

Chapter 2

THERMODYNAMICS OF A β -HAIRPIN TO COIL TRANSITION: APPLICATION OF FREE ENERGY DECOMPOSITION AND CONSTRAINT THEORY

Donald J. Jacobs and Michael J. Fairchild
Department of Physics & Optical Science
The University of North Carolina at Charlotte
Charlotte, NC 28223

Abstract

A novel Distance Constraint Model (DCM) that combines concepts of free energy decomposition (FED) and constraint theory has been found successful in predicting thermodynamic stability in proteins and polypeptides. The DCM represents microscopic interactions as distance constraints having enthalpy and entropy contributions that define a FED. We present new results for polypeptides that undergo a β -hairpin to coil transition using a minimal DCM that accounts for crosslinking hydrogen bonding and two types of torsion states along the polypeptide chain. The topological arrangement of these constraints defines a mechanical framework, from which independent constraints are determined and used to estimate conformational entropy. Here, we identify independent constraints using a mean field approximation, which simplifies the mathematical analysis greatly while retaining the essential physics. The free energy landscape and partition function are calculated exactly for two types of Gibbs ensembles: (i) Restricted conformations allowing only native hydrogen bond contacts to form, and (ii) an ensemble allowing both native and non-native hydrogen bond contacts. We compare heat capacity curves, phase diagrams and free energy landscapes for both of these ensembles. We also demonstrate the importance of employing constraint theory within a FED scheme to account for non-additivity effects in conformational entropy.

1. Introduction

Hairpin conformations represent the dominant nucleic acid secondary structure, and β -hairpin conformations are the third most common secondary structure in proteins. Besides their naturally occurring abundance, it has been found that non-native folding of hairpins (for

proteins and nucleic acids) parallel the onset of a variety of human diseases [15, 5]. Due to their biological importance, studies of thermodynamic stability of hairpin structures and the hairpin-coil transition have received much attention recently [43, 50]. We present a novel phenomenological approach that employs a *Distance Constraint Model* (DCM) [18], using free energy landscapes, with the goal of better understanding the essential mechanisms responsible for the thermodynamic stability and folding/unfolding of a β -hairpin.

The DCM is a coarse-grained approach that combines a *free energy decomposition scheme* [7] with *constraint theory* [40]. Constraint theory is employed to determine the rigid and flexible parts of a molecule, and we refer to this detailed mechanical information as *network rigidity* [46]. In some respects, the DCM appears similar to Gō-like models [10, 4, 6] or Ising-like models [37, 17, 38, 51, 32, 28]. In reality, the DCM is distinctly different because network rigidity is calculated explicitly, and this information is used as an underlying mechanical interaction to account for the effect of non-additivity of entropy [18, 27]. The DCM resolves the problem of non-additivity in component entropies with regards to *conventional* free energy decomposition schemes [36, 7] by accounting for correlation between conformational entropy components [2]. The DCM has been used previously to investigate the alpha-helix to coil transition for homogeneous polypeptide chains [18], and for heterogeneous chains [31]. The DCM describes both heat and cold denaturation [25], and it has been applied to study protein stability [19]. In particular, the DCM is robust in reproducing experimental protein folding heat capacity curves [33], and it elucidates thermodynamic and mechanical properties of proteins important for their function by providing quantitative stability and flexibility relationships [34, 21].

In this article, we first examine fundamental aspects about free energy decomposition schemes and show how non-additivity of entropy is accounted for using network rigidity. Second, the *minimum Distance Constraint Model* (mDCM) is defined in the context of investigating the β -hairpin to coil transition. Not being concerned about the properties of a particular sequence, we investigate generic properties of a homogeneous polypeptide. In addition, we employ a mean field approximation known as *Maxwell constraint counting* to simplify the mathematical analysis for conceptual clarity and to arrive at exact solutions that retain the essential elements of the β -hairpin to coil transition. A comparison of the thermodynamic response is made between two cases: (1) only native contacts are accessible (i.e. no misfolding is possible), and (2) non-native contacts are accessible. Within the context of the mDCM, we also consider switching off the network rigidity aspect of the model, to see the effect of non-additivity of entropy.

2. Hidden Thermodynamics Revealed

In the late eighties continuing through the early to middle nineties there was excitement about the prospect of explaining free energy changes in protein mutations by way of a *linear* free energy decomposition scheme [39, 35, 41, 11]. The idea was to make systematic transfer measurements similar to what is done in measuring the partition coefficient. The measured changes in free energy, enthalpy and entropy were recorded into lookup tables for standard conditions. Assuming a linear dependence, the idea was to predict the total changes in free energy, enthalpy and entropy of a protein by summing over the individual changes in these quantities. Unfortunately, this approach did not consistently work well,

and there were many large (and unexplained) discrepancies, which at the time were very surprising. Therefore, quite early on, the phrase “hidden thermodynamics” was introduced to suggest that a non-trivial reason could be found to explain all the data [9]. Despite many ingenious attempts, no linear free energy decomposition scheme could be constructed that worked well over a comprehensive dataset.

The collection of failed attempts to decompose the free energy of a protein into sums over specific interactions revealed that additivity principles will almost always break down. Mark and Gunsteren [36] rigorously showed that the observed non-additivity is due to the intrinsic property of entropy, and concluded:

“In regard to the detailed separation of free energy components, we must acknowledge that the hidden thermodynamics of a protein will, unfortunately, remain hidden.”

The importance of finding hidden thermodynamics within a free energy decomposition scheme is best appreciated by considering how to calculate conformational entropy. It was realized that if one could track correlations, perhaps one could utilize a free energy decomposition [2]. But how to do this, besides using brute force Molecular Dynamics (MD) simulations [14] was left as an open problem. Due to the enormous complexity of this problem, coarse grained models are needed [29, 16]. In this work, we reveal the hidden thermodynamics in a free energy decomposition scheme through network rigidity [46].

2.1. Conformational Entropy

Conformational entropy is of central importance for the thermodynamic stability and function of proteins [8, 3]. It often happens that a protein folds into a well-defined native structure, stabilized by crosslinking interactions, such as hydrogen-bonds (H-bonds). A *structural transition* involving the loss of H-bonds occurs at elevated temperatures due to an increase in conformational entropy, S_c , which is related to atomic motions on all time scales. Consequently, it is difficult using all-atom MD simulations with explicit solvent to ascertain thermodynamic properties. For example, a typical time step of $\sim 10^{-15}$ sec is used in MD simulations to numerically integrate $\sim 10^4$ to 10^6 second order differential equations simultaneously for small to large proteins respectively. Even with enormous computational power (say 10,000 processors) it is impractical if not impossible to reach a millisecond (10^{-3} sec) for moderate size proteins. Moreover, biologically relevant time-scales often exceed milliseconds. Thus, much more than $\sim 10^{12}$ steps are necessary to predict thermodynamic behavior, and this is just for one thermodynamic condition (e.g. one temperature). Of course, for a small polypeptide undergoing the β -hairpin to coil transition, full scale all-atom MD simulations can reach millisecond time scales [50].

Interestingly, only statistical weights obtained from averaging over long-time MD trajectories are necessary to obtain thermodynamic properties [44]. After painstaking computational expense, the fine details about atomic motions are unused. It is beneficial, therefore, to throw away these unwanted details, allowing faster computation times through *coarse graining*, which leads to many types of *reduced models* [29, 6]. Frequently, reduced models involve treating solvent molecules implicitly (with effective potentials) and grouping atoms together as effective units. A coarse grained approach allows one to investigate thermodynamic properties more efficiently by ignoring many non-essential details from the

start. Coarse grained Ising-like models [37] have been used to study the helix-coil transition [51, 32], the hairpin-coil transition [28, 38] and protein thermodynamics [17]. Unfortunately, all of these Ising-like models suffer from assuming an additive free energy decomposition scheme [36, 7], which *fundamentally* limits their generality. How can these limitations be resolved? Relying on strong supporting evidence [18, 25, 33, 19, 34, 21], which has been summarized in a concise review [27], we claim with high confidence that application of constraint theory resolves the problem of non-additivity in entropy (and free energy). The approach of the DCM has the potential to be extremely fast while retaining high accuracy, both of which are needed in high throughput computational biology applications [47].

2.2. Constraint Theory

Intuitively, one may expect that as constraints are added to a system, its number of *degrees of freedom* (DOF) will decrease. *Constraint theory* makes this intuitive process mathematically precise. For a molecular system containing a certain set of distance constraints between pairs of atoms, some groups of atoms will be mutually rigid, while others will form flexible mechanisms that allows for continuous relative motion between the atoms [40, 46]. Now consider adding a distance constraint to the network.

Adding a distance constraint between a pair of atoms that allow for relative motion (i.e. variable distance), will indeed reduce the number of DOF by one. This constraint is said to be *independent*. If a new constraint is added between a pair of atoms that are already mutually rigid, the constraint will not reduce the number of DOF, and this constraint is said to be *redundant*. Constraint theory deals with the problem of determining an algorithm (numerical, combinatorial counting or otherwise) to test whether a constraint is *independent* or *redundant*. Once such an algorithm is available, one can apply this test many times to completely determine all rigid regions (groups of atoms) and all flexible regions throughout the network. Within (rigid, flexible) regions, the *number* of (redundant constraints, DOF) is also determined. The specification of all (rigid,flexible) regions with their numbers of (redundant constraints, DOF) is referred to as *network rigidity*. Note, however, that labeling of any individual constraint as redundant or independent is *not* unique, and depends on the order that the constraints are tested.

Network rigidity plays an important role in determining the allowed motions for a polypeptide conformation [20], and therefore it is intimately related to conformational entropy. In particular, for each constraint in the network, we show how knowledge of that constraint as redundant or independent can be used in calculations to obtain an accurate estimate of the conformational entropy. Before this procedure is described, we elaborate on some important aspects of network rigidity using simple two dimensional examples. Bear in mind that the same concepts presented here will be applied to the three dimensional structures of a polypeptide undergoing a β -hairpin to coil transition.

2.2.1. Graph Rigidity in Two Dimensions

In the plane, there are three *trivial DOF* for a rigid object: two corresponding to its Center of Mass (CM) coordinates and one DOF for rotation about its CM. For point-like particles

(non-extended objects), there are only two DOF, one for each CM coordinate. Thus for three particles in the plane, there are six DOF needed to specify their configuration, two for each particle. If we now introduce constraints, such as fixed-length bars between each pair of particles, then the number of DOF is reduced. As a first example, consider the triangle illustrated in Figure 1.

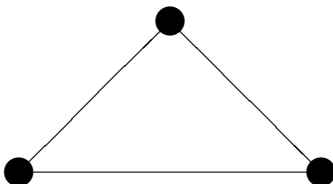


Figure 1. A triangle in the plane is rigid.

Normally, we disregard trivial DOF (corresponding to the location of the CM and rotation angle(s) about the CM), and consider internal DOF that govern relative motions between particles (atoms). The graph in Figure 1 has zero internal DOF: $3(2) - 3 - 3 = 0$, because there are three particles (with two DOF each), three bars (each being an independent distance constraint), and three trivial DOF. This answer was obtained by simple counting, and did not require knowledge about the precise coordinates of the particles.

In our work with peptides and proteins, the atomic structure is represented as a connected graph, and we only calculate rigidity properties that can be determined directly from the graph [22, 20, 49].

As an example of a flexible graph, consider Figure 2. This square has one internal DOF,

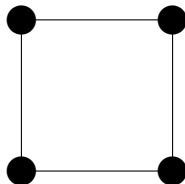


Figure 2. A square in the plane is flexible.

corresponding to a shearing motion. To see this, one can hold the bottom bar fixed and slide the top bar left-and-right. Constraint counting gives the number of internal DOF as $4(2) - 4 - 3 = 1$. Although four angles change, the motion of all four particles (atoms) are correlated, fully determined by a single internal DOF, which can be selected to be *any one of the four angles*. This example illustrates the important point that identification of a flexible region and its *number* of internal DOF is unique, but the choice of which angle (or other suitable generalized coordinate) is not.

We now consider the case that certain subgraphs of a graph are rigid while other subgraphs are flexible. This case is illustrated in Figure 3. With six particles and nine constraints, simple constraint counting gives $6(2) - 9 - 3 = 0$, predicting a rigid graph. This count is obviously *wrong*. By inspection, the right square has a shearing motion with one internal DOF, while the left square is overconstrained with one redundant constraint. If we *move* a diagonal constraint from the left square over to the right square, then the whole

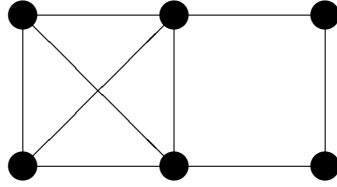


Figure 3. A case where constraint counting fails.

graph would be rigid. This illustrates the importance of where constraints are placed. Regions of high constraint density are likely to be rigid and contain redundant constraints, while regions of low constraint density tend to be flexible. For all but the smallest graphs, the method of inspection fails because of the long-range character of network rigidity [12].

Fortunately, several efficient graph algorithms for calculating network rigidity of large graphs (in both two and three dimensions) are available to determine exact network rigidity properties [30, 12, 22, 24, 45, 23, 49, 26, 20]. These methods are based on the core calculation of testing whether a constraint is independent or redundant. Applying this constraint-test recursively, a given network is built up one constraint at a time. Although algorithms for testing constraints as independent or redundant are systematic, they are not unique because they depend on the order in which constraints are placed (i.e. the way the network is built).

2.2.2. Maxwell Constraint Counting as a Mean Field Approximation

Given N particles embedded in a m -dimensional space, and knowing that there are I *independent* distance constraints between various pairs of particles, it follows that the number of *internal* DOF, \mathcal{D} , is simply given by

$$\mathcal{D} = mN - I - \frac{m(m+1)}{2}, \quad (1)$$

where $m(m+1)/2$ is the number of trivial rigid body DOF in an m -dimensional space. Although Eq. 1 is exact, it happens to be totally useless, because we do not know *a priori* the number of *independent* constraints! The non-trivial part of constraint theory is finding the way to exactly calculate I . In this work, we follow the genius of Maxwell¹ who assumed every constraint is independent until the entire network is globally rigid, at which point all additional constraints are redundant. The constraint counting of Maxwell (now commonly called Maxwell constraint counting) is schematically illustrated in Fig. 2.2.2..

Maxwell constraint counting turns out to be a much better approximation than one may initially think possible. To see why, notice that as long as the constraint density is *uniform* (with virtually no fluctuations), Maxwell constraint counting is essentially exact! In modern language, this represents a mean field approximation, because the actual constraint density is replaced by a constant average value with zero fluctuations. In practice, some mistakes will be made, such as the prediction that the graph in Fig. 3 is globally rigid with no redundant constraints. In Fig. 3, Maxwell constraint counting goes wrong because there are

¹Incidentally, this is the same Scottish theoretical physicist and mathematician James Clerk Maxwell (1831-1879) famed for Maxwell's equations of electrodynamics.

regions of high and low constraint density. In this work, we will apply Maxwell constraint counting and accept all mistakes in the same spirit one employs in any type of mean field approximation. We note that this simplification is made for convenience to make the mathematical analysis simple, while retaining the essential physics in the problem. Whereas in all prior works exact constraint counting was implemented, our main objective here is to highlight the importance of invoking constraint theory when working with free energy decomposition schemes.

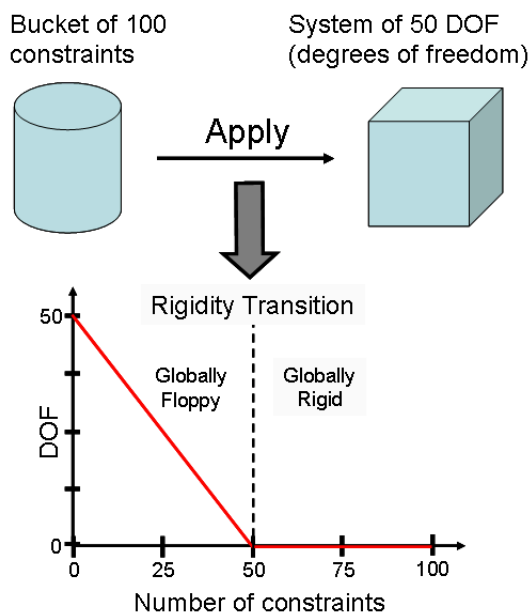


Figure 4. Schematic showing Maxwell Constraint Counting.

2.3. Free Energy Decomposition Schemes

The basic idea behind a free energy decomposition scheme is to decompose a large system into *independent* (uncoupled) parts. Then we assume each part (or subsystem) is in thermodynamic equilibrium with the rest of the system. A partition function can then be constructed for each subsystem *separately*, and in turn, a free energy function for the subsystem is well-defined. For distinguishable subsystems, the total partition function of the system is given by: $Z_{\text{total}} = \prod_i Z_i$, where Z_i is the partition function for the i -th subsystem. When this separation is possible, it follows that the free energy decomposes into a linear function, where the total free energy is given by: $G_{\text{total}} = \sum_i G_i$, where $G_i = -RT \ln Z_i$ is the free energy of the i -th subsystem. When this linear decomposition is possible, it follows that the total enthalpy and entropy are each separately additive functions over the subsystems. This approach is commonly used in small molecules where, for example, it is often written $Z_{\text{total}} = Z_{\text{trans}} Z_{\text{rot}} Z_{\text{vib}}$, where the translational, rotational and vibrational DOF are assumed to be decoupled. This approximation is commonly used in physical chemistry because it is very good when applied to small molecules. However, it generally does not

work for macromolecules [7].

Often, a physics problem can be solved much easier when using a very well-chosen coordinate system. The entire argument given above about finding a linear free energy decomposition scheme rests upon finding an appropriate coordinate system. For example, if one is given a solid material (not necessarily a crystal), one can approximate the interactions between atoms as springs. Obviously, all atoms are coupled to one another, and a direct attempt to apply the above ideas will lead to errors at low temperature (e.g. the Einstein model for heat capacity in solids). However, finding the normal modes of vibration allows one to transform into the normal coordinates, which describe complex collective motions among the atoms, but each mode is independent of all others. In this case, one can apply a linear free energy decomposition scheme that is essentially exact (e.g. phonon theory). Without working all the tedious details, an intermediate theory that is quite accurate can be established by assuming some simple dispersion relationships for lattice vibrations (e.g. the Debye model for heat capacity in solids).

In simple Ising-like models such as the classic Lifson-Roig model [32] for the alpha-helix to coil transition, the coordinates (interaction types) were well-chosen to make the problem amenable to a linear free energy decomposition scheme. However, in general, for polypeptides that fold into structures with many native contacts compared to unfolded structures with much fewer contact interactions, any well-chosen coordinate system for one conformational state will be a poor choice for a different conformation. Indeed, the main reason why non-additivity in free energy decomposition schemes is detected more readily in macromolecules compared to small molecules is because of the diversity in their accessible conformational states. As a result, the Lifson-Roig model is limited in applicability, and its phenomenological parameters are not transferable between different polypeptides [1]; even worse, for the same polypeptide, they depend on the length of the polypeptide [18]. Other Ising-like models [17] express free energy as a linear sum of components that are directly related to solvent exposed surface areas. These models assume, knowingly or unknowingly, the local exposed surfaced areas define a good generalized coordinate system that allows linear free energy decomposition to hold for all types of conformational states. Despite many successes that have been made under the guise of using additive free energy decomposition schemes, bad consequences also follow. Some of these include non-transferability of parameters, violation of the second law of thermodynamics when dealing with entropy of hydration [13], and oversimplifications that greatly limit the applicability of the model.

An alternate approach is to first establish a well-defined set of interactions to model, and realize that these interactions compete with one another differently in different conformational states. Second, for any given conformation the coupling that exists between subsystems will be accounted for in a computationally efficient way that provides reasonable accuracy. Both of these objectives are met by combining the concepts of free energy decomposition with constraint theory. A Distance Constraint Model (DCM) was introduced [18] to resolve the problem of non-additivity within a free energy decomposition that appears in the process of estimating conformational entropy. Consequently, the DCM deals with non-additivity effects by explicitly regarding network rigidity as an underlying mechanical interaction.

2.4. Connecting Thermodynamics to Network Rigidity

In the DCM, contact interactions (i.e. H-bonding, torsion interactions, hydrophobic interactions, etc. . .), are modeled as distance constraints, each of which is assigned an enthalpy and nominal entropy contribution. The entropy contribution reflects the number of microstates (atomic configurations) for which the energy is nearly a constant within a coarse-grained energy bin. The conformation of the molecule determines the placement of constraints (of various types) leading to a particular constraint topology, forming a mechanical framework, denoted \mathcal{F} . For any given mechanical framework (a connected graph) the Gibbs free energy of the framework is given by Eq. 2.

$$G(\mathcal{F}) = H(\mathcal{F}) - TS_c(\mathcal{F}) \quad (2)$$

This says the Gibbs free energy of this particular framework equals the enthalpy of that framework minus a term arising on account of the conformational entropy. This conformational entropy is related to all-atomic configurations consistent with a fixed constraint topology having *limited* wiggle room.

The enthalpy of the framework is considered additive over individual enthalpy contributions, and it is given in Eq. 3,

$$H(\mathcal{F}) = \sum_i h_i N_i(\mathcal{F}), \quad (3)$$

where the sum is taken over different types of enthalpy contributions, h_i is the enthalpy contribution of the i th type, and N_i is the number of constraints of type i . Under a linear free energy decomposition scheme, the conformational entropy of the framework would be considered additive and simply expressed as

$$S_c(\mathcal{F}) = \sum_i s_i N_i(\mathcal{F}) \quad (\text{over-estimates the entropy}). \quad (4)$$

Equation 4 is similar to Eq. 3, except it sums over s_i , which is the entropy contribution of the i th type. Unfortunately, this approach overestimates the conformational entropy because it “double counts” configuration space, as it ignores correlated motions between atoms. Applying constraint theory, however, the *independent constraints* are identified, and a more accurate estimate of the conformational entropy is obtained as a sum over *only independent constraints*, given by

$$S_c(\mathcal{F}) = \sum_i s_i I_i(\mathcal{F}) \quad (5)$$

where $I_i(\mathcal{F})$ is the number of *independent* constraints of type i .

The *novel* aspect of the DCM is its ability to *account for non-additivity of entropy*. To illustrate this idea, consider a simple two-dimensional example shown in Figure 5.

In the reference state (5a), the system has one DOF. Adding one diagonal constraint (5b or 5c) gives both an enthalpy and entropy reduction, making the system rigid, with no remaining DOF. It doesn't make any difference which diagonal we put the constraint across, but adding a second diagonal (5d) is redundant, and it does not further decrease the entropy

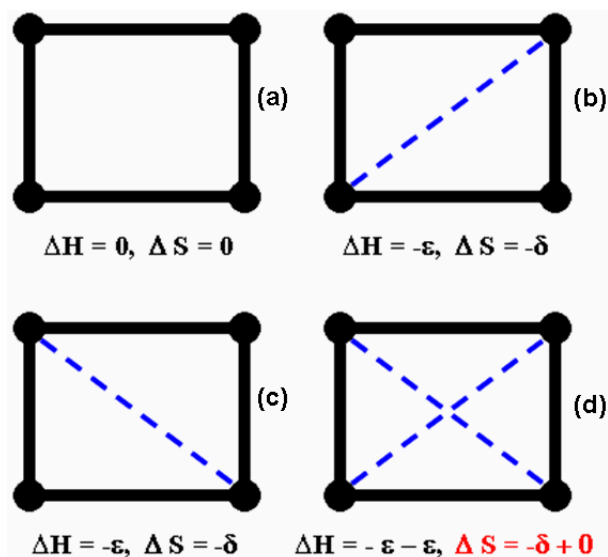


Figure 5. Non-Additivity of Entropy on a 2-D Graph

(to first order). Thus, adding the s_i contribution for the second diagonal is incorrect because the distance is already predefined by other interactions (i.e. modeled as constraints) within the network. This example shows how network rigidity is used explicitly to account for non-additivity effects. This approach is in sharp contrast to other models, where entropy components are simply considered additive. The importance of non-additivity has been pointed out by Ken Dill [7], who says:

“Perhaps some of our models in computational biology are based on flawed assumptions. Thermodynamic additivity principles are the foundation of chemistry, but few additivity principles have yet been found successful in biochemistry.”

It should be noted that, on the one hand, the total conformational entropy of a framework is expressed in Eq. 5 in terms of contributions from constraints (not DOF), since the interactions are properly represented as constraints. On the other hand, the energy functions for these interactions are described in terms of DOF (not constraints). For example, a two-body central force interaction between a pair of atoms will depend on the distance between these atoms, which is a DOF. Another example is the four body interactions involving the φ and ψ angles along the backbone conformation of a polypeptide chain. In all cases, the natural DOF used in describing a particular interaction are coarsened into regions of near constant energy, and the entropy contribution is related to the configuration space volume of the low energy basins. All internal DOF that describe an interaction are in some way limited to parts of configuration space during a coarse graining procedure that fixes the energy of a group of atoms to be within a certain tolerance. In general, the number of distance constraints used to represent an interaction is equal to the number of DOF it takes to express its energy function.

The role of a constraint is to fix the otherwise free DOF to be within a particular energy basin, which is characterized by its depth and width. The depth of the energy basin is reflected by the enthalpy or energy assigned to the constraint. Likewise, its width is reflected by the assigned *nominal* entropy. The nominal entropy is counted only when the constraint is found to be independent. For an accurate estimate of the conformational entropy for a given constraint topology (or mechanical framework), the DCM requires a complete set of constraint types to be defined. In this case, if one ignores entropy assignments, the entire framework will be rigid. However, the constraints are ranked ordered from lowest to highest by their nominal entropy, corresponding from the “*strongest*” to “*weakest*” constraints respectively. Conceptually, an extremely weak constraint could be thought of as a DOF, whereas an extremely strong constraint approaches a perfect distance constraint. Employing entropy assignments provides for a *quantitative* way to calculate all entropic contributions, large or small, using Eq. 5.

Since choosing how to label the constraints as dependent or independent is not unique, there is more than one correct sum for Eq. 5. This multivalued “function” is not immediately useful, since a unique solution is required to obtain a physical result. However, it is important to note that any way of carrying out that sum provides an *upper bound* estimate to the conformational entropy. This is because some constraints, even though they are independent, will restrict conformational entropy in a direction of configuration space that is already partially restricted by other constraints previously placed in the network. The lack of complete “orthogonality” thus leads to “double counting” some parts of configuration space. This residual double counting of configuration space among independent constraints is obviously much less than considering all constraints as independent when using Eq. 4. What we need then, is a rule (or procedure) to tell us how to label the constraints as dependent or independent so the sum is well defined. This procedure is established by simply *sorting* the constraints by their nominal entropy contributions from lowest to highest. We then follow this *preferential order* as constraints are placed one at a time to build a network as described above. Preferentially ordering the constraints provides a rigorous lowest upper bound estimate for the conformational entropy, while having no effect on Eq. 3. The lowest upper bound is a mathematically well-defined quantity that is unique for each constraint topology.

Within the DCM the lowest upper bound estimate for conformational entropy is considered to be the exact answer. However, this result is only an intermediate step. To study thermodynamics of a system, the partition function needs to be constructed over an ensemble of all accessible constraint topologies. To make these calculations tractable, further approximations are usually required that include formulating a simple free energy decomposition scheme. Although the DCM (and often the method used to solve it) is an approximation, numerous prior studies [18, 25, 33, 19, 34, 21] found the DCM to be in remarkably good agreement with experimental measurements. At the very least, the DCM performs overall no worse than any other coarse grained model.

3. Model: β -Hairpin to Coil Transition

We employ the DCM to investigate the β -hairpin to coil transition. We invoke Maxwell constraint counting to obtain exact results with minimal mathematical complication, while

retaining essential elements of the problem. Applying Maxwell constraint counting also demonstrates the versatility in available methodologies to solve the DCM. In this section, we address the remaining issues of specific model details, how to calculate the partition function, free energy landscapes and the associated thermodynamic response for the β -hairpin.

3.1. Minimal Distance Constraint Model (mDCM)

We present a minimal DCM (mDCM) that mirrors previous work on protein thermodynamics [33]. Our free energy decomposition scheme consists of four interaction types listed in Table 3.1., where it should be noted that (i) energy is used in place of enthalpy because no pressure dependence is being considered in this work, (ii) R is the ideal gas constant, (iii) NDC is the number of distance constraints used to model a particular interaction, and (iv) the three variables γ , δ_{nat} , and δ_{dis} are dimensionless “pure entropy” parameters. In addition to these interactions, the covalent bonded chain of amino acids that define the polypeptide is also accounted for, but no energy and entropy parameters are required because covalent bonds will not break and reform (fluctuate) at the temperatures of interest. Compared to proteins, we further simplify the problem by considering all intramolecular H-bonds as equivalent, and we neglect all sidechain interactions. Consequently, only mainchain-mainchain intramolecular H-bonds are considered within the β -hairpin.

Table 1. Free Energy Decomposition Scheme

Constraint Type	Energy	Entropy	NDC
Intramolecular H-bond	E	$R\gamma$	3
Native torsion	v	$R\delta_{\text{nat}}$	1
Disordered torsion	0	$R\delta_{\text{dis}}$	1
Solvent H-bond	u	N/A	0

The free energy decomposition scheme concerns itself only with molecular interactions of various types and is not tied to any particular structure. Different conformational states of the polypeptide will support different numbers of intramolecular H-bonds. Despite the name, a native-torsion constraint is not defined in reference to a “native” structure. Simply, a native-like torsion constraint tolerates only a narrow range of φ or ψ backbone dihedral angle variance within an amino acid residue that is energetically favorable, where $v < 0$. In contrast, a disordered torsion constraint represents a local conformational state poised to sweep a broad range of allowed φ or ψ backbone dihedral angles that are energetically unfavorable. All rotatable covalent bonds (via the φ and ψ dihedral angles) are coarse-grained into either native-like or disordered states, which are *similar* to the coarse-grained “helix” and “coil” states, respectively, employed in the Lifshitz-Roig model [32]. The difference is that the “native-like” state is not tied to any specific low energy structure. Rather, any *local* conformation that looks native-like is counted, regardless if the overall conformation has no resemblance to a native structure with lowest energy.

As listed in Table 3.1. there are 6 free parameters. In general, an energy and entropy parameter must be specified for each interaction type. With four interaction types, 8 parameters can be expected, however, two parameters can be quickly eliminated. Without loss of generality, a zero energy reference is assigned to the disordered torsion state. Next, the H-bond to solvent entropy is not applicable, because it is assumed² that H-bonding to solvent does not reduce conformational entropy more effectively than any of the other interactions.

In view of the interaction types utilized in the free energy decomposition, the macrostate of the polypeptide (or protein) is best characterized by the number of intramolecular H-bonds, which we denote N_{hb} , and number of native-like torsion angles, which we denote N_{nt} . In prior work on proteins, the total energy, U , was given as $U = U_{\text{ihb}} - uN_{\text{hb}} + vN_{\text{nt}}$, where the quantity U_{ihb} is the total intramolecular H-bond energy. The parameter u is an effective energy (a negative quantity) characterizing the average H-bond energy between a polypeptide to solvent. This term directly competes with the intramolecular energy, U_{ihb} , in such a way that for each intramolecular H-bond that forms, there is a loss of a H-bond to solvent, and vice versa. In proteins, U_{ihb} was modeled to be dependent on atomic-structure, but here, $U_{\text{ihb}} = EN_{\text{hb}}$, where E is the energy of a single H-bond. With this simplification, the total energy is given as

$$U(N_{\text{hb}}, N_{\text{nt}}) = \epsilon N_{\text{hb}} + vN_{\text{nt}} \quad (6)$$

where the new energy parameter, ϵ , defined as $\epsilon = E - u$, represents the difference in energy between intramolecular H-bonds to that of H-bonds to solvent. The remaining 5 parameters, $\{\epsilon, v, \gamma, \delta_{\text{nat}}, \delta_{\text{dis}}\}$ will be left open.

If we assume an additive decomposition scheme in entropy components using Eq. (4), the estimated conformational entropy for a polypeptide of N amino acid residues is simply

$$S_c = R[3\gamma N_{\text{hb}} + \delta_{\text{nat}} N_{\text{nt}} + \delta_{\text{dis}}(2N - N_{\text{nt}})] \quad (\text{over-estimates entropy}). \quad (7)$$

Instead, the non-additivity of conformational entropy will be reflected through the global criteria imposed by Maxwell constraint counting. The estimate for conformational entropy using Eq. (5) with Maxwell constraint counting depends on the preferential rank ordering of the entropy parameters. Given that $\delta_{\text{dis}} > \delta_{\text{nat}}$, this leaves only three possible preferential entropy rank orderings, where the total conformational entropy is given as

$$S_c = \begin{cases} S_{c1} & \text{if } \gamma < \delta_{\text{nat}} < \delta_{\text{dis}} \\ S_{c2} & \text{if } \delta_{\text{nat}} < \gamma < \delta_{\text{dis}} \\ S_{c3} & \text{if } \delta_{\text{nat}} < \delta_{\text{dis}} < \gamma \end{cases}, \quad (8)$$

and S_{c1} , S_{c2} , and S_{c3} are given by

$$\frac{S_{c1}}{R} = 3\gamma N_{\text{hb}} + \delta_{\text{nat}} \min(N_{\text{nt}}, 2N - 3N_{\text{hb}}) + \delta_{\text{dis}} \max(2N - N_{\text{nt}} - 3N_{\text{hb}}, 0) \quad (9)$$

$$\frac{S_{c2}}{R} = 3\delta_{\text{nat}} N_{\text{nt}} + \gamma \min(3N_{\text{hb}}, 2N - N_{\text{nt}}) + \delta_{\text{dis}} \max(2N - N_{\text{nt}} - 3N_{\text{hb}}, 0) \quad (10)$$

$$\frac{S_{c3}}{R} = \delta_{\text{nat}} N_{\text{nt}} + \delta_{\text{dis}}(2N - N_{\text{nt}}). \quad (11)$$

²This assumption can be easily lifted, and in previous work [25] we included affects of solvent in the form of hydration and local clathrate structures to model cold denaturation.

For all possible cases of preferential entropy rank orderings, S_c can be expressed as a function of the macrostate (N_{hb}, N_{nt}) , because Maxwell constraint counting is a *global* counting procedure. Specifically, the above equations for S_c were derived from global counting considerations to be discussed next.

We define the state of the network before any of the constraint types that are defined by the free energy decomposition scheme are placed. From this, it is easy to see that the total number of *independent* constraints from H-bonding, native and disordered torsions must sum up to $2N$ DOF. This result is understood by recalling that sidechain dihedral angles are not considered (we are working only with the backbone, and are neglecting sidechain interactions), and the covalent bonding forms a template structure. This template structure is just the polypeptide backbone that has 2 DOF (φ and ψ angles) per residue (N of them). A total of $2N + 3N_{hb}$ constraints are then placed in the network, since each dihedral angle is labeled as a native-like or disordered torsion constraint (always totaling to $2N$), and there are three distance constraints for each intramolecular H-bond. Whenever intramolecular H-bonds are present (i.e. $N_{hb} > 0$) there will be more constraints present in the network than $2N$ DOF, indicating there will be $3N_{hb}$ redundant constraints. In all cases, there are always enough constraints to *rigidify* the template structure, verifying the free energy decomposition is complete. Maxwell constraint counting, combined with the preferential entropy rank ordering, throws away $3N_{hb}$ constraints having the largest entropy contributions. The difference between the three cases given above is caused by the *relative strengths* of the interactions present in reducing conformational flexibility, which is quantified by the nominal entropy assignment.

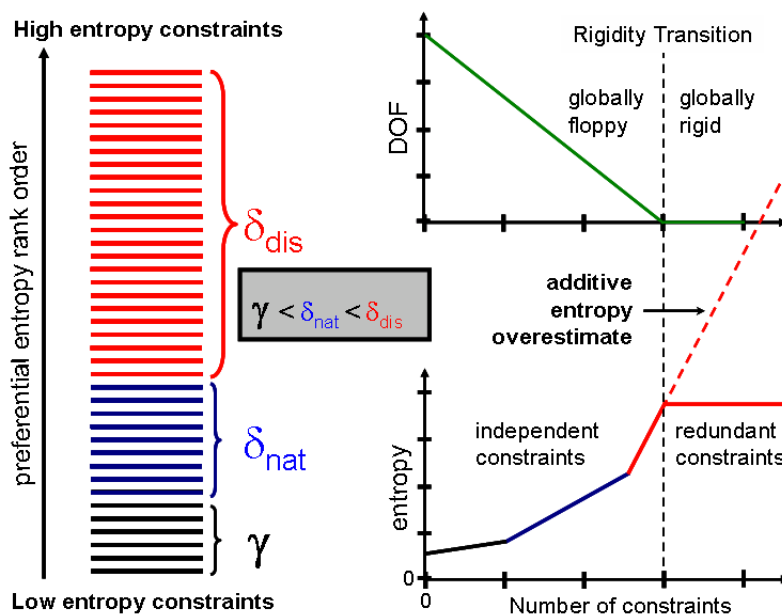


Figure 6. Maxwell Counting in the mDCM - Preferential Ordering.

It is worth mentioning that the alternative exact calculation of network rigidity would not allow S_c to be expressed as a function of N_{hb} and N_{nt} , because the exact answer would

depend on the specific way constraints are *distributed*. It is precisely at this point where employing Maxwell constraint counting dramatically simplifies the mathematics. Application of Maxwell constraint counting on the list of preferentially ordered constraints is effectively defining a dispersion relation that is analogous with the Debye theory for heat capacity in solids. As schematically shown in Fig. 3.1. there is, *a priori*, a fixed number of independent constraints that will contribute to S_c . In the Debye model the number of vibrational modes (each an independent contribution) sum to the known number of DOF in the solid.

3.2. Partition Function

Within the distance constraint model, the partition function is given by

$$Z_{\text{DCM}} = \sum_{\{\mathcal{F}\}} e^{-\beta G(\mathcal{F})} = \sum_{\{\mathcal{F}\}} e^{\frac{S_c(\mathcal{F})}{R}} e^{-\beta H(\mathcal{F})} \quad (12)$$

where respectively $G(\mathcal{F})$, $H(\mathcal{F})$ and $S_c(\mathcal{F})$ are given by Eqs. (2, 3, 5), and the summation is over an ensemble of all accessible constraint topologies $\{\mathcal{F}\}$. Central to the DCM, the non-trivial factor $e^{\frac{S_c(\mathcal{F})}{R}}$ appearing in Eq. 12 is the *geometrical degeneracy*. For a *fixed constraint topology* \mathcal{F} , the energy is approximately constant, and the geometrical degeneracy factor accounts for all accessible atomic motions consistent with the given (fixed) constraint topology. In many applications the number of distinct constraint topologies is astronomical in number. Nevertheless, it is insightful to consider the two most extreme constraint topologies, as shown in Fig. 3.2. for a polypeptide that undergoes a β -hairpin to coil transition.

Schellman's original two-state analysis [42] using two extreme states for the helix to coil transition, and then for proteins is also applicable here, as Fig. 3.2. shows schematically. The lowest energy state corresponds to the maximum number of H-bonds forming a ladder, with all torsion interactions being in the native state. The maximum entropy state is when there are no H-bonds present, and all the torsion interactions are disordered, yielding a random coil. Note that a single constraint topology (all disordered torsion constraints and no H-bond constraints) represents all possible conformations the polypeptide can take, which is quantified through the geometrical degeneracy factor. From these two extremes, one can estimate ΔH and ΔS for the transition between them, from which one can estimate the melting temperature as $T_m = \frac{\Delta H}{\Delta S}$, assuming $\Delta G = 0$ at the transition point. Of course, considering only two states is a very crude approximation. The exact solution is to include all possible constraint topologies.

3.2.1. Class of β -Hairpin Constraint Topologies

An illustration of a β -hairpin structure is shown in Fig. 8.

A mechanical framework for the β -hairpin consists of N residues each having a φ and ψ angle, and possibly crosslinking H-bonds. The three dimensional structure (or mechanical framework) is then represented by a graph. The graph characterizes different types of possible constraint topologies, where Fig. 9 shows four examples. In all four cases, different H-bond crosslinking is given, but the φ and ψ backbone dihedral angles are assumed to be native-like.

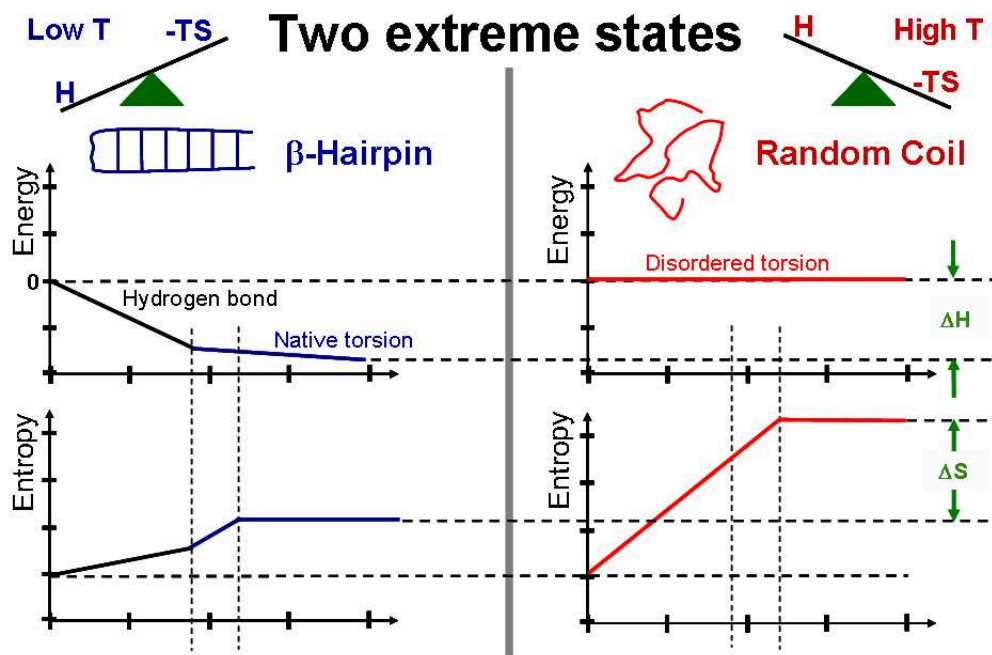


Figure 7. Schematic: The see-saws at top to the (left, right) illustrate tradeoff between energy and entropy. At (low, high) temperatures the β -hairpin will be in a (low energy, high entropy) state. The other four figures show a graphical representation of how the energy and entropy for the two extreme states are calculated.

In the context of the mDCM, there are 2^{2N} distinct constraint topologies for each H-bond placement (c.f. Fig. 9) because each residue has a φ and ψ angle, which are coarse grained into two possible states. Each constraint topology will have an energy and conformational entropy that are affected by the number of native-torsion constraints, as well as the number of H-bond constraints. The four examples of Fig. 9 give an indication of the type of diversity the native only and non-native contact ensembles have.

Referring back to the expressions for the total energy and conformational entropy (See Eqs. 6, 9, 10, 11) it is noticed that all of them are strictly functions of the macrostate (N_{hb} , N_{nt}), and hence they are independent of the particular way the constraints are dis-

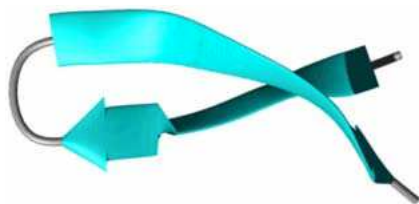


Figure 8. The β -Hairpin Structural Motif

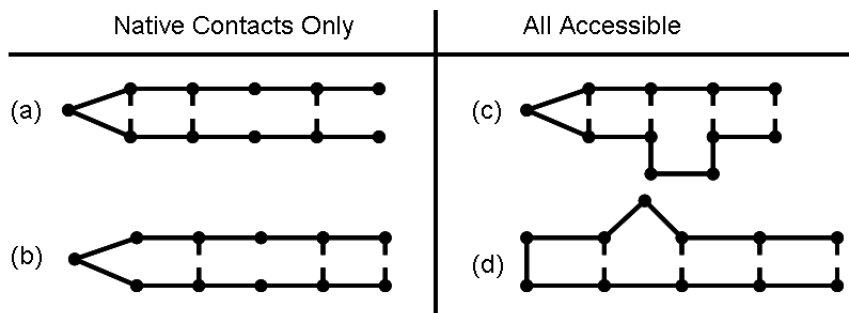


Figure 9. Four example constraint topologies for the β -hairpin with $N = 10$ residues are shown. In these graphs, each edge represents a residue having φ and ψ backbone dihedral angles, and they are all assumed to be in their native-state. The nodes of the graph (dots) represent the peptide bond part of the polypeptide chain that allows crosslinking H-bonds to form. Panels (a) and (b) show two examples of a “ladder” configuration that is part of the native contact only ensemble with two residues at the turn. Both of these graphs have 3 H-bonds and share the same macrostate. Panels (c) and (d) show two examples of configurations that have non-native crosslinking H-bonds, which are out of register (described in more detail below). Both of these graphs have 4 H-bonds and share the same macrostate. All four graphs are represented in our all-contact ensemble, which includes native and non-native contacts.

tributed. In general, there are many different constraint topologies consistent with a given macrostate, and this defines a topological degeneracy, denoted $\Omega(N_{\text{hb}}, N_{\text{nt}})$. In principle, the ensemble of constraint topologies being summed over in Eq. (12) can be arbitrarily diverse. However, we are only considering a certain class of constraint topologies that are representative of the β -hairpin. Constraint topologies that are included in the ensemble range from being in the lowest possible energy state to the high energy random coil state, as well as an entire spectrum of conformational states that have β -hairpin character somewhere in between these two extremes. Excluded constraint topologies are those that are representative of structural motifs other than a β -hairpin, such as an α -helix for example.

3.2.2. Method of Calculation

The partition function given by Eq. (12) can be re-expressed as a summation over all accessible macrostates by grouping all topologically degenerate terms with the same statistical weight (i.e. $e^{-\beta G(N_{\text{hb}}, N_{\text{nt}})}$). This procedure is summarized in Fig. 10 that highlights individual steps of the calculation to the very end where thermodynamic response is calculated. The partition function for a N -residue polypeptide capable of undergoing a β -hairpin to coil transition is given by:

$$Z(N) = \sum_{N_{\text{nt}}=0}^{2N} \sum_{N_{\text{hb}}=0}^{M_{\text{hb}}} \Omega(N, N_{\text{hb}}, N_{\text{nt}}) e^{-\beta G(N, N_{\text{hb}}, N_{\text{nt}})} \quad (13)$$

where N has been explicitly written as a reminder that the functions depend on the length of the chain. In regards to the limits on the two summations: Note $2N$ is the maximum number of native-torsion constraints that can be present, and $M_{\text{hb}} = \lfloor N/2 \rfloor$ is the maximum number of H-bond constraints that can be present (within the class of ensembles considered here), which occurs when all the H-bonds are in register. The function $\lfloor N/2 \rfloor$ rounds down to the nearest integer (i.e. the floor function) when the chain consists of an odd number of residues.

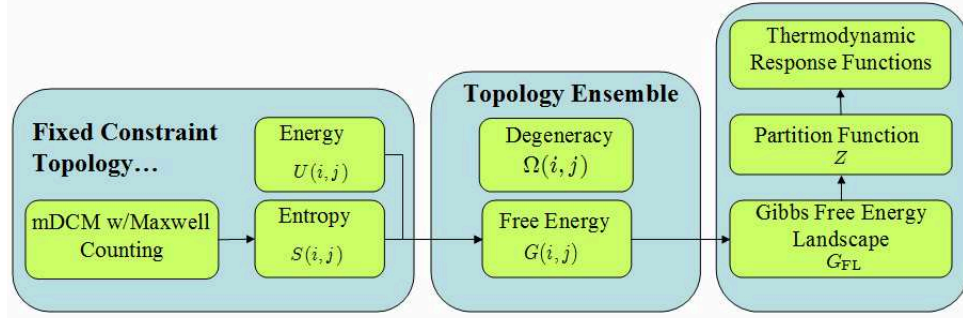


Figure 10. Flowchart of Calculations

At this point, only the functional form of the topological degeneracy factor remains to be determined. The derivation of the topological degeneracy factor $\Omega(N_{\text{hb}}, N_{\text{nt}})$ is given below in Section 3.4. for two different types of ensembles of constraint topologies. In the first case, only native β -hairpin contacts (and all possible ways they can be broken) are considered accessible, while in the second case both native and a restricted type of non-native contacts are allowed. For the moment, we leave the precise functional form for the topological degeneracy factor unspecified, and we focus on the relationships between the full partition function obtained as a sum over macrostates and the free energy landscape in “constraint space.”

3.3. Free Energy Landscape in Constraint Space

The partition function can also be expressed in terms of a free energy landscape as a summation over all accessible macrostates, given by

$$Z(N) = \sum_{N_{\text{nt}}=0}^{2N} \sum_{N_{\text{hb}}=0}^{M_{\text{hb}}} e^{-\beta G_{\text{FL}}(N, N_{\text{hb}}, N_{\text{nt}})} \quad . \quad (14)$$

By matching Eq. (14) to Eq. (13) it follows that the function G_{FL} describing the free energy landscape is defined as

$$e^{-\beta G_{\text{FL}}(N, N_{\text{hb}}, N_{\text{nt}})} = \Omega(N, N_{\text{hb}}, N_{\text{nt}}) e^{-\beta G(N, N_{\text{hb}}, N_{\text{nt}})} \quad . \quad (15)$$

Notice that the free energy landscape, G_{FL} , includes the topological degeneracy factor. For conceptual clarity, G_{FL} is expressed in terms of the free energy of a particular constraint

topology $G(N, N_{\text{hb}}, N_{\text{nt}})$, and a topological entropy term. This relationship is given in Eq. 16

$$G_{\text{FL}}(N, N_{\text{hb}}, N_{\text{nt}}) = G(N, N_{\text{hb}}, N_{\text{nt}}) - \frac{\ln \Omega(N, N_{\text{hb}}, N_{\text{nt}})}{\beta} \quad (16)$$

where a *mixing entropy* is defined in Eq. (17).

$$S_{\text{mix}}(N, N_{\text{hb}}, N_{\text{nt}}) = k_B \ln \Omega(N, N_{\text{hb}}, N_{\text{nt}}). \quad (17)$$

The mixing (or topological) entropy is related to the number of accessible constraint topologies having the same macrostate.

The folded state (i.e. a β -hairpin structure) and unfolded state are identified as local low free energy basins. The unfolded state corresponds to relatively few crosslinking H-bonds and a large number of disordered torsion constraints. The folded state has many crosslinking H-bonds, and a large number of native-like torsion constraints. These folded and unfolded basins are illustrated in Figure 11, labeled as **F** and **U** respectively. The numbers of H-bonds and native-torsion constraints are nothing more than order parameters that define the macrostate. However, in the context of the DCM, the free energy landscape is embedded in constraint space. As the number of *independent* disordered torsion constraints increase, the overall flexibility of the structure increases.

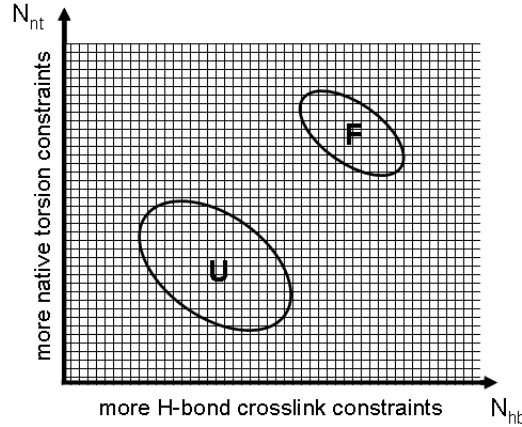


Figure 11. Macrostates of the mDCM

The expression for the free energy landscape given in Eq. (16) has the same form as the phenomenological Gibbs free energy given by Eq. (18) in terms of five parameters, u , v , γ , δ_{nat} and δ_{dis} that reflect the free energy decomposition scheme used to describe protein thermodynamics [33].

$$G_{\text{FL}}(N_{\text{hb}}, N_{\text{nt}}) = U_{\text{ihb}} - uN_{\text{hb}} + vN_{\text{nt}} - T(S_c(\gamma, \delta_{\text{nat}}, \delta_{\text{dis}}) + S_{\text{mix}}). \quad (18)$$

We briefly comment on some aspects of the protein DCM that connect to this work.

For proteins, the quantity U_{ihb} is the total intramolecular H-bond energy based on a H-bond energy function that depends on atomic positions. Starting from the native protein structure as determined (usually) from X-ray crystallography, the first step was to define the

native H-bonds. Then the ensemble of constraint topologies were generated by considering permutations of removing various H-bonds. For each of these H-bond configurations, the torsion-constraints were allowed to be either native-like or disordered as done here. For proteins, a Monte Carlo method was used to generate a representative set of degenerate constraint topologies with the same macrostate corresponding to a single node in constraint space as shown in Fig. 11. Thus each node represents a sub-ensemble of constraint topologies. Exact network rigidity calculations were performed within each node for each constraint topology in the sub-ensemble, and *statistical average* properties for each node were determined. Thus in the protein case, Eq. 18 is a mean field expression.

An interesting comparison is that the *exact expression* for the free energy landscape given by Eq. (16) under the mean field approximation using Maxwell constraint counting is of the same form as the mean field treatment of solving the DCM for proteins. In contrast to the Monte Carlo method used for proteins, in the β -hairpin problem at hand, every term in the expression for the free energy landscape is calculated exactly, which means there are no sampling errors.

The thermodynamic response of the β -hairpin will readily follow from the partition function in terms of G_{FL} . The mDCM computes G_{FL} on a discretized mesh of macrostates (c.f. Fig. 11 for a fixed temperature, repeated over different temperatures. The heat capacity is predicted through the relation $C_V = (1/k_B T^2)(\langle E^2 \rangle - \langle E \rangle^2)$.

3.4. Ensemble of Accessible Constraint Topologies

In this section we derive the topological degeneracy for the β -hairpin under two different types of ensembles, from which S_{mix} will be determined exactly. The first type of ensemble will be restricted to constraint topologies that have *native-contacts only*. The second ensemble will include non-native contacts that are restricted so as not to allow for (unphysical) crossing of H-bonds and will not allow for structural motifs other than structures resembling a β -hairpin. In other words, the ensemble for which non-native contacts are allowed will represent possible misfolds of the β -hairpin. Comparing these two cases will allow the effects that arise from using more restricted native-contact ensembles to be identified.

3.4.1. Native Contacts Only

For native contacts only, we make the following restrictions on the constraint topology: (a) The β -hairpin is “centered.” To explain this, refer back to Fig. 9. By centered, we mean the β -turn is located so that the number of backbone residues on the top side of the β -hairpin equals the number of residues on the bottom side to within ± 1 . (b) The crosslinking H-bonds must join two residues directly across from one another in register (i.e. they cannot join diagonally between residues).

These two restrictions comprise the native contacts only constraint topology, which does not account for the possibility of misfolding, such as when a residue or two are kinked upward with others sliding in to take their place. This approximation is just like that done with proteins, where all permutations of breaking these native contacts are considered by removing the crosslinking H-bonds. Nevertheless, we are considering many more constraint topologies than that of the classic *zipper* model [28]. In the zipper model, the native structure is thought of as a ladder, where rungs on the ladder must be removed starting from the

furthest rung from the β -turn. The rungs must be removed one at a time moving toward the β -turn consecutively. In contrast, all possible defects of missing a native-contact is considered within our ensemble.

Since the residues must line up in register for a H-bond to form, we are left only with the question of how to place the H-bonds in well defined slots. Recall that we have N residues in total, and suppose that we have N_{hb} crosslinking H-bonds. For N_s slots, the number of ways of placing N_{hb} H-bonds is simply

$$\binom{N_s}{N_{\text{hb}}} = \frac{N_s!}{N_{\text{hb}}!(N_s - N_{\text{hb}})!}. \quad (19)$$

But how many slots N_s are there? If there are N_c residues at the β -turn, then there are $N - N_c$ remaining residues along each of the two parallel sides. According to restrictions (a) and (b) above, and from Fig. 9, it can be seen that N_s is simply given by

$$N_s = \left\lfloor \frac{N - N_c}{2} \right\rfloor, \quad (20)$$

where $\lfloor \cdot \rfloor$ is the least integer (or floor) function. Therefore, our expression for the number of ways of placing the crosslinking H-bonds becomes

$$\frac{\lfloor (N - N_c)/2 \rfloor!}{N_{\text{hb}}!(\lfloor (N - N_c)/2 \rfloor - N_{\text{hb}})!} \quad (21)$$

In this work, we place two residues at the β -turn (i.e., $N_c = 2$).

Next, recall that each backbone residue corresponds to two torsion constraints, one for each of the angles in the $\{\varphi, \psi\}$ pair. Thus there are $2N$ torsion constraints, and each of these torsion constraints must be labeled as native-like or disordered. Since there are two options, the number of ways of choosing N_{nt} of them to be native-like is simply

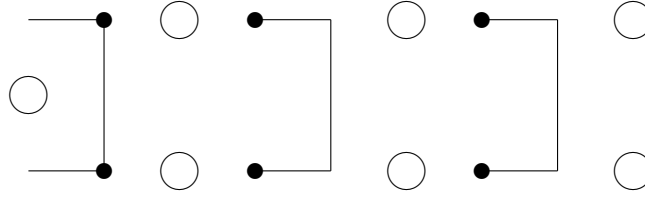
$$\binom{2N}{N_{\text{nt}}} = \frac{(2N)!}{N_{\text{nt}}!(2N - N_{\text{nt}})!} \quad (22)$$

The labeling of torsion constraints as native-like or disordered is independent of the placement of the H-bonds, so their multiplicities multiply. Combining Eq. (22) with Eq. (21), then, for N residues, N_{nt} native-like torsion constraints, N_{hb} crosslinking H-bonds, and with $N_c = 2$, the multiplicity is

$$\Omega(N, N_{\text{hb}}, N_{\text{nt}}) = \frac{\lfloor (N - 2)/2 \rfloor!}{N_{\text{hb}}!(\lfloor (N - 2)/2 \rfloor - N_{\text{hb}})!} \cdot \frac{(2N)!}{N_{\text{nt}}!(2N - N_{\text{nt}})!}. \quad (23)$$

3.4.2. Non-native Contacts Allowed

We now consider the case where non-native contacts are allowed, but under the following restrictions: (a) No (unphysical) crossing of H-bonds are allowed, and (b) we only consider H-bonds forming that connect two opposite sides of the hairpin structure. Thus, the H-bonds can still be considered as rungs on a ladder, but now the entire concept of all register positions is lost, allowing for misfolded structures. To understand the counting for this case, consider Figure 12.

Figure 12. Binning Diagram for Multiplicity Ω

If there are N_{hb} crosslinking H-bonds, we divide the β -hairpin into bins. The bins will be filled up with a certain number of residues that fall between pairs of residues connected by a crosslinking H-bond. By first focusing on connected pairs of residues, it is found that there will be $2N_{\text{hb}} + 1$ bins. Essentially there are two bins per crosslinking H-bond, and the $+1$ comes from the β -turn at the left side of Figure 12). Since each H-bond has two residues attached to it (one at top and bottom), there are $N - 2N_{\text{hb}}$ remaining residues to place in the $2N_{\text{hb}} + 1$ bins. The number of ways to place the remaining residues into the bins is given by the multichoose function in combinatorics:

$$\left(\binom{2N_{\text{hb}} + 1}{N - 2N_{\text{hb}}} \right) = \binom{2N_{\text{hb}} + 1 + (N - 2N_{\text{hb}}) - 1}{N - 2N_{\text{hb}}} = \binom{N}{N - 2N_{\text{hb}}} \quad (24)$$

The labeling of the backbone residues as native-like or disordered is the same as above given in Eq. 22. Again, these choices are independent, so Eq.(24) and Eq. (22) multiply, giving

$$\Omega(N, N_{\text{hb}}, N_{\text{nt}}) = \frac{N!}{(2N_{\text{hb}})!(N - 2N_{\text{hb}})!} \cdot \frac{(2N)!}{N_{\text{nt}}!(2N - N_{\text{nt}})!} \quad (25)$$

4. Results and Discussions

In this work, we selected a reasonable set of model parameters based on prior work with proteins. For the results presented here, the specific values we used were $\epsilon = -5$ kcal/mol, $v = -0.7$ kcal/mol, $\gamma = 2$, $\delta_{\text{nat}} = 3$ and $\delta_{\text{dis}} = 5$. Note that value of 5.092 for δ_{dis} has been treated as a transferable parameter across a diverse set of proteins in the work involving proteins with marked success. The other values for the entropy parameters are in line with typical values used in proteins. Note that the disordered torsion constraint has the most entropy, and H-bonds have the least.

For this set of entropy parameters, we note that the preferential rank ordering considers H-bonds as more effective in reducing conformational flexibility compared to the native-torsion constraint. Therefore, to calculate S_c , Eq. (9) is used. We did not try to optimize these five free parameters, nor did we try to predict experimental data or compare to MD simulation data. We can, however, with considerable confidence, obtain interesting qualitative generic properties of the β -hairpin for which we explore. Our selected set of parameters provide physically reasonable results for both types of ensembles considered, and this is an indication that the parameterization used in the mDCM has physical meaning that is somewhat transferable — as is intended when employing a free energy decomposition scheme.

4.1. Free Energy Landscapes

We present free energy landscapes, heat capacity plots, and phase diagrams using the mDCM for two ensembles of constraint topologies. We compare and contrast the results for the *native only constraint topology ensemble* (N-CTE) and the *all accessible constraint topology ensemble* (A-CTE) that includes a restricted set of non-native contacts in addition to native contacts (see Fig. 9).

Figure 13 illustrates four snapshots of the free energy landscape with $N=32$ at various temperatures; $T = 250, 310, 350,$ and 449 K for the A-CTE case using constraint theory. The Gibbs free energy landscape G_{FL} is plotted over the 2-D constraint space determined by the number of native-like torsion constraints N_{nt} and the number of H-bonds N_{hb} . The region near the (0,0) corner in constraint space indicates the unfolded coil state of the hairpin. The opposite diagonal corner of constraint space corresponds to the folded state, as expected. At lower temperatures, the free energy is a minimum in the region with high N_{nt} and N_{hb} , indicating a folded state. As the temperature increases, the free energy minimum shifts toward a state with small N_{nt} and N_{hb} , indicating an unfolded state.

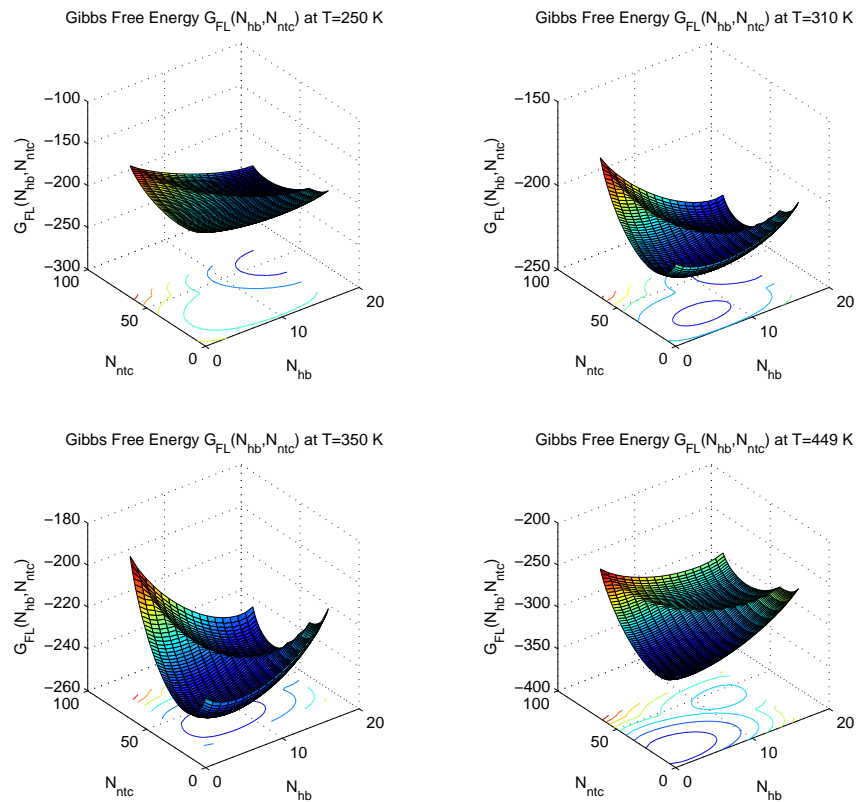


Figure 13. Free Energy Landscapes with $N = 32$ at $T = 250, 310, 350,$ and 449 K for the A-CTE case

Figure 14 shows the same information as Fig. 13, but as macrostate probability contour maps. The contour maps show the corresponding probability of being found in a given macrostate based on the Gibbs free energy landscape plots shown in Fig. 13. As the temperature increases from 250 K to 449 K, the most-probable macrostate of the β -hairpin clearly shifts from a folded macrostate to an unfolded macrostate.

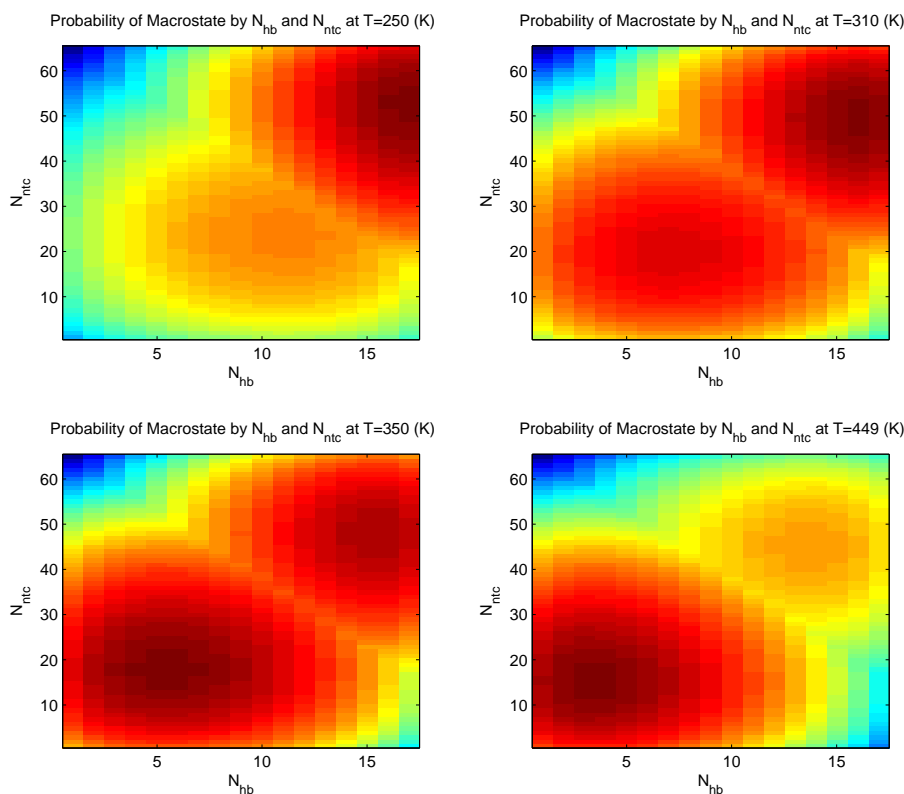


Figure 14. Macrostate Probability Contour Maps with $N = 32$ for $T = 250, 310, 350,$ and 449 K for A-CTE case. A qualitative color-scale is used, where red is high probability, yellow is moderate level, and blue is low probability.

Repeating the same calculations, the free energy landscapes are shown in Fig. 15 for the N-CTE case. Comparing these results to those shown in Fig. 13, we find the free energy landscapes are qualitatively the same. The A-CTE and N-CTE cases both indicate a transition around 340-350 K. However, more curvature is detected for the ensemble that includes non-native contacts (A-CTE). The additional curvature in A-CTE provides a wider basin in contrast to a more nearly constant slope toward a corner that appears for N-CTE. These differences in geometric form of the free energy landscape is an indication that there will be competitive missfolded structures present in the low free energy basin for the A-CTE.

Figure 16 shows the calculated heat capacity curves for a series of chains having either

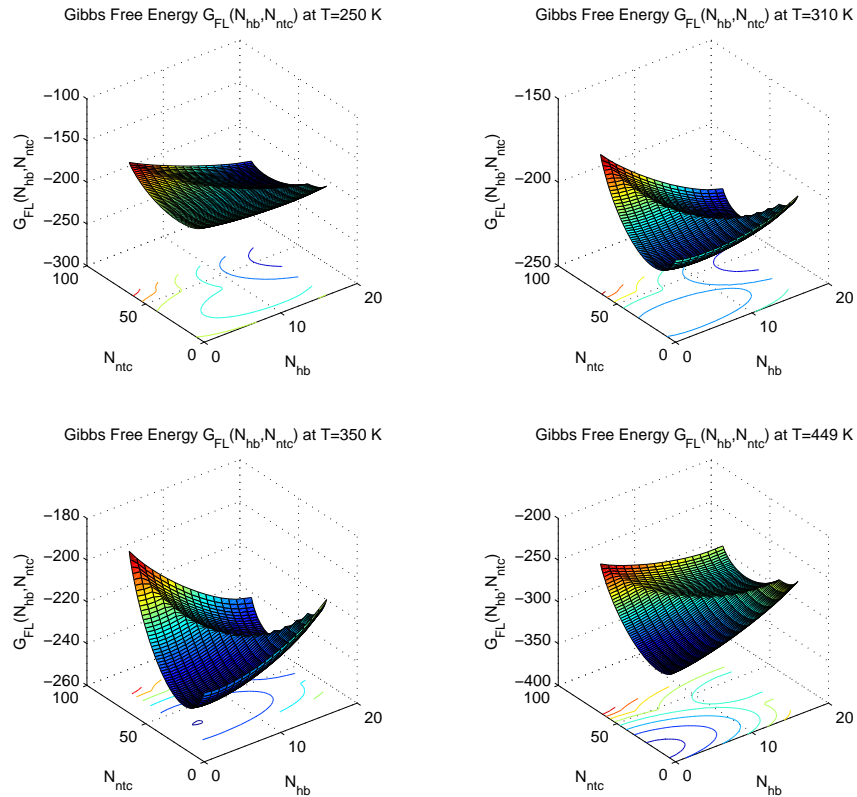


Figure 15. Free Energy Landscapes with $N = 32$ at $T = 250, 310, 350,$ and 449 K for the native contacts only constraint topology

even or odd number of N residues, over the temperature range $250 \leq T \leq 450$ K, for the A-CTE case. The transition temperature T_m is clearly identified at the peak of each curve, with a sharpening of the peak as N increases. It should be noted that these plots are *not* normalized by the number of residues. In a non-cooperative system, the peak in the heat capacity at the melting temperature should be directly proportional to the size of the system (i.e. an extensive variable). We find that there is no super-linear or sub-linear variation, suggesting that the transition is not cooperative. When normalized by the number of residues, the peak heights are essentially the same, but the plot is visually obfuscated. The key feature is the narrowing of the peak width as the number of residues increases. This result indicates the presence of a sharper transition as the chain length increases, and therefore, the system does exhibit *weak* cooperativity.

The heat capacity plots for the native contacts only constraint topology ensemble are shown in Fig. 17. The peak heights divided by the number of residues is also nearly a constant for different length chains. Differences in the heat capacity plots are observed between the native contacts only and the non-native contacts ensembles. It is clear that

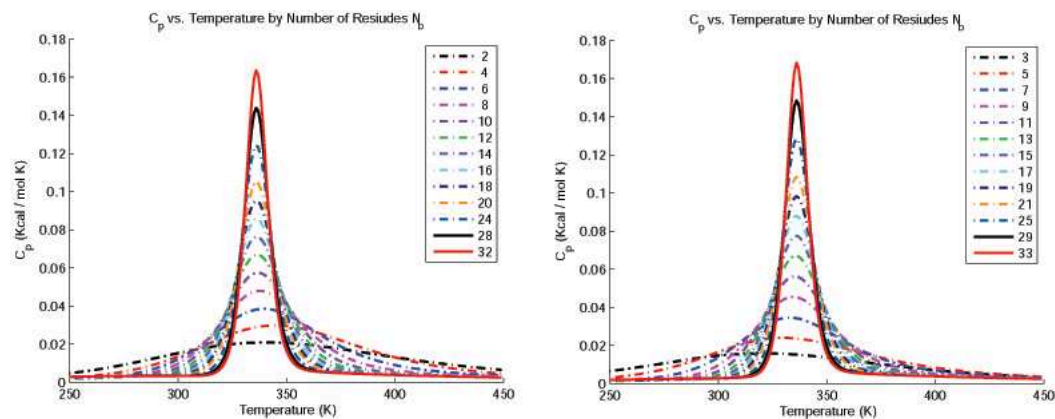


Figure 16. Heat Capacity Calculation by Even and Odd N : A-CTE case.

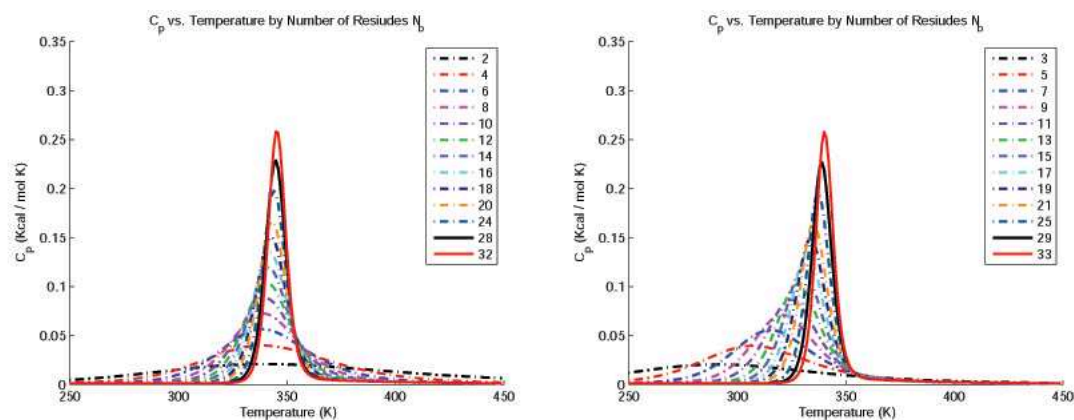


Figure 17. Heat Capacity Calculation by Even and Odd N : N-CTE.

the width of the heat capacity curves at half-maximum for N-CTE are wider than their A-CTE counterparts. Moreover, the peak position that defines T_m shifts more as a function of chain length, which indicates stronger cooperativity. We suggest this result is somewhat expected, because non-native contacts broaden the so called “native” free energy basin, having many misfolded structures sharing nearly the same minimal free energy values. We suspect that in general, native contacts only ensembles will produce sharper transitions compared to more complete ensembles. Upon further comparison and examination of Figs. 16 and 17, it is clear that there are differences between odd and even number residue chains. Moreover, the differences are more pronounced for shorter chains, where it is found that the melting temperature is lower for odd number of residues, compared to an even number of residues. This makes intuitive sense, as the odd number of residues forces one H-bond to be unsatisfied.

The dependence of T_m on chain length N is established more clearly by constructing a phase diagram. In Fig. 18, the results for T_m vs. N are shown for both the A-CTE and

N-CTE cases.

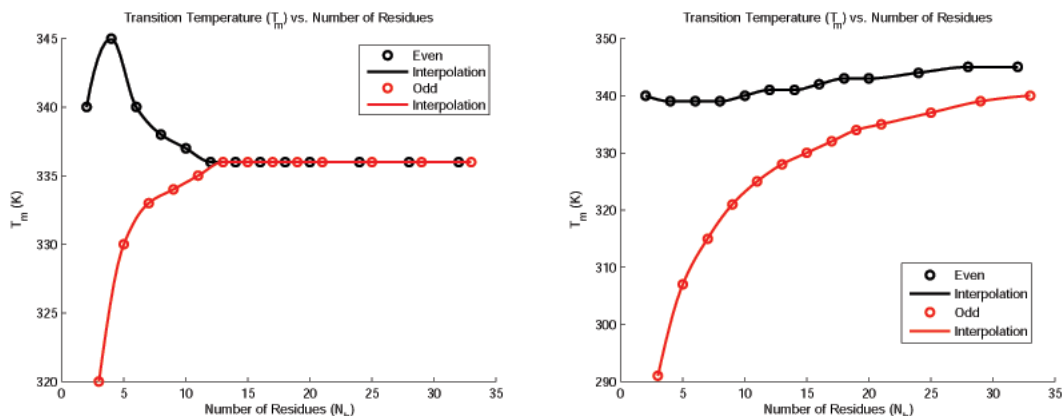


Figure 18. Transition Temperature T_m by Number of Residues N . A-CTE case at left, and N-CTE at right.

To make the effect of even and odd residues more apparent, the number of residues are split into even and odd categories. Interestingly, this even vs. odd dichotomy goes away for long chains for the A-CTE case, and moreover, the odd and even results converge to the same large N limit for T_m lines as seen in Fig. 18. In contrast, this effect remains for long chains for the N-CTE case, where the even/odd T_m lines seem to separately converge onto limiting asymptotic values for long chains. For the case of N-CTE, this separation of limiting values is physically explained by the unpaired residue that is not crosslinked by a H-bond for the odd cases. Mathematically, this is explained easily by recalling Eq. 23. We see that if the odd (leftover) residue is truncated off, *only the degeneracy factor will be modified*, while the remainder of the calculation is not effected. The convergence to the *single* limiting value of T_m is natural for the A-CTE because the number of residues making up the turn is no longer fixed at two (i.e. not centered), and the loss of one H-bond is more easily compensated.

4.1.1. Additivity of Entropy

We repeated the calculations again, but now with the rigidity calculations “turned off.” This means additivity of entropy was used (i.e. replacing the use of Eq. (9) to calculate S_c with Eq. (7)), which greatly over-estimates the conformational entropy when there is a dense number of constraints present in the network. The results with additivity of entropy were uninteresting for both types of ensembles. As an example, the free energy landscape is more-or-less a plane (see Fig 20), with no double well behavior. In fact, the minimum at the edge of constraint space is in the unfolded state. We now consider in more detail the heat capacity for the A-CTE case (see Fig 19) and simply note the same characteristics were also found for the N-CTE case.

We find an additive free energy decomposition scheme does not produce a transition and thus no structural transformation. Since we are considering generic properties of the

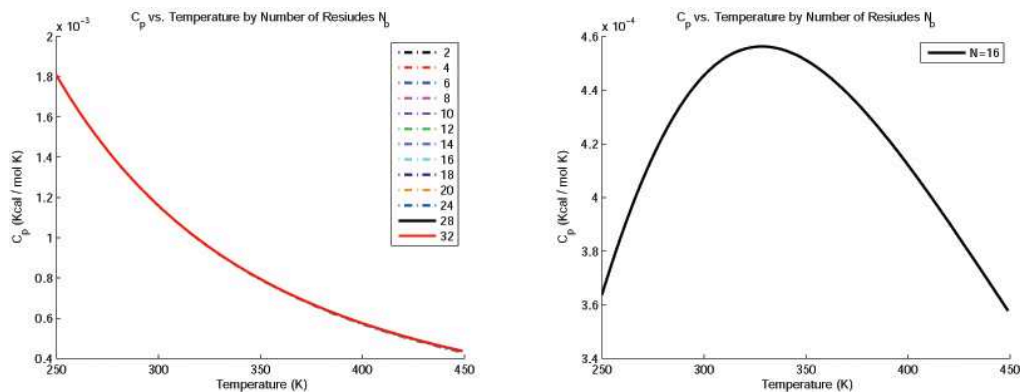


Figure 19. Heat Capacity Predictions using an Additive Free Energy Decomposition. (Left panel) *Without* network rigidity, no transition is predicted using additivity of entropy for the same parameters used for the mDCM. (Right panel) The single parameter v was lowered to -2.2 Kcal/mol in order to produce a double minimum in the free energy landscape to obtain two separate states. However, creating a native state basin was not possible, although a peak is formed. Because the scale of values for C_p are insignificantly low, no transition is observed.

β -hairpin to coil transition using a simple model (and knowing from experiments that a transition exist), our expectation is that we should be able to find a peak in the heat capacity for some modified parameters. However, we could not find a transition using an additive formula for the conformational entropy for the same set of parameters, nor by adjusting the parameter v (holding all others fixed) — nor uniformly changing the entropy parameters by a uniform multiplicative scaling factor.

The null result of no transition with network rigidity ignored is actually not too surprising once it is realized that the additive model predicts that as more H-bonds are added to a network (polypeptide), the conformational entropy will increase. Network rigidity in the context of a DCM is essential.

5. Conclusion

We modeled the β -hairpin to coil transition using a minimal distance constraint model (mDCM). The mDCM allows a free energy decomposition scheme to be utilized while accounting for non-additivity in conformational entropy using constraint theory. Considering two types of ensembles of constraint topologies, an exact partition function under a mean-field approximation was calculated by employing Maxwell constraint counting. The main conclusions of this work are:

- A β -hairpin to coil transition is predicted using Maxwell constraint counting for two different ensembles of constraint-topologies that consider (a) only native contacts, and (b) native and non-native contacts. For both types of constraint topology ensembles, the unfolding transitions are predicted at physically realistic temperatures using

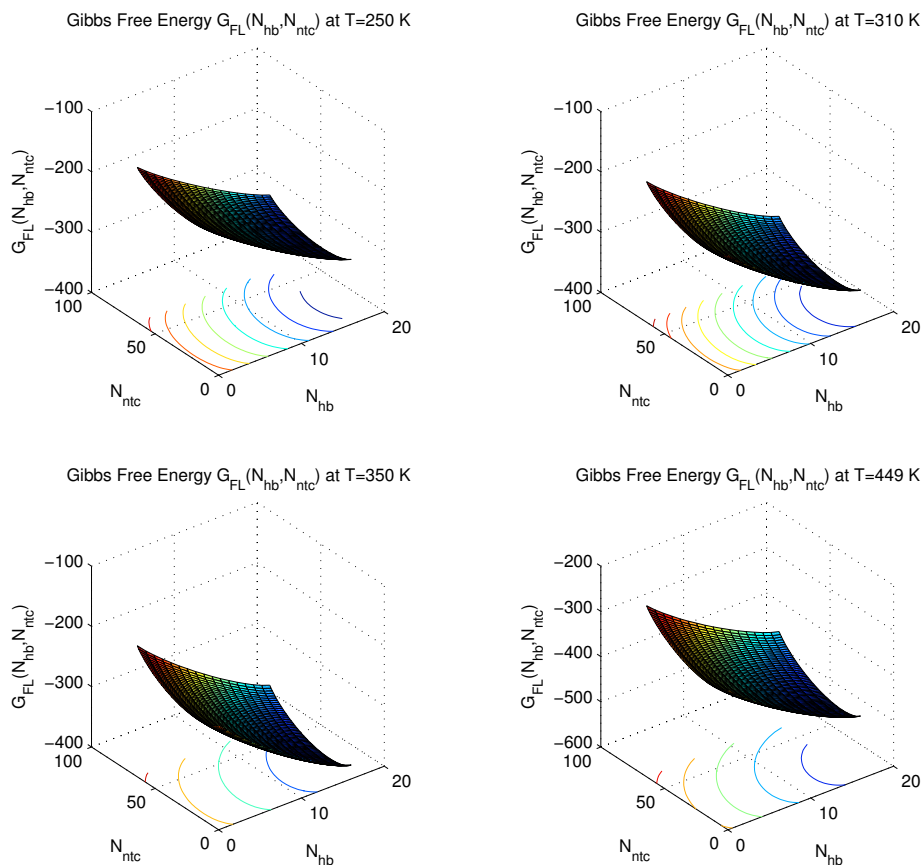


Figure 20. Free Energy Landscapes with $N = 32$ at $T = 250, 310, 350,$ and 449 K for the A-CTE case using additivity of entropy

physically reasonable values for the model parameters that were obtained from work describing protein thermodynamics.

- The transition appears not to be cooperative, or very weakly cooperative. The native contact only ensemble tends to produce a sharper transition overall, but is qualitatively similar to the case when non-native contacts are allowed. These results show that there are many other non-native structures with nearly the same free energy as the native-state. Therefore, this model is consistent with experiments showing that missfolding events are indeed populated.
- The hairpin-coil transition is not predicted for the same set of realistic model parameters when employing additivity in the model by ignoring constraint theory, and treating all constraints as independent. This exercise suggests that accounting for non-additivity in conformational entropy may be essential in coarse grained models that invoke free energy decomposition schemes. Application of constraint theory appears to work well, and is likely to resolve the mystery behind the so-called “hidden

thermodynamics” in protein chemistry. At the very least, regarding network rigidity as an underlying mechanical interaction provides a novel modeling paradigm that has been markedly successful over many different situations, including describing the β -hairpin to coil transition as demonstrated here.

5.0.2. Future and Related Work

Future work includes comparing our heat capacity predictions to experimental curves or MD simulations and exploring the effects of different model parameters. In addition, we plan to apply exact rigidity theory using recursion relations in order to determine the accuracy of the mean-field approximation employed here. Motivated by prior works and the results presented above, a much more detailed DCM using a more detailed free energy decomposition scheme that includes interaction types to explicitly model residue variations, hydration effects and hydrophobic interactions is in progress. We note that the DCM is currently under-utilized, probably because it is based on a new and unfamiliar paradigm (free energy decomposition combined with constraint theory) that is mathematically more complicated than other coarse grained models. In a related work [48], the DCM is being carefully formalized and developed as an *ab initio* theory from classical statistical mechanics to place our phenomenological approach on firm ground.

Acknowledgement

We wish to thank Jim Mottonen, Andrei Istomin, Dennis Livesay and especially Oleg Vorov for useful discussions. This work is supported by the National Institutes of Health grant: NIH RO1-GM073082-01A1.

References

- [1] Bierzynski, A. and K. Pawlowski *Acta Biochim Pol.* 1998, 45, 228.
- [2] Brady, G.P. and K.A. Sharp *J. Mol. Biol* 1995, 254, 77.
- [3] Chelvaraja, S. and H. Meirovitch *PNAS* 2004, 101, 9241.
- [4] Clementi C., H. Nymeyer, J. Onuchic. *J. Mol. Biol.* 2000, 298, 937.
- [5] Daidone I., F. Simona, D. Roccatano, R.A. Broglia, G. Tiana, G. Colombo and A.D. Nola *Proteins* 2004, 57, 198204
- [6] Das, P., S. Matysiak and C. Clementi *PNAS* 2005, 102, 10141.
- [7] Dill, K.A. *J. Biol. Chem.* 1997, 272, 701.
- [8] Fitter J. *Biophys. J.* 2003, 84, 3924.
- [9] Gao J., K. Kuczera, B. Tidor and M. Karplus *Science* 1989, 244, 1069.
- [10] Gō, N. *Ann. Rev. Biophys. Bioeng.* 1983, 12, 183.

-
- [11] Gomez, J., V.J. Hilser, D. Xie and E. Freire *Proteins* 1995, 22, 404.
- [12] Guyon, E., S. Roux, A. Hansen, D. Bideau, J-P. Troadec and H. Crapo *Rep. Prog. Phys.* 1990, 53, 373.
- [13] Hallerbach, B. and H.J. Hinz *Biophys. Chem* 1999, 76, 219.
- [14] Hansson, T., C. Oostenbrink, and W.F. van Gunsteren *Cur. Opin. Struct. Bio* 2002, 12, 190.
- [15] Hartenstine M.J., M.F Goodman and J. Petruska *J. Biol. Chem.* 2000, 275, 18382.
- [16] Head-Gordon, T. and S. Brown *Curr. Opin. Struct. Biol.* 2003, 13, 160.
- [17] Hilser, V.J. and E. Freire *J. Mol. Biol.* 1996, 262, 756.
- [18] Jacobs, D.J., S. Dallakyan, G.G. Wood, and A. Heckathorne *Phys. Rev. E* 2003, 68, 061109.
- [19] Jacobs, D. J. and S. Dallakyan *Biophys. J* 2005, 88, 903.
- [20] Jacobs, D.J., A.J. Rader, L.A. Kuhn, and M.F. Thorpe *Proteins* 2001, 44, 150.
- [21] Jacobs, D.J., D. R. Livesay, J. Hules, and M.L. Tasayco *JMB* 2006, 358, 882.
- [22] Jacobs, D.J. and M.F. Thorpe *Phys. Rev. Lett.* 1995, 75, 4051.
- [23] Jacobs, D.J. *J. Phys. A Math. Gen* 1998, 31, 6653.
- [24] Jacobs, D.J. and Hendrickson, B. *J. Comput. Phys.* 1997, 137, 346-65
- [25] Jacobs, D.J. and G.G. Wood *Biopolymers* 2004, 75, 1.
- [26] Jacobs, D.J., and M.F. Thorpe, U.S. Patent # 6014449 (2000)
- [27] Jacobs, D.J. *Recent Res. Devel. Biophys.* 2006, 5, 71.
- [28] Kittel C. *Am. J. Phys.* 1969, 37, 917.
- [29] Kolinski, A. and J. Skolnick *Polymer* 2004, 45, 511.
- [30] G. Laman *J. Eng. Math.* 1970, 4, 331.
- [31] Lee, M.S., G.G. Wood and D.J. Jacobs *J. Phys. Cond. Mat.* 2004, 16, S5035.
- [32] Lifson, S. and Roig, A. *J. Chem. Phys.*, 34, 1963.
- [33] Livesay, D. R., S. Dallakyan, G.G. Wood, and D.J. Jacobs *FEBS Lett.* 2004, 576, 468.
- [34] Livesay, D. R. and D.J. Jacobs *Proteins* 2006, 62, 130.
- [35] Makhatadze, G.I. and P.L. Privalov *J. Mol. Biol* 1993, 232, 639.
- [36] Mark, A.E. and W.F. van Gunsteren *J. Mol. Biol* 1994, 240, 167.

- [37] Munoz, V. *Curr. Opin. Struct. Biol.* 2001, 11, 212.
- [38] Munoz, V., R. Ghirlando, F.J. Blanco, G.S. Jas, J. Hofrichter, W.A. Eaton *Biochemistry* 2006, 45, 7023.
- [39] Murphy, K.P. and Gill, S.J. *J. Mol. Biol* 1991, 222, 699.
- [40] Phillips J.C., and M.F. Thorpe *Solid State Comm.* 1985, 53, 699.
- [41] Privalov, P.L. and G.I. Makhatadze *J. Mol. Biol.* 1993, 232, 660.
- [42] Schellman, J.A. *J. Phys. Chem.* 1958, 62, 1485.
- [43] Stotz C.E. and E.M. Topp *J. Pharm. Sci.* 2004, 93, 2881.
- [44] Tai, K. *Biophysical Chem.* 2004, 107, 213.
- [45] Tay, T-S. and W. Whiteley *Structural Topology* 1984, 9, 31.
- [46] *Rigidity Theory and Applications*; Thorpe, M.F. and Duxbury, P.M.; Plenum, NY, 1999.
- [47] Vendruscolo, M. *TRENDS in Biotechnology* 2002, 20, 1.
- [48] Vorov, O.K., A.Y. Istomin, D.R. Livesay and D.J. Jacobs, *Phys. Rev. Lett.* In review.
- [49] Whiteley, W. *Phys. Biol.* 2005, 2, S116.
- [50] Yoda, T., Sugita Y. and Okamoto Y. *Proteins* 2007, 66, 846.
- [51] Zimm, B.H. and J.K. Bragg *J. Chem. Phys.* 1959, 31, 526.