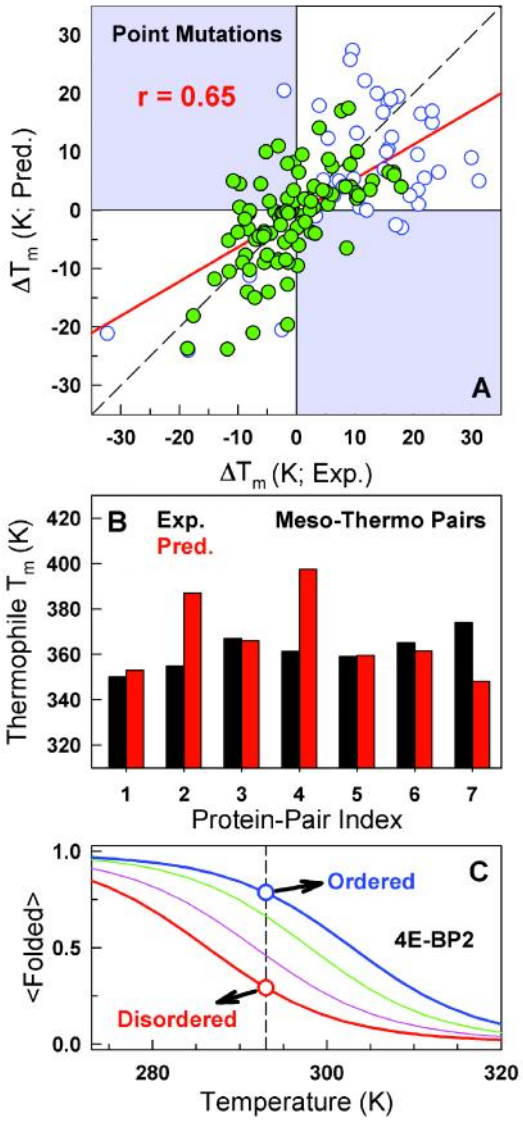

Fig. 3: (A) Experimental versus predicted changes in stability of mutations involving charged residues. Filled and open circles represent single and multiple mutations and the dashed line is the 1:1 correlation line. Mutants in the shaded quadrants are false positives. (B) Predicted versus experimental changes in the melting temperature (T_m) of thermophilic proteins employing the respective mesophilic proteins' thermodynamic behavior as reference. The pairs are CspB (1, 2), L30e (3), Hpr (4), RNase H (5), Cyt C (6) and Che Y (7). Adapted with permission from(Naganathan, 2013). Copyright 2013, American Chemical Society. (C) Predicted changes in stability employing the fully phosphorylated WT 4E-BP2 protein (pT20pT29, blue) as reference. Singly phosphorylated pT29, pT20 and non-phosphorylated variants are shown in green, magenta and red, respectively. The dashed line represents the experimental temperature. Adapted from (Gopi *et al.*, 2015), with permission from the PCCP Owner Societies

database of 7 mesophile-thermophile protein pairs, whose melting temperatures are available under identical experimental conditions. We followed a similar procedure to that outlined for Fig. 3A: adjust ξ to reproduce the mesophilic proteins' T_m and then predict the stability of the corresponding thermophilic variant using identical parameters.We obtained a remarkable agreement between the experimental and predicted melting temperatures of the thermophilic proteins (Fig. 3B) with the latter consistently exhibiting a higher T_m in all cases (Naganathan, 2013).

Modeling PTMs Involving Changes in Charge Status

The WSME model with electrostatics is applicable not just to simple point mutations but even post-translational modifications (PTMs) like methylation, phosphorylation *etc.* that change the charge-status of residues. A recent extreme example is the phosphorylation of the mammalian protein 4E-BP2 involved in translation initiation. 4E-BP2 is disordered and functional in the absence of phosphorylation, while step-wise phosphorylation at two specific threonine residues (T20 and T29) increases the thermodynamic stability in a graded manner that simultaneously abrogates binding (Bah *et al.*, 2015). To understand this behavior, we used an approach similar to before while assigning a charge of -2 for the phoshoryl groups at pH 7.0. Since experimental unfolding curves are not available, it was assumed that the doubly phosphorylated form (pT20pT29) exhibits at least 80% folded population at the experimental temperature of 293 K. With this unfolding curve as a reference (blue in Fig. 3C), we systematically removed the phosphoryl groups and simulated the unfolding curves in each case while employing identical parameters for all variants. We identified a trend with the variant pT29 exhibiting a folded population of ~67%, pT20 ~46% and finally, the non-phosphorylated variant ~26%, in tune with experimental observations. The work also highlights how the phosphorylation status affects not just the folding thermodynamics, but also the nature of intermediate states populated during folding and in its equilibrium ensemble (Gopi *et al.*, 2015).

A Rapid Method for Identifying Stabilizing Mutations

The analyses detailed until now involved the characterization of known WT-mutant or mesophile-thermophile protein pairs. However, an important requirement is to identify mutations that can specifically enhance protein stabilities and that can be exploited in industrial or pharmaceutical applications. To do so, we have developed a simple and highly parallelizable methodology in which the charges on the native structure are randomly shuffled and the net electrostatic interaction energy calculated in each case using the Debye-Hückel model with an effective dielectric constant of 29 (Naganathan, 2013). The charge shuffling procedure includes charge neutralization or charge reversal on existing charged residues and addition of charges on large polar side chains (Q and N). The calculations are not computationally demanding as up to 10^5 4-point substitutions can be performed in just over 40 seconds on a 2.8 GHz Intel Core i7 processor. The charge-charge interaction energies are ranked and the best performing mutants' structures can be generated with PyMol. The corresponding unfolding curves can then be generated in a matter of seconds with the mutant structures and the WT parameters. The advantage of this approach over conventional procedures is multi-fold: (a) we employ a small parameter set to characterize the WT and hence no extra parameters are needed for predicting the mutant behavior, (b) highly parallelizable, (c) output is provided in the experimentally accessible scale of melting temperatures, which is not provided in any other atomistic- or coarse-grained approaches know to us, and (d) ensemble nature of folding is taken into consideration with 2^N microstates. While most procedures employ just the native state or a small collection of structures to predict mutational effects (Guerois *et al.*, 2002; Yin *et al.*, 2007; Gribenko *et al.*, 2009).

Capturing Ionic-Strength Effects

An advantage of the DH-approximation employed in the WSME model is that the effect of changes in solution ionic strength on thermodynamic stabilities of proteins can be directly predicted without additional assumptions. It should be noted that the DH-approximation is applicable only in the low ionic-strength regime (<200 mM) and as an avenue to model ion-shielding effects. It does not account for effects of ions on water structure (chaotropic/kosmotropic effects) or the non-trivial outcomes of ion-binding to either the folded, unfolded or partially structured states.

The folding thermodynamics of Fyn-SH3 presents an interesting case in this regard. While most SH3 domains are super-stable (T_m>373 K at pH 7.0), Fyn-SH3 uniquely displays reduced stability under the same conditions (de Los Rios and Plaxco, 2005). It has been shown before that this feature arises from a unique spatially close distribution of negatively charged residues on the protein surface that is functionally critical. In fact, simple thermodynamic measurements have shown that the stability change in Fyn-SH3 can be well approximated by the DH electrostatics (de Los Rios and Plaxco, 2005). Since the latest version of the WSME modelexplicitly accounts for DH effects, we test its performance by systematically increasing the ionic strength value (I). We find a continuous increase in the T_m of Fyn-SH3 until ~0.5 M beyond which it plateaus out (Fig. 4A). A good correlation is accordingly observed between the experimental chemical midpoint (C_m) and T_m with the same plateauing effect observed only for high salt concentrations (Fig. 4B). A related observation is the remarkably tunable behavior observed in the one-state downhill folding protein BBL (Garcia-Mira *et al.*, 2002). BBL is completely unfolded under acidic and low ionic-strength conditions, but gradually attains structure upon increase in ionic strength of the medium (Desai *et al.*, 2010).The model again captures this observation (Fig. 4C) corroborating the experimental interpretation that screening of the destabilizing positive charges on the protein surface is the primary origin of this unique behavior.

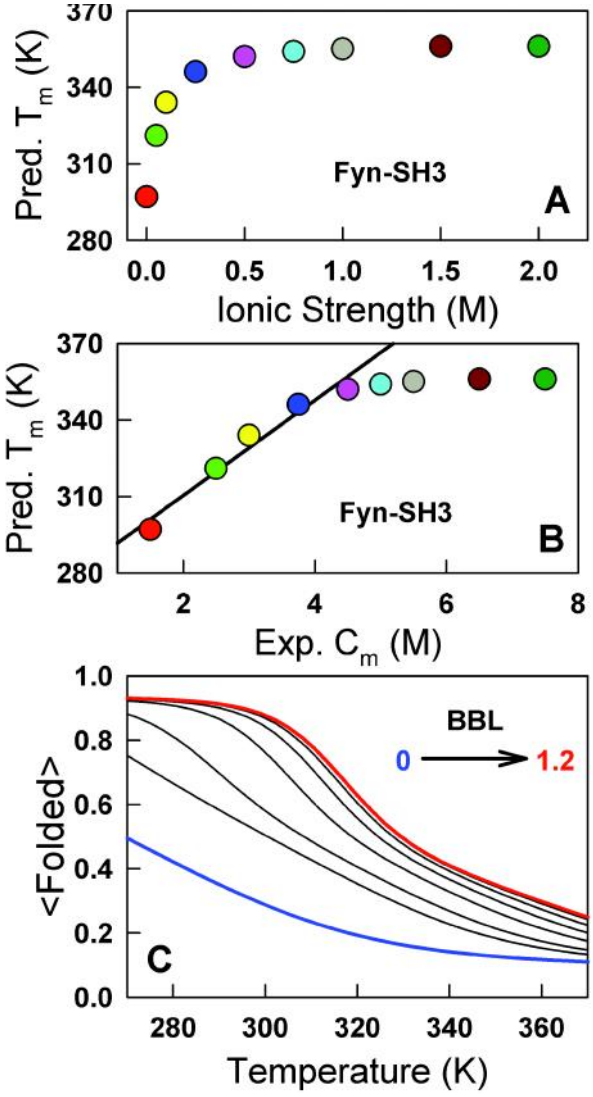

Fig. 4: (A) Predicted changes in the melting temperature (T_m) of Fyn-SH3 as a function of ionic strength. (B) Same as in panel A, but by directly comparing the experimental chemical midpoint (C_m) with the predicted thermal midpoint. The line highlights the good agreement between experiment and predictions until 0.5 M ionic strength. (C) Blue represents the unfolding curve of BBL at pH 3.0 and in the absence of salt. The folded population and stability of BBL increase continuously upon increasing the ionic strength from 0 to 1.2 M (red)

Intermediate States in Protein Folding

Charge-charge interactions have traditionally been seen as influencing merely the folding thermodynamics and manipulation of these interactions for stabilization of proteins and enzymes without loss of function has a long history (Loladze *et al.*, 1999; Sanchez-Ruiz and Makhatadze, 2001). However, the work of Halskau *et al.* on HEWL/BLA demonstrated that they have a wider role to play in determining the nature of partially structured states populated during folding and the effective thermodynamic barrier separating the unfolded and folded states (Halskau *et al.*, 2008). The WSME model with electrostatics and solvation is able to account for the experimentally observed differences in the HEWL/BLA family providing computational evidence for the same (Naganathan, 2012). There is a possibility that this could simply be an isolated observation. However, we have now established through intensive research over the past few years that electrostatic interactions can play a dominant role in determining the nature of intermediates populated during folding. For example, the protein Barstar has been shown to fold through multiple intermediate states through varied experimental techniques and from

more than two decades of work (Lakshmikanth *et al.*, 2001; Bhuyan and Udgaonkar, 1999; Sridevi and Udgaonkar, 2002; Sarkar *et al.*, 2013).There is however little computational evidence to the structural features of the intermediate states and most importantly, the underlying physico-chemical origin of the complex folding behavior.

We have recently addressed this issue through a combination of electrostatic calculations, statistical mechanical modeling and all-atom MD simulations (Naganathan *et al.*, 2015). Briefly, Barstar is highly frustrated electrostatically primarily due to 4 residues D35, D39, E76 and E80, all of which are spatially close providing a large acidic surface that is critical for binding with Barnase (Fig. 5A, 5B). The WSME model with electrostatics is able to capture the changes in stability of Barstar upon mutations of these charged residues quite well highlighting the robustness of the calculation (Fig. 5C). The only outliers (mutants in the fourth quadrant) arise from non-trivial non-native interactions, which are not represented in the model. The effective one-dimensional free energy profile in the presence of electrostatic terms results in 2-3 intermediate states in agreement with experiments (blue in Fig. 5D). Interestingly, upon switching off the electrostatic term a near-perfect two-state-like free energy profile was obtained, convincingly demonstrating that the intermediates states arise from destabilizing electrostatic energetics (black in Fig. 5D). To identify the residues that contribute the most to the destabilization energetics, we generated the structure of all single point mutational variants of charged residues; there are 24 charged residues in Barstar, resulting in 48 charge reversal and deletion variants.A clear correlation between stabilizing interactions and population of intermediates was observed from which residues E76 and E80 were identified as critical for the population of intermediates, thus addressing one of the long-standing questions on the structural-thermodynamic origins of Barstar folding complexity (Naganathan *et al.*, 2015).

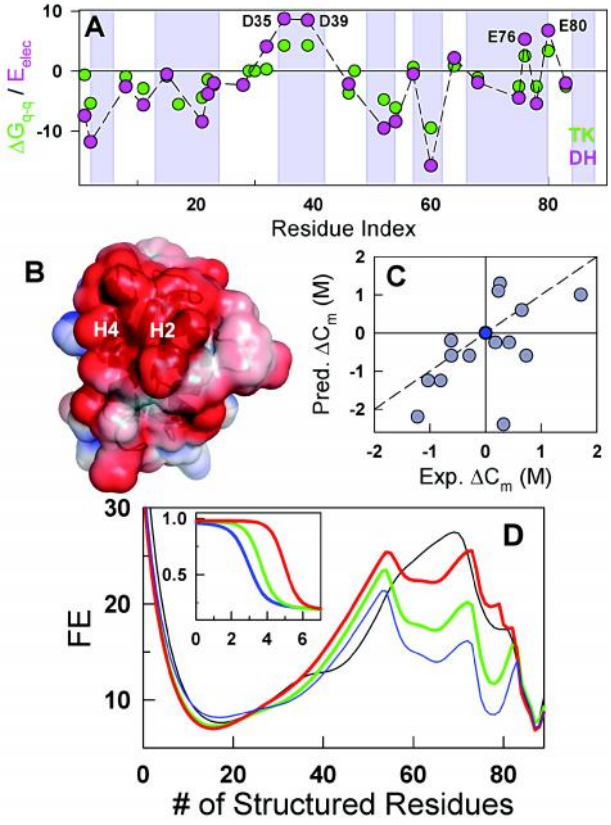

Fig. 5: The relevant energy units are in kJ mol^{-1}. (A) Residue-wise charge-charge interaction energy as calculated from the Debye-Hückel (DH) approximation employed in the WSME model (magenta) or from the Tanford-Kirkwood (TK) algorithm (green). The shaded areas represent the secondary structure elements of Barstar. (B) Electrostatic potential energy surface highlighting the large negative potential on the Barnase-binding face of Barstar. Helices H2 and H4 are highlighted. (C) Experimental versus predicted changes in chemical midpoints. Blue represents the WT and the dashed line is the 1:1 correlation line. (D) Free energy profiles of Barstar in the presence and absence of electrostatics in its energy function (blue and black, respectively). The free energy profiles of E76Q and E76K are shown in green and red, respectively. (Inset) Equilibrium unfolding curves following the same color code.Adapted with permission from(Naganathan *et al.*, 2015). Copyright 2015, American Chemical Society

RNase H is another prototypical example of a protein folding through multiple intermediate states (Chamberlain *et al.*, 1996; Raschke *et al.*, 1999). Recent hydrogen-exchange combined with mass spectrometry (HX-MS) experiments indicate that the protein folds through a single dominant macroscopic pathway involving the condensation of specific secondary structure elements (foldons) (Hu *et al.*, 2013). To test this, we calibrated the energetic terms of the WSME model with electrostatics by simultaneously reproducing the thermal and chemical denaturation profiles of RNase H (Narayan and Naganathan, 2014). Once calibrated, the various

experimental observations directly fall out of the predictive model without additional parameterization. The presence of intermediates can be directly inferred from the one-dimensional free energy profiles that indicate at least three on-pathway partially structured states (I_1, I_2 and N*; Fig. 6A). Since the intermediate states are high in free energy, they are only minimally populated in equilibrium thermal/chemical denaturation experiments (Fig. 6B), highlighting the fact that the observation of sigmoidal unfolding curves is no evidence for two-state folding.

Though specific intermediates are observed in the 1D free-energy profiles of RNase H, the free energies of I_1, I_2 and N* are the effective averages over the statistical weights of millions of microstates. Such a representation does not address the issue of pathway complexity – a single dominant macroscopic pathway? - or the structural features of the intermediates. To do so, we constructed the so-called single sequence approximation (SSA) landscape (Garcia-Mira *et al.*, 2002). In this approach, the statistical weights of only those microstates that have a single-stretch of folded residues (i.e. two or more islands of *1*s separated by *0*s is not allowed) are calculated, using identical parameters as the exact-solution. The advantage of this approach is that it reduces the conformational complexity from 2^N to $N^* (N+1)/2$, while allowing for a simple way to visualize them. Figure 6C plots the projected free energy of the SSA microstates on to two coordinates *m* and *n* in a spectral color-coding scale (low and high free energies correspond to blue and red, respectively), where, *m* is the starting residue and *n* is the number of structured residues. For example, the coordinate (42, 79) corresponds to the intermediate I_2 with 79 residues structured from and including 42, and, whose structure can be directly obtained from the PDB file. Importantly, most of the landscape is disallowed (red) with a single valley of blue starting from the unfolded state through I_1, to I_2 and then finally through two alternate paths to the native state. The path corresponding to the dashed line from I_2 to N is not observed experimentally and this is possibly because of the slightly higher free energy associated with this transition (~10 kJ mol^{-1}). In effect, this representation provides strong computational evidence for the sequential folding mechanism in RNase H from the perspective of an ensemble-based model (Narayan and Naganathan, 2014).

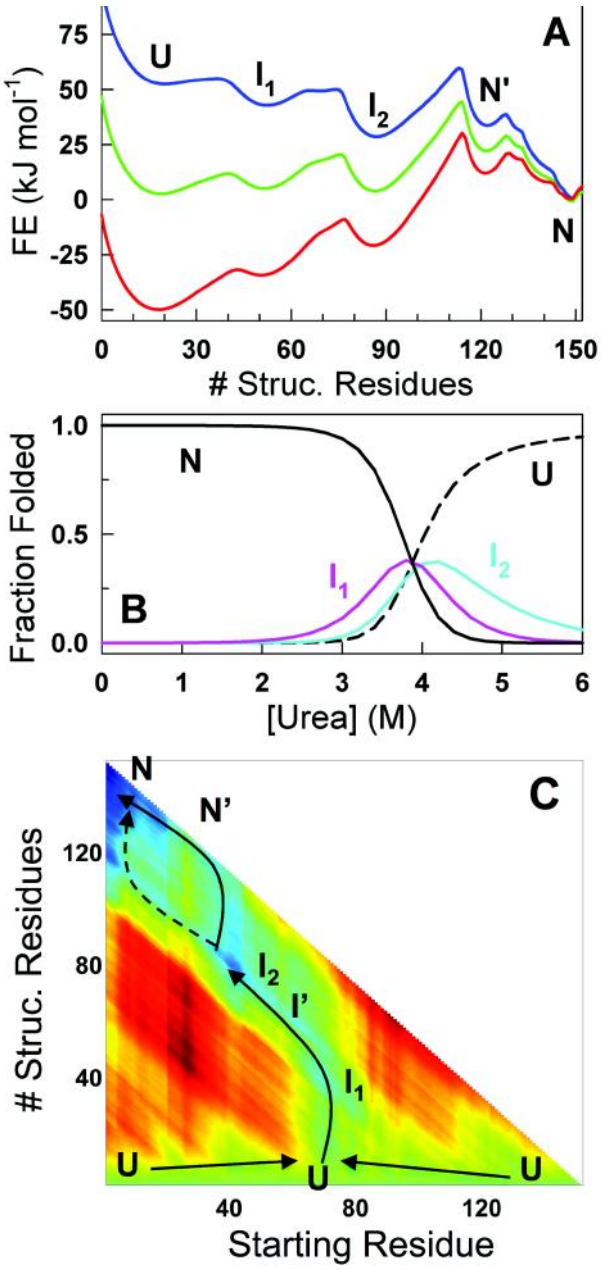

Fig. 6: (A) One-dimensional free energy profiles of RNase H at 0 M (blue), 3.8 M (green) and 6 M (red) urea, respectively. (B) Changes in populations of different macroscopic states as a function of urea. The populations of I_1 and I_2 have been multiplied by a factor of 3 for visual clarity. (C) The SSA landscape as a function of the coordinates, the starting residue (m) and the number of structured residues (n), at 298 K and 0 M urea concentration. A spectral color-coding is employed going from low to high free energy (blue to red). The arrows highlight the only possible path for a protein molecule to fold. The dashed line represents an alternate pathway that is not observed in experiments. Adapted with permission from (Narayan and Naganathan, 2014). Copyright 2014, American Chemical Society

Discussion

The WSME model can be rapidly parameterized employing real experimental data (heat capacity profiles, unfolding curves, HX data, NMR-derived population of states), thus allowing for a physically reasonable energy function. The latest version of the WSME model (Naganathan, 2012; Naganathan, 2013) now includes experimentally derived empirical terms for solvation, a robust and highly predictive electrostatic energy function and the conventional vdW term for packing interactions. It therefore resembles a force field in itself, but with only a minimal subset of tunable parameters. Additionally, a simple sequence dependent entropic penalty can be included from the calorimetrically derived estimates of Freire and co-workers (Daquino *et al.*, 1996) or secondary-structure dependent entropy from the statistical analysis of Ramachandran maps of proteins (Muñoz and Serrano, 1994).

The magnitude of the parameters themselves is very reasonable and can be compared to various experimental or empirical estimates. For example, the entropic penalty associated with fixing a residue in a ordered conformation falls in the range between -11 to –20 J mol^{-1} K^{-1} per residue and varies slightly from one protein to another, compared to the expected values of –15 to –19 J mol^{-1} K^{-1} per residue derived from various empirical estimates. The vdW interaction energy per contact, falls in the range between -50 to –70 J mol^{-1}at 6 Å distance cut-off and is again comparable to –46.1 J mol^{-1} calculated form Amber force field parameters (Cornell *et al.*, 1995). The effective dielectric constant term of 29 for charged residues on protein surface, as discussed before, is also in agreement with the expectations from independent works (Vicatos *et al.*, 2009; Li *et al.*, 2013).

We propose that a direct quantification of equilibrium unfolding curves provides a simple and robust avenue to parameterize models and that such data carries as much, if not more, information as kinetic rate constants. It is important to note that this avenue is not conventionally exploited in all-atom MD simulations or other coarse-grained treatments. This is simply because it is still challenging to rapidly generate multiple unfolding curves from such simulations and then systematically parameterize them by adjusting any one of the several parameters. The WSME model is advantageous in this regard, as it is not only rapid enough to characterize unfolding curves but also detailed enough to include the various energetic terms and account for the statistical nature of the folding process (i.e. an ensemble treatment). Also, to our knowledge, this is the only ensemble-based treatment that can also reproduce experimental kinetic amplitudes and cold denaturation. However, artifacts can also arise from this approach because of its 'local interactions form first and non-local later' principle, despite the fact that this is intuitively expected and is also consistent with independent observations from large-scale MD simulations (Lindorff-Larsen *et al.*, 2011), Gō-models (Naganathan and Orozco, 2011) and mutational analysis (Naganathan and Munõz, 2010). The implication is that, the model is not applicable to single- or multi-domain proteins, whose folding involves long-range condensation of structure, with little local structure formation; but it should be possible to accurately estimate the statistical weights of the various partially structured states even in such proteins by appropriately accounting for the excess entropy of disordered regions.

Apart from the examples provided above, the WSME model with electrostatics and simplified solvation has been employed in conjunction with experiments to quantify the dynamic and thermodynamic effect of protein surface electrostatics in 4 homologous families (Naganathan, 2012), map the conformational landscape of an intrinsically disordered protein (Naganathan and Orozco, 2013), model the effect of disorder in a repeat protein termed the 'domino-like destabilization mechanism' (Sivanandan and Naganathan, 2013), decipher the effect of functional constraints on folding (Naganathan *et al.*, 2015; Munshi and Naganathan, 2015), understand the subtle effect of post-translational modifications on disorder-order equilibrium (Gopi *et al.*, 2015), and even generate fitness landscape of small proteins (Gopi *et al.*, 2015). The ability of the model to capture the folding features of proteins with various topologies and in agreement with experiments is possibly an indication that the underlying assumptions (particularly the ensemble composition) are more realistic than previously thought. We believe that the WSME model is now poised to address numerous questions in the field of protein folding including the origins, magnitude and temperature

dependence of pathway heterogeneity, structural-energetic relations, quantifying disorder in statistical thermodynamic terms and the effect of point mutations on folding mechanism, function and hence disease.

Acknowledgment

The author thanks the past and present graduate and undergraduate students, summer trainees, and research fellows who have made this work possible. Work in our lab is supported by the Department of Biotechnology (DBT), Govt. of India, award number BT/06/IYBA/2012, Science and Engineering Research Board (SERB), Department of Science and Technology (DST), Govt. of India, award number YSS/2014/000011 and New Faculty Seed Grant from IIT Madras. A. N. N. is a Wellcome Trust / DBT India Alliance Intermediate Fellow.

References

Abkevich V I, Gutin A M and Shakhnovich E (1995) Impact of local and non-local interactions on thermodynamics and kinetics of protein folding *J Mol Biol* **252** 460-471

Akmal A and Muñoz V (2004) The nature of the free energy barriers to two-state folding *Proteins* **57** 142-152

Bah A, Vernon R M, Siddiqui Z, Krzeminski M, Muhandiram R *et al.* (2015) Folding of an intrinsically disordered protein by phosphorylation as a regulatory switch *Nature* **519** 106-109

Baldwin R L (2007) Energetics of protein folding *J Mol Biol* **371** 283-301

Baldwin R L (2008) The search for folding intermediates and the mechanism of protein folding *Ann Rev Biophys* **37** 1-21

Baxa M C, Haddadian E J, Jumper J M, Freed K F and Sosnick T R (2014) Loss of conformational entropy in protein folding calculated using realistic ensembles and its implications for NMR-based calculations *Proc Natl Acad Sci USA* **111** 15396-15401

Best R B (2012) Atomistic molecular simulations of protein folding *Curr Opin Struct Biol* **22** 52-61

Best R B, Hummer G and Eaton W A (2013) Native contacts determine protein folding mechanisms in atomistic simulations *Proc Natl Acad Sci USA* **110** 17874-17879

Bhuyan A K and Udgaonkar J B (1999) Observation of Multistate Kinetics during the Slow Folding and Unfolding of Barstar *Biochemistry* **38** 9158-9168

Brooks C L (1998) Simulations of protein folding and unfolding *Curr Opin Struct Biol* **8** 222-226

Bruscolini P and Naganathan A N (2011) Quantitative Prediction of Protein Folding Behaviors from a Simple Statistical Model *J Am Chem Soc* **133** 5372-5379

Bruscolini P and Pelizzola A (2002) Exact solution of the Muñoz-Eaton model for protein folding *Phys Rev Lett* **88** 258101

Bryngelson J D, Onuchic J N, Socci N D and Wolynes P G (1995) Funnels, Pathways, and the Energy Landscape of Protein-Folding - a Synthesis *Proteins* **21** 167-195

Chamberlain A K, Handel T M and Marqusee S (1996) Detection of rare partially folded molecules in equilibrium with the native conformation of RNaseH *Nat Struct Biol* **3** 782-787

Chan H S, Zhang Z, Wallin S and Liu Z (2011) Cooperativity, Local-Nonlocal Coupling, and Nonnative Interactions: Principles of Protein Folding from Coarse-Grained Models *Ann Rev Phys Chem* **62** 301-326

Clementi C, Garcia A E and Onuchic J N (2003) Interplay among tertiary contacts, secondary structure formation and side-chain packing in the protein folding mechanism: All-atom representation study of protein L *J Mol Biol* **326** 933-954

Cooper A (1976) Thermodynamic Fluctuations in Protein Molecules *Proc Natl Acad Sci USA* **73** 2740-2741

Cooper A (2010) Protein heat capacity: An anomaly that maybe never was *J Phys Chem Lett* **1** 3298-3304

Cornell W D, Cieplak P, Bayly C I, Gould I R, Merz K M *et al.* (1995) A 2nd Generation Force-Field for the Simulation of Proteins, Nucleic-Acids, and Organic-Molecules *J Am Chem Soc* **117** 5179-5197

Daquino J A, Gomez J, Hilser V J, Lee K H, Amzel L M *et al.* (1996) The magnitude of the backbone conformational entropy change in protein folding *Proteins* **25** 143-156

de Los Rios M A and Plaxco K W (2005) Apparent Debye-Huckel electrostatic effects in the folding of a simple, single domain protein *Biochemistry* **44** 1243-1250

Desai T M, Cerminara M, Sadqi M and Muñoz V (2010) The Effect of Electrostatics on the Marginal Cooperativity of an Ultrafast Folding Protein *J Biol Chem* **285** 34549-34556

Dill K A and Chan H S (1997) From Levinthal to pathways to funnels *Nat Struct Biol* **4** 10-19

Doshi U and Hamelberg D (2015) Towards fast, rigorous and efficient conformational sampling of biomolecules:

Advances in accelerated molecular dynamics *Biochim Biophys Acta* **1850** 878-888

Editorial (2005) So much more to know *Science* **309** 78-102

Englander S W, Mayne L and Krishna M M G (2007) Protein folding and misfolding: mechanism and principles *Q Rev Biophys* **40** 287-326

Felitsky D J and Record M T (2003) Thermal and urea-induced unfolding of the marginally stable lac repressor DNA-binding domain: A model system for analysis of solute effects on protein processes *Biochemistry* **42** 2202-2217

Ferreiro D U, Hegler J A, Komives E A and Wolynes P G (2007) Localizing frustration in native proteins and protein assemblies *Proc Natl Acad Sci USA* **104** 19819-19824

Ferreiro D U, Hegler J A, Komives E A and Wolynes P G (2011) On the role of frustration in the energy landscapes of allosteric proteins *Proc Natl Acad Sci USA* **108** 3499-3503

Ferreiro D U, Komives E A and Wolynes P G (2014) Frustration in biomolecules *Q Rev Biophys* **47** 285-363

Finkelstein A V and Shakhnovich E I (1989) Theory of cooperative transitions in protein molecules. II. Phase diagram for a protein molecule in solution *Biopolymers* **28** 1681-1689

Freire E (1995) Protein stability and folding. Totowa, New Jersey: Humana Press

Freire E and Biltonen R L (1978) Statistical Mechanical Deconvolution of Thermal Transitions in Macromolecules.1. Theory and Application to Homogeneous Systems *Biopolymers* **17** 463-479

Garcia-Mira M M, Sadqi M, Fischer N, Sanchez-Ruiz J M and Muñoz V (2002) Experimental identification of downhill protein folding *Science* **298** 2191-2195

Ghosh K, Ozkan S B and Dill K A (2007) The ultimate speed limit to protein folding is conformational searching *J Am Chem Soc* **129** 11920-11927

Godoy-Ruiz R, Henry E R, Kubelka J, Hofrichter J, Munõz V *et al.* (2008) Estimating free-energy barrier heights for an ultrafast folding protein from calorimetric and kinetic data *J Phys Chem B* **112** 5938-5949

Gomez J, Hilser V J, Xie D and Freire E (1995) The Heat-Capacity of Proteins *Proteins* **22** 404-412

Gopi S, Rajasekaran N, Singh A, Ranu S and Naganathan A N (2015) Energetic and topological determinants of a phosphorylation-induced disorder-to-order protein conformational switch *Phys Chem Chem Phys* **17** 27264-27269

Gosavi S (2013) Understanding the Folding-Function Tradeoff in Proteins *PLoS One* **8** e61222

Gribenko A V, Patel M M, Liu J, McCallum S A, Wang C Y *et al.* (2009) Rational stabilization of enzymes by computational redesign of surface charge-charge interactions *Proc Natl Acad Sci USA* **106** 2601-2606

Guerois R, Nielsen J E and Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations *J Mol Biol* **320** 369-387

Halskau O, Perez-Jimenez R, Ibarra-Molero B, Underhaug J, Muñoz V *et al.* (2008) Large-scale modulation of thermodynamic protein folding barriers linked to electrostatics *Proc Natl Acad Sci USA* **105** 8625-8630

Halskau O, Underhaug J, Froystein N A and Martinez A (2005) Conformational flexibility of alpha-lactalbumin related to its membrane binding capacity *J Mol Biol* **349** 1072-1086

Henry E R and Eaton W A (2004) Combinatorial modeling of protein folding kinetics: free energy profiles and rates *Chem Phys* **307** 163-185

Henry E R, Best R B and Eaton W A (2013) Comparing a simple theoretical model for protein folding with all-atom molecular dynamics simulations *Proc Natl Acad Sci USA* **110** 17880-17885

Hilser V J and Freire E (1996) Structure-based calculation of the equilibrium folding pathway of proteins. Correlation with hydrogen exchange protection factors *J Mol Biol* **262** 756-772

Hu W B, Walters B T, Kan Z Y, Mayne L, Rosen L E *et al.* (2013) Stepwise protein folding at near amino acid resolution by hydrogen exchange and mass spectrometry *Proc Natl Acad Sci USA* **110** 7684-7689

Hyeon C and Thirumalai D (2011) Capturing the essence of folding and functions of biomolecules using coarse-grained models *Nat Commun* **2** 487

Ibarra-Molero B, Loladze V V, Makhatadze G I and Sanchez-Ruiz J M (1999) Thermal versus guanidine-induced unfolding of ubiquitin. An analysis in terms of the contributions from charge-charge interactions to protein stability *Biochemistry* **38** 8138-8149

Ikegami A (1981) Statistical thermodynamics of proteins and protein denaturation *Adv Chem Phys* **46** 363-413

Jones C M, Henry E R, Hu Y, Chan C K, Luck S D *et al.* (1993) Fast Events in Protein-Folding Initiated by Nanosecond Laser Photolysis *Proc Natl Acad Sci USA* **90** 11860-11864

Kubelka G S and Kubelka J (2014) Site-specific thermodynamic stability and unfolding of a de novo designed protein structural motif mapped by 13C isotopically edited IR spectroscopy *J Am Chem Soc* **136** 6037-6048

Kubelka J, Henry E R, Cellmer T, Hofrichter J and Eaton W A (2008) Chemical, physical, and theoretical kinetics of an ultrafast folding protein *Proc Natl Acad Sci USA* **105** 18655-18662

Kumar S, Tsai C J and Nussinov R (2000) Factors enhancing protein thermostability *Protein Eng* **13** 179-191

Lakshmikanth G S, Sridevi K, Krishnamoorthy G and Udgaonkar J B (2001) Structure is lost incrementally during the unfolding of barstar *Nat Struct Biol* **8** 799-804

Leone V, Marinelli F, Carloni P and Parrinello M (2010) Targeting biomolecular flexibility with metadynamics *Curr Opin Struct Biol* **20** 148-154

Levinthal C (1968) Are There Pathways for Protein Folding? *Journal De Chimie Physique Et De Physico-Chimie Biologique* **65** 44

Levy Y, Onuchic J N and Wolynes P G (2007) Fly-casting in protein-DNA binding: Frustration between protein folding and electrostatics facilitates target recognition *J Am Chem Soc* **129** 738-739

Li L, Li C, Zhang Z and Alexov E (2013) On the Dielectric "Constant" of Proteins: Smooth Dielectric Function for Macromolecular Modeling and Its Implementation in DelPhi *J Chem Theory Comput* **9** 2126-2136

Lindorff-Larsen K, Piana S, Dror R O and Shaw D E (2011) How Fast-Folding Proteins Fold *Science* **334** 517-520

Loladze V V, Ibarra-Molero B, Sanchez-Ruiz J M and Makhatadze G I (1999) Engineering a thermostable protein via optimization of charge-charge interactions on the protein surface *Biochemistry* **38** 16419-16423

Lumry R, Biltonen R L and Brandts J F (1966) Validity of the "Two-State" Hypothesis for Conformational Transitions of Proteins *Biopolymers* **4** 917-944

Ma H R and Gruebele M (2005) Kinetics are probe-dependent during downhill folding of an engineered lambda (6-85) protein *Proc Natl Acad Sci USA* **102** 2283-2287

Ma J, Pazos I M, Zhang W, Culik R M and Gai F (2015) Site-specific infrared probes of proteins *Ann Rev Phys Chem* **66** 357-377

Mirny L and Shakhnovich E (2001) Protein folding theory: From lattice to all-atom models *Ann Rev Biophys Biomol Struct* **30** 361-396

Moffitt J R, Chemla Y R, Smith S B and Bustamante C (2008) Recent advances in optical tweezers *Ann Rev Biochem* **77** 205-228

Moody C L, Tretyachenko-Ladokhina V, Laue T M, Senear D F and Cocco M J (2011) Multiple Conformations of the Cytidine Repressor DNA-Binding Domain Coalesce to One upon Recognition of a Specific DNA Surface *Biochemistry* **50** 6622-6632

Munõz V and Eaton W A (1999) A simple model for calculating the kinetics of protein folding from three-dimensional structures *Proc Natl Acad Sci USA* **96** 11311-11316

MunõzV and Sanchez-Ruiz J M (2004) Exploring protein folding ensembles: a variable barrier model for the analysis of equilibrium unfolding experiments *Proc Natl Acad Sci USA* **101** 17646-17651

Munõz V and Serrano L (1994) Intrinsic Secondary Structure Propensities of the Amino-Acids, Using Statistical Phi-Psi Matrices - Comparison with Experimental Scales *Proteins* **20** 301-311

Munshi S and Naganathan A N (2015) Imprints of function on the folding landscape: functional role for an intermediate in a conserved eukaryotic binding protein *Phys Chem Chem Phys* **17** 11042-11052

Naganathan A N (2012) Predictions from an Ising-like Statistical Mechanical Model on the Dynamic and Thermodynamic Effects of Protein Surface Electrostatics *J Chem Theory Comput* **8** 4646-4656

Naganathan A N (2013) A Rapid, Ensemble and Free Energy Based Method for Engineering Protein Stabilities *J Phys Chem B* **117** 4956-4964

Naganathan A N (2013) Coarse-grained models of protein folding as detailed tools to connect with experiments *WIREs Comput Mol Sci* **3** 504-514

Naganathan A N and Munõz V (2010) Insights into protein folding mechanisms from large scale analysis of mutational effects *Proc Natl Acad Sci USA* **107** 8611-8616

Naganathan A N and Orozco M (2011) The protein folding transition-state ensemble from a G (o) -like model *Phys Chem Chem Phys* **13** 15166-15174

Naganathan A N and Orozco M (2013) The conformational landscape of an intrinsically disordered DNA-binding domain of a transcription regulator *J Phys Chem B* **117** 13842-13850

Naganathan A N, Doshi U and Munõz V (2007) Protein folding kinetics: Barrier effects in chemical and thermal denaturation experiments *J Am Chem Soc* **129** 5673-5682

Naganathan A N, Doshi U, Fung A, Sadqi M and Munõz V (2006) Dynamics, energetics, and structure in protein folding *Biochemistry* **45** 8466-8475

Naganathan A N, Sanchez-Ruiz J M, Munshi S and Suresh S (2015) Are Protein Folding Intermediates the Evolutionary Consequence of Functional Constraints? *J Phys Chem B* **119** 1323-1333

Narayan A and Naganathan A N (2014) Evidence for the sequential folding mechanism in RNase H from an ensemble-based model *J Phys Chem B* **118** 5050-5058

Onuchic J N, LutheySchulten Z and Wolynes P G (1997) Theory of protein folding: The energy landscape perspective *Ann Rev Phys Chem* **48** 545-600

Perez A, Morrone J A, Simmerling C and Dill K A (2016) Advances in free-energy-based simulations of protein folding and ligand binding *Curr Opin Struct Biol* **36** 25-31

Piana S, Klepeis J L and Shaw D E (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations *Curr Opin Struct Biol* **24** 98-105

Ramachandran G N, Ramakrishnan C and Sasisekharan V (1963) Stereochemistry of polypeptide chain configurations *J Mol Biol* **7** 95-99

Raschke T M, Kho J and Marqusee S (1999) Confirmation of the hierarchical folding of RNase H: a protein engineering study *Nat Struct Biol* **6** 825-831

Robertson A D and Murphy K P (1997) Protein structure and the energetics of protein stability *Chem Rev* **97** 1251-1267

Sadqi M, de Alba E, Perez-Jimenez R, Sanchez-Ruiz J M and Muñoz V (2009) A designed protein as experimental model of primordial folding *Proc Natl Acad Sci USA* **106** 4127-4132

Sanchez-Ruiz J M (2011) Probing free-energy surfaces with differential scanning calorimetry *Ann Rev Phys Chem* **62** 231-255

Sanchez-Ruiz J M and Makhatadze G I (2001) To charge or not to charge? *Trends Biotech* **19** 132-135

Sarkar S S, Udgaonkar J B and Krishnamoorthy G (2013) Unfolding of a Small Protein Proceeds via Dry and Wet Globules and a Solvated Transition State *Biophys J* **105** 2392-2402

Schuler B and Hofmann H (2013) Single-molecule spectroscopy of protein folding dynamics—expanding scope and timescales *Curr Opin Struct Biol* **23** 36-47

Sekhar A and Kay L E (2013) NMR paves the way for atomic level descriptions of sparsely populated, transiently formed biomolecular conformers *Proc Natl Acad Sci USA* **110** 12867-12874

Shea J E, Onuchic J N and Brooks C L (2000) Energetic frustration and the nature of the transition state in protein folding *J Chem Phys* **113** 7663-7671

Sinha K K and Udgaonkar J B (2008) Barrierless evolution of structure during the submillisecond refolding reaction of a small protein *Proc Natl Acad Sci USA* **105** 7998-8003

Sivanandan S and Naganathan A N (2013) A Disorder-Induced Domino-Like Destabilization Mechanism Governs the Folding and Functional Dynamics of the Repeat Protein IêBá *PLOS Comp Biol* **9** e1003403

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

Southall N T, Dill K A and Haymet A D J (2002) A view of the hydrophobic effect *J Phys Chem B* **106** 521-533

Sridevi K and Udgaonkar J B (2002) Unfolding Rates of Barstar Determined in Native and Low Denaturant Conditions Indicate the Presence of Intermediates *Biochemistry* **41** 1568-1578

Taketomi H, Ueda Y and Go N (1975) Studies on Protein Folding, Unfolding and Fluctuations by Computer-Simulation1 Effect of Specific Amino-Acid Sequence Represented by Specific Inter-Unit Interactions *Inter J Prot Pep Res* **7** 445-459

Tanford C and Kirkwood J G (1957) Theory of Protein Titration Curves I. General Equations for Impenetrable Spheres *J Am Chem Soc* **79** 5333-5339

The PyMOL Molecular Graphics System Version 1.2r1 Schrödinger, LLC

Thirumalai D (1995) From Minimal Models to Real Proteins - Time Scales for Protein-Folding Kinetics *J Phys* I **5** 1457-1467

Udgaonkar J B (2008) Multiple routes and structural heterogeneity in protein folding *Ann Rev Biophys***37** 489-510

Uversky V N (2013) A decade and a half of protein intrinsic disorder: Biology still waits for physics *Protein Sci* **22** 693-724

Uversky V N, Gillespie J R and Fink A L (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins: Struct Funct Genet* **41** 415-427

Vendruscolo M (2007) Determination of conformationally heterogeneous states of proteins *Curr Opin Struct Biol* **17** 15-20

Vicatos S, Roca M and Warshel A (2009) Effective approach for calculations of absolute stability of proteins using focused dielectric constants *Proteins: Struct Func Bioinf* **77** 670-684

Vu D M, Myers J K, Oas T G and Dyer R B (2004) Probing the folding and unfolding dynamics of secondary and tertiary structures in a three-helix bundle protein *Biochemistry* **43** 3582-3589

Wako H and Saito N (1978) Statistical Mechanical Theory of Protein Conformation.2. Folding Pathway for Protein *J Phys Soc Japan* **44** 1939-1945

Walters A L, Deka P, Corrent C, Callender D, Varani G *et al.* (2007) The highly cooperative folding of small naturally occurring proteins is likely the result of natural selection *Cell* **128** 613-624

Yin S Y, Ding F and Dokholyan N V (2007) Eris: an automated estimator of protein stability *Nat Methods* **4** 466-467.