4

An Interfacial Thermodynamics Model for Protein Stability

Donald J. Jacobs
University of North Carolina at Charlotte
USA

1. Introduction

Proteins are important macromolecules that exhibit thermodynamic and kinetic properties that are highly tuned to facilitate biological function within limited ranges of environmental conditions. Despite having a wealth of understanding of the interactions that affect protein stability [Dill, 1990; Pace, et al. 2004], such as the hydrophobic effect, hydrogen bonding, packing, solvation and electrostatic effects: Predicting thermodynamic properties of proteins is difficult because these interactions simultaneously work together within the molecular structure comprising of heterogeneous microenvironments that change dynamically as the conformational state of the protein changes. Consequently, a protein is truly a *complex system* [Bar-Yam, 1997] where thermodynamic and other emergent physical properties are sensitive to small perturbations in protein structure or its environment. It remains an open problem to develop models that can accurately predict protein stability, ligand-protein binding affinities and allosteric response, all of which are critical to the function of a protein [Petsko & Ringe, 2004; Klepeis, et al. 2004; Bray & Duke, 2004].

1.1 Available computational approaches

A rigorous brute force method that can in principle computationally predict thermodynamic properties of proteins is through all-atom molecular dynamics (MD) simulation in explicit solvent [Lindorff-Larsen, et al. 2009]. In this approach, the equations of motions for all atoms must be integrated over femtosecond time-steps out to timescales extending to hours or more. Then the emergent properties governing how a protein functions must be extracted from the massive amount of atomic coordinates contained in the MD trajectory. Invoking the ergodic hypothesis to determine thermodynamic averages of physical quantities by time averaging over many such long trajectories makes this approach painstakingly slow, especially when one would like to scan over large numbers of possible what-if scenarios for the purpose of finding conditions that yield a desired "engineered" response, such as identifying specific mutations, ligands or solvent composition. Unfortunately, it is not yet possible to reduce statistical errors by employing MD simulations to a point where they are negligible, especially in the context of high-throughput applications.

Despite practical limitations in sampling, MD has proven indispensible for gaining insight into protein function over limited timescales [Gunsteren, et al. 2006]. Furthermore, a coarse-

grained MD approach with implicit solvent is faster, and multiscale modelling can be fast while achieving good accuracy [Nielsen, et al. 2010]. While multiscale MD is promising, substantial computer resources are required that preclude high-throughput applications, including studies that systematically vary temperature, pressure, pH, and concentration of co-solvents. Therefore, it is not yet feasible in engineering design applications to calculate free energy and other thermodynamic properties of proteins from MD simulations.

There are alternative approaches that represent the fold of a protein in terms of an Ising-like model involving discrete "spin" variables assigned to residues [Hilser & Freire, 1996; Muñoz 2001; Bakk & Hoye 2002; Zamparo & Pelizzola, 2006]. By discretizing macrostates, conformational ensembles that include native, unfolded and partially unfolded states can be generated efficiently [Jacobs, 2010]. Moreover, calculating thermodynamic properties is feasible as a result of the approximations used to reduce *degrees of freedom* (DOF). One such critical approximation common to Ising-like models is that the three-dimensional native structure is used as a *template*. Spin variables decorate the template to partition the protein at the residue level into native-like (spin up) and disordered (spin down) regions. For N residues, the 2^N possible spin configurations retain only topological significance because geometrical information beyond the native state structure is not considered. The models are simple enough that for practical purposes exact thermodynamic properties (of the model) can be calculated. However, because non-native interactions are not accessible, the ensemble of conformational states generated (expressed by microscopic "spin" configurations) is not complete physically. Consequently, Ising-like model predictions can be made *rapidly*, they are *precise*, but *accuracy* becomes questionable because the models tend to be oversimplified.

1.2 Identifying a fundamental problem

Most Ising-like models that describe protein stability are based on the concept of *free energy decomposition* (FED), where the free energy of a system is partitioned into parts by assigning enthalpy and entropy contributions to subsystems. Assuming transferability, a ledger is created to account for the gain or loss of enthalpy and entropy relative to a reference state. The naïve method is to sum over all enthalpy and entropy contributions to arrive at the total free energy of the protein in a specific macrostate. This approach is extremely fast, and it is accurate when all subsystems are essentially independent of one another. Unfortunately, errors occur when DOF within subsystems couple. Since cooperative behaviour is typically found in proteins, the assumption of additivity generally fails [Mark & van Gunsteren, 1994; Dill, 1997], causing inaccurate predictions and/or model parameters to be non-transferable.

The application of FED and assumption of additivity is commonly employed to interpret single site mutations and ligand binding affinities. One reason why it is not obvious that the assumption of additivity is flawed is because non-transferability of model parameters also derives from a lack of completeness in modelling interactions, and, Ising-like models are notoriously incomplete. Nevertheless, the fundamental problem is in treating subsystems as independent, which is tantamount to assuming all internal DOF are independent, and this leads to a dramatic overestimate of conformational entropy in the native-state relative to the unfolded state. Fortunately, this problem can be largely overcome by keeping track of the correlations between DOF using concepts of *network-rigidity* (also called *graph-rigidity*).

1.3 The distance constraint model

The challenge of accurately predicting protein stability lies in developing a model that can account for all essential types of interactions while demanding the model is computationally tractable for high-throughput applications. Toward this goal, I describe a *Distance Constraint Model* (DCM) that is an Ising-like model that employs a FED, but the assumption of additivity is not used. The total free energy is calculated through the process of *Free Energy Reconstitution* (FER) to account for coupling of DOF between subsystems. The critical component of the FER is to employ network-rigidity as a long-range mechanical interaction to govern the non-additive nature of conformational entropy.

The DCM has been employed in various forms (differences in model details and methods to solve the DCM) to describe the helix-coil transition [Jacobs, et al. 2003; Jacobs & Wood, 2004; Lee, et al. 2004; Vorov, et al. 2009; Wood, et al. 2011], the hairpin-coil transition [Jacobs & Fairchild, 2007a] and protein stability [Livesay, et al. 2004; Jacobs & Dallakayan, 2005; Jacobs, et al. 2006a; Vorov, et al. 2011]. Moreover, the DCM predicts substructures within a protein that are rigid or flexible, and identifies sets of atoms that are co-rigid or co-flexible within a correlated motion. Many studies on proteins using a *minimal* DCM (mDCM) have elucidated stability/flexibility relationships important to function [Livesay & Jacobs, 2006; Livesay, et al. 2008; Mottonen, et al. 2009; Verma, et al. 2010] including the study of allostery [Mottonen, et al. 2010]. The DCM provides a good estimate for conformational entropy in simple loop systems compared to exact calculations [Vorov, et al. 2008]. This body of work has been reviewed previously [Jacobs, 2006b; Jacobs, et al. 2012]. The success of the DCM across disparate systems indicates that it is well suited for high-throughput applications that include macromolecular design, large-scale comparative analysis and drug discovery.

1.3.1 Advantages of the DCM and its limitations

The DCM offers several advantages over other models for protein stability, listed in order of importance: 1) Network-rigidity is employed to account for the coupling of DOF between subsystems for better estimates of conformational entropy, thereby restoring the utility of a FED. 2) Structural characteristics are linked to thermodynamic properties by associating mechanical constraints to enthalpy-entropy compensation mechanisms. 3) Molecular structure is represented at the all-atom level. 4) The DCM is not restricted to use template structures, although use of templates allows rapid calculation of the partition function. 5) Relationships between flexibility and stability are quantified, which gives insight into the mechanisms of protein function [Luque, et al. 2002]. 6) The DCM can be solved efficiently in multiple ways. 7) The DCM is a general approach that is not restricted to proteins.

Limitations of the DCM include: 1) Calculation of conformational entropy is *approximate*. Errors are introduced when the geometrical problem is simplified to a topological one (explained below), and because loop corrections are neglected. 2) Long-range electrostatic interactions are not considered. To acknowledge these limitations, the DCM is formulated as an empirical spin-model. 3) Moreover, in the mDCM, mean-field approximations are used to replace many essential enthalpy-entropy compensation mechanisms that are not explicitly modelled, especially in regards to solvation effects. Thus, effective parameters are required to compensate for mDCM inadequacies, suggesting that non-transferability in parameters observed across diverse proteins largely derives from discretionary oversimplifications.

1.3.2 Generalization of the DCM

Despite several studies indicating the mDCM is useful for predicting flexibility and stability of proteins, the merits and limitations of the DCM paradigm remain to be assessed. Many limitations can be substantially reduced by generalizing the form of the DCM to provide a more accurate estimate for conformational entropy without much more computational cost. With the goal of establishing an accurate high-throughput empirical approach, adding terms to model solvation effects offers the least amount of effort for the greatest improvement in accuracy, parameter transferability and retaining rapid calculations. Here, I will redevelop the DCM in the context of interfacial thermodynamics, which has not been done before.

The DCM has four key elements each of which will be discussed separately. First, network-rigidity is explicitly considered as a long-range mechanical interaction to account for the non-additive property of conformational entropy. Second, the FED provides a complete set of elementary subsystems and interaction types that are uniquely identified based on the three dimensional structure of the protein and its macrostate. Third, order parameters are used to define the macrostate of a protein that include the composition of protein-solvent interactions and the number of native-like intramolecular interactions. Forth, applying the FER to each macrostate allows the *free energy landscape* (FEL) to be calculated.

2. Linking network-rigidity to conformational entropy

As the amplitude of motion of a flexible molecular structure increases, the conformational entropy will increase accordingly. By ascribing an entropic measure to distance constraints, the DCM posits a quantitative link between network-rigidity and conformational entropy [Jacobs, et al. 2003]. This link requires assigning and characterizing tolerances to constraints.

2.1 Draconian view of network-rigidity

A draconian view of network-rigidity in proteins is that *some* interactions can be modelled by placing a distance constraint between certain pairs of atoms, while the distances between all other pairs of atoms are not fixed. Given a network of distance constraints, the program FIRST (Floppy Inclusion and Rigid Substructure Topography) gives a detailed mechanical analysis of protein structure [Jacobs, et. al. 2001] that includes rigid cluster decomposition (RCD). A RCD defines all rigid substructures where the distance between all pairs of atoms is fixed within a substructure. As such, a protein is modelled as a collection of rigid bodies, where conformational change is through relative motions between rigid substructures.

The RCD depends on the set of distance constraints modelling various types of interactions. Covalent bonds are always modelled as distance constraints, while other interactions may or may not contribute to distance constraints. For example, the hydrogen bond (H-bond) has a wide variation of strength. In FIRST, a *dilution* analysis is employed to represent a H-bond as 5 distance constraints when its energy is lower than some cut-off energy. By *lowering* this cut-off, more H-bonds are identified as *weak*, which do not contribute distance constraints. As such, a protein will undergo a mechanical transition from being mostly a rigid structure with flexible pockets (all H-bonds contribute to distance constraints whether weak or strong) to a globally floppy structure interconnecting many small rigid clusters (only the strongest H-bonds contribute distance constraints). This dilution idea [Jacobs, et al. 1999] was later interpreted as a *kinetic* mechanism for protein unfolding [Rader, et al. 2002].

The notion of a rigid substructure is an idealization. For example, FIRST often identifies a long alpha-helix as a rigid substructure, but an alpha-helix actually bends and twists (just like a metal bar can do!). Nevertheless, the RCD is useful to understand long-time scales, where small amplitude conformational deviations in substructures within a protein are neglected, such as the compression, elongation, bending or twisting of an alpha-helix. The rapid calculations for the RCD by FIRST (requiring tiny fractions of a second) has proved to be useful in making comparative studies across protein families, and to elucidate common structural features regarding flexibility important to function [Hespenheide, et al. 2002; Rader, et al. 2004; Fuxreiter, et al. 2005; Costa, et al. 2006; Radestock & Gohlke 2008; Mamonova et al. 2008; Rader, 2010; Heal, et al. 2011; Radestock & Gohlke 2011]. It has also been shown there is a statistically significant correlation between the propagation of rigidity between two mutation sites within a protein to non-additive effects in free energy cycles describing double mutant studies [Istomin, et al. 2008].

If all weak interactions are allowed to contribute to distance constraints, FIRST will predict a completely rigid protein, failing to be of any use. Instead, excluded volume effects due to van der Waals interactions are included in geometrical simulation that allows rigid clusters to wiggle about without violating any distance constraint, or without atoms passing through one another. FRODA (Framework Rigidity Optimized Dynamics Algorithm) is one method [Wells, et al. 2005; Farrell, et al. 2010], among others [Lei, et al. 2004; Thomas, et al. 2007; Jimenez-Roldan, et al. 2011; Yao, et al. 2012] that uses FIRST to identify a native RCD that is preserved during the simulation. FRODA efficiently explores the native state ensemble of conformations [Jacobs, 2010; David & Jacobs 2011]. The main limitation is that the native state structure defines the distance constraints, and once set, they never break. This *athermal* mechanical description of a protein cannot account for non-native contacts, making large-scale conformational change between different conformational states impossible if native-contacts in either state must be broken along the pathway. However, pathways between conformational states can be achieved by using a common collection of distance constraints between native conformations [Farrell, et al. 2010].

2.2 Liberated view of network-rigidity

The draconian view suffers from three awkward problems. There is no prescription for (1) *how* to determine the proper number of distance constraints to model an interaction, and for (2) *when* an interaction will contribute distance constraints. Also, (3) the selection threshold that determines whether distance constraints are placed in the network causes artificial discontinuities. A discontinuity is illustrated by two nearly identical H-bonds having a small energy difference slightly above and lower than the cut-off energy. The H-bond with higher energy is modelled as *infinitely weak* (not present) while the other is *infinitely strong*!

By assigning tolerances to configuration *variables*, the accessible range for each variable can be quantified. The distinction between a DOF and a constraint is reflected in the process of *before and after* a tolerance is assigned. That is, a variable *acts* as a DOF before its tolerance is assigned. Once the range of a variable is restricted, it acts as a generic distance constraint. Applying graph-rigidity algorithms determine if a generic distance constraint is redundant or independent. A redundant distance constraint has zero tolerance because its length is determined by independent constraints. The length of independent distance constraints can vary within their tolerances. After all variables are analysed, a system of N atoms is

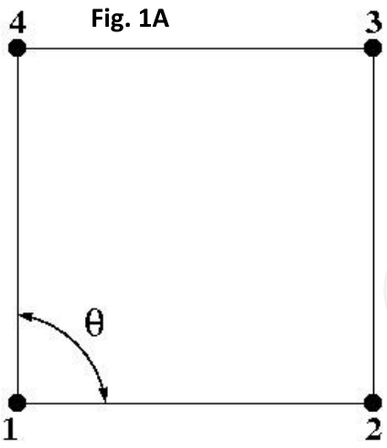
generically rigid (having 0 internal DOF) consisting of $3N-6$ independent distance constraints all with finite tolerances that quantify the accessible geometrical embedding of the constraint topology. In other words, for a given constraint topology there will be an entire ensemble of geometrical realizations that are consistent with the accessible tolerances. The conformational entropy is then related to the logarithm of this geometrical degeneracy.

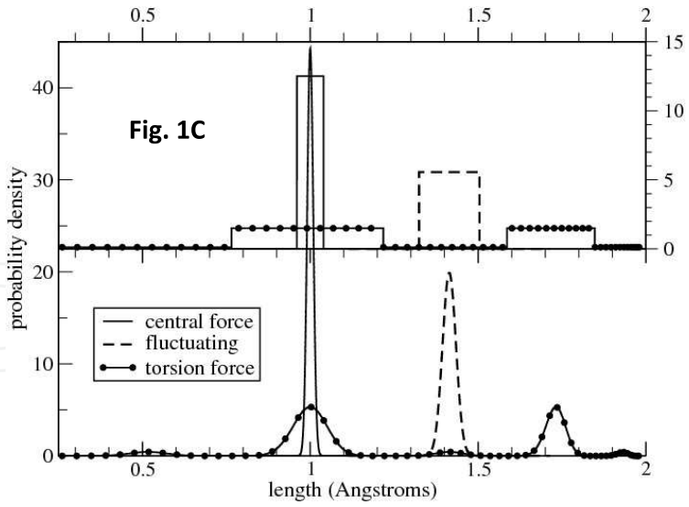
In this view, the three awkward problems mentioned above are eliminated. A system of N atoms consists of $3N-6$ internal DOF. Similarly, a subsystem with n atoms for $n \geq 3$ has $3n-6$ internal DOF, and for $n=2$ has 1 internal DOF. A proper description of a subsystem should only be in terms of independent configuration variables that need to be assigned tolerances. Therefore, the number of contributing constraints for a subsystem of n atoms is just equal to the number of its internal DOF. For example, regarding a H-bond as a 3 atom subsystem (the donor, hydrogen and acceptor) requires 3 internal DOF to specify its configuration. Thus, a H-bond will contribute 3 generic distance constraints whenever it forms, and it will randomly form or break based on a probability that is appropriate for the system to be in thermodynamic equilibrium, which removes arbitrary thresholds.

2.2.1 Illustration: A two dimensional quadrilateral

To illustrate the points discussed above, consider 4 particles confined to a plane as shown in Fig. 1A. Quenched central-force interactions are between particles (1,2), (2,3), (3,4) and (4,1) with respective relative distances: ℓ_{12}, ℓ_{23}, ℓ_{34} and ℓ_{41}. The word "quenched" indicates that the interaction is always present within the network. Fluctuating interactions can also form between particles (1,3) and (2,4) with respective distances ℓ_{13}, ℓ_{24}. Quenched torsion-force interactions are between particles (4,1,2), (1,2,3), (2,3,4) and (3,4,1) with respective angles θ_1, θ_2, θ_3, θ_4. The potential energy for the *central* (c), *fluctuating* (f) and *torsion* (t) interactions are given as: $V_c = E_c + g\left(\frac{\ell-a}{\Delta L_c}\right)^2$, $V_f = E_f + g\left(\frac{\ell-b}{\Delta L_f}\right)^2$ and $V_t = E_t^k + g\left(\frac{\theta-\theta_k}{\Delta\theta}\right)^2$ where E_c, E_f, E_t^k are reference energies, g is a scaled spring constant, ΔL_c, ΔL_f, $\Delta\theta$ are configuration variable tolerances, a and b are equilibrium lengths, θ_k is an equilibrium angle, and the index, k, labels bins that partition the angle range. Four cases are compared below that differ only in g and b, each assigned two distinct values with common parameters held fixed. Common parameters are set to: E_c = -20 kcal/mol, E_f = -5 kcal/mol, $E_t^1 = E_t^3 = E_t^5$ = 2 kcal/mol, $E_t^2 = E_t^4$ = 0, $a = 1$ Å, ΔL_c = 0.04 Å, ΔL_f = 0.09 Å, $\Delta\theta$ = 15° and k = 1, 2, 3, 4, 5 to define five bins for the respective angle ranges (15°, 45°), (45°, 75°), (75°, 105°), (105°, 135°), (135°, 165°) and for corresponding θ_k = 30°, 60°, 90°, 120°, 150°.

The four cases to be considered are: g = 8 kcal/mol or 0, and $b = \sqrt{2}$ Å or 1.5321 Å. For the *harmonic* potential where g = 8 kcal/mol; the quadrilateral defines an elastic network. A *flat* potential (g = 0) has a constant energy as lengths or angles vary within a tolerance, but an infinite energy outside the tolerance. A sketch of how much the quadrilateral can deform for a flat potential is shown in Fig. 1B. Shaded regions define accessible geometries for each angular range (5 rows) and distinct constraint topology (4 columns) due to the fluctuating





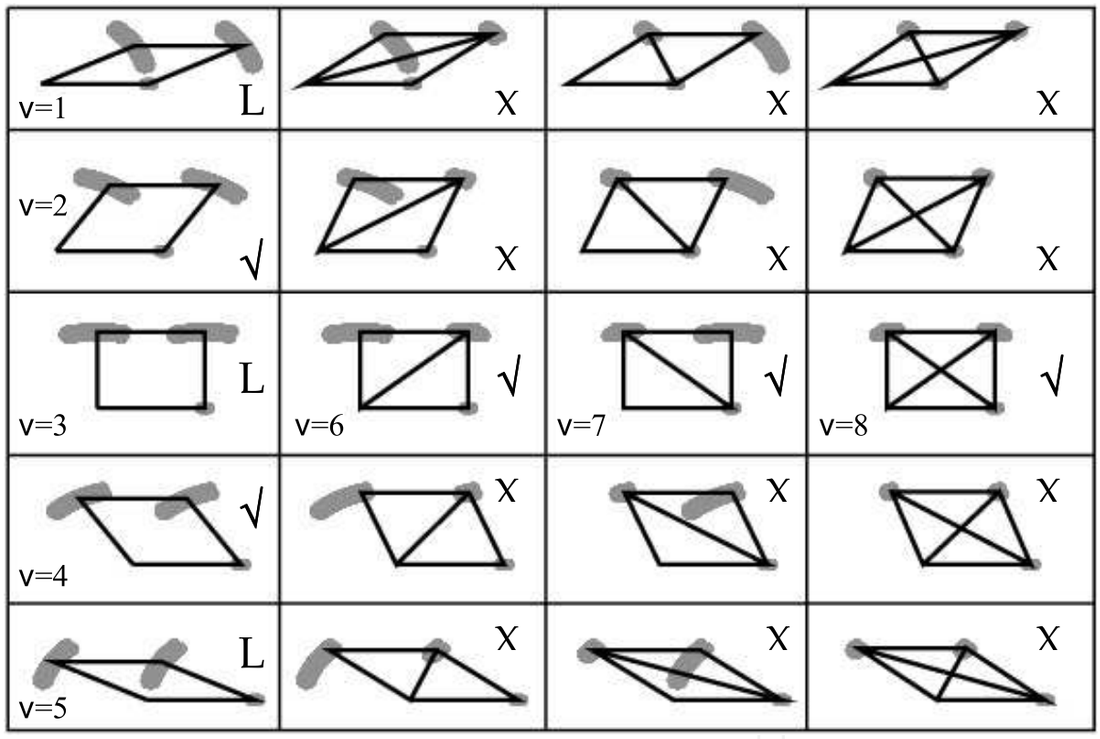


Fig. 1. (A) Definition of the quadrilateral network. (B) The accessible range of motion for each coarse-grained configuration is shown in grey. The middle two columns show how the more limited range of motion along a diagonal interaction compared to the range allowed by the torsion-forces is the determining factor restricting the motion. The symbols (√, L, X) indicate which configurations are (very probable, improbable, inaccessible). The index, ν, is used to label all accessible configurations. In the example considered here, the inaccessible configurations essentially have infinite energy, but in a more refined model they could have a finite energy. (C) Treating interactions as *independent*, the probability density, $\rho(\ell)$, for the length of a distance constraint is plotted for central-, fluctuating- and torsion-interactions, where the lower and upper graphs correspond to the cases: g = 8 kcal/mol and g = 0.

interaction along the diagonals. Fluctuating interactions break and form to reflect hidden DOF in the system not modelled. For example, a fluctuating interaction along a diagonal of the quadrilateral could model a H-bond that may appear or not depending on the details of the electronic structure of the donor, hydrogen and acceptor atoms.

2.2.2 Free energy decomposition and reconstitution

The configuration integral, $Q_\nu = \delta^{-5}\int\cdots\int \exp(-\beta V_\nu)\, dx_2 dx_3 dy_3 dx_4 dy_4$ is calculated where V_ν is the potential energy of the system in the ν–th coarse-grained configuration, $\beta = (RT)^{-1}$, where R is the ideal gas constant, T is absolute temperature, and $\delta = 0.002$ Å is a length scale factor (in classical statistical mechanics entropy is defined up to an arbitrary constant). Since rigid body DOF are not of interest, the quadrilateral is translated and rotated so that particle 1 is at the origin while particle 2 is along the x-axis, requiring a five dimensional integral. From Q_ν, the total internal free energy is given as $F_\nu = -RT\ln(Q_\nu)$, thermal energy is $U_\nu = \langle V_\nu \rangle$, and entropy is $R\tau_\nu$, where $\tau_\nu = \beta(U_\nu - F_\nu)$ is the total *pure* entropy. Similarly, the FED defines $\{Q_x, F_x, U_x, \sigma_x\}$ for each interaction type when *treated* as an independent subsystem. For the quadrilateral: $x \to c, f, t1, t2, t3, t4, t5$ denotes central-, fluctuating-, and torsion interactions for bin 1,2,3,4,5 respectively. Taking care to maintain the spatial length scale, $Q_c = \delta^{-1}\int \exp(-\beta V_c)\, d\ell$, $Q_f = \delta^{-1}\int \exp(-\beta V_f)\, d\ell$, and $Q_{tk} \approx \delta^{-1}\int \exp(-\beta V_k)\, \frac{d\ell}{d\theta}\, d\theta$ where a Jacobian is inserted to convert from an angle to a length measure using an approximation that constrains $\ell_{12} = \ell_{23} = \ell_{34} = \ell_{41} = a$. Consequently, a slight underestimate of the entropy is incurred because the angle is actually defined without imposing these length constraints. The probability density, $\rho(\ell)$, for finding the distance between a pair of particles (the plot is for T=400K) for a particular interaction is shown in Fig. 1C, which is given by the Boltzmann factor normalized by the corresponding partition function.

For each interaction type, x, the thermal energy, U_x, and pure entropy, σ_x, are plotted in Fig. 2A and Fig. 2B respectively for a flat potential energy, and in Fig. 2C and Fig. 2D for a harmonic potential energy. Notice that σ_x decreases when the peak width in the probability density, $\rho(\ell)$, decreases as shown in Fig. 1C. Interestingly, transforming from an angle to a length variable to describe torsion interactions causes σ_{tk} to decrease as the angle increases (larger k) because a smaller length variation results from the same angular deviation. This difference is important when network rigidity is used to calculate conformational entropy.

For an *additive* FER: The free energy for the ν–th configuration is: $F_1 = -RT\ln(Q_c^4 Q_{t1}^2 Q_{t5}^2)$, $F_2 = -RT\ln(Q_c^4 Q_{t2}^2 Q_{t4}^2)$, $F_3 = -RT\ln(Q_c^4 Q_{t3}^4)$, $F_4 = F_2$, $F_5 = F_1$, $F_6 = F_7 = -RT\ln(Q_f Q_c^4 Q_{t3}^4)$ and $F_8 = -RT\ln(Q_f^2 Q_c^4 Q_{t3}^4)$. Recall that a product of partition functions corresponds to adding the free energies over independent subsystems. Thus, in the additive approach, the total free energy of a configuration is simply a sum over all free energy components, such as $F_1 = 4F_c + 2F_{t1} + 2F_{t5}$ or $F_8 = 2F_f + 4F_c + 4F_{t3}$.

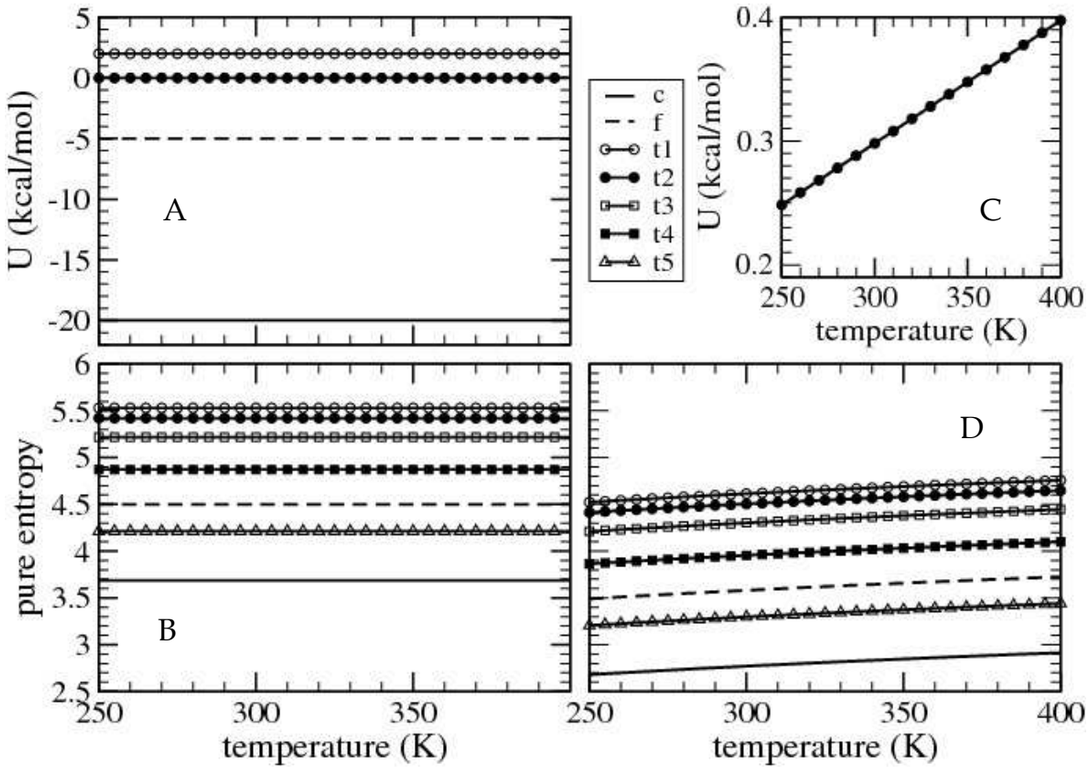


Fig. 2. The legend (middle-top) applies to all four graphs. For a flat potential: (A) Thermal energy has no temperature dependence. Note $U_{t1} = U_{t3} = U_{t5}$ and $U_{t2} = U_{t4}$. (B) The pure entropy also has no temperature dependence. For a harmonic potential: (C) Thermal energy has temperature dependence according to the equipartition theorem, but this cannot be seen using the energy scale of graph-A. Therefore, only U_{t2} is shown on a magnified scale. (D) The pure entropy monotonically increases, but it has a limiting plateau of the flat potential. Note that graph-D uses the same vertical axis as graph-B to assist in direct comparisons.

To proceed further, a more general indexing is needed to label *specific interactions*. For the quadrilateral network, there are four torsion-force, two fluctuating and four central-force interactions. Let j=1,2,3,4 label the angle variables $\theta_1, \theta_2, \theta_3, \theta_4$. Let j=5,6 label the fluctuating interactions between particles (1,3) and (2,4). Let j=7,8,9,10 label the central-force interactions between particles (1,2), (2,3), (3,4) and (4,1). The j index identifies a particular interaction within a configuration that is labelled by the ν index. In application to proteins, the ν index labels a particular accessible coarse-grained geometry that a subsystem can explore. Thus, the index-pair, νj, is a general indexing scheme used to uniquely label energy and entropy contributions. In this example, each νj maps to the x index, where $x \to c, f, t1, t2, t3, t4, t5$.

For a *non-additive* FER: The free energy of configuration, ν, is given as:

$$F_\nu = U_\nu^{cnf} + U_\nu^{vib} - T R \tau_\nu \tag{1}$$

where U_ν^{cnf} is the lowest possible energy of the selected basin, and U_ν^{vib} describes the energy associated with vibrations within the basin. Furthermore, the conformational entropy, $R\tau_\nu$,

is associated with any continuous deformation that is able to take place within a basin over a constant (or nearly constant) energy surface. The three functions are explicitly given as:

$$U_\nu^{cnf} = \sum_{\{j\}} n_{\nu j} E_{\nu j} \quad U_\nu^{vib} = \sum_{\{j\}} n_{\nu j} q_{\nu j} \langle \delta E \rangle_{\nu j} \quad \tau_\nu = \sum_{\{j\}} n_{\nu j} q_{\nu j} \sigma_{\nu j} \tag{2}$$

where the non-additive contributions can be identified by the terms containing the $q_{\nu j}$ variables. The $E_{\nu j}$ define reference energies, $\sigma_{\nu j}$ define pure entropies as plotted in Figs. 2B and 2D, and, $\langle \delta E \rangle_{\nu j}$ define the difference between thermal and reference energies. For a flat potential energy, $\langle \delta E \rangle_{\nu j}$ = 0, and for the harmonic potential with g = 8 kcal/mol, the *equipartition theorem* applies, which yields $\langle \delta E \rangle_{\nu j} = \frac{1}{2} RT$. Quantities for individual interactions, such as $E_{\nu j}$, $\sigma_{\nu j}$ and $\langle \delta E \rangle_{\nu j}$ are to be worked out in advance and stored in *lookup tables* to define the FED. The variable $n_{\nu j}$ can equal (1 or 0) when the j-th interaction is (*present* or *not present*) in the ν–th configuration. For quenched interactions, $n_{\nu j}$ = 1 $\forall \nu$. The variable $q_{\nu j}$ can equal (1 or 0) when the j-th interaction is represented by (an *independent* or *redundant*) distance constraint. It is important to notice that the assignment of which distance constraints are independent or redundant is not unique. Therefore, τ_ν will also not be unique! However, because distance constraints with smaller tolerances restrict motion more than those with greater tolerances (see Fig. 1B and view the middle two columns), the *lowest* value that can be obtained for τ_ν yields the best estimate for the *net* conformational entropy in the basin. Therefore, $q_{\nu j}$ are determined by augmenting a *preferential rule* to the graph-rigidity algorithm that manifest as building the network by placing one distance constraint at a time in the order defined by the sorted *set* of $\sigma_{\nu j}$ from *smallest to largest*.

As can be seen from Fig. 2B and Fig. 2D, the *rank ordering* of pure entropies from smallest to largest values when considering all interaction types does not depend on temperature. The same rank ordering is obtained whether flat or harmonic potentials are considered. The values for $n_{\nu j}$, $q_{\nu j}$ and corresponding $\nu j \to x$ labels are listed in Table 1 to enable explicit hand calculation of F_ν. Note that an additive FER is obtained by setting all $q_{\nu j}$ = 1, so that all interactions are considered independent, despite being *inconsistent* with network rigidity. Regardless of the FER employed, the free energy of the system with n fluctuating interactions *present* is given by: $F(n) = -RT \ln[Z(n)]$, where $Z(0) = \sum_{\nu=1}^{5} Q_\nu$, $Z(1) = Q_6 + Q_7$, $Z(2) = Q_8$ and of course $Q_\nu = \exp(-\beta F_\nu)$. The $F(n)$ are calculated in three different ways: (1) Exact answers are obtained by numerically performing the 5 dimensional configuration integral for each Q_ν , and approximate answers are obtained by employing an (2) additive and (3) non-additive FER. In Fig. 3, $F(n)$ are plotted for the four cases g = 8 or 0 kcal/mol, and $b = \sqrt{2}$ or 1.5321 Å. Table 2 summarizes the relative errors for the additive and non-additive FER predictions for the thermal energy and entropy of the system separately.

Comparing to the exact answers using Fig. 3 and Table 2, it is seen that the predictions of the non-additive FER are good for the harmonic and flat potential energy cases when b= $\sqrt{2}$ a. This is because the geometry of the fluctuating-interaction is *commensurate* with the torsion-interaction. Conversely, when b=1.5321a, the predictions breakdown because geometrical *frustration* between these two interactions cause large *strain energy* in the network. The additive and non-additive FER procedures are now juxtaposed. The results for this simple quadrilateral network example highlight the key concepts that are applied to proteins.

j=	1	2	3	4	5	6	7-10
ν	(n,q) - x	(n,q) - x	(n,q) - x	(n,q) - x	(n,q) - x	(n,q) - x	(n,q) - x
1	(1,0) - $t1$	(1,1) - $t5$	(1,0) - $t1$	(1,0) - $t5$	(0,0)	(0,0)	(1,1) - c
2	(1,0) - $t2$	(1,1) - $t4$	(1,0) - $t2$	(1,0) - $t4$	(0,0)	(0,0)	(1,1) - c
3	(1,1) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(0,0)	(0,0)	(1,1) - c
4	(1,1) - $t4$	(1,0) - $t2$	(1,0) - $t4$	(1,0) - $t2$	(0,0)	(0,0)	(1,1) - c
5	(1,1) - $t5$	(1,0) - $t1$	(1,0) - $t5$	(1,0) - $t1$	(0,0)	(0,0)	(1,1) - c
6	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,1) - f	(0,0)	(1,1) - c
7	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(0,0)	(1,1) - f	(1,1) - c
8	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,0) - $t3$	(1,1) - f	(1,0) - f	(1,1) - c

Table 1. The variables $n_{\nu j}$, $q_{\nu j}$ and corresponding x - indices are specified for all allowed configurations (for ν=1-8) and interactions (for j=1-10). The νj pair-index for the $n_{\nu j}$ and $q_{\nu j}$ variables is suppressed. Note that $q_{\nu j}$ can be determined by first placing the four x=c type of interactions in the network. Once the very next constraint is added, it will reduce one more DOF to make the network rigid. Therefore, the total number of independent distance constraints is 5 for all configurations, which can be checked by inspection.

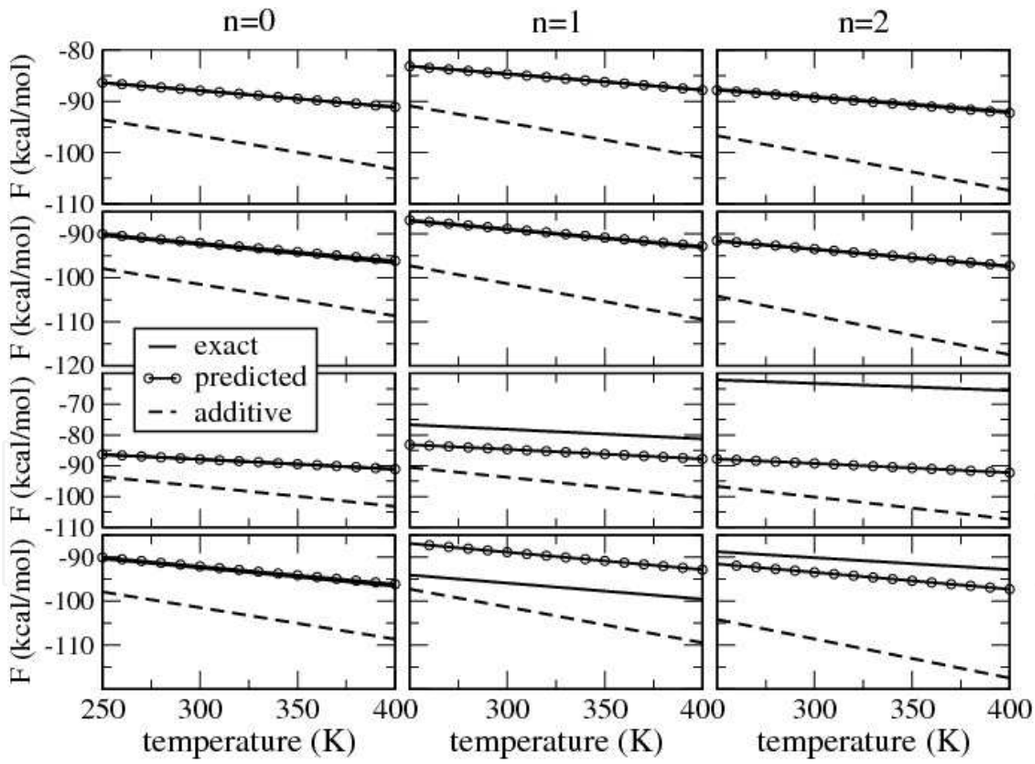


Fig. 3. The legend (left column and between middle rows) applies to all 12 graphs. The rows from top to bottom correspond to $b=\sqrt{2}$ Å for harmonic and flat potential energies, and then for b=1.5321 Å again for harmonic and flat potential energies. The columns show the free energy, $F(n)$, for n=0, 1, 2 fluctuating interactions present in the system. Note that the predicted and exact free energies for the $b=\sqrt{2}$ Å cases are nearly on top of one another.

Specs:		b= $\sqrt{2}$ a				b=1.5321a			
g	# fip	a-E-%e	a-S-%e	p-E-%e	p-S-%e	a-E-%e	a-S-%e	p-E-%e	p-S-%e
8	0	-1.8	77.8	0.0	-0.5	-1.8	77.8	0.0	-0.5
8	1	1.6	112.8	0.0	0.4	-7.6	112.9	-9.4	0.4
8	2	1.9	147.3	0.0	2.3	-39.9	213.0	-42.5	29.5
0	0	0.0	72.1	0.0	-3.0	0.0	72.1	0.0	-3.0
0	1	0.0	101.0	0.0	-1.7	9.4	122.3	9.4	8.7
0	2	0.0	129.4	0.0	-1.0	0.0	226.7	0.0	40.9

Table 2. The relative percent errors (denoted as %e) are given for energies (denoted as E–) and entropies (denoted as S–) for the additive FER (denoted as a–) and the non-additive FER prediction (denoted as p–) for when the system has 0, 1, 2 fluctuating interactions present (denoted as # fip) for $g =$ 8 kcal/mol and 0 on the top and bottom three rows respectively. Also the ($b = \sqrt{2}a$ and $b = 1.5321a$) cases are shown on the (left and right) sides of the table. These percent errors apply to T=300K, and reflect typical values at all other temperatures.

The additive approach is *completely wrong* for predicting conformational entropy because it predicts conformational entropy of a network increases as more constraining interactions appear. Thus: *The conformational entropy of a protein in its native state will always be predicted to be greater than that of the unfolded protein when using an additive model.* Although energy estimates from the additive FED are good, there are two sources of errors that occur for harmonic potentials (not for flat potentials). Part of this discrepancy is directly caused by over counting thermal energy contributions (i.e. $\frac{1}{2}RT$) for all quadratic DOF, rather than just the independent ones. This problem is more severe with respect to entropy estimates, and these errors cause relative statistical weighting of the various configurations in the ensemble to error, thereby indirectly leading to errors in average energies.

A non-additive FED *naturally* models energy-entropy compensation mechanisms that link atomic structure characteristics to thermodynamic response. As interactions form within a *protein*, such as from H-bonds or packing, more constraints are placed on conformational motions, leading to a decrease in conformational flexibility and entropy. In particular, if an energetically favourable interaction forms in an *otherwise flexible region*, it will constrain motions and decrease conformational entropy. However, if the interaction forms within an *otherwise rigid region*, no entropic cost is incurred because the constraint that is imposed is redundant. This effect is the underlying source of non-additivity. Thus, network rigidity is an interaction between entropic contributions that provides a simple mechanism for groups of interactions that constrain conformational flexibility to form and break cooperatively.

Errors appear in the non-additive FER primarily due to three reasons: (1) The Jacobian that *should* be part of the reconstitution of free energy components to account for *how subsystems interface geometrically is ignored* by assuming independent constraints are orthogonal. For example, the very skewed configurations shown in Fig. 1B have smaller conformational entropy than the configurations with all θ_k approximately 90 degrees. The coarse-graining into restricted angular bins was required to capture this difference. (2) *The energy landscape of*

a subsystem cannot change as it interacts with other subsystems. The two harmonic potential energy examples violate this assumption, but this approximation fails badly for a frustrated network. (3) *Information is lost by coarse-graining structure into local configurations* (identified by the ν index). For example, when $b = 1.5321a$ the flawed prediction in energy (~ ±10% relative error as listed in Table 2) occurs because when *just one* fluctuating-interaction is present it can stretch between multiple configurations. In addition, the building blocks are not commensurate, leading to strain energy. When *both* fluctuating interactions are present in the system, only the problem of strain energy remains.

Partitioning local structure into *finer* coarse-grained bins to define accessible configurations with *more restricted range of motion* systematically diminishes all three sources of error. A high degree of accuracy can be obtained by *finely* coarse-graining the energy landscape (such as a harmonic well) by a *large* number of flat energy tiers differing by tiny increments (say 0.01 kcal/mol). As the quantities $\{Q_x, F_x, U_x, \sigma_x\}$ are further refined for each interaction type, x, this will create more binning labels that comprise the FED. Although defeating the objective of *rapidly* calculating absolute free energies, it is important to note that errors can be reduced to low levels in principle. Also important is that the computational cost for the non-additive FER scales linearly with respect to the number of atoms in the network, N, with a pre-factor that is proportional to the number of coarse-grained bins used. Since exact integration scales as δ^{-3N}, consideration of a sophisticated FED may be worth the effort.

The errors caused by large strain energies in frustrated configurations can be identified and *removed* from the ensemble in applications to proteins based on the empirical justification that proteins exhibit folding funnels because they are minimally frustrated [Onuchic & Wolynes, 2004]. In practice, this is *accomplished* by considering only native contacts with respect to a specified template structure that is obtained experimentally (say from X-ray crystallography) or from a model structure that is fully relaxed. This leads to a FED scheme that classifies protein structure in terms of a finite number of local energy basins such as accessible backbone conformations within a Ramachandran plot and sidechain rotamers for residues. Moreover, a variety of different types of H-bonds can be classified. The complete classification of local structure defines the set of all possible subsystems that can appear within a protein. Then, the minimum of the potential energy of a basin is used to obtain the conformational part of the free energy, with the free energy contributions from modes of vibration augmented. Consequently, the U_x and σ_x parameters for the various basins are *temperature independent*. Notice that for the quadrilateral example, the temperature dependence in U_x and σ_x appears because of the harmonic potential energy, as seen in Fig. 2C and 2D. Thus, the free energy of a subsystem is separated into conformational and vibrational parts, such that $F_x^{net} = F_x^{cnf} + F_x^{vib}(\omega_x)$ where ω_x is the frequency of oscillation. Only when $\hbar\omega_x << RT$ does the equipartition theorem apply. More generally, $F_x^{vib}(\omega_x)$ is empirically modelled as the free energy of a quantum harmonic oscillator with natural frequency, ω_x. An observation that can be seen by comparing the harmonic and flat potential energy example cases is that the free energy contributions from vibration originate only from independent modes. A subsystem (in three dimensions) with n atoms for $n \geq 3$ has $3n-6$ independent modes of vibration, and one independent mode when $n = 2$. At this point, an empirical interfacial thermodynamic model can be developed.

3. Interfacial thermodynamics model for protein stability

The previous section showed how to reconstitute conformational free energy for a given constraint topology or *framework* using network rigidity. The staring information is that all energy and entropy components for accessible subsystems are stored in lookup tables. The focus was on *internal* DOF of the system. Now the affect of solvent on a protein will be included. Solvent DOF are *external* to the protein and need not be kept track of because they are part of a *reservoir*. Therefore, treating all free energy components from a reservoir as additive is consistent with a thermodynamic hypothesis. The problem that is at hand now is to determine the partition function of a protein, which takes the generic form:

$$Z = \sum_{\psi} Q_{\psi}^{slv} Q_{\psi}^{cnf} Q_{\psi}^{vib} = \sum_{\psi} Q_{\psi}^{slv} \exp\left(-\beta \sum_{\{\nu j\}_{\psi}} \left(n_{\nu j} E_{\nu j} - RT q_{\nu j} \sigma_{\nu j} \right) \right) \exp\left(-\beta \sum_{\{\nu j\}_{\psi}} q_{\nu j} F^{vib}(\omega_{\nu j}) \right) \quad (3)$$

The index, ψ, defines an accessible configuration that is generated from a template structure decorated with Ising-like spin variables to specify the local environment of each residue. Through coupling, the template decoration helps define a mechanical framework, which is specified by energy basins, labelled by ν, and its member distance constraints labelled by j, across all subsystems. The placement of distance constraints is specified by the $n_{\nu j}$ values. Topological information about the mechanical framework that is contained in ψ serves as input to a graph-rigidity analysis that yields the $q_{\nu j}$ values. Because the *number* of distance constraints and modes of vibration are equal within a subsystem, the j index is reused to label modes[1]. In Eq. (3) involving *random variables*, it is understood that whenever $n_{\nu j}$ =0, so does $q_{\nu j}$ =0, since if a distance constraint is not present, it cannot be independent. The $q_{\nu j}$ in the last term are necessary because the free energy of vibration is reconstituted by adding only *independent* modes of vibration. The term, Q_{ψ}^{slv}, takes on a form similar to many FED schemes commonly employed in the literature that relate transfer free energies to estimate changes in free energy of residues and other designated chemical groups based on whether they are exposed or buried in the protein through solvent exposed surface area.

3.1 Free Energy Decomposition (FED)

The FED accounts for enthalpy[2] and entropy contributions from solvent, conformation and vibration. The geometry of a protein is defined by one or more template structure(s). Given any template structure, all its atoms are partitioned into contiguous groups of atoms that are classified and parameterized by the FED. As such, each atom within a system must map to one and only one *molecular constituent*, which also serves as a primary subsystem. Molecular constituents in proteins define the residues, as illustrated in Fig. 4.

[1] Each mode of vibration can be represented by a distinct set of distance constraints to better capture the local atomic motions within a subsystem, but this requires another sub-index for distance constraint labeling that is suppressed for this discussion.

[2] Enthalpy is considered to be a function of pressure using standard pressure as a reference point, about which a Taylor expansion is employed. At standard conditions, the terms enthalpy and energy are considered synonymous in this discussion.

The FED considered here consists of 1) residues, 2) covalent bonds that link residue pairs, 3) H-bonds and 4) hydration interactions that together constitute a constraint network, and, the additive contributions consist of 5) residue solvation and 6) hydrophobic interactions that together model solvent effects. To account for protonation states on titratable residues, an additional solvent dependent partition function, Q_{ψ}^{ion}, must be inserted in Eq. (3), but this is not discussed here because it introduces technical complexity without offering anything more conceptually. Packing interactions are implicitly included. Because the FED can divide a system up in different ways, and because of the empirical assignments, different effects can be lumped together in various terms. Here packing effects are included in the residue states that are identified as *native-like* or *disordered*, corresponding to good or poor atomic packing with respect to the strain free template. With exception of long-range electrostatic interactions, and electing to work with a fixed protonation state, the six listed types of contributions encompass all essential enthalpy-entropy compensation mechanisms.

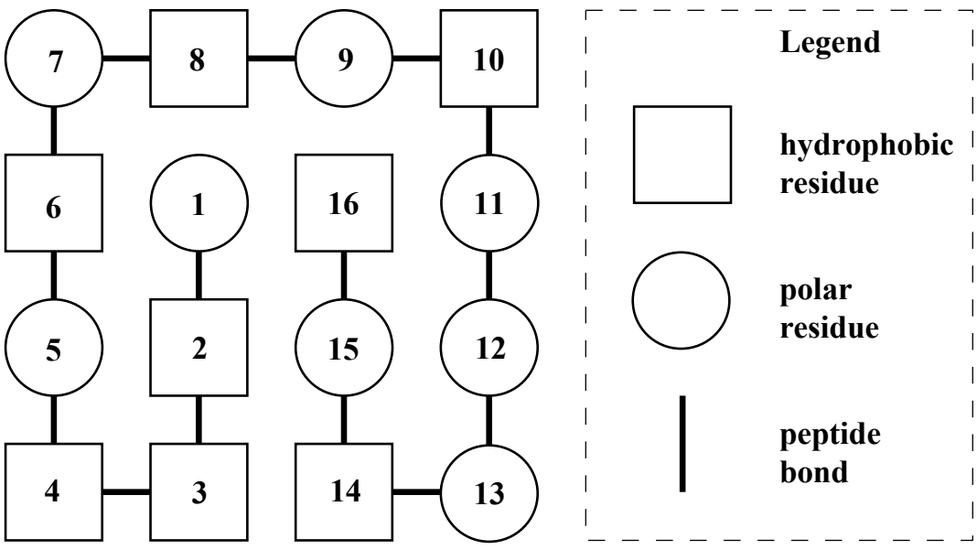


Fig. 4. A schematic of a protein template structure is shown consisting of 16 residues. Each residue defines a subsystem that includes all atoms from its backbone and sidechain. The model can include a sufficient number of probable low energy basins for each residue type covering the most frequently occurring conformations identified in the Ramachandran plot and the sidechain rotamers. The degree of coverage depends on the level of coarse graining, which ultimately controls the accuracy of the model and speed of the calculations.

3.2 The free energy functional

Solving Eq. (3) poses insurmountable challenges. Therefore, Eq. (3) is reformulated as a *free energy functional* (FEF) that can be efficiently solved numerically using self-consistent mean field theory. The FEF in generic form is written in a format that is germane to the concept of a FED, where various types of contributions can be identified.

$$G_{FEF} = \left(G_{slv}^{res} + G_{slv}^{hph} + G_{slv}^{shb} + G_{slv}^{etc}\right) + \left(G_{cnf}^{res} + G_{cnf}^{lnk} + G_{cnf}^{ihb} + G_{cnf}^{hyd} + G_{cnf}^{etc}\right) + G_{vib} \qquad (4)$$

Terms are grouped together to reflect contributions that involve solvent, conformation and vibration. Terms involving *solvent* effects (subscripted by *slv*) are: G_{slv}^{res} for residue solvation; G_{slv}^{hph} for the hydrophobic effect, G_{slv}^{shb} for solvent-protein H-bonds, and G_{slv}^{etc} indicates the model can be extended if needed, such as including a G_{slv}^{ion} contribution. Similarly, terms involving the *conformation* contributions (subscripted by *cnf*) are: G_{cnf}^{res} for the set of primary subsystems where residues define molecular constituents; G_{cnf}^{lnk} for peptide bonds linking residues together along the backbone and crosslinking disulphide bonds when present; G_{cnf}^{ihb} for intramolecular H-bonds within the protein; G_{cnf}^{hyd} for conformational constraints that are *externally imposed* on the protein structure due to solvent molecules --- often described as forming a clathrate-structure, and G_{cnf}^{etc} indicates that the model can be extended if needed, such as modelling packing interactions explicitly.

The FEF is expressed in terms of a set of a priori unknown probability functions describing the microstates of the protein. The exact nature of what the microstates are will depend on the FED. In addition to the various FED terms that make up the FEF, order parameters are employed to define the macrostate of a protein that reflect sub-ensembles of microstates. By minimizing the FEF under the global constraints imposed by the order parameters, a *free energy landscape* (FEL) is calculated. The first step is to define the FED based on microscopic mechanisms deemed important to model, which naturally leads to defining variables and their associated probability functions. The second step is to define the order parameters that will be used to define the FEL. The third step is to solve the FEF. How to solve the FEF will be explained below in a specific context of the FED. The task at hand now is to define the FED in terms of enthalpy-entropy compensation mechanisms essential to protein stability.

3.2.1 Solvent related enthalpy-entropy compensation mechanisms

Residue solvation: A residue can be *buried* (b) in the core of a protein without solvent contact, or it can be *exposed* to solvent. When exposed, the solvent molecules surrounding the residue might be *mobile* (m) or structured *clathrate* (c). Each residue is assigned a solvation state, s, to characterize its local environment, where $s = \{b, m, c\}$. Residue solvation together with the given template structure is used to specify a microstate of the protein. The ensemble of all accessible solvent states for a given template with n residues consists of 3^n configurations. The solvent state *decorates* the template structure. For example, Fig. 5 illustrates a decoration of the template structure shown in Fig. 4 by one such solvent state configuration.

Let p_{rs} be the probability that residue, r, is in solvation state, s. Then G_{slv}^{res} is given by:

$$G_{slv}^{res} = \sum_{r=1}^{n} \sum_{s} \left[\varepsilon_{rs}^{slv} + v_{rs}^{slv} \left(P - P_o \right) - TR\, \alpha_{rs}^{slv} + TR \ln\left(p_{rs} \right) \right] p_{rs} \tag{5}$$

where the parameters, $(\varepsilon_{rs}^{slv}, \alpha_{rs}^{slv})$, give the (energy, entropy) contributions for residue, r, in solvation state, s, which are scaled by solvent accessible surface area as determined from the template. The parameters, v_{rs}^{slv}, are first order Taylor expansion coefficients for the solvation enthalpy with respect to pressure, P. For purposes of simplicity, the parameters are treated as constant over the temperature range of interest. Moreover, by limiting the calculations to the reference pressure, $P = P_o$, the v_{rs}^{slv} parameters are not needed. The model parameters

can be expanded in terms of pressure and co-solvent concentrations. Notice that with the exception of the extra $-RT\,\alpha_{rs}^{slv}$ term, the form of Eq. (5) is the standard expression for the free energy of a system comprised of independent subsystems, where the mixing entropy is accounted for in the last term. At this point, each residue is able to independently explore three solvation states in thermodynamic equilibrium. However, as more terms are added to the FEF, these states will become coupled in the same way spin-spin coupling occurs in Ising or Potts models. In the FED considered here, the set of functions $\{\,p_{rs}\,\}$ will form the basis for completely representing the FEF as an Ising-like model with generalized spin-spin coupling terms. The coupling terms that are described next account for interactions at the interface between subsystems, which are the molecular constituents defined by the residues.

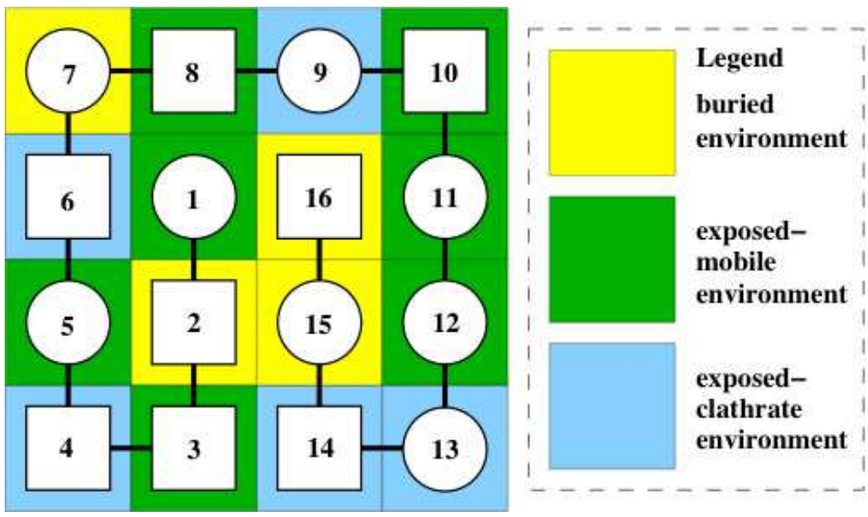


Fig. 5. A schematic of a protein template that is decorated by specifying one of three possible solvation states for each residue. Buried means that some part of a residue is not in contact with solvent. The contributions of free energy, enthalpy and entropy in the buried state are proportional to the solvent accessible surface area of the residue. Exposed solvation states have maximum solvent accessible surface area. The difference in these quantities between buried and exposed states will be less for corner residues {4,7,10,13} compared to the other surface residues {3,5,6,8,9,11,12,14}. A maximum difference occurs for core residues {1,2,15,16}, which can become exposed to solvent due to solvent penetration. Recall the template defines a fixed topology, but not a fixed geometry. This type of coarse-grained description of solvation is common to Ising-like models.

Hydrophobic interaction: The change in free energy to transfer water from an *interface* that separates neighboring molecular constituents into bulk solvent is *how* the hydrophobic effect is modeled in the FEF. This interface term is illustrated in Fig. 6A. Then G_{slv}^{hph} is given by:

$$G_{slv}^{hph} = \sum_{r1=1}^{n} \sum_{r2=1}^{n} \left[\varepsilon_{r1,r2}^{hph} - TR\,\alpha_{r1,r2}^{hph} \right] p_{r1,b}\; p_{r2,b} \tag{6}$$

where $\varepsilon^{hph}_{r1,r2} = n^{wat}_{r1,r2}\,\varepsilon^{hph}$, $\alpha^{hph}_{r1,r2} = n^{wat}_{r1,r2}\,\alpha^{hph}$ and $n^{wat}_{r1,r2}$ is an estimate for the number of water molecules that could reside at the interfacial surface between residues $r1$ and $r2$ based on the geometry of the template structure. Note that $n^{wat}_{r1,r2}$ is the interfacial surface area divided by the *specific area* covered by a single water molecule. The parameters, $(\varepsilon^{hph}, \alpha^{hph})$, represent the (energy, entropy) contributions to the free energy for transferring one water molecule from a generic non-polar reference environment to bulk solvent.

An interesting property of Eq. (6) is that the accumulated *strength* of the hydrophobic effect is proportional to the total surface area of the buried-buried interfaces that snake through a protein. The nature of these interfaces depends on the solvation state of the protein. Also a significant part of the overall strength of the hydrophobic effect is due to the *chemical nature* of the bulk solvent (i.e. affecting chemical potential), which is reflected in the parameters $(\varepsilon^{hph}, \alpha^{hph})$ by expressing them as *functions of co-solvent concentrations*. In aqueous solution a thermodynamic force is generated to expel water from the core of a globular protein, thereby resisting water penetration. The hydrophobic effect competes against the desire for residues of all types (hydrophobic or polar) to be solvated. The nuanced details of the solvation properties of each residue type combined with where residues are located in the template structure determines the amount of "dry" or "wet" interfacial surface area, and this directly relates to water penetration pathways associated with partial unfolding events.

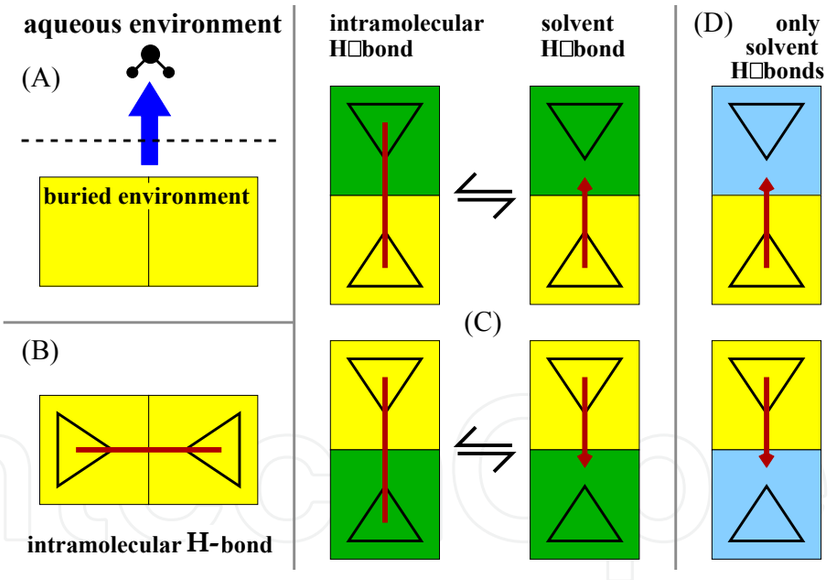


Fig. 6. (A) Schematic of a water molecule transferring from a non-polar environment to bulk solvent. (B) Intramolecular H-bonds (red line) identified in the template are present when both residues are buried (yellow block). Triangles represent residues and indicate a relative orientation. (C) Neighboring pair of residues in buried and exposed-mobile (green block) states is shown. When mobile solvent surrounds an exposed residue; a *fluctuating* intramolecular H-bond (red line) can form between it and a buried residue, or a *fluctuating* solvent-protein H-bond can form (short red arrow) between the buried residue and solvent. (D) An exposed residue surrounded by clathrate water (blue block) prevents a H-bond to form between it and a neighbouring buried residue because the immobile water molecules cannot properly rearrange, and thereby shields the residue from intramolecular H-bonding.

Solvent-protein H-bond: H-bonds appear in the FEF in both the solvation and conformational parts of the FED. When a H-bond forms between two neighboring residues in a template structure, the model must account for nine cases corresponding to each of the residues being in one of its three possible solvation states. First, the intramolecular H-bond will be present when both residues are buried, making it impossible for another H-bond to form between solvent and to that particular buried region of the protein. If both residues are exposed, the question about solvent-protein H-bonding to these residues is irrelevant because the residue solvation free energy fully accounts for theses interactions. The cases that generate a solvent-protein H-bond are when one residue is exposed to solvent, while its neighboring residue is buried. The discontinuity in local environment creates a surface term at a wet-dry interface between the two subsystems. In particular, the intramolecular H-bond can remain in tact, or be replaced by a solvent-protein H-bond that forms between solvent and the buried residue.

In all, Fig. 6B, 6C and 6D summarizes 7 cases. The three cases that involve intramolecular H-bonding will be addressed below when considering the conformational part of the FED. However, because two cases are coupled dealing with fluctuations between solvent-protein and intramolecular H-bonds (see Fig. 6C), another probability function must be introduced to determine if a solvent-protein H-bond or an intramolecular H-bond will form when both options are accessible. Let p_h^{ihb} be the probability that the *h*-th *intramolecular H-bond* (ihb) identified in the template structure is present. Although it is not difficult to associate the *h*-index of an identified intramolecular H-bond to its residues that provide donor and acceptor atoms, it does require *cumbersome notation* to explicitly show this correspondence. Therefore, in formulas that involve the *h*-index and two residue indices (r_1 and r_2) this correspondence is *implied*. Then, as another *interface* term in the FEF, G_{slv}^{shb} is given by:

$$G_{slv}^{shb} = \left(\varepsilon^{shb} - TR\,\alpha^{shb}\right)\left\langle N_{shb}\right\rangle + RT\sum_h \left(\ln(p_h^{ihb})p_h^{ihb} + (1-p_h^{ihb})\ln(1-p_h^{ihb})\right) \tag{7}$$

$$\left\langle N_{shb}\right\rangle = \sum_h p_h^{shb} \text{ where } p_h^{shb} = \left(p_{r1,b}\,p_{r2,c} + p_{r1,c}\,p_{r2,b}\right) + \left(p_{r1,b}\,p_{r2,m} + p_{r1,m}\,p_{r2,b}\right)\left(1-p_h^{ihb}\right) \tag{8}$$

In Eq. (7) the parameters $(\varepsilon^{shb}, \alpha^{shb})$ characterize the energy and entropy contributions from solvent-protein H-bonds. These parameters *should be* dependent on local structural details of the solvent and protein, but the model is based on an *implicit solvent*. Including myriad structural details would be intractable, but as effective parameters, they depend only on the *chemical nature* of the bulk solvent. Thus, all that matters is the total number of H-bonds between solvent and the protein, given by $\left\langle N_{shb}\right\rangle$. This average is over the ensemble of microstates that represent all fluctuations taking place between intramolecular H-bonds and solvent-protein H-bonds. Therefore, the second term is the mixing entropy associated with intramolecular H-bonds forming and breaking. The average number of H-bonds in Eq. (8) is expressed as a simple sum over the probability of finding a solvent-protein H-bond, which is then expanded out in detail corresponding to the four terms shown in Fig. 6C.

3.2.2 Entropy spectrum for molecular constituents and interfacial subsystems

As outlined in section 2, subsystems are coarse-grained into configurations corresponding to low energy basins, each with the same number of distance constraints but with *particular*

characteristics. The number of distance constraints is just enough for the subsystem to be *isostatically rigid*, meaning no distance constraint is redundant when the subsystem is isolated. Suppose subsystems 1 and 2 with n_1 and n_2 atoms are connected together to form a larger rigid system with $(n_1 + n_2)$ atoms through an interaction at their interface involving n atoms from each subsystem. This *interfacial interaction* is modelled using $3(2n) - 6$ distance constraints[3] so it too is *isostatic* when isolated. Within the combined system, there will be $3(n_1 + n_2) - 6$ independent distance constraints. However, the two subsystems together with the interfacial interaction produce $3(n_1 + 2n + n_2) - 18$ constraints, leading to $6n - 12$ redundant constraints in the combined system. Therefore, *interfacial subsystems will generally create redundant constraints* as molecular constituents are coupled through interactions. Interestingly, peptide bonds along the backbone that join residues together do not generate redundant constraints. Thus, a random coil (no crosslinks) has no redundant constraints.

A *complete entropy* assignment to all distance constraints can be made for each basin by performing a local all-atom sampling using a quasi-harmonic approximation. This means that absolute entropies are estimated from the Schlitter entropy formula based on the covariance matrix of atomic fluctuations (Andricioaei & Karplus, 2001) that is obtained from an all-atom MD simulation using an accurate molecular mechanics force field. Afterwards, by considering contributions from all accessible basins, the intrinsic free energy of a subsystem can be reconstituted. This method applies to residues (Wang, et al. 2011) and all interfacial subsystems. While it is important that a *robust procedure* to determine entropy and energy parameters for the conformational part of the FED has been established, for my discussions here it suffices to know that the parameters concretely exist with entropy values on an absolute scale! Furthermore, it is possible to model each mode of vibration within an energy basin of a subsystem using a specific distribution of distance constraints. However, a surprisingly simple description can be made without invoking any details about *how* the distance constraints are distributed.

For the purpose of explaining the interfacial thermodynamics paradigm, network rigidity will be described in terms of *Maxwell constraint counting* (Whitely, 2005). Maxwell constraint counting (MCC) is a mean field approximation to graph-rigidity. As shown previously, the application of MCC to solve the mDCM yields results that capture the correct qualitative thermodynamic response in beta-hairpins (Jacobs & Fairchild, 2007a), alpha-helices (Vorov, et al. 2009) and proteins (Vorov, et al. 2011). The advantage of using MCC is that concepts can be calculated easily without obfuscation because all topological details about how the constraints are distributed in the network are ignored. The only specification required is that a subsystem with n atoms has $3n - 6$ distance constraints, each assigned an entropy value of σ_j for j=1 to n, such that $\sigma_j \leq \sigma_{j+1}$. This ordered set of values from lowest to greatest defines an *entropy spectrum* for a subsystem. All of the different types of subsystems that make up the conformational part of the FED are now described.

3.2.3 Conformation related enthalpy-entropy compensation mechanisms

Molecular constituents: Residues are molecular constituents that define subsystems involving a certain number of atoms. Each residue, r, in a specific energy basin, ν, has its own entropy

[3] The special case n=1 is not considered in this discussion because no subsystem is employed that contains less than 3 atoms.

spectrum. While it is possible to have a detailed description of each residue by using a large number of energy basins, an Ising-like model similar to the mDCM is described here, where each residue is classified as *native-like* or *disordered*. A conditional dependence as to whether a residue is native-like (ν=n) or disordered (ν=d) is tied to its solvation state. If a residue is exposed to solvent, it is modelled as disordered. However, a buried residue can be native-like or disordered. Native-like implies the local geometry of a residue will be similar to the template structure, and disordered implies poor atomic packing that is reflected by higher energy and entropy. Respecting conditional dependences on solvation, G_{cnf}^{res} is given by:

$$G_{cnf}^{res} = \begin{cases} \sum_{r=1}^{R} \left(\varepsilon_{rd}^{cnf} - TR \sum_{j} q_{rdj}\, \sigma_{rdj}^{cnf} \right) \left((1 - p_{rb}) + p_{rb}\,(1 - p_r^{nat}) \right) + \\ \sum_{r=1}^{R} \left(\left[\varepsilon_{rn}^{cnf} - TR \sum_{j} q_{rnj}\, \sigma_{rnj}^{cnf} \right] p_r^{nat} + TR \left[p_r^{nat} \ln\left(p_r^{nat} \right) + (1 - p_r^{nat}) \ln\left(1 - p_r^{nat} \right) \right] \right) p_{rb} \end{cases} \tag{9}$$

In Eq. (9) the top term contributes when residue, *r*, is exposed with probability $(1 - p_{rb})$ or when it is buried *and* disordered with probability, $p_{rb}(1 - p_r^{nat})$. The energy contribution is ε_{rd}^{cnf} and the *maximum* possible entropy contribution is given by $R\sum_j \sigma_{rdj}^{cnf}$ when all distance constraints are independent (i.e. $q_{rdj} = 1$). The set of pure entropies, $\{\sigma_{rdj}^{cnf}\}$, define the entropy spectrum for residue, *r*, when it is in the disordered state, *d*. Once residue, *r*, couples to other residues within the protein, some of its distance constraints may become redundant. Recall redundant constraints do not contribute to conformational entropy. Therefore, q_{rdj} is the probability that distance constraint, *j*, in residue, *r*, and in its disordered state, *d*, is independent. The probability for residue, *r*, to be native-like when it is buried, is given as p_r^{nat}. Similarly, the bottom term in Eq. (9) contributes ε_{rn}^{cnf} energy and $R\sum_j q_{rnj}\sigma_{rnj}^{cnf}$ is the net entropy when the residue is buried *and* native-like with probability, $p_r^{nat} p_{rb}$. The native-like entropy spectrum is given by $\{\sigma_{rnj}^{cnf}\}$, and q_{rnj} is the probability that distance constraint, *j*, in residue, *r*, and in its native-state, *n*, is independent. The last term involving the square brackets is the mixing entropy for the buried residue, *r*, to be either native-like or buried.

Taken together, G_{cnf}^{res} supports the following possibilities: A large number of buried residues that are mostly disordered correspond to a collapsed state driven by the hydrophobic effect. As more native-like residues form, but with high variance, the protein will transition to a molten globular state with fluctuating secondary structure. When the majority of buried residues are native-like, the protein will be in its native-state with some degree of flexibility. Thus, all common known phases can be described by the accessible microenvironments.

Covalent bond linkers: When a covalent bond links two residues, it involves two atoms in each of the two residues. Therefore, the covalent bonds are modelled as an interfacial subsystem containing 4 atoms connected by 6 distance constraints. The parameterization of the distance constraints for a covalent bond with a flexible dihedral angle such as a

disulphide bond, or a peptide bond with a fixed dihedral angle within a *trans* or *cis* basin will have distinct energy and entropy parameters. Since chemical reactions involving the breaking or forming of a covalent bond is not considered here, these interactions are *quenched*. Moreover, only one energy basin for each type of covalent bond is considered here. Then, G_{cnf}^{lnk} is given by:

$$G_{cnf}^{lnk} = \sum_{k=1}^{K} \left(\varepsilon_k^{cnf} - TR \sum_j q_{kj} \, \sigma_{kj}^{cnf} \right) \tag{10}$$

Here ε_k^{cnf} is the energy, $\{\sigma_{kj}^{cnf}\}$ is the entropy spectrum, and, q_{kj}^{cnf} is the probability that the *j*-th distance constraint for the *k*-th covalent bond in the protein is independent. It is worth noting that for peptide bonds, the entropy spectrum will consist of such low entropy values that a graph-rigidity analysis augmented by the preferential rule of placing lowest entropy distance constraints first, yields q_{kj}^{cnf} = 1 *always*. In this case, a constant contribution will always come from peptide bonds, rendering its affect on thermodynamic response. As such, the parameters for peptide bonds are unnecessary to specify. In contrast, for a rotatable covalent bond, usually 5 out of 6 distance constraints can be "frozen" out. However, the sixth distance constraint characterizes tolerances in the torsion-angle. Note that even using one energy basin to model a rotatable covalent bond can affect thermodynamic response because depending on other distance constraints in the network, the q_{k6}^{cnf} probability need not be 1 always, and thus not frozen out. For more accuracy, more than one basin can be considered of course, which would also provide variation in the energy if the energy basins correspond to frequent and rare angular ranges.

Intramolecular H-bonds: An intramolecular H-bond (IHB) identified in the template structure between residues r_1 and r_2 (labeled by the *h*-index) will be present when both residues are buried. When one residue is buried and the other is exposed to mobile solvent there is a probability p_h^{ihb} for this IHB to be present, otherwise it will be broken as shown in Fig. 6C. When both residues are buried as shown in Fig. 6B, there are four cases to consider because each buried residue may be native-like or disordered. Consolidating the four cases into two leads to either both residues are native-like, which defines a native IHB, otherwise a disordered IHB forms. The *h*-th native IHB will have energy, ε_{hHB}^{cnf}, and entropy spectrum, $\{\sigma_{hHBj}^{cnf}\}$, characteristic of the local geometry of the template structure. When one or both of the residues are disordered, the native geometry is disrupted, creating a disordered IHB. Therefore, it is natural to define two basins: A basin reflecting properties specific to the native geometry of the *h*-th IHB, and a basin with common properties for all disordered IHB. A disordered IHB is modeled with an energy, ε_{dHB}^{cnf}, and entropy spectrum, $\{\sigma_{dHBj}^{cnf}\}$, independent of location in the template. Since a H-bond has 3 atoms (donor, hydrogen acceptor) and there are *3n-6* distance constraints, all IHB subsystems will have j=1,2,3. Taking the conditional dependencies into account, G_{cnf}^{ihb} is given by:

$$G_{cnf}^{ihb} = \begin{cases} \sum_{h=1}^{H} \left[\left(\varepsilon_{hHB}^{cnf} - TR \sum_j q_{hnj}\, \sigma_{hHBj}^{cnf} \right) p_{r1}^{nat}\, p_{r1}^{nat} + \left(\varepsilon_{dHB}^{cnf} - TR \sum_j q_{hdj}\, \sigma_{hHBj}^{cnf} \right) \left(1 - p_{r1}^{nat}\, p_{r1}^{nat} \right) \right] p_{r1,b}\, p_{r2,b} \\ + \sum_{h=1}^{H} \left(\varepsilon_{dHB}^{cnf} - TR \sum_j q_{hdj}\, \sigma_{dHBj}^{cnf} \right) \left(p_{r1,b}\, p_{r2,m} + p_{r1,m}\, p_{r2,b} \right) p_h^{ihb} \end{cases} \quad (11)$$

The form of Eq. (11) follows a *general pattern* that applies to all conformational interactions. That is, there is a probability for the interaction to be present, possibly in a specific state, and under this local condition modelled by a particular basin, it contributes energy that adds to the system. Moreover, the conformational entropy that it contributes is non-additive due to network rigidity, which is used to calculate the probabilities for distance constraints to be independent. In the top term of Eq. (11), the q_{hnj} and q_{hdj} give the probabilities for the *h*-th IHB in the native and disordered basins respectively. Notice that the q_{hdj} in the lower term is the same as in the top term. When Eq. (11) is considered together with Eq. (7) and Eq. (8), we see the protein structure must balance enthalpy-entropy compensation in the hydrogen bond network of the protein versus forming H-bonds to solvent.

Hydration interaction: The hydration interaction is introduced to model the affect of aqueous solvent on a residue when it is exposed to a clathrate environment. In this case, the water molecules surrounding a residue form a *rigid motif* that manifests itself as a *mechanical clamp* on the residue. Many distinct molecular configurations can lead to a rigid motif. As such, a large number of basins are needed to fully characterise the structure of water molecules around a residue. However, this detailed information is difficult to obtain, and considering the level of coarse-graining that has been made in regards to residue solvation, and to the native-like and disordered classifications of a residue with respect to a template structure, it is appropriate to employ a single basin. Following the general pattern, G_{cnf}^{hyd} is given by:

$$G_{cnf}^{hyd} = -TR\, \sigma_{hyd} \left(\sum_{r=1}^{R} \sum_j q_{rcj} \right) p_{r,c} \quad (12)$$

In Eq. (12) the energy term is not included because it is already accounted for in the clathrate solvation energy parameter. The additive part of the solvation entropy parameter accounts for the solvent DOF having a reduction in conformational entropy. But *mechanical clamping* from the clathrate structure reduces conformational flexibility and entropy of the residue. This clamping is modelled by *3n-6* distance constraints to ensure the residue of *n* atoms is isostaticly rigid. A degenerate conformational entropy parameter is given by σ_{hyd}, and q_{rcj} gives the probability for the *j*-th distance constraint to be independent, where the *c*-subscript denotes clathrate. A reduction in conformational entropy at low temperatures is a critical mechanism to understanding cold denaturation, as shown in previous work [Jacobs & Wood, 2004]. Technically, this clamping effect could also be present for an exposed-mobile environment, and using the same modelling scheme but with different parameters would allow for different types of clamping depending on the details of the local water structure. Again, as found previously, breaking up this complicated many body interaction of water molecules on the protein into two simple states (mobile verses clathrate) proves sufficient.

Network rigidity: The non-additive reconstitution of total conformational entropy involving all entropy components is a critical aspect of the FEF. First note that the labelling scheme for component entropies across various interaction types is quite cumbersome, although useful for distinguishing their roles. In regards to network rigidity calculations, it is convenient to define a *parallel indexing scheme*. Let σ_c be the pure entropy of distance constraint, *c*, which runs from *1* to *C* to account for all accessible interactions. Note that the total number of distance constraints in a protein is *not equal* to *C*. Rather, if all identified subsystems were present *simultaneously*, which is not physically possible, *C* is the sum of the numbers of distance constraints in all subsystems defined by the template structure. The labelling of all these accessible entropies fall in sorted order such that $\sigma_c \le \sigma_{c+1}$. That is, $\{\sigma_c\}$ will play the role of an entropy spectrum for a template structure, but not all *levels* can be occupied.

As part of a general pattern, there is a probability for an interfacial subsystem to be present, and tracing over a chain of conditions as products of probabilities translates to an *occupation probability*, p_c , for an individual distance constraint. Similarly, q_c , is the probability that a distance constraint is independent given it is present. With this generic labelling scheme, the network rigidity part of the FEF is *compactly* stated by the following three formulas:

$$N_c = \sum_{c=1}^{C} p_c \quad I_c = \sum_{c=1}^{C} q_c p_c = 3n - 6 \quad \tau = \sum_{c=1}^{C} \left(q_c \sigma_c \right) p_c \tag{13}$$

The average number of constraints within a protein is N_c , which will generally be greater than the *3n – 6* DOF required to position all atoms in a protein. The minimum value that N_c can be is *3n – 6* corresponding to a protein in an *ideal* random coil state without any disulphide bonds (i.e. only residues and peptide bonds). Starting from a random coil, as intramolecular H-bonds crosslink the backbone, the number of constraints will increase. Yet the total number of independent constraints, I_c , that are present in the protein is *conserved* across all accessible constraint topologies to maintain a *rigid* protein. The condition of a rigid protein is required because it means that everything that can be specified is specified within tolerances. This property reflects the *completeness* of the FED. With the liberated view of rigidity, $R\tau$ provides a lowest upper bound *estimate* for the total conformational entropy.

3.2.4 Vibrational contributions

The contribution to the FEF from G_{vib} *would be* straightforward once the frequency is known for all 3n – 6 modes. Unfortunately, determining vibrational frequencies for the entire protein is a task that must be avoided to retain computational efficiency in the interfacial thermodynamics model. Therefore, the frequency spectrum is modelled *empirically* to capture two important features. Vibrational mode frequencies will (increase, decrease) as redundant distance constraints are (added, removed) to a network as it (stiffens, softens). Furthermore, the lowest frequency decreases as the size of the system increases. Assuming a vibrational frequency can be defined for each *c*-index, G_{vib} is given by:

$$G_{vib} = RT \sum_{c=1}^{C} q_c \, p_c \, \ln\left[1 - \exp\left(-\beta\hbar\omega_c\right)\right] \tag{14}$$

To account for the first feature, the entropy spectrum for the system is employed to define a frequency spectrum where $\omega_c = \omega_{\max}\sigma_{\min}/\sigma_c$ to reflect the generic property that frequency is inversely proportional to pure entropy. This relation would be sufficient if the entire protein was employed as the sole constituent, but the pure entropy spectrum is comprised from many small subsystems resulting in a range for ω_c not reaching sufficiently low frequency. By generalizing the Debye model [Kittel, 1996] for vibrations in a crystalline solid, system size dependence can be better accounted for. In the Debye model, the frequency spectrum is given as: $\omega_m = \omega_{\min} m$ for mode *m*, where $\omega_{\max} = \omega_{\min}(3n-6)$ is the maximum frequency possible assuming a *linear dispersion*, and knowing the total number of vibrational modes in the system. Inverting these relationships leads to $\omega_m = \omega_{\max} m/(3n-6)$. Combining both of these general aspects to model the frequency spectrum yields the empirical model:

$$\omega_c = \omega_{\max}\frac{\sigma_{\min}}{\sigma_c}\frac{\left(3n-5-\sum_{b=1}^{c} q_b\, p_b\right)}{(3n-6)} \tag{15}$$

In Eq. (15), the right-most fraction parallels the Debye model, where the numerator defines an *effective mode number* that ranges from *1* to *3n – 6*. Mode (*3n – 6*) corresponds to *c=1*, or the lowest *c*-index present in the network. As the mode index decreases with increasing *c*-index, the frequency of vibration lowers in the same way as the Debye model. The front scale factor $\sigma_{\min}/\sigma_c$ (i.e. $\sigma_{\min} = \sigma_1$) modifies the dispersion, which is required to span a disparate range of vibrational frequencies typical of proteins. Consequently, as more redundant constraints appear in the network, the lowest frequency mode will be reached sooner on the entropy ladder (smaller *c*-index) so that the lowest lying vibrational frequencies shift higher without affecting high frequencies. The spectrum for a protein in a random coil will reach the lowest possible frequency. Note that upon protein-ligand binding, frequency shifts that take place due to changes in the constraint network and size of the system are roughly accounted for.

4. Self-consistent constraint theory

The essence of constraint theory applied to an interfacial thermodynamics model is to determine how a system responds under certain *global* conditions that are consistent with *microscopic* heterogenous environments. The previous section constructed a FEF to describe protein stability by building up a hefty collection of free energy terms representing specific enthalpy-entropy mechanisms. This process lead to conjuring up the probability functions $\{p_{rs}, p_h^{ihb}, p_r^{nat}, p_c, q_c\}$ that will be determined self-consistently when solving the FEF. In doing so, it is critically important that *heterogeneous micoenvironments* throughout the protein are taken into account. An efficient way to solve the FEF is to introduce a number of *constitutive equations* to transform the functional into a variational problem in parametric form that allows certain global constraints to be imposed. Specifically, a trial function with a number of variables is substituted into the functional to reduce the problem to finding the minimum of a function. These variables correspond to Lagrange multipliers that control the values of selected order parameters. Actually, this method builds upon a previous method used to solve the mDCM (Jacobs & Dallakayan, 2005), but there are critical differences.

First, I highlight key points about the initial method (Jacobs, 2006b) for solving the mDCM. Occupation probabilities $\{p_c\}$ are calculated using Lagrange multipliers to enforce the amount of intramolecular H-bonds and native-like character within a protein. Given $\{p_c\}$, a constraint network is randomly generated, and a rigidity algorithm is applied with the preferential entropy rule to identify the distance constraints as independent ($\tilde{q}_c = 1$) or redundant ($\tilde{q}_c = 0$). After collecting N_s *random samples*, an estimate is given as $q_c = \langle \tilde{q}_c \rangle$. By focusing on distance constraints, all the formulas in Eq. (13) apply, and N_s should be at least 200 to obtain a *useful* estimate for conformational entropy, $R\tau$. The employed trial function for p_c does not depend on whether a distance constraint is independent or not. For this information, p_c must be a function of q_c. In this latter case, the calculation scheme would then be to assume $q_c = 1$, calculate p_c, determine $q_c = \langle \tilde{q}_c \rangle$ from sampling, then recalculate p_c, and iterate this process until the latest values of p_c are nearly equal to the previous values of p_c within some tolerance to obtain self-consistency.

To simply implement this self-consistent approach using averaging does not work because convergence will not occur when the precision in q_c is low. To ensure convergence, N_s *should be at least 500,000*, suggesting a self-consistent calculation is not tractable! Recently, a new algorithm to calculate average network rigidity properties as a *probability flow* problem describing independent DOF and where DOF absorb onto constraints has been successfully developed [Gonzalez, et al. 2011a, Gonzalez, et al. 2011b]. Now q_c is calculated to within numerical precision, making the self-consistent calculation feasible. Also different from the mDCM: The FEF described above has almost all conformational subsystems that involve distance constraints coupled to the solvation states of residues, as well as other interfacial surface terms between the residues that reflect local microenvironments. Tracking these additional details goes far beyond the mean field treatment invoked in the mDCM. Consequently, a very different set of order parameters need to be considered.

4.1 Order parameters and the free energy landscape

The free energy of a protein is numerically calculated while subjected to global constraints imposed by order parameters that define a specific macrostate. Scanning over the entire range of order parameters produces the FEL. Since a protein is of finite size, the minimum free energy is not the only point on the FEL of interest. Rather, the entire FEL is of interest because it maps out all the low-lying basins and saddles. The natural variables of the FEF that describe microstates dictate the form of the macrostates, and this determines what order parameters need to be considered. The solvent environment of the residues and whether they are native-like when buried is the only information needed to completely define the microstate of a protein. Therefore, order parameters B, M and N are introduced to specify the macrostate of a protein, giving the total number of residues that are respectively *Buried*, exposed to *Mobile* solvent and *Native-like*. Note that the number of residues in the exposed clathrate state[4], H, is not an independent variable, since the total number of residues, R, is given by: $R = B + M + H$. Furthermore, there is a restriction on N, such that $0 \le N \le B$. As such, the FEL is expressed as $G(B, M, N \mid T, \cdots)$ where the triple dots are a reminder that in

[4] The more natural symbol of C is already used, and the symbol H better reflects the idea of a local hydration shell.

addition to temperature, pressure and pH can be directly considered, although not here, and, many parameters in the FEF depend on solvent composition.

4.2 Hierarchical application of global constraints

The macrostate (B, M, N) is associated with the following three constraint equations:

$$B = \sum_{r=1}^{R} p_{r,b} \quad M = \sum_{r=1}^{R} p_{r,m} \quad N = \sum_{r=1}^{R} p_r^{nat}(1 - p_{r,b}) \tag{16}$$

Instead of solving these three equations simultaneously, the calculations are simplified by assuming B can be solved for first, then M, and finally N. To reflect this hierarchical chain, the constitutive equations for the probability functions are expressed in terms of conditional probabilities when necessary. Let $Z_{rs}^{slv} = \exp(\alpha_{rs}^{slv} - \beta\varepsilon_{rs}^{slv})$ and λ_B, λ_M and λ_N be the Lagrange multipliers for the B, M and N order parameters respectively. With $p_{r,c} = 1 - p_{r,b} - p_{r,m}$, the constitutive equations for $p_{r,b}$ and $p_{r,m}$ are given as:

$$p_{r,b} = \frac{Z_{rb}^{slv}\exp(\chi(r)\lambda_B)}{Z_{rb}^{slv}\exp(\chi(r)\lambda_B) + Z_{rm}^{slv} + Z_{rc}^{slv}} \quad p_{r,m} = \frac{Z_{rm}^{slv}\exp(\phi(r)\lambda_M)}{Z_{rm}^{slv}\exp(\phi(r)\lambda_M) + Z_{rc}^{slv}}(1 - p_{r,b}) \tag{17}$$

The Lagrange multipliers λ_B and λ_M are coupled to propensity functions $\chi(r)$ and $\phi(r)$ respectively. Propensity functions characterize physical and/or chemical properties relevant to their conjugate order parameters in the local environment surrounding residue *r*, based on the template structure. The local propensity for being buried is $\chi(r)$, defined as the number of nearest neighbour contacts to residue, *r*. The greater number of nearest neighbours a residue has in the template structure, the greater resistance to solvent penetration irrespective of its intrinsic solvation character quantified by the Z_{rs}^{slv} factors. For $\phi(r)$, it is set to *1*, so that no differentiation is made between clathrate and mobile exposed microenvironments. The template structure is used to define local microenvironments that individual residues will experience. By adjusting the Lagrange multipliers, the total number of residues buried and exposed to mobile solvent is controlled (at least on average). Note that fluctuations are accounted for in the FEF through the mixing entropy terms. Constraining a protein to macrostate (B, M, N) corresponds to selecting a sub-ensemble of microstates that share the common property that the total numbers of buried and exposed to mobile residues are B and M. It is seen from Eq. (17) that the calculation for $p_{r,b}$ based on the global constraint B is independent of M and N. Then the result for $p_{r,b}$ hierarchically feeds into the next level of calculation for $p_{r,m}$ involving the global constraint M.

The next hierarchical step is to calculate p_r^{nat} with N fixed. The two constitutive equations that come into play are:

$$p_r^{nat} = \frac{\exp(-\beta\varepsilon_{rn}^{cnf} + \sum_j q_{rnj}\,\sigma_{rnj}^{cnf} + \lambda_N)}{\exp(-\beta\varepsilon_{rn}^{cnf} + \sum_j q_{rnj}\,\sigma_{rnj}^{cnf} + \lambda_N) + \exp(-\beta\varepsilon_{rd}^{cnf} + \sum_j q_{rdj}\,\sigma_{rdj}^{cnf})} \tag{18}$$

$$p_h^{ihb} = \frac{\exp(-\beta\varepsilon_{dHB}^{cnf} + \sum_j q_{hdj}\ \sigma_{dHBj}^{cnf})}{\exp(-\beta\varepsilon_{dHB}^{cnf} + \sum_j q_{hdj}\ \sigma_{dHBj}^{cnf}) + \exp(-\beta\varepsilon^{shb} + \alpha^{shb})} \tag{19}$$

The form of Eq. (18) is that it has a Boltzmann factor for a native-like state in the numerator, divided by the sum of Boltzmann factors for the native-like and disordered states. This ratio gives the probability that residue, *r*, will be in the native-state. The Boltzmann factor for the native-like state contains the Lagrange multiplier, λ_N, to enforce the number of native-like residues in the system to be, N, in accordance with Eq. (16). The form of Eq. (19) is similar, except the numerator is the Boltzmann factor for a disordered IHB, and the denominator is the sum of Boltzmann factors for a disordered IHB and a protein-solvent H-bond. These two possibilities compete head to head, but there is no additional Lagrange multiplier, as this process is not tied to an order parameter.

It is clear from Eqs. (18 and 19) that knowing the probability certain distance constraints in the constraint network are independent is necessary. However, these *q*-values are initially unknown, and therefore, an iterative self-consistent calculation must be invoked. Note that the probability for a distance constraint to be present is equal to the probability for the basin that it is a member of to be present in the network. For example, all distance constraints used to model the conformational part of the free energy for residue, *r*, when it is native-like is equal to p_r^{nat}. Conversely, the probability of $(1-p_r^{nat})$ is assigned to all the distance constraints for this residue when it is in a disordered state.

The occupation probabilities, p_c, is 1 for quenched constraints, or it is straightforward to get the probabilities from $\{p_{rs}, p_h^{ihb}, p_r^{nat}\}$ once they are known. There is, of course, a chicken and egg problem because q_c is determined after p_c is known, but q_c must be known before p_h^{ihb} and p_r^{nat} can be calculated. The procedure is to guess the initial values of q_c, calculate p_c, apply a rigidity analysis to obtain q_c and then recalculate p_c. Iterate this process until the values for both p_c and q_c converges. Note that these equations converge to the unique solution independent of initial guess. The type of guesses tried include, for each *c*-index q_c set to 1, set to 0, or set to a value between 0 and 1, or independently assign a random number between [0,1]. Notice that q_c and p_c imply one-dimensional arrays for *c=1* to *C*. In fact, any variable that has one index or more than one index implies an array of values. It is worth mentioning here that convergence is reached typically within 15 iterations using the new rigidity algorithm [Gonzalez, et al. 2011a]. However, MCC yields qualitatively similar results, and as described next, captures the essential features about the role of rigidity.

4.3 The entropy spectrum and maxwell constraint counting

The entropy spectrum for an example set of subsystems that can be found within a protein is schematically shown in Fig. 7. Applying MCC with the preferential entropy rule is equivalent to filling the available levels of the entropy spectrum of the system starting from the bottom until the system becomes isostatically rigid. All the distance constraints that are placed in the network before the protein has the minimum number of constraints to become isostatically rigid are considered independent. As more distance constraints are added to the network, they are all redundant. This global and uniform transition point between where

the constraints are independent and redundant defines the Maxwell level. This means, that $q_c = 1$ for $1 \le c < M_L$ and $q_c = 0$ for $M_L < c \le C$ and $0 < q_c \le 1$ at $c = M_L$. Let q_M be the value of q_c at the Maxwell level. Then the self-consistent calculations described above amounts to finding a solution in the form of a step function, where the only unknown is where the step is located on the entropy spectrum of the system.

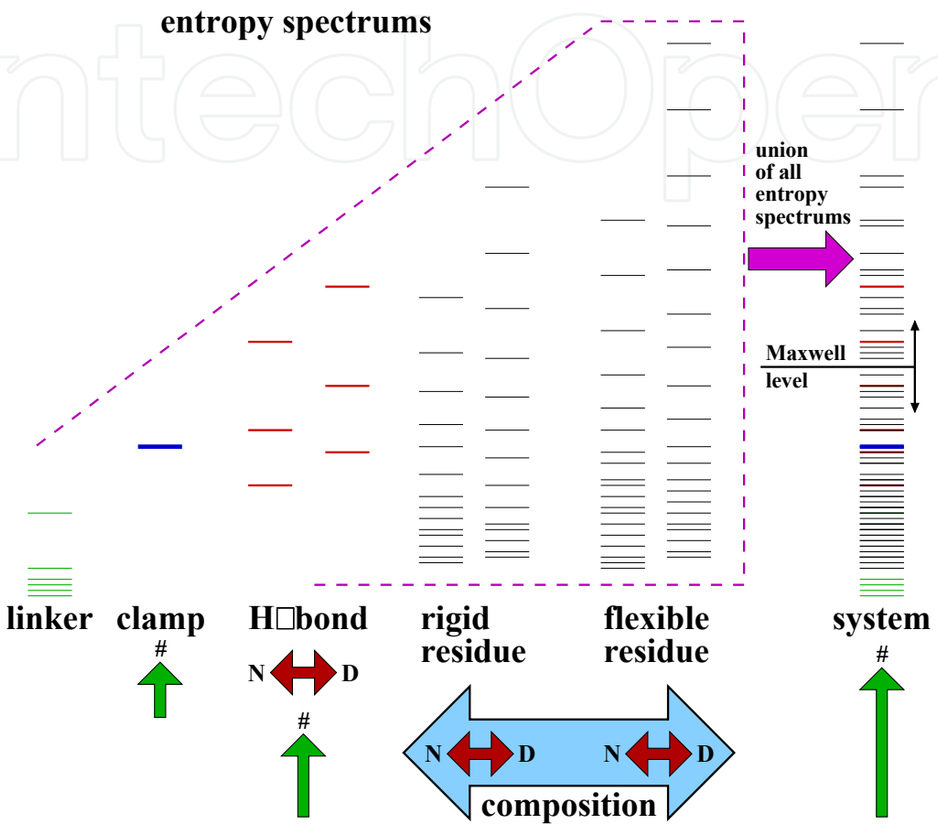


Fig. 7. Schematic of how the entropy spectrums associated with subsystems combine into a single entropy spectrum for the entire system. Green vertical arrows pointing to the # sign indicate that the number of interactions depend on the decorated microstate. The "clamp" interaction derives from clathrate water. The horizontal dark red double arrow indicates a certain number of interactions can be native like (N) or disordered (D), and their respective entropy spectrums are shown. Each residue has a unique entropy spectrum. Residues with (smaller, larger) entropy values are more (rigid, flexible). The protein sequence defining the residue composition is represented as the large horizontal blue double arrows. The system entropy spectrum is characterized by the variables (p_c, q_c). From Eq. (20) distance constraints with entropy less than the Maxwell level are independent, and this level slides up and down the spectrum depending on numbers and types of interactions present in the system.

For a given specification of p_c the Maxwell level is determined by solving the equation:

$$3n - 6 = \sum_{c=1}^{M_L - 1} p_c + D_{M_L} q_{M_L} \tag{20}$$

In Eq. (20) the variable D_{M_L} is the degeneracy for the number of distance constraints having the same entropy value at the Maxwell level. Notice that Eq. (20) reflects the step nature of q_c. Despite the simplicity of the global constraint expressed in Eq. (20), the self-consistent solution results in a dramatic impact on the thermodynamic response of the system. This is because in general every constraint in the network is competing against all other constraints that are present in the network. What changes is the number of distance constraints that appear within the protein for different macrostates. As more constraints are added to the network, the entropy drops.

4.4 Response of local environments to global demands

The type of intramolecular interactions and their locations within a protein depends on the solvation state of the residues and local environments encoded by the amino acid sequence and template structure. For example, using the fold architecture shown in Fig. 4, and for the solvation decoration shown in Fig. 5, a hydrophobic homo-polymer (HH), a heterogeneous protein (HP) and a polar homo-polymer (PH) are shown in Fig. 8.

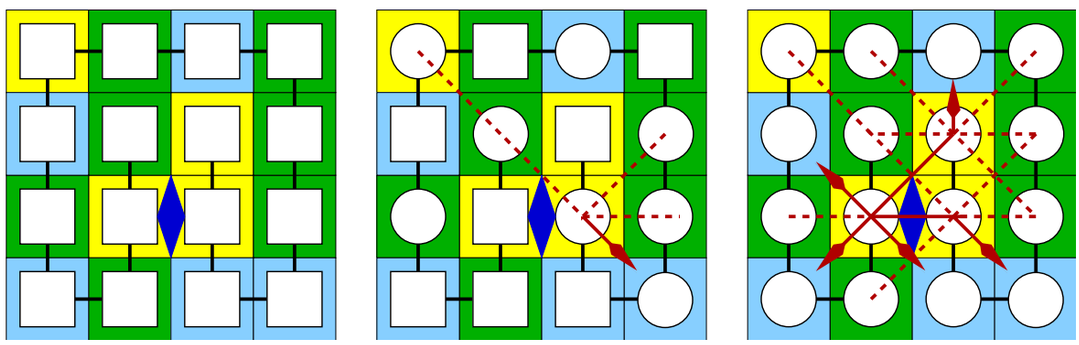


Fig. 8. Schematic illustration of three sequences with the same architecture and solvent decoration. Hydrophobic residues (squares) do not participate in H-bonding. H-bonding is only allowed between polar residues (circles) that are nearest or next nearest neighbours not linked by covalent bonds. Because of the solvent decoration there is only one hydrophobic interaction, shown as a dark blue diamond. Non-fluctuating intra-H-bonds are shown as solid red lines, and red dashed lines indicate fluctuations between intra- and solvent-protein H-bonds. From left to right the three panels show the HH, HP and PH cases, respectively exhibiting (no, limited, many) intramolecular H-bonds for this solvation microstate. The hydrophobic interactions are identical across the three sequences because they only depend on the transfer of water from buried regions to bulk solvent, which is the same in all cases.

In water, HH will form a collapsed state much like an oil droplet. PH will be soluble and will resemble a random coil because few crosslinks form. HP can potentially produce a rich phase diagram. In the next section, stability curves for all three of these *toy* polymers will be shown based on model parameters that were adjusted to produce heat and cold denaturation in HP for the purpose to facilitate general discussions. The same parameters are used for all three cases, and they are in a physically reasonable range. However, the toy models are not structurally realistic, therefore, the parameter values (not given here) are not important. Rather, the *critical issue* at hand is developing a tractable paradigm that can

accurately model the complexity of all the coupled interactions within/on a protein, and do this in a computationally efficient way.

Self-consistent constraint theory applied to the FEF determines the microenvironments that emerge as *most probable*. Although Eq. (3) and Eq. (4) *at face value* appear to be additive, the myriad coupling between interactions will cause two interactions of identical type placed in different microenvironments and/or under different solvent and thermodynamic conditions to respond differently. In particular, the coupling through network rigidity *renormalizes* the conformational entropic contributions (via the q_c values), and this strongly affects where and how solvent penetrates the protein; thus changing the properties of local environments, which impacts the constraint network. For example, a cluster of H-bonds can form a strong nucleation barrier causing a localized buried region to be highly resistant to solvent penetration compared to a similar buried region without a H-bond cluster. Consequently, *non-additive response derives from a chemicophysical feedback loop*. This is captured in the *process of minimizing the free energy* to determine the optimal constraint topology and solvent decoration under the specified thermodynamic and solvent conditions for a given template structure while satisfying Eq. (16) and Eq. (20) for a particular macrostate (B, M, N).

4.5 The Gibbs triangle

The algorithm that is applied to solve the FEF takes the following steps.

1. Scan over the temperature, T, and other thermodynamic and solvent conditions.
2. Calculate $p_{r,b}$ while the variable B is looped over.
3. For given B: Calculate $p_{r,m}$ and $p_{r,c}$ while the variable M is looped over.
4. For given B and M:
   a. Self consistently solve for all p_c and q_c while the variable N is looped over.
   b. Finalize calculation for all probability functions: $\{p_{rs}, p_h^{ihb}, p_r^{nat}, p_c, q_c\}$.
   c. Calculate the free energy: $G(B, M, N \mid T, \cdots)$.
5. Finish all nested loops over N, M and B.

A high dimensional FEL is obtained once the algorithm finishes. In the example considered here, the FEL is four dimensional, consisting of temperature and the three order parameters, (B, M, N). When the free energy value is included to perform an exploration of the FEL, a five-dimensional space is required! Basins for stable and metastable states and free energy barriers between these states can be identified. Within a basin, information about flexibility and its relationship to stability can be obtained. However, protein stability can largely be understood in terms of its solvation properties. Therefore, it proves convenient to construct a two-dimensional version of the FEL specified only by (B, M) to describe how a protein is solvated. At fixed T, this construction is given as:

$$G_2(B,M) = -RT\ln(Z_2) \quad Z_2(B, M) = \sum_{N=0}^{B} \exp\left[-\beta G(B,M,N)\right] \tag{21}$$

The free energy $G_2(B,M \mid T)$ describes the stability of a protein based on a macrostate that characterizes the solvation property of the protein. For fixed T, it is convenient to look at a

phase diagram in terms of how much the protein is buried or exposed to solvent in different forms. Because of the constraint that $1 = (B + M + H) / R$, a Gibbs triangle is employed so that all solvation states can be viewed simultaneously in terms of percentages, as shown in Fig. 9.

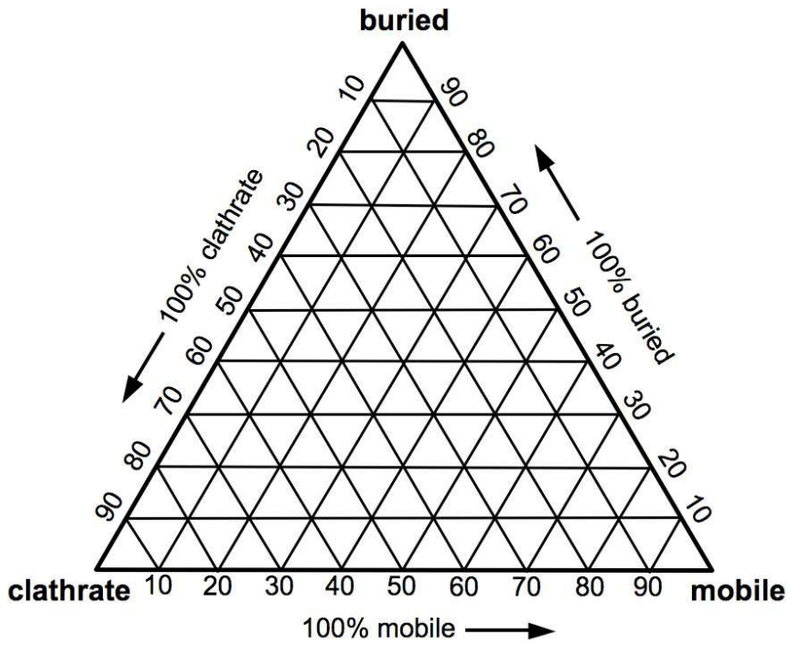


Fig. 9. The Gibbs triangle serves as the base of a three-dimensional FEL that looks like a wedge when viewed at fixed temperature and using the B, M and N order parameters. That is, $N = 0$ on the bottom horizontal line when $B = 0$, and the tip of the wedge is at $B = 1$.

Three stable basins will generically appear in the Gibbs triangle. At low temperatures, it is possible that a free energy minimum will appear when the percent of clathrate structure is dominant. At intermediate temperatures, the number of buried residues in the protein will be dominant, but they can be native-like or disordered. Therefore, a protein may be in either the molten globular[5] or native-fold states. At high temperatures, a molten globular state may remain, or further unfold to allow a majority of residues to become exposed to solvent. Next, the free energy of three representative macrostates is plotted as a function of temperature. Structural phase transitions corresponding to changing free energy basins are like first order transitions, because they occur when stability curves for different states cross one another.

5. Stability curves: Heat and cold denaturation

The stability curves for the toy polymer cases (HH, HP, PH) shown in Fig. 8 are calculated for three macrostates in the Gibbs triangle that correspond to a dominant characteristic of

[5] The degree to which a protein is native-like versus disordered is lost during the process of summing over the native-like order parameter in Eq. (21). However, this information is known from the original 3D free energy landscape.

clathrate, buried or mobile, and these curves are plotted in Fig. 10 A, B, C respectively. In fact, the particular decoration shown in Fig. 8 has a much higher free energy than the lowest free energy state, and is for practical purposes a state of measure zero. Nevertheless, the FEF determines the statistical weights for all macrostates in the high dimensional FEL, and the high free energy states around saddles are important in describing transition states.

As expected, the HH case shows that the buried state is the most stable form over the entire temperature range, and thus there is no phase transition. For the PH case, the structure is always exposed to solvent, but it is interesting to note that there can be a structural phase transition in a protein between low and high temperature without it involving a compact folded structure. This result suggests there is a difference between structural properties in the conformational ensembles of a cold- and heat-denatured polymer. Interestingly, the HP case exhibits both cold and heat denaturation. In all likelihood, a protein of this size (16 residues) would not exhibit a phase transition, however, the parameters were optimized to make this situation occur for the HP case. Although the same set of parameters is used for the HH and PH cases, the competing enthalpy-entropy compensation mechanisms within a heterogenous protein (modelled by HP) make it possible for such a rich phase diagram.

The phenomenon of cold denaturation often does not occur because the temperature at which it *would* take place is too low to be observed. Osmolytes can be used to modify bulk solvent properties to control where the crossing points of the triangle shown in Fig. 10B occur. It is seen in Fig. 10D that the total entropy of a protein increases as a function of temperature. Fig. 10E shows that the order parameter for clathrate solvation content is a monotonically decreasing function of temperature, so that at higher temperatures a greater competition between buried and exposed-mobile states occur. Other order parameters can be easily calculated, such as the number of intramolecular H-bonds or hydrophobic contacts, which are shown in Fig. 10F as tracking one another. Although not shown here, tracking the native contact order parameter allows one to determine if a compact structure is native-like or that of a molten globular. In general, detailed information about solvent penetration and mechanical response of a protein is predicted at fine resolution, and this interplay is very important to protein function [Purkiss, et al. 2001], and these relationships have been more recently been probed experimentally [Kamerzell, et al. 2008; Pais, et al. 2009].

5.1 Conformational ensembles in the native and denatured states

Experiments indicate that native structure *persist* in the denatured states of proteins at low temperature [Shan, et al. 2010] and high temperature in the molten globular state [Shortle, 1999]. Established many years ago, the converse is true: There is appreciable solvent penetration into the native state [Woodward, et al. 1982], while buried secondary structure regions can be very resistant to solvent penetration [DeFlores & Tokmakoff, 2006]. These experimental results suggest to me that using template structures is justified, although this is not to say non-native contacts are negligible. For these cases, multiple templates should be used and these other templates can be computationally generated.

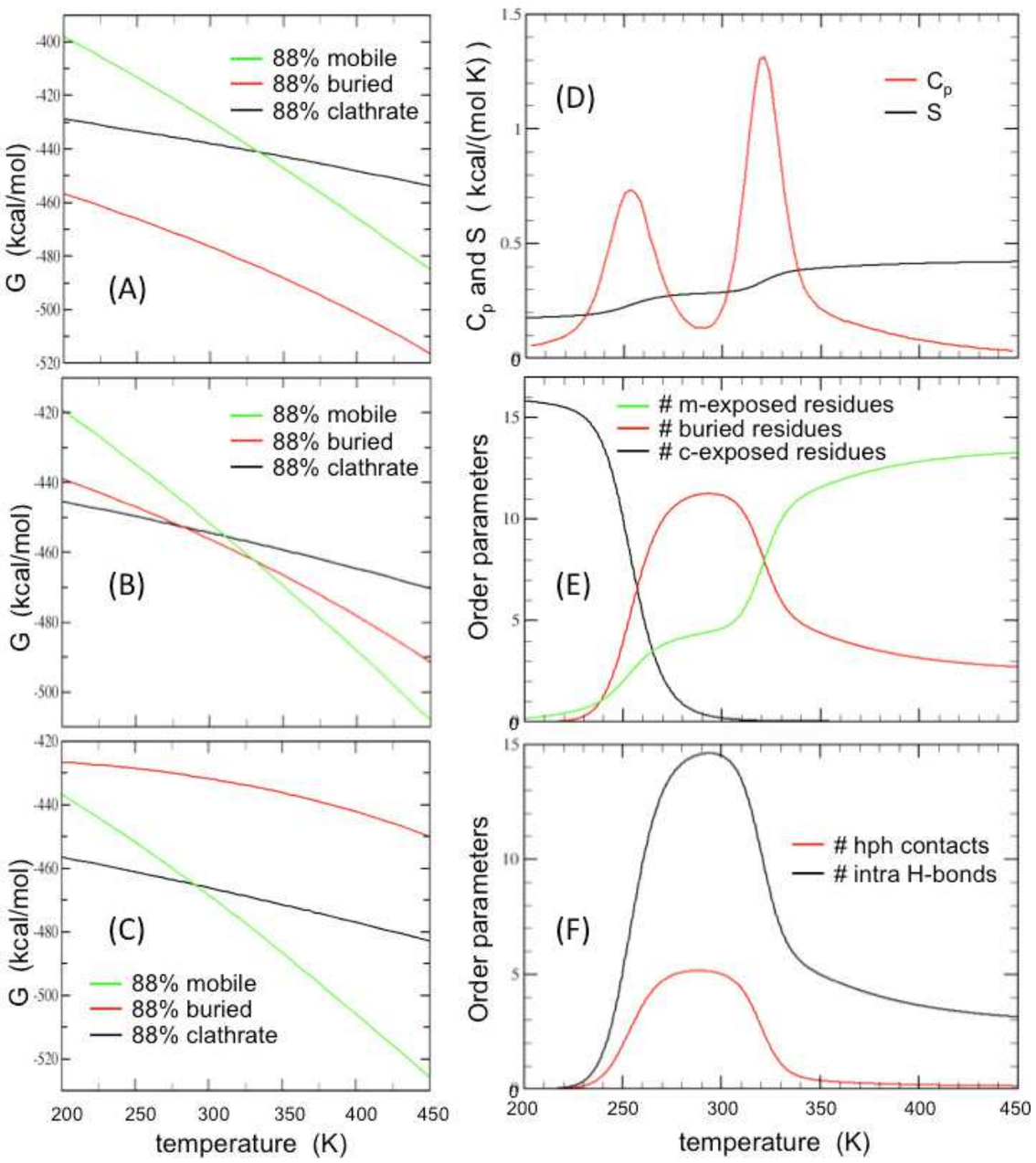


Fig. 10. Left column: The stability curves are shown for macrostates corresponding to a majority of residues that are clathrate (black), buried (red) and mobile (green). In all cases the majority contributor is 88%, and the two minority contributors are each 6%. Thus, the macrostates selected are located at the apex of each corner in the Gibbs triangle. In A, B, C, results are given for HH, HP and PH respectively. Right column: For the HP case only, D) shows the heat capacity and entropy, and E & F show the temperature dependence on five different types of order parameters that respectively correspond to the residue solvation states (clathrate, buried, mobile) and numbers of H-bonds and hydrophobic contacts.

5.2 Phenomenological modeling of protein stability

The standard thermodynamic analysis for protein stability assumes two states describe an unfolded (U) and folded (F) structure. The *mechanism of protein unfolding is hidden* in the temperature dependence of $\Delta G = G_U - G_F$. Specifically, ΔG is a concave quadratic function of temperature with two solutions for $\Delta G(T) = 0$ that yields $T = T_C$ for *cold* denaturation and $T = T_H$ for *heat* denaturation. However, the fact that two transition temperatures exist at all implies a minimum of two order parameters are required to describe the phenomena. The order parameters characterize the emergent behaviour of microscopic properties, which inspired the idea of modelling residue solvation states as three possible states. In a similar way, the molten globular state is distinct from the native state with respect to the degree of disorder, which inspired applying the native-like order parameter to define the FEL. Three independent order parameters appear to me as the minimum number of descriptors to describe protein stability consisting of the exposed unfolded state at low temperature, the compact native and molten globular states, and, the exposed unfolded state with little to no residual secondary structure at high temperature. The results presented in Fig. 10 produced three distinct states[6], indicating the ensemble of conformations of the unfolded state is structurally distinct at low and high temperatures, as demonstrated by the PH example in Fig. 10C. However, the *controversial* clathrate mechanism is invoked and it appears essential.

5.3 Clathrate mechanism for cold denaturation: Fiction or reality?

Models often invoke the clathrate mechanism to describe cold denaturation [Hansen, et al. 1998; Widom, et al. 2003]. The notion of this mechanism is based on *interpretations* of indirect measurements, which has been scrutinized [Graziano, 2004; Lopez, et al. 2009; Oshima, et al. 2009]. However, the critical review on protein hydration dynamics in solution [Halle, 2004] appears to me to dispel paradoxes that result from the controversy. Although the name clathrate may or may not be misleading, the interfacial thermodynamics model is based on general principles of statistical mechanics, which does not depend on a name. All that matters is the affect on the protein from solvent. The interfacial thermodynamics model accounts for native-like and disordered structure, and solvent penetration due to structural deformations. The model requires a partition function for the ensemble of water configurations around a residue, and an empirical contact term representing the affect on the protein's flexibility. This partition function is surely difficult to calculate from first principles, but a partition function can always be partitioned (hence the name) into a sum of terms. Dividing all configurations into two groups that classify the solvent in contact with a residue based on whether there is a small or large reduction in flexibility, no matter how small the difference may be is a valid *mathematical exercise* that does not change the physics because no Boltzmann factors are dropped.

The *exact* partition function, Z, is written as: $Z = Z_c + Z_m$ where Z_c sums over all terms that reduces the flexibility in the residue more than the terms summed in Z_m, no matter how small of a difference there may be. Therefore, Z_m describes a *more-mobile* water-residue

[6] Actually four distinct states are possible: Unfolded at low or high temperature, a native-like fold and a molten globular.

system, and Z_c describes a *less-mobile* water-residue system. Because "more-mobile" and "less-mobile" is cumbersome, I prefer to use the names *mobile* and *clathrate*. However, the important point is that it is not the property of water that is more or less mobile, it is the interaction between the water and residue that cause the residue to be more or less mobile, which is the basis of the classification. In fact, this description has been done for the twenty amino acids found in proteins to arrive at all the necessary solvation parameters [Du, et al. 2011]. Therefore, the notion of a clathrate mechanism is a *reality*, but it may not correspond to the original notion, and it must be calculated in the context of a large ensemble of water-residue configurations, where the flexibility of the residue must be quantified and assessed. Furthermore, a more refined classification scheme is in principle possible.

6. Unifying different perspectives on protein stability

Recent work on the thermodynamic response of a system subjected to geometrical constraints [Chen, et al. 2009] suggests that indirect intermolecular correlations, rather than geometric constraints, are the key to achieving a first-order phase transition. Although this latter conclusion was obtained in a different context, it appears the same idea is implied in the interfacial thermodynamics model. In particular, the template structure provides the native-state geometrical constraints, but all subsequent calculations only involve molecular correlations. That is, terms in the FEF couple intramolecular interactions to residue solvation properties as the protein conformation changes, albeit no new geometries are generated.

I was surprised that the "standard model" for protein stability and folding was criticized with much scrutiny recently [Ben-Naim, 2011] with several misconceptions highlighted. The main concerns raised were: 1) Free energy landscapes must be used to quantify protein stability, not energy landscapes; 2) non-additivity is an inherit property of entropy; 3) stability differences due to the trade off between intramolecular and protein-solvent H-bonds are too weak to drive protein folding; 4) hydrophobic interactions are also too weak to be the *dominant* driving force; whereas 5) the hydrophilic interactions are the strong driving forces that fold a protein. These five points and the clathrate mechanism controversy provide an opportunity to exam the assumptions of the interfacial thermodynamics model.

All solvation effects appear either as volume terms for the primary constituents (residues) or as surface terms between the interfaces of these constituents. The volume and surface terms taken together represent hydrophilic and hydrophobic interactions, the clathrate mechanism and protein-solvent H-bonds interactions using implicit solvent. Non-additivity in entropy components is a primary concern of the approach, and it directly deals with the free energy landscape. Moreover, all parameters have thermodynamic interpretations. For example, the parameters $(\varepsilon^{hph}, \alpha^{hph})$ for the hydrophobic interaction combine into the chemical potential to transfer a water molecule from a buried region in the protein to bulk solvent. In short, the modelling scheme put forth is complete, with the exception of long-range electrostatics.

7. Future direction

The particular FED that has been described above to define all the terms in the FEF, and using the simple MCC in the self-consistent constraint theory calculations is but one possible implementation of the interfacial thermodynamics model for protein stability. Similar to the

strategy described here, but with finer coarse-graining [Jacobs, 2007b], new software called *FAST* is being finalized. *Flexible to Flexibility* proteins to predict myriad thermodynamic and mechanical properties for high-throughput applications. With my collaborator Prof. Dennis Livesay at UNC Charlotte, and our research associates, Dr. Hui Wang and Dr. Chuanbin Du, the software has been designed and coded in C++ from scratch to provide a stable platform to support calculations similar to those described here. The FEF is of a general form that includes accounting for protonation states on titratable residues and explicit packing interactions. Moreover, the FEF is self consistently solved using an accurate rigidity algorithm for which Dr. Luis Gonzalez has helped develop, and he has documented its accuracy over the course of his Ph.D. studies. Many results to be published have been reported at several conferences, such as how model parameters are determined. With the Herculean effort spent on computational methods and optimization by Dr. Wang, *FAST* exceeds the speed of the mDCM and has greater accuracy. Going forward, the main concern is to find support to finish *FAST* so it can be released as free software to academic users.

8. Conclusion

From conception, the interfacial thermodynamics model for protein stability was designed to balance accuracy with computational cost such that it can be applied in high-throughput applications. To meet this *pragmatic objective*, the fundamental problem of non-additivity of conformational entropy that plagues free energy decomposition schemes has been tackled directly by employing the Distance Constraint Model. In particular, network rigidity is invoked as an underlying long-range interaction that couples entropy components between intramolecular subsystems comprising a protein. The problem is formulated as a free energy functional, and it is numerically solved using constitutive equations and self-consistent constraint theory. While the model makes many approximations, it is able to retain essential elements that describe protein thermodynamics and mechanical properties. Perhaps the best aspect of the interfacial thermodynamics model is that every term is intuitive physically and chemically. Different types of enthalpy-entropy compensation mechanisms can be modeled, and their competing effects can be simultaneously calculated with high efficiency.

9. Acknowledgement

This work has been supported by NIH R01 GM073082.

10. References

Andricioaei, I. & Karplus, M. (2001). On the calculation of entropy from covariance matrices of the atomic fluctuations, *J. Chem. Phys.*, Vol.115, pp. 6289-6292

Bakk, A. & Hoye, J. (2003). One-dimensional Ising model applied to protein folding. *Physica A*, Vol.323, pp. 504-518

Ben-Naim, A. (2011). Some aspects of the protein folding problem examined in one-dimensional systems, *J. Chem. Phys.*, Vol.135, No.085104, pp. 1-15

Bar-Yam, Y. (1997). *Dynamics of complex systems*. Addison Wesley Longman, Inc., ISBN 0-201-55748-7

Bray, D. & Duke T., (2004). Conformational Spread: The Propagation of Allosteric States in Large Multiprotein Complexes. *Annu. Rev. Biophys. Biomol. Struct.* Vol.33, pp. 53–73

Chen, Y., Kilburg, R. & Donohue, M. (2009). Thermodynamics of systems with different geometric constraints and intermolecular correlations, *J. Phys. Chem. B,* Vol.113, pp. 12530-12535

Costa, J. & Yaliraki, S. (2006). Role of rigidity on the activity of proteinase inhibitors and their peptide mimics, *J. phys. Chem. B,* Vol.110, pp. 18981-18988

David, C. & Jacobs, D. (2011). Characterizing protein motions from structure, *J. Mol. Graph Model,* Vol.31, pp.41-56

Dill, K. (1990). Dominant Forces in Protein Folding. *Biochemistry,* Vol.29, pp. 7133-7155

Dill, K. (1997). Additivity principles in biochemistry, *J. Biol. Chem*., Vol.272, pp. 701-704

DeFlores, L. & Tokmakoff, A. (2006). Water penetration into protein secondary structure revealed by hydrogen-deuterium exchange two-dimensional infrared spectroscopy, *J. Am. Chem. Soc.,* Vol.128, pp. 16520-16521

Du, C., Wang, H., Livesay, D. and Jacobs, D. (2011). *A Three-State Model of Amino Acid Solvation that Accounts for Hydration at Low Temperatures*, (in review)

Farrell, D., Speranskiy, K., & Thorpe, M. (2010). Generating stereochemically acceptable protein pathways, *Structure, Function, and Bioinformatics,* Vol.78, pp. 2908–2921

Fuxreiter, M., Magyar, C., Juhasz, T., Szeltner, Z., Polgar, L. & Simon, I. (2005). Flexibility of Prolyl Oligopeptidase: Molecular dynamics and molecular framework analysis of the potential substrate pathways, *Proteins: Structure, Function, and Bioinformatics,* Vol.60, pp. 504-512

Gonzalez, L., Wang, H., Livesay, D. and Jacobs, D. (2011a). *A virtual pebble game to ensemble average graph rigidity* (in review)

Gonzalez, L., Wang H., Livesay, D. &Jacobs ,D. (2011b). *Calculating Ensemble Averaged Descriptions of Protein Rigidity without Sampling,* in press, Plos One.

Graziano, G. (2004). Comment on "The hydrophobic effect" by B. Widom, P. Bhimalapuram and K. Koga, *Phys. Chem. Chem. Phys*., 2003, Vol.5, pp.-3085, *Phys. Chem. Chem. Phys*., Vol.6, pp. 4527-4528

Gunsteren, W., Bakowies, D., Baron, R., Chandrasekhar, I., Christen, M., Daura, X., Gee, P., Geerke, D., Glattli, A., Hunenberger, P., Kastenholz, M., Oostenbrink, C., Schenk, M., Trzesniak, D., Vegt, N., & Yu, H. (2006). Biomolecular Modeling: Goals, Problems, Perspectives. *Angew. Chem. Int. Ed.,* Vol.45, pp. 4064-4092

Halle, B. (2004). Protein hydration dynamics in solution; a critical survey, *Phil. Trans. R. Soc. Lond. B*., Vol.359, pp. 1207-1224

Hansen, A., Jensen, M., Sneppen, K. & Zocchi G. (1998). Statistical mechanics of warm and cold unfolding in proteins, *Eur. Phys. J. B*, Vol.6, pp. 157-161

Heal, J., Wells, S., Jimenez-Roldan, J., Freedman, R. & Romer, R. (2011). Rigidity analysis of HIV-1 protease, *J. Phys. CMMP10 conference series,* Vol.286, pp. 012006

Hespenheide, B., Rader, A., Thorpe, M. & Kuhn L. (2002). Identifying protein folding cores from the evolution of flexible regions during unfolding, *J. Mol. Graphics and Modelling,* Vol.21, pp. 195-207

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

Istomin, A., Gromiha, M., Vorov, O., Jacobs, D. &. Livesay, D. (2008). Insight into Long-Range Nonadditivity within Protein Double-Mutant Cycles, *Proteins: Structure, Function, and Bioinformatics,* Vol.70, pp. 915-924

Jacobs, D., Kuhn, L. & Thorpe, M. (1999). *Flexible and Rigid Regions in Proteins*, Rigidity Theory and Applications, pp. 357-84, Eds: M.F. Thorpe & P.M. Duxbury, Plenum Publishing , New York, USA

Jacobs, D., Rader, A., Kuhn, L. & Thorpe, M. (2001). Graph Theory Predictions of Protein Flexibility, *Proteins: Structure, Function, and Genetics*, Vol.44, No.2, pp. 150-65

Jacobs D., Dallakayan, S., Wood G. & Heckathorne A. (2003). Network rigidity at finite temperature: Relationships between thermodynamic stability, the nonadditivity of entropy, and cooperativity in molecular systems, *Physical Reviews E*, Vol.68, pp. 061109 1-21

Jacobs, D. & Wood, G. (2004). Understanding the α-Helix to Coil Transition in Polypeptides Using Network Rigidity: Predicting Heat and Cold Denaturation in Mixed Solvent Conditions, *Biopolymers,* Vol.75, pp. 1-31

Jacobs, D. & Dallakayan, S. (2005). Elucidating Protein Thermodynamics from the Three Dimensional Structure of the Native State Using Network Rigidity, *Biophysical Journal*. Vol.88, pp. 1-13

Jacobs, D., Livesay, D., Hules, J. & Tasayco, M. (2006a). Elucidating quantitative stability-flexibility relationships within thioredoxin and its fragments using a distance constraint model, *Journal of Molecular Biology*, Vol.358, pp. 882-904

Jacobs, D. (2006b). *Predicting Protein Flexibility and Stability using Network Rigidity: A new Modeling Paradigm*, Recent Research Developments in Biophysics Publisher: Transworld Research Network, Vol.5, pp.71-131 ISBN: 81-7895-215-7, Trivandrum, India

Jacobs, D. & Fairchild, M. (2007a). *Thermodynamics of a beta-hairpin to coil transition elucidated by Constraint Theory*, Biopolymer Research Trends, Ed: Pablo C. Sánchez. Nova Publishers, ISBN: 1-60021-984-5 45-76, New York, USA

Jacobs, D. (2007b). *Computer Implemented System for Quantifying Sability and Flexibility Relationships in Macromolecules*. US Patent Pending, Application: 12232008.

Jacobs, D. (2010). Ensemble-Based methods for Describing Protein Dynamics, *Current Opinion in Pharmacology,* Livesay, D. (Ed.) Vol.10, pp. 760-769

Jacobs, D., Livesay, D., Mottonen, J., Vorov, O. Istomin, A. & Verma, D. (2012). *Ensemble Properties of Network Rigidity Reveal Allosteric Mechanisms, Allostery: Methods and Protocols*, Methods in Molecular Biology, Vol.796, Aron W. Fenton (ed.), Humana Press, Springer Science and Business Media, LLC

Jimenez-Roldan, J., Wells, S., Freedman, R. & Romer, R. (2011). Integration of FIRST, FRODA and NMM in a coarse grained method to study protein disulphide isomerase conformational change, *J. Phys. CMMP10 conference series,* Vol.286, pp. 012002

Kamerzell, T. & Middaugh, C. (2008). The complex inter-relationships between protein flexibility and stability, *J. Pharm. Sci.*, Vol.97, pp. 3494-3517

Kittel, C. (1996). *Introduction to Solid State Physics*, 7th Ed., Wiley

Klepeis, J., Gunasekaran, K., Ma, B. & Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins,* Vol.57, pp 433–443

Lee, M., Wood, G. & Jacobs, D. (2004). Investigations on the alpha-helix to coil transition in HP-heterogeneous polypeptides using network rigidity, *Journal Physics, Condensed Matter*, Vol.16, pp. S5035-46

Lei, M., Kuhn, A., Zavodszky, M. & Thorpe, M. (2004). Sampling protein conformations and pathways, *J. Comput. Chem.,* Vol.25, pp. 1133-1148

Lindorff-Larsen, K., Dror, R. & Shaw, D. (2009). Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.,* Vol.19, pp. 120-127

Livesay, D., Dallakyan, S., Wood, G. & Jacobs, D. (2004). A flexible approach for understanding protein stability, *FEBS Letters,* Vol.576, pp. 468-76

Livesay, D. & Jacobs, D. (2006). Conserved quantitative stability/flexibility relationships (QSFR) in an orthologous RNase H pair, *Proteins: Structure, Function, and Bioinformatics*, Vol.62, pp. 130-43

Livesay, D., Huynh, D., Dallakyan, S. & Jacobs, D. (2008). Hydrogen bond networks determine emergent mechanical and thermodynamic properties across a protein family, *Chemistry Central Journal*, Vol.2, No.17, pp. 1-20

Lopez, C., Darst, R. & Rossky, P. (2009). Mechanistic elements of protein cold denaturation, *J. Phys. Chem. B.*, Vol. 112, pp.5961-5967

Luque, I., Leavitt, S. & Freire, E. (2002). The linkage between protein folding and functional cooperativity: Two sides of the same coin? *Annu. Rev. Biophys. Biomol. Struct*, Vol.31, pp. 235-56

Mamonova, T., Speranskiy, K. & Kurnikova, M. (2008). Interplay between structural rigidity and electrostatic interactions in the ligand binding domain of GluR2., *Proteins: Structure, Function, and Bioinformatics,* Vol.73, pp.656-671

Mark, A. & van Gunsteren, W. (1994). Decomposition of the free energy of a system in terms of specific interactions. Implications for theoretical and experimental studies, *J. Mol. Biol.*, Vol.240, pp. 167-176

Mottonen, J., Xu M., Jacobs, D. & Livesay, D. (2009). Unifying mechanical and thermodynamic descriptions across the thioredoxin protein family, *Proteins: Structure, Function, and Bioinformatics,* Vol.75, pp. 610-627

Mottonen, J., Jacobs, D. & Livesay, D. (2010). Allosteric Response is Both Conserved and Variable Across Three CheY Orthologs, *Biophysical Journal.*, Vol.99, pp. 2245-54

Muñoz, V. (2001). What can we learn about protein folding from Ising-like models? *Curr Opin Struct Biol*, Vol.11, 212–216

Nielsen, S., Bulo, R., Moore, P. & Ensing, B. (2010). Recent progress in adaptive multiscale molecular dynamics simulations of soft matter. *Phys. Chem. Chem. Phys.*, Vol.12, pp. 12401–12414

Onuchic, J. & Wolynes, P. (2004). Theory of protein folding, *Cur. Opin. Struct. Biol.*, Vol.14, pp. 70-75

Oshima, H., Yoshidome, T., Amano, K., & Kinoshitab, M. (2009). A theoretical analysis on characteristics of protein structures induced by cold denaturation, *J. Chem. Phys.*, Vol.131, pp. 205102

Pace, N., Trevino S., Prabhakaran, E. & Scholtz, M. (2004). Protein structure, stability and solubility in water and other solvents. *Phil. Trans. R. Soc. Lond. B*, Vol.359, pp. 1225–1235

Pais, T., Lamosa, P., Garcia-Moreno B., Turner, D. & Santos, H. (2009). Relationship between protein stabilization and protein rigidification induced by mannosylglycerate, *J. Mol. Biol.,* Vol.394, pp.237-250

Petsko, G. & Ringe, D. (2004). *Protein structure and function.* New Science Press Ltd., ISBN 0-9539181-4-9

Purkiss, A., Skoulakis, S. & Goodfellow J. (2001). The protein-solvent interface: a big splash, *Phil. Trans. R. Soc. Long. A*, Vol.359, pp. 1515-1527

Radestock, S. & Gohlke, H. (2008). Exploiting the link between protein rigidity and thermostability for data-driven protein engineering, *Eng. Life Sci.,* Vol.8, pp. 507-522

Radestock S. & Gohlke H. (2011). Protein rigidity and thermophilic adaptation. *Proteins: Structure, Function, and Bioinformatics,* Vol.79, pp. 1089-1108

Rader, A., Hespenheide, B., Kuhn, L., & Thorpe, M. (2002). Protein unfolding: rigidity lost. *Proc. Natl. Acad. Sci*., USA, Vol.99, pp. 3540-3545

Radar, A., Anderson, G., Isin, B., Khorana, H., Bahar, I. & Klein-Seetharaman K. (2004). Identification of core amino acids stabilizing rhodopsin, *Proc. Nat. Acad. Sci.,* Vol.101, pp.7246-7251

Rader, A. (2010). Thermostability in rubredoxin and its relationship to mechanical rigidity. *Phys. Biol.,* Vol.7, pp. 016002

Shan, B., McClendon, S., Rospigliosi, C., Eliezer, D. & Raleigh D. (2010). The cold denatured state of the C-terminal domain of protein L9 is compact and contains both native and non-native structure, *J. Am. Chem. Soc.,* Vol.132, pp. 4669-4677

Shortle, D. (1999). Protein folding as seen from water's perspective Nature Struct, *Biol. news and views*, Vol.6 pp. 203-205

Thomas, S., Tang, X., Tapia, L. & Amato, N., (2007). Simulating Protein Motions with Rigidity Analysis, *J. Comp. Bio.*, Vol.14, pp. 839-855

Verma, D., Jacobs, D. & Livesay, D. (2010). Predicting the Melting Point of Human C-Type Lysozyme Mutants, *Current Protein and Peptide Science*, Vol.11, pp. 562-572

Vorov, O., Livesay, D. & Jacobs, D. (2008). Conformational entropy of an ideal cross-linking polymer chain, *Entropy*, Vol.10, pp. 285-308

Vorov, O., Livesay, D. & Jacobs, D. (2009). Helix/coil nucleation: A local response to global demands, *Biophysical Journal*, Vol.97, pp. 3000-3009

Vorov, O., Livesay, D. & Jacobs, D. (2011). Nonadditivity in Conformational Entropy Upon Molecular Rigidification Reveals a Universal Mechanism for Affecting Folding Cooperativity, *Biophysical Journal*, Vol.16, pp. 1129-38

Wang, H., Fairchild, M., Livesay, D. & Jacobs, D. (2011). *Intrinsic molecular partition functions for amino acids* (unpublished, to be submitted)

Wells, S., Menor, S., Hespenheide, B. & Thorpe, M. (2005). Constrained geometric simulation of diffusive motion in proteins*, Phys. Biol.,* Vol.2, pp. S127–S136

Whiteley, W. (2005). Counting out to the flexibility of molecules, *Phys. Biol*., Vol.2, pp. S116-126

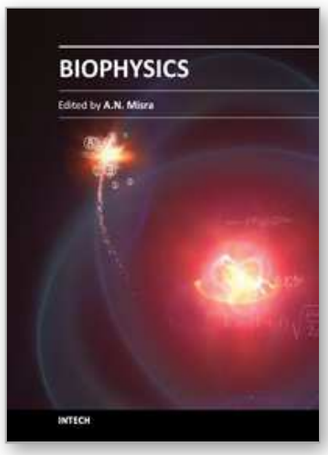
Widom, B., Bhimalapuram, P. & Koga, K. (2003). The hydrophobic effect, *Phys. Chem. Phys*., Vol.5, pp. 3085-3093

Wood, G., Clinkenbearda, D. & Jacobs, D. (2011). Nonadditivity in the alpha-helix to coil transition, *Biopolymers*, Vol.95, pp. 240–253

Woodward, C., Simon, I. & Tuchsen, E. (1982). Hydrogen Exchange and the Dynamic Structure of Proteins, *Molecular and Cellular Biochemistry*, Vol. 48, pp. 135-160

Yao, P., Zhang, L. & Latombe J. (2012). Sampling-based exploration of folded state of a protein under kinematic and geometric constraints, *Structure, Function, and Bioinformatics,* Vol.80 pp.25-43

Zamparo M. & Pelizzola, A. (2006). Kinetics of the Wako-Saitô-Muñoz-Eaton model of protein folding. *Phys. Rev. Lett.*, Vol.97, pp. 068106 (1-4)



Biophysics

Edited by Dr. Prof. Dr. A.N. Misra

ISBN 978-953-51-0376-9
Hard cover, 220 pages
Publisher InTech
Published online 21, March, 2012
Published in print edition March, 2012

Biophysics is a vast cross-disciplinary subject encompassing the fields of biology, physics and computational biology etc in microbes, plants, animals and human being. Wide array of subjects from molecular, physiological and structural are covered in this book. Most of these chapters are oriented toward new techniques or the application of techniques in the novel fields. The contributions from scientists and experts from different continents and countries focuss on major aspects of biophysics. The book covers a wide range of topics reflecting the complexity of the biological systems. Although the field of biophysics is ever emerging and innovative, the recent topics covered in this book are contemporary and application-oriented in the field of biology, agriculture, and medicine. This book contains mainly reviews of photobiology, molecular motors, medical biophysics such as micotools and hoemodynamic theory.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Donald J. Jacobs (2012). An Interfacial Thermodynamics Model for Protein Stability, Biophysics, Dr. Prof. Dr. A.N. Misra (Ed.), ISBN: 978-953-51-0376-9, InTech, Available from: http://www.intechopen.com/books/biophysics/an-interfacial-thermodynamics-model-for-protein-stability

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen