A semi-analytical description of protein folding that incorporates detailed geometrical information

Yoko Suzuki,^{1,a),b)} Jeffrey K. Noel,^{2,b)} and José N. Onuchic²

¹*Department of Physics, School of Sciences and Engineering, Meisei University, 2-1-1 Hodokubo, Hino-shi, Tokyo, 191-8506, Japan*

²*Center for Theoretical Biological Physics and Department of Physics, University of California at San Diego, La Jolla 92093-0374, USA*

(Received 15 March 2011; accepted 20 May 2011; published online 22 June 2011)

Much has been done to study the interplay between geometric and energetic effects on the protein folding energy landscape. Numerical techniques such as molecular dynamics simulations are able to maintain a precise geometrical representation of the protein. Analytical approaches, however, often focus on the energetic aspects of folding, including geometrical information only in an average way. Here, we investigate a semi-analytical expression of folding that explicitly includes geometrical effects. We consider a Hamiltonian corresponding to a Gaussian filament with structure-based interactions. The model captures local features of protein folding often averaged over by mean-field theories, for example, loop contact formation and excluded volume. We explore the thermodynamics and folding mechanisms of beta-hairpin and alpha-helical structures as functions of temperature and Q , the fraction of native contacts formed. Excluded volume is shown to be an important component of a protein Hamiltonian, since it both dominates the cooperativity of the folding transition and alters folding mechanisms. Understanding geometrical effects in analytical formulae will help illuminate the consequences of the approximations required for the study of larger proteins. © 2011 American Institute of Physics. [doi:10.1063/1.3599473]

I. INTRODUCTION

The energy landscape theory of protein folding is strongly based on the idea of minimal frustration and the funnel concept.^{1–5} This theory states that evolution has achieved folding robustness by selecting for sequences where the interactions present in the native state are mutually consistent and therefore have sufficiently small energetic frustration. In this case, while precise energetic features of folding are sensitive to minor changes in the protein sequence, the geometrical features of the folding mechanism are robust to many mutational and environmental changes. Because of this aspect of the folding landscape, simplified models that neglect energetic frustration, called structure-based models,^{6–8} have been successful in reproducing experimentally observed folding trends^{9,10} and mechanisms.^{7,8,11,12} These numerical simulations have enormously aided the interpretation and design of experiments that seek to provide a quantitative understanding of protein folding.¹³ A full understanding of these simulation results, however, would be impossible without the aid of much simpler analytical expressions that have been developed simultaneously.^{14–32}

Analytical treatments, including free energy functional approaches,^{14–20} microscopic theories employing the variational method,^{21–31} and elegant simplifications of these,^{34,35} add an important complement to the understanding of protein folding. Plotkin and Onuchic developed a free energy

functional model based on polymer theory^{16–19} to investigate interplay between energetics and geometric constraints in protein folding. This theory was generalized by some of us to include a phenomenological cooperativity into the configurational entropy and the internal energy²⁰ so that the folding free energy barrier agreed well with simulations. This result showed that the global entropy can be well estimated in the free energy functional model. However, due to the mean-field approximation, it lacks local geometrical information. It is difficult to improve the estimation and understanding of the local information with a phenomenologically derived entropy. A sensible approach to a more detailed theory is to develop a microscopic treatment of the entropy, which can help guide and validate useful analytic approximations. Microscopic, variational Hamiltonians have been shown to reproduce protein transition states remarkably well.^{25–31}

To this end, Suzuki *et al.* introduced an integrable Hamiltonian model of a peptide described by a Gaussian chain (GC) and a structure-based potential with a many-body effect.³² This initial study showed that, unlike previous studies, geometrical information can be taken into account without making additional assumptions. While this benefit came at the cost of increasing levels of complexity, the study showed the full temperature dependence of energy and entropy of a β -hairpin on the model's parameters. One intriguing result was that in order to fit the analytic results to simulation, a non-zero many-body effect was needed to achieve the same level of cooperativity. Since the simulation Hamiltonian did not include a many-body effect, the origin of the needed cooperativity was likely due to the enforcement of excluded volume in the simulation. Excluded volume is a well known contributor to

^{a)} Author to whom correspondence should be addressed. Electronic mail: youko@phys.meisei-u.ac.jp.

^{b)} These authors contributed equally to this work.

nonadditive effects in protein folding such as the Ramachandran map and protein desolvation.

Towards a fuller description of the microscopic entropy in peptides, a detailed treatment of excluded volume over the full folding transition is warranted. With a quantitative study of the configuration space available to a peptide with hard-sphere repulsive residues, one can better understand the consequences of treating excluded volume through more analytically tenable forms, such as, soft-core Gaussian repulsions²⁸ or simply the Gaussian chain.³³ Of particular interest is comparing energetic cooperativity introduced through phenomenological many-body effects with cooperativity imparted simply as a consequence of geometric constraints.

In this paper, we quantitatively explore how the geometry affects the dynamics of the peptide. The Gaussian chain model is extended to include the effects of excluded volume by adding hard-sphere repulsion between all non-nearest-neighbor residues in the Hamiltonian. We systematically study the effects of excluded volume on the temperature dependence of the energy, entropy, and specific heat by exploring different folded structures (β -hairpin/ α -helix) and varying the size of the beads. To understand the folding mechanisms, a Legendre transformation is employed to study the thermodynamic variables as a function of the traditional reaction coordinate Q , the fraction of native contacts formed.

II. MODEL AND METHODS

A. Partition function with microscopic Hamiltonian

We adopt a bead-on-a-filament model to describe a peptide chain, in which positions of N beads are given by $\mathbf{x}_i$ where $i = 1, 2, \dots, N$. Each bead corresponds to a residue of a peptide chain. By introducing a conditional probability, $\tau(\mathbf{x}_{i+1}|\mathbf{x}_i)$, the probability of finding the $i + 1$ th bead at the point $\mathbf{x}_{i+1}$ provided the i th bead is fixed at the point $\mathbf{x}_i$, the partition function is given by

$$Z(T) = \int d^3\mathbf{x}_1 d^3\mathbf{x}_2 \cdots d^3\mathbf{x}_N e^{-\beta V} \prod_{j=1}^{N-1} \tau(\mathbf{x}_{j+1}|\mathbf{x}_j), \quad (1)$$

where $\beta = 1/(k_B T)$ and k_B is the Boltzmann constant. In Eq. (1), V is the potential energy characterizing the native structure. V is composed of a sum between a repulsive part

FIG. 1. The Hamiltonian. (a): The nonspecific excluded volume interaction takes the form of an impenetrable wall at a distance σ_R . (b): The specific native interactions I are included through attractive square wells with width δ and centered at σ .

V_R and a native contacts part V_{NC} (see Fig. 1),

$$V = - \sum_{I=1}^M \epsilon M_I + \sum_{i=3}^N \sum_{j=1}^{i-2} R_{i,j} \equiv V_{NC} + V_R, \quad (2)$$

$$M_I = \theta \left(\sigma_I + \frac{\delta}{2} - |\mathbf{x}_{i_l} - \mathbf{x}_{j_l}| \right) \cdot \theta \left(|\mathbf{x}_{i_l} - \mathbf{x}_{j_l}| - \sigma_I + \frac{\delta}{2} \right). \quad (3)$$

Index I means I th native contact and M_I , the interaction between native contacts, is a square well of depth $\epsilon (> 0)$, width δ , and centered at the native distance σ_I . δ and ϵ could in principle vary with I , but for simplicity they are taken as constants in the present study. In our model, σ_I , δ , σ_R , and the contact map are “geometrical constraints” and ϵ (and later $e^{(2)}$) is an “energetic parameter.” We will justify these assignments by factorizing the partition function derived from the potential energy V in Sec II D. M is the total number of native contacts. θ is the Heaviside step function. $R_{i,j}$ is the repulsive force between a pair of non-neighboring residues, i.e., $|i - j| \geq 2$. We set the repulsive force as a hard sphere interaction,

$$R_{i,j} = V_0 \theta(\sigma_R - |\mathbf{x}_i - \mathbf{x}_j|), \quad (4)$$

where σ_R is the radius of the repulsive core and $V_0 \rightarrow \infty$. Throughout the paper, σ_R is varied to probe the effect of excluded volume. The pair correlation depends only on distance, $\tau(\mathbf{x}_{i+1}|\mathbf{x}_i) = \tau(\mathbf{x}_{i+1} - \mathbf{x}_i)$, and is given by

$$\tau(\mathbf{x}_{i+1} - \mathbf{x}_i) = \left(\frac{3}{2\pi a^2} \right)^{\frac{3}{2}} e^{-\frac{3(\mathbf{x}_{i+1} - \mathbf{x}_i)^2}{2a^2}}. \quad (5)$$

Thus, the peptide chain is described as a Gaussian chain where a is the mean square distance between residues i and $i + 1$. The model parameters are $a = 3.8 \text{ \AA}$ and $\delta = 0.5 \text{ \AA}$. Thermodynamic variables can be calculated by obtaining the free energy F from Z :

$$F(T) = -\frac{1}{\beta} \log Z(T). \quad (6)$$

Entropy S , energy E , and specific heat C_v are obtained from the free energy in the usual way.³⁶ $\langle A \rangle$ is defined as a thermal average of A given by

$$\langle A \rangle = \frac{1}{Z} \int \Omega(\vec{\mathbf{x}}) A e^{-\beta V}, \quad (7)$$

where $\Omega(\vec{\mathbf{x}})$ is the weight of the Gaussian chain for configuration $\vec{\mathbf{x}} \equiv (\mathbf{x}_1, \dots, \mathbf{x}_N)$,

$$\Omega(\vec{\mathbf{x}}) = d^3\mathbf{x}_1 d^3\mathbf{x}_2 \cdots d^3\mathbf{x}_N \prod_{j=1}^{N-1} \tau(\mathbf{x}_j - \mathbf{x}_{j+1}). \quad (8)$$

Except for the introduction of excluded volume V_R , the model so far presented has been introduced and studied previously.³² In the current study, we include excluded volume, which will be shown to have significant effects on cooperativity and folding mechanisms. The folding mechanism, another novelty in this study, are determined by following the formation probability for the I th contact Q_I . Contact formation is defined as

$$Q_I(T) \equiv \langle M_I \rangle, \quad (9)$$

the probability that contact I is within the square well interaction region M_I . The global reaction coordinate Q , with $0 \leq Q \leq 1$, is given as

$$Q(T) \equiv \frac{1}{M} \sum_{I=1}^M Q_I(T). \quad (10)$$

B. Solving the model

One technique often used in problems of the statistical mechanics of many-body systems is the Mayer expansion. The partition function $Z(T)$ is expanded into terms containing increasing numbers of interacting particles. For protein systems, if the contact density is low enough, $Z(T)$ can be estimated by truncating the series. The problem of a folding peptide necessarily includes the term with all particles interacting, so the series cannot be truncated. Nevertheless, in the previous paper, this technique was employed to obviate a symmetry that reduced the dimension of the partition integral.³² In the present model, the excluded volume breaks this symmetry and so the partition function is calculated by brute force using Monte Carlo integration. The Mayer expansion is still useful in this case to differentiate energetic and geometric effects (detailed in Sec. II D). Substituting the model Hamiltonian, Eqs. (2) and (3), into Eqs. (1) and (9), some thermodynamic variables are written as

$$Z(T) = \int \Omega(\vec{x}) \left[\prod_{I=1}^M (1 + (e^{\beta\epsilon} - 1)M_I) \right] \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (11)$$

$$Q_I(T) = \frac{1}{Z(T)} \int \Omega(\vec{x}) e^{\beta\epsilon} \cdot M_I \times \left\{ \prod_{J=1 (J \neq I)}^M [1 + (e^{\beta\epsilon} - 1)M_J] \right\} \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (12)$$

$$E(T) = -\epsilon \sum_{I=1}^M Q_I(T). \quad (13)$$

Another approach to solving a polymer model with a microscopic Hamiltonian is to use a variational method. This method has been extensively explored by Portman and Wolynes and coworkers.²⁵⁻³¹ In their variational approach, the interactions between the specific native contact pairs is given by a sum of Gaussian functions and the interaction between the non-native contact pairs is introduced by a repulsive core in terms of a Gaussian function.²⁸ The variational parameters in the reference Hamiltonian are set as a nonuniform external field that constrains the residues to lie near their native locations. This variational approach shows striking agreement

between theory and experiment,^{28,30,31} for proteins with well separated folded and unfolded basins. Study of peptides and proteins absent a free energy folding barrier, as the peptides in the present study are, is not well defined in the variational approach. Here, in order to study the complete folding process, i.e., including the contributions of configurations far from native, we estimate in full the terms in Eqs. (11)–(13)

C. Low temperature limit and entropy crisis

As $T \rightarrow 0$, the thermodynamical functions given by Eqs. (10)–(13) have the following limits:

$$\lim_{T \rightarrow 0} Z(T) = \int \Omega(\vec{x}) \left(\prod_{I=1}^M e^{\beta\epsilon} M_I \right) \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (14)$$

$$\lim_{T \rightarrow 0} E(T) = -\epsilon M, \quad (15)$$

$$\lim_{T \rightarrow 0} Q_I(T) = 1, \quad (16)$$

$$\lim_{T \rightarrow 0} Q(T) = 1, \quad (17)$$

$$\lim_{T \rightarrow 0} S(T) = -k_B \log \int \Omega(\vec{x}) \left(\prod_{I=1}^M M_I \right) \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right). \quad (18)$$

The expression for entropy, Eq. (18), is independent of temperature and only depends on the geometrical information of the native state. This results from treating the contact potential as a square well. If the contact potential, $-\epsilon(\mathbf{x}_{i_l} - \mathbf{x}_{j_l})$, is not a step function we have to use $e^{\beta\epsilon(\mathbf{x}_{i_l} - \mathbf{x}_{j_l})} - 1$ instead of $(e^{\beta\epsilon} - 1)M_I$ in Eq. (11). In this case, the temperature dependence will remain in the entropy $S(T)$, and will be sensitive to the shape of the contact potential at low temperature. We highlight this effect because when performing classical molecular dynamics simulations of these systems a continuous differentiable function is chosen as the contact potential, e.g., a Gaussian well. In this case, $S(T) \rightarrow -\infty$ as $T \rightarrow 0$; there is an entropy crisis. As temperature is lowered the contacts become increasingly constrained by their curved minima, the allowed area goes to zero. It is necessary to include quantum effects into such a model in order to avoid such problems.

D. Partition function factors into energetic and geometric parts

In our previous work, cooperativity was introduced with an energetic many-body interaction. In this new model, cooperativity is introduced by changing the geometry to include

excluded volume. By factoring the partition function we can show that the excluded volume can affect the order of contact formation, which we call a “folding mechanism,” while the many-body interaction cannot. The Hamiltonian with both excluded volume and the many-body interaction is given by

$$V' = - \sum_{I=1}^M \epsilon M_I + \sum_{i=3}^N \sum_{j=1}^{i-2} R_{i,j} - \sum_{I,J (I < J)} e^{(2)} M_I M_J, \quad (19)$$

where M_I and $R_{i,j}$ are defined by Eqs. (3) and (4). The third term represents a many-body interaction that favors configurations with multiple contacts formed.³² By using the Hamiltonian given by Eq. (19),

$$\begin{aligned} e^{-\beta V'} &= \prod_{K=1}^M [1 + (e^{\beta \epsilon} - 1) M_K] \\ &\times \prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \\ &\times \prod_{I=1}^M \prod_{J=1}^{I-1} [1 + (e^{\beta e^{(2)}} - 1) M_I M_J], \end{aligned} \quad (20)$$

substituting Eq. (20) into Eq. (1) and enumerating all configurations with γ contacts, we can factorize $Z(T)$ into a geometric part and an energetic part;

$$Z(T) = \sum_{\gamma=0}^M \mathcal{E}_{\gamma}(T, \epsilon, e^{(2)}) \sum_{\{\Gamma(\gamma)\}} \mathcal{G}(\Gamma(\gamma)), \quad (21)$$

where $\mathcal{G}(\Gamma(\gamma))$ is a geometric part given by

$$\begin{aligned} \mathcal{G}(\Gamma(\gamma)) &\equiv \int \Omega(\tilde{\mathbf{x}}) \\ &\times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right) \left(\prod_{I \in \Gamma(\gamma)} M_I \right), \end{aligned} \quad (22)$$

and $\mathcal{E}_{\gamma}(T, \epsilon, e^{(2)})$ is an energetic part. $\Gamma(\gamma)$ is a combination of γ objects selected from $\{1, 2, \dots, M\}$, giving a configuration with γ native contacts formed. $\{\Gamma(\gamma)\}$ is the set of all combinations of γ objects that can be selected from $\{1, 2, \dots, M\}$. Note that $\mathcal{E}_{\gamma}(T, \epsilon, e^{(2)})$ is only a function of the energetic parameters, not a function of $\tilde{\mathbf{x}}$ or the native contact map.

The order of contact formation is completely determined by the relative weight $\mathcal{G}(\Gamma(\gamma))$ of the different configurations of contacts, since the energetic part $\mathcal{E}_{\gamma}(T, \epsilon, e^{(2)})$ does not depend on the configuration $\Gamma(\gamma)$ for a fixed γ . A configuration meaning, for example, native contacts 1, 2 and 4 formed. Since the many-body effect is completely contained within $\mathcal{E}_{\gamma}$, it cannot change the folding mechanism while the excluded volume may.

E. Elucidating the dependence on the order parameter Q

Thermodynamical variables in terms of a structural order parameter such as Q are convenient for studying geometri-

cal information, e.g., overlap between an area allowed in the native state and area forbidden by the excluded volume. In order to study the dependence of the thermodynamical variables upon $Q = (1/M) \sum_I Q_I$, we introduce a generating function by adding a source term $\mu \sum_I M_I$ to the potential energy,

$$Z(T, \mu) \equiv \int \Omega(\tilde{\mathbf{x}}) e^{-\beta(V - \mu \sum_I M_I)}. \quad (23)$$

The free energy as the function of μ is then defined as

$$F(T, \mu) = -\frac{1}{\beta} \log Z(T, \mu). \quad (24)$$

The fraction of native contact is written as

$$Q(T, \mu) = \frac{1}{M} \sum_{I=1}^M \langle M_I \rangle_{\mu} = -\frac{1}{M} \frac{\partial F(T, \mu)}{\partial \mu}, \quad (25)$$

where $\langle A \rangle_{\mu}$ is the thermal average under the existence of μ ,

$$\langle A \rangle_{\mu} \equiv \frac{1}{Z(T, \mu)} \int \Omega(\tilde{\mathbf{x}}) A e^{-\beta(V - \mu \sum_I M_I)}. \quad (26)$$

By performing a Legendre transformation,

$$\tilde{F}(T, Q) \equiv F(T, \mu) - \frac{\partial F(T, \mu)}{\partial \mu} \cdot \mu = F(T, \mu) + M Q \mu, \quad (27)$$

we obtain the free energy as a function of T and Q . Taking a partial derivative with respect to Q and setting $\mu = 0$ to recover the original system gives the stationary condition,

$$\frac{\partial \tilde{F}(T, Q)}{\partial Q} = 0. \quad (28)$$

The solution of the stationary equation, Eq. (28), gives the temperature dependence of $Q(T)$ in Eq. (10), $Q(T) = Q_0(T)$. From an identity of the Legendre transformation, the entropy is given by

$$S(T, \mu) = - \left(\frac{\partial F(T, \mu)}{\partial T} \right)_{\mu} = - \left(\frac{\partial \tilde{F}(T, Q)}{\partial T} \right)_Q = S(T, Q). \quad (29)$$

To calculate $S(T, Q)$, we use the following equation:

$$\begin{aligned} S(T, Q) &= S(T, \mu(T, Q)) = -\frac{1}{T} F(T, \mu(T, Q)) \\ &+ \frac{1}{T} \left\langle \left\{ V - \mu(T, Q) \cdot \left(\sum_{I=1}^M M_I \right) \right\} \right\rangle_{\mu(T, Q)} \\ &= -\frac{1}{T} F(T, \mu(T, Q)) \\ &+ \frac{1}{T} (\langle V \rangle_{\mu(T, Q)} - \mu(T, Q) \cdot M Q). \end{aligned} \quad (30)$$

The second equation in Eq. (30) is derived from the second equation of Eq. (29). By using Eqs. (27) and (30), we can write the energy as a function of Q ,

$$\tilde{E}(S(T, Q), Q) = \langle V \rangle_{\mu(T, Q)}, \quad (31)$$

which satisfies the relation,

$$\tilde{F}(T, Q) = \tilde{E}(S(T, Q), Q) - T S(T, Q). \quad (32)$$

Hereafter, we represent $\tilde{E}(S(T, Q), Q)$ as $\tilde{E}(T, Q)$. By using an identity of the Legendre transformation, $Q_I(T, Q)$

$= Q_I(T, \mu)$, and by definition, $Q_I(T, \mu) = \langle M_I \rangle_\mu$. Therefore, $Q_I(T, Q)$ is given by

$$Q_I(T, Q) = \langle M_I \rangle_{\mu(T, Q)} = \frac{1}{Z(T, \mu(T, Q))} \times \int \Omega(\vec{x}) M_I e^{-\beta(V - \mu(T, Q) \sum_I M_I)}. \quad (33)$$

The temperature dependence of the transformed thermodynamical variables can be connected back to the untransformed dependencies through $Q_0(T)$; $F(T) = \tilde{F}(T, Q_0(T))$, $E(T) = \tilde{E}(T, Q_0(T))$, $S(T) = \tilde{S}(T, Q_0(T))$, $Q_I(T) = Q_I(T, Q_0(T))$.

Substituting the potential, Eq. (2), into the transformed thermodynamical variables (Eqs. (23), (31), (33) and (25)), we obtain explicit μ dependencies.

$$Z(T, \mu) = \int \Omega(\vec{x}) \prod_{I=1}^M (1 + X(T, \mu) M_I) \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (34)$$

$$\tilde{E}(T, Q(T, \mu)) = -\frac{1}{Z(T, \mu)} \int \Omega(\vec{x}) \sum_{I=1}^M \epsilon(X(T, \mu) + 1) M_I \times \left[\prod_{J=1}^M (1 + X(T, \mu) M_J) \right] \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (35)$$

$$Q_I(T, \mu) = \frac{1}{Z(T, \mu)} \int \Omega(\vec{x}) (X(T, \mu) + 1) M_I \times \left[\prod_{J=1}^M (1 + X(T, \mu) M_J) \right] \times \left(\prod_{i=3}^N \prod_{j=1}^{i-2} \theta(|\mathbf{x}_i - \mathbf{x}_j| - \sigma_R) \right), \quad (36)$$

$$Q(T, \mu) = \frac{1}{M} \sum_{I=1}^M Q_I(T, \mu). \quad (37)$$

As a check, note that from Eqs. (35) and (37) we obtain,

$$\tilde{E}(T, Q) = -\epsilon M Q. \quad (38)$$

In each equation, the T and μ dependence is contained within $X(T, \mu)$ given by

$$X \equiv e^{\beta(\epsilon + \mu)} - 1. \quad (39)$$

In order to determine the dependence on Q , first write Eq. (25) as

$$Q(X) = \frac{1}{M} (X + 1) \frac{\partial (\log Z(X))}{\partial X}. \quad (40)$$

Since X is a single-valued function of $\beta\mu$, X can be obtained graphically by inverting Eq. (40) to obtain $X(Q)$. This X is then used as input for the above thermodynamic relations (Eqs. (34)–(36)) to obtain the Q dependence. A further consequence of the X dependence is that the transformed thermodynamical variables do not depend on T :

$$Z(X) = Z(X(Q)) = Z(Q), \quad (41)$$

$$\begin{aligned} S(T, Q) &= \frac{1}{T} (\tilde{E}(T, Q) - \tilde{F}(T, Q)) \\ &= \frac{1}{T} \left(-\epsilon M Q - \frac{1}{\beta} \log Z(Q) - M Q \mu \right) \\ &= -k_B [M Q \log(X(Q) + 1) + \log Z(Q)] = S(Q), \end{aligned} \quad (42)$$

$$Q_I(T, Q) = Q_I(T, \mu) = Q_I(X) = Q_I(Q). \quad (43)$$

The lack of T dependence is a special property of the step function contact potential. If the attractive potential is not given by a step function, the T dependence remains in the functions $\tilde{E}(T, Q)$, $S(T, Q)$, and $Q_I(T, Q)$.

III. RESULTS AND DISCUSSION

The characterization of the model comes in three sections. The first section explains and explores the cooperative effect derived from excluded volume. In the second section, the temperature dependence of the thermodynamical functions is explored in detail for a host of different native state geometries, including a β -hairpin and an α -helix. Lastly, courtesy of the analytic approach, a Legendre transformation is employed to look at the dependence of thermodynamical functions on a global order parameter Q , the average number of contacts formed. In these sections, we define $\mathcal{M}_X$ as a contact map for the native state geometry denoted by the subscript X . Some examples are presented in Fig. 2. Two maps, $\mathcal{M}_S$ and $\mathcal{M}_D$ representing either a dense or a sparse hairpin turn. The other two maps $\mathcal{M}_{4\beta}$ and $\mathcal{M}_{6\alpha}$ present a β -hairpin with 4 contacts and an α -helix with 6 contacts. Here, $\mathcal{M}_{M\alpha}$ and $\mathcal{M}_{M\beta}$ denote an α -helix and β -hairpin with M contacts, respectively. The contact distances are taken to mimic actual protein structures. All the symmetric β -hairpin contacts in $\mathcal{M}_S$, $\mathcal{M}_D$, and $\mathcal{M}_{M\beta}$ are 5 Å. $\mathcal{M}_D$ introduces two crossing contacts whose distance is determined from trigonometry: $\sqrt{5^2 + 3.8^2} = 6.3$ Å.

FIG. 2. Schematic representation of contact maps discussed throughout the paper. Red lines represent native contact.

FIG. 3. Excluded volume in a Gaussian chain. Curves show the effect of adding excluded volume to the Gaussian chain on root mean squared separation d_{rms} between different pairs of residues. At $\sigma_R = 0$, d_{rms} is $\sqrt{|i-j|} \times a$ with $a = 3.8$ Å. There is a slope change at $\sigma_R \approx 2$ Å showing that the Gaussian chain, at least for small sequence separation, has an implicit excluded volume around 2 Å. The blue curves show the behavior of a 13 residue segment and two smaller subsegments. The nonspecific interactions of the surrounding residues slightly alter d_{rms} of subsegments of length 4 (5,9) and length 8 (3,11).

The helical contacts are between residues i and $i + 4$ and are each 6.2 Å.

A. Excluded volume and cooperativity

1. Implicit repulsive effect of the Gaussian chain

A Gaussian chain is a common choice to represent a polymer since it is amenable to analytical treatment.^{33,37} Each residue in the polymer is correlated to the position of its nearest neighbors by a spherical Gaussian distribution of their displacements $\mathbf{r}_i = \mathbf{x}_{i+1} - \mathbf{x}_i$, $P(r_i) \propto e^{-3r_i^2/(2a^2)}$. This distribution is peaked at a separation of a and does not allow nearest neighbors to collapse onto each other. The distribution satisfies $\langle |\mathbf{x}_i - \mathbf{x}_j| \rangle = \sqrt{|i-j|} \times a$. The chain resists collapse, but is likely to closely overlap somewhere as the chain length grows. By adding a hard sphere repulsive core between all residues, the chain is strictly forbidden from overlapping at any point.

To illustrate the effect of adding a repulsive core to a Gaussian chain, the root mean squared separation d_{rms} between residues in a Gaussian chain with no native contacts is plotted as a function of the repulsive core size σ_R in Fig. 3. d_{rms} increases as σ_R increases with a slope change at $\sigma_R \approx 2$ Å. Excluded volume has a relatively larger effect on longer chains. In a standard Gaussian chain, any two residues separated by the same sequence length will have the same d_{rms} . Through the addition of excluded volume, the sequence position will matter, since the accessibility of configurations will depend on the local density of neighbors.

The slope change at $\sigma_R \approx 2$ Å implies an implicit excluded volume, at least for short chains, of 2 Å. It is instructive to look ahead to Fig. 6 where thermodynamic functions are compared for $\sigma_R = (0, 2, 4)$ Å. The curves are only slightly perturbed at $\sigma_R = 2$ Å compared to $\sigma_R = 0$ Å, whereas $\sigma_R = 4$ Å shows large deviations. Typical coarse-grained models of proteins used in molecular dynamics have a repulsive core size of 4 Å (Ref. 7) and the results here would predict that this repulsive core has a significant role.

FIG. 4. Excluded volume reduces available configuration space. (a) D_{config} is all of configuration space. D_{ns} is the subset of D_{config} belonging to the native state, where all contacts are formed and D_{ex} is the space forbidden by excluded volume. (b) Entropy $S(T)$ and (c) the specific heat $C_v(T)$ as functions of temperature are shown for $\mathcal{M}_S$ with three different distributions of excluded volume: no repulsion $\sigma_R = 0$ Å, full repulsion $\sigma_R = 4$ Å, and repulsion only between native contacts $\sigma_{\text{NC}} = 4$ Å. Compared to $\sigma_R = 0$ Å, $\sigma_{\text{NC}} = 4$ Å has the same folded state entropy, while the entropy of the unfolded ensemble is reduced. This shifts the C_v peak to the right. Adding full repulsion reduces the entropy of the folded state more than the unfolded states, which shifts the C_v to the left for $\sigma_R = 4$ Å. As can be seen both by the slope of $S(T)$ and by the height of the peak in $C_v(T)$, the cooperativity is higher with full repulsion than with native repulsion only or with no repulsion.

2. Excluded volume results in increased cooperativity

Excluded volume reduces the allowed configuration space of the chain. Figure 4(a) defines three domains: D_{config} is configuration space, all possible configurations, D_{ns} the configurations belonging to the native state where all contacts are formed, and D_{ex} the configurations forbidden by excluded volume. D_{ex} is a subset of D_{config} and intersects D_{ns} . One effect of D_{ex} is to increase the average size of the chain as seen in Fig. 3.

The reduction of allowed D_{config} by D_{ex} can directly be seen in the entropy S . The entropy versus temperature $S(T)$ is plotted for three different realizations of contact map $\mathcal{M}_S$ in Fig. 4(b). Consider first the case where excluded volume is added for only the native contacts $\sigma_{\text{NC}} = 4$ Å. The lowest entropy state is at $T = 0$ and corresponds to D_{ns} . D_{ns} is unchanged since the shell where a contact is formed does not intersect with σ_{NC} and hence does not change $S(T \rightarrow 0)$. Now as $T \rightarrow \infty$, native-only repulsion shows a small reduction in entropy since D_{ex} now overlaps relevant area of D_{config} . This behavior forces the gradient dS/dT to lessen if only native repulsion is included.

If excluded volume is included between all non-neighbors, D_{ex} overlaps area from D_{ns} . This is because configurations with all native contacts satisfied can have non-native residues overlapping. This reduces the number of allowed native state configurations and hence $S(T \rightarrow 0)$ decreases. As $T \rightarrow \infty$, repulsion between all non-neighbors shows a reduction in entropy. Opposite to native-only repulsion above, dS/dT increases. The overall entropy change ΔS decreases for $\sigma_{\text{NC}} > 0$ and increases for $\sigma_{\text{R}} > 0$ compared to the $\sigma_{\text{R}} = 0$ case.

The gradient of the entropy dS/dT corresponds to the height of the specific heat peak C_v as shown in Fig. 4(c). This correspondence can be seen directly from the Hamiltonian description. $C_v = dE/dT = T(dS/dT)$ and so the gradient in entropy directly controls the height of the specific heat peak. Using the relation $E(T) = -\epsilon M Q(T)$ (derived from Eqs. (12) and (13)), that the energy is equal to the number of formed contacts, the specific heat can be written as

$$\frac{dE(T)}{dT} = C_v = \frac{\epsilon^2}{k_B T^2} \sum_{I=1}^M \sum_{J=1}^M (\langle M_I M_J \rangle - \langle M_I \rangle \langle M_J \rangle). \quad (44)$$

The C_v is the sum of the pair correlations between the contacts. Highly correlated contact formation is the essence of cooperativity and in this model the correlation directly corresponds to the gradients dE/dT and dS/dT or the height of the peak in C_v .

The peak in C_v not only changes its height, but also its position T_0 (Fig. 4(c)). The shift is controlled by whether D_{ex} has a larger relative effect on D_{ns} or D_{config} . If the entropy of the native state is reduced relatively more than the unfolded ensemble by excluded volume, the native state is destabilized and the peak in C_v moves towards lower T . If instead the entropy of the unfolded states is relatively more reduced, the unfolded ensemble is destabilized and the peak in C_v moves towards higher T . This simple model parallels the observed C_v shifts in protein folding models under different excluded volume conditions as seen in Lammert *et al.*³⁸ Inclusion of repulsion between all non-neighbors tends to increase cooperativity and shift C_v peak lower while native only repulsion tends to decrease cooperativity and shift C_v peak higher.

3. Comparison between energetic and entropic cooperativity

Section III A 2 showed that excluded volume, a purely entropic effect, can increase cooperativity. This cooperative effect came from a nonspecific many-body interaction imparted by the hard spheres in V_{R} . In the previous paper³² cooperativity was imparted by a specific many-body interaction contained in V_{NC} . This many-body interaction created a weighted contact interaction that rewards configurations with many contacts formed and penalizes configurations with few contacts formed. This kind of interaction has been used to model nonadditive phenomena such as solvation.³⁹ The weighting encourages contacts to break or form in concert and hence modulates cooperativity. With the many-body effect, if you consider a configuration with m contacts formed, the energy ϵ_m required to break a contact is given explicitly by

FIG. 5. Energetic versus entropic cooperativity. The energy E , entropy S and specific heat C_v are shown as functions of temperature for three models: simple GC, Gaussian chain with excluded volume (EV), and Gaussian chain with a many-body effect (MB). The parameters are (GC, green) $\sigma_{\text{R}} = 0 \text{ \AA}$, (EV, red) with excluded volume $\sigma_{\text{R}} = 4 \text{ \AA}$, and (MB, blue) $\sigma_{\text{R}} = 0 \text{ \AA}$ and including a many-body effect $e^{(2)}$ that adds an energetic correlation between native contacts. The protein geometry is $\mathcal{M}_{5\beta}$ shown in Fig. 2. (a): The transition shifts to lower energy for $\sigma_{\text{R}} > 0$ and higher energy for $e^{(2)} > 0$. σ_{R} has a destabilizing effect while $e^{(2)}$ has a stabilizing effect. (b): Unlike excluded volume, the geometrical factor $\mathcal{G}$ (see Eq. (22)) is unaffected by $e^{(2)}$ so $S(T = 0)$ and $S(T = \infty)$ does not shift. Both σ_{R} and $e^{(2)}$ increase the slope of S . (c): The C_v clearly shows the increased cooperativity upon introduction of σ_{R} or $e^{(2)}$.

$\epsilon_m = \epsilon + (m - 1)e^{(2)}$. ϵ is the pairwise energy giving a total contribution to the energy of $m\epsilon$ and $e^{(2)}$ is the many-body term giving a contribution of $[m(m - 1)/2]e^{(2)}$ since there are $m(m - 1)/2$ pairwise correlations among m contacts.

We compare the thermodynamics of geometry $\mathcal{M}_{5\beta}$ with cooperativity introduced by excluded volume to that imparted by a many-body effect in Fig. 5. The parameters $e^{(2)}$ and ϵ in the many-body case are fit to give the same total binding energy 5 kcal/mol and the same width ΔT of the specific heat as in the excluded volume case ($e^{(2)} = 0$ and $\sigma_{\text{R}} = 4 \text{ \AA}$). These conditions fix $e^{(2)} = 0.31 \text{ kcal/mol}$ and $\epsilon = 0.37 \text{ kcal/mol}$. Obtaining an equivalent cooperativity from the many-body interaction requires a stronger many-body contribution $[M(M - 1)/2]e^{(2)} = 3.1 \text{ kcal/mol}$ than pairwise contribution $M\epsilon = 1.9 \text{ kcal/mol}$.

As was detailed in Sec. II D the partition function can be factored (see Eq. (21)) into a geometrical part $\mathcal{G}(\Gamma(\gamma))$ and an energetic part $\mathcal{E}_\gamma(T, \epsilon, e^{(2)})$. The many-body term only effects $\mathcal{E}_\gamma$, while the excluded volume only effects $\mathcal{G}$. The entropy is determined by the relative weightings of $\mathcal{G}(\Gamma(\gamma))$, so for $T \rightarrow 0$ and $T \rightarrow \infty$ the many-body effect has an unchanged value of the entropy, i.e., ΔS is constant. The many-body effect stabilizes the protein since it becomes more difficult to break a contact in the folded state, while the excluded volume destabilizes the folded state, as shown by effects on the specific heat (Fig. 5(c)).

B. Excluded volume in simple protein-like native state geometries

In this section, the temperature dependence of various thermodynamical functions is studied for several native state

FIG. 6. Native contact map $\mathcal{M}$ changes repulsive core effect. Thermodynamical variables energy E , entropy S , and specific heat C_V are shown as functions of temperature. Top: Sparse contact map $\mathcal{M}_S$. Bottom: Dense contact map $\mathcal{M}_D$. $\sigma_R = 2 \text{ \AA}$ is nearly the same as the no excluded volume case, illustrating the implicit excluded volume of the Gaussian chain. The C_V sharpens, and therefore becomes more cooperative, with increasing excluded volume for both maps. The C_V peak shifts in opposite directions for sparse and dense contact maps, i.e., repulsive core destabilizes $\mathcal{M}_S$ and stabilizes $\mathcal{M}_D$.

geometries. First, two maps representing either a dense or a sparse hairpin turn, $\mathcal{M}_S$ and $\mathcal{M}_D$, are compared. Second, several maps representing different lengths of protein secondary structure are compared. For clarity, schematics of the contact maps are presented near their respective thermodynamic results again, $\mathcal{M}_S$ and $\mathcal{M}_D$ in Fig. 6 and $\mathcal{M}_{M\alpha}$ and $\mathcal{M}_{M\beta}$ in Fig. 8.

In our previous exploration of this model,³² there was no detailed look at the folding mechanism. To look at the ordering of the contact formation, $Q_I(T)$, the formation of contact I , is introduced and satisfies $Q(T) = (1/M) \sum_{I=1}^M Q_I(T)$ and $E(T) = -\epsilon M Q(T)$. A contact is considered formed if $Q_I = 1$, i.e., the two residues are within the square well interaction (see Eq. (3)).

1. Repulsive effects depend on native contact map

The temperature dependence of thermodynamic functions for the two maps $\mathcal{M}_S$ and $\mathcal{M}_D$ are shown in the top and bottom rows of Fig. 6, respectively. Their geometries are shown in Fig. 6. $\mathcal{M}_S$ has a two β -hairpin type contacts. $\mathcal{M}_D$ adds two “crossing” contacts which would correspond to side-chain interactions in the β -hairpin structure.

The high temperature behavior is very similar for the two maps. Certainly, as $T \rightarrow \infty$ the native contact map no longer differentiates the geometries, since the contacts provide little weight. The behavior near the transition temperature T_0 and as $T \rightarrow 0$ though is different for the two maps. The difference stems from D_{ex} differentially affecting D_{ns} in $\mathcal{M}_S$ compared with $\mathcal{M}_D$. Figure 6 shows that for the no excluded volume case, $S(T=0)$ for $\mathcal{M}_S$ is larger than $\mathcal{M}_D$. In $\mathcal{M}_D$, the native state is highly constrained by the existence of more contacts, thus the area D_{ex} removes from D_{ns} is small. $\mathcal{M}_S$ has many more possible configurations that satisfy D_{ns} and therefore has a much larger loss to D_{ex} . Adding excluded volume (increasing D_{ex}) relatively shifts $S(T=0)$ lower in $\mathcal{M}_S$ than $\mathcal{M}_D$. Since the two maps must have the same $S(T \rightarrow \infty)$, an

increase in excluded volume will necessarily destabilize $\mathcal{M}_S$ more than $\mathcal{M}_D$. In fact, T_0 even shifts to higher temperature in $\mathcal{M}_D$ with increasing excluded volume. This feature is explained in Sec. III C 3.

The excluded volume increases cooperativity in both maps. This is seen in the gradients dE/dT , dS/dT , and dQ_I/dT which all steepen and in the height of the C_V peaks. The curves for $\sigma_R = 2 \text{ \AA}$ are only slightly shifted relative to the no excluded volume case, reiterating the implicit excluded volume of the Gaussian chain. The trend is the same for σ_R of $2 \text{ \AA}$ as $4 \text{ \AA}$.

An interesting effect of the excluded volume is that it can influence the folding mechanisms of the peptide as shown in Fig. 7. Consider $Q_I(T)$ for the two maps shown in Fig. 6. The mechanism is as expected for $\mathcal{M}_S$, the order is determined by the loop entropies of the contacts. The loop length

FIG. 7. Repulsive core effect can change contact ordering. Contact ordering Q_I is shown as a function of temperature. Top: Sparse contact map $\mathcal{M}_S$. Bottom: Dense contact map $\mathcal{M}_D$. The ordering of the contacts changes with the addition of excluded volume. In $\mathcal{M}_D$, crossing type contacts (1,6 and 2,7) become more stable than the shortest loop length contact (2,6).

FIG. 8. Comparison of helical and hairpin geometry. Thermodynamical variables energy E , entropy S , and specific heat C_v are shown as functions of temperature. Top: β -hairpin geometry $\mathcal{M}_{M\beta}$ for $M = (3, 4, 5)$ contacts with $(8, 10, 12)$ residues, respectively. Bottom: α -helical geometry $\mathcal{M}_{M\alpha}$ for $M = (4, 5, 6)$ contacts with $(8, 9, 10)$ residues, respectively. The cooperative effect of the excluded volume is seen in the steepening of the curves for $\sigma_R = 4 \text{ \AA}$.

between (2,6) is $4a$ and between (1,7) is $6a$, where (i, j) means a contact between residue i and residue j and $a = 3.8 \text{ \AA}$. A longer loop gives up more entropy when constrained, therefore (2,6) forms before (1,7). The introduction of excluded volume steepens $Q_I(T)$ but does not change the order between these two contacts. The mechanism in $\mathcal{M}_D$ shows more sensitivity. The ordering of contact formation naturally follows loop length for the no excluded volume case, but upon the inclusion of excluded volume, (2,6) is destabilized relative to the symmetric crossing type contacts (1,6) and (2,7), even though (2,6) has a shorter loop length. This is because (1,6) and (2,7) have a longer native distance σ_I and so are less sensitive to excluded volume.

2. α -helix versus β -hairpin

The most ubiquitous geometries in proteins are the α -helix and the β -strand. Figure 8 shows a comparison of the thermodynamics of an α -helix and β -hairpin with and without excluded volume. Schematics of the maps are shown in the figure. Continuing the theme, addition of excluded volume results in increased cooperativity as seen from increases in the gradients dE/dT , dS/dT and dQ_I/dT and the rising peaks in C_v . The cooperativity can be quantified with a unitless measure $\sigma^C = \Delta T/T_0$. ΔT , the C_v peak half-width, and T_0 , the peak temperature, are defined in Fig. 4. Smaller σ^C implies higher cooperativity, since a tighter width implies a more abrupt transition.

Addition of excluded volume has a relatively larger cooperative effect on the β -hairpin structures. Consider the two geometries with 10 residues, $\mathcal{M}_{6\alpha}$ and $\mathcal{M}_{4\beta}$. For the helix $\mathcal{M}_{6\alpha}$, $\sigma^C = 115/131 = 0.88$ for $\sigma_R = 4 \text{ \AA}$ and $\sigma^C = 150/144 = 1.04$ for $\sigma_R = 0 \text{ \AA}$. The line sharpening can be quantified by the ratio of $\sigma^C \rightarrow 1.04/0.88 = 1.2$. For the hairpin $\mathcal{M}_{4\beta}$, $\sigma^C = 70/130 = 0.54$ for $\sigma_R = 4 \text{ \AA}$ and $\sigma^C = 135/140 = 0.92$ for $\sigma_R = 0 \text{ \AA}$. The line sharpening is given by $0.92/0.54 = 1.7$. Therefore, the relative C_v line sharpening of the hairpin is stronger than that of the helix since $1.7 > 1.2$.

The differential impact of excluded volume on the two geometries lies in the distribution of loop entropies of the contacts. In the helix all contacts have the same loop entropy and so the same local density of residues. In the hairpin, the loop entropies vary, getting larger farther from the turn. This variation means the excluded volume acts differently on each contact, destabilizing the shorter loop length contacts more than the long loop length contacts. Since the long loop length contacts are the least stable, this enhances the cooperativity of the transition. The ordering of the contacts can be seen in Fig. 9. In the hairpin, the contacts clearly fold short to long, analogous to a zippering motion. The helix has all contacts equally stable at any temperature. The slight asymmetry seen in the helix is caused by edge effects, a contact made by the first or last residue in the helix has a slightly different neighborhood than a contact in the interior of the helix.

FIG. 9. Loop entropy and contact ordering. Contact ordering Q_I is shown as a function of temperature. Top: β -hairpin geometry $\mathcal{M}_{4\beta}$. Bottom: α -helical geometry $\mathcal{M}_{6\alpha}$.

FIG. 10. Interplay between native specificity and nonspecific interactions. (a) Two different contact maps for a β -hairpin are considered, a five contact case $\mathcal{M}_{5\beta}$ and a three contact case $\mathcal{M}_{3\beta}^*$, where the green contacts are omitted. Both maps lead to the same native structure. $\mathcal{M}_{5\beta}$ has more cooperative and highly peaked specific heat than $\mathcal{M}_{3\beta}^*$. The relaxed specificity in $\mathcal{M}_{3\beta}^*$ can be compensated by including the nonspecific excluded volume interaction as seen in the green curve. (b) Two different contact maps for a helix, $\mathcal{M}_{6\alpha}$ and, with the two green contacts omitted, $\mathcal{M}_{6\alpha}^*$. The helix behaves similarly to the hairpin. Therefore, the Hamiltonian's specificity may be relaxed without losing structural specificity and cooperativity through the combined effect of many nonspecific interactions.

3. Relaxation of constraints

In a structure-based model of a protein, the Hamiltonian is highly specific to a particular native structure. In the real protein, the residues are not *a priori* aware of their specific partners, instead the specificity emerges from the complicated combination of many nonspecific interactions, interactions both attractive and repulsive. While a structure-based model describes the folding process very well, it is clear that the specificity is overstated. Understanding how to relax this specificity while maintaining a sharp transition is one of the current challenges in protein folding. Here, with this simple model, excluded volume is shown to be an important component of a less specific Hamiltonian.

Figure 10 shows the behavior of a less specific Hamiltonian for an α -helix and a β -hairpin. The specificity is relaxed by removing contacts but keeping the overall energy bias the same by scaling up the strength of the remaining contacts. For example in the β -hairpin, the three remaining contacts in $\mathcal{M}_{3\beta}^*$ have a depth of $5\epsilon/3$ to maintain the overall bias of 5ϵ . For both geometries, when the contacts are removed, the transition becomes less cooperative as seen in the broadening of the C_v . When excluded volume is introduced, the sharpness of the transition is recovered. The requirement of forming the remaining native contacts while satisfying excluded volume reduces the configuration space of the transition region sufficiently to have cooperative folding.

C. Q dependence of thermodynamical functions

In Secs. III A and III B, the dependence of the thermodynamical variables was determined as a function of T . Temperature is the standard variable in the canonical ensemble, the ensemble probed by molecular dynamics. With a Legendre transformation, the thermodynamical dependence can be switched from temperature to a reaction coordinate Q . Here Q is defined as the fraction of native contacts formed, $Q = (1/M) \sum_{I=1}^M \langle M_I \rangle$, see Eq. (25). Q is of course directly related to the energy when M_I is a square well, $\tilde{E}(Q) = -\epsilon M Q$. Neither entropy $S(Q)$ nor contact probability $Q_I(Q)$ depend on temperature, and so give purely geometrical information, see Eqs. (42) and (43). In Secs. III C 1–III C 3, we study the Q dependence of contact ordering, entropy and free energy for the contact maps considered previously, $\mathcal{M}_D$ (dense), $\mathcal{M}_S$ (sparse), $\mathcal{M}_{6\alpha}$ (helix), and $\mathcal{M}_{4\beta}$ (hairpin).

1. Contact ordering with Q

A folding mechanism is the sequence of structures a protein traverses during folding. In Sec. II $Q_I(T)$ is shown and provides a description of the ensemble of structures at each temperature. In the canonical ensemble, the relevant configurations at each temperature are weighted by energy. In a cooperative system, i.e., one with an abrupt transition with temperature, there will be very little information about the transition state in $Q_I(T)$, since the ensemble is dominated by the folded and unfolded states. In the transformed ensemble with Q as an independent variable, all structures at a particular Q have the same energy. Therefore, they are only weighted by entropy and the contact ordering can be seen for the whole range of unfolded to folded states calibrated by Q .

Figure 11 shows the probability of each contact Q_I along Q for three native geometries. $Q_I(Q)$ satisfies the constraint, $Q = (1/M) \sum_{I=1}^M Q_I(Q)$. Comparison with Fig. 7 and Fig. 9 shows that, as expected, $Q_I(T)$ and $Q_I(Q)$ have the same ordering. $Q_I(Q)$ though, can be compared with free energy functional (FEF) approaches where Q is used as the order parameter. When a mean field approach is used to solve the FEF model, the only determinant of the ordering is the loop entropy of the contacts: $Q_I(Q) < Q_J(Q)$ for the loop lengths $\ell_I > \ell_J$.²⁰ While loop entropy has been seen to be a strong determinant of ordering in the current model, excluded volume also plays a role. The effect is most dramatic in the dense contact map $\mathcal{M}_D$ where the ordering is altered upon the addition of excluded volume (Fig. 11, top row). The orderings in $\mathcal{M}_{4\beta}$ and $\mathcal{M}_{6\alpha}$ are more subtly altered by excluded volume. In the middle row, excluded volume enhances the sequential folding of the β -hairpin, already present due to differences in loop entropy. In the bottom row, the loop entropy of the contacts are all equal in the helix $\mathcal{M}_{6\alpha}$. Therefore, the contacts form nearly in unison. Though addition of excluded volume introduces only a small overall shift, it removes symmetry from the ordering. (1,5)/(6,10) are less stable than the four inner contacts for $\sigma = 0$ Å. At $\sigma = 4$ Å all three symmetric pairs of contacts separate, with the inner pair (3,7)/(4,8) less stable than the middle pair (2,6)/(5,9). Also, before the

FIG. 11. Folding mechanisms (native contact ordering) for simple geometries. The top, middle, and bottom rows show the average formation of contact I Q_I versus the global reaction coordinate Q for the contact maps M_D , $M_{4\beta}$, and $M_{6\alpha}$. The second and third columns show the contact ordering for $\sigma_R = 0$ Å and $\sigma_R = 4$ Å. Top: The ordering in the dense contact map M_D is changed by excluded volume. Middle: Excluded volume enhances the sequential folding of the β -hairpin, already present due to differences in loop entropy. Bottom: The loop entropy of the contacts are all the same in the helix $M_{6\alpha}$. Therefore, the contacts form nearly in unison. Though addition of excluded volume introduces only a small overall shift, it introduces a crossing. This crossing is highlighted in the insets. The missing legend is the same as the other helical figure's legend.

curves converge at $Q = 0$, there is a crossing at $MQ \approx 1.5$ where the order reverses.

To emphasize the importance of excluded volume in models of protein folding, data from an all-atom structure-based model of a two-state protein, SH3, is presented in Fig. 12. At folding temperature, SH3 populates both a folded and an unfolded ensemble. The folding mechanism can change dramatically with excluded volume as shown by the contact ordering in Figs. 12(c) and 12(d). Hairpins H1 and H3 in SH3 have similar loop entropies and fold in unison in the low excluded volume limit. With increased excluded volume H3 forms before the transition state and H1 forms during the transition state. The ordering of several contacts' formation changes with different excluded volume.

2. Geometric properties are a robust feature of the landscape

This simple model highlights the generally held notion that geometric properties such as the folding mechanism are robust features of the energy landscape, while energetic properties such as cooperativity and thermal stability are sensitive to the details of the model.^{8,38,40} Drastic changes in excluded volume only slightly perturbs the contact formation ordering

FIG. 12. Excluded volume can affect the cooperativity and folding mechanism in all-atom structure-based simulations. (a) Native contact map for SH3 domain. Hairpin 1 (H1) and hairpin 3 (H3) are highlighted. (b) Specific heat C_V shown for a standard atomic excluded volume size $\sigma_R = 1.6$ Å and a reduced atomic excluded volume $\sigma_R = 1.0$ Å. Loss of excluded volume reduces the cooperativity. The inset shows the change in the free energy as a function of Q , $F(Q)$, upon reduction of excluded volume. The barrier falls from $7k_B T$ to $3k_B T$. (c) Contact ordering is shown for H1 and H3 in the low excluded volume limit $\sigma_R = 1.0$ Å. (d) Ordering for the standard excluded volume. In the low excluded volume limit, both hairpins can fold simultaneously, whereas standard excluded volume introduces constraints that bias H3 to order before H1. The simulation parameters are described in detail in Ref. 43.

seen in Figs. 7, 9, and 11. At the same time, the cooperativity and, to a lesser degree, the thermal stability, are altered under the same parametric changes (Figs. 5, 6, 8, and 13). This feature of the energy landscape led us, in Sec. II D, to factorize the partition function into terms able/unable to affect the folding mechanism (geometric/energetic parts). The many-body effect was completely contained within the energetic part, able to affect energetic features (Fig. 5) while leaving the folding mechanism unperturbed. Other investigators have included a many-body effect by including “three-body interactions.”^{41,42} These studies show, as in the current model, a transition state structure insensitive to the strength of the many-body effect.

This property of factorability exposes the origin of robust geometric features in the current model. This factorization though depends on the choice of square-well potentials for the native contact interactions. While a strict factorization is impossible with more common potential shapes such as the Lennard-Jones potential, perhaps approximate factorizations can illuminate similar geometrical features.

3. Entropy and geometrical information

As previously noted, $S(Q)$ is a purely geometrical quantity, independent of temperature. Its value depends only on the available configuration space D_{config} at a certain nativeness Q . Using this quantity we can make a simple argument to show whether a peptide with a particular native contact map will

FIG. 13. The Q dependence of free energy F for β -hairpin $\mathcal{M}_{4B}$ (top row) and helix $\mathcal{M}_{6A}$ (bottom row) for $\sigma_R = 0 \text{ \AA}$ (left) and $\sigma_R = 4 \text{ \AA}$ (right). These free energy profiles indicate that, under the present model isolated secondary structure has only a single stable ensemble that moves smoothly (barrierlessly) with temperature. As the temperature increases, the minimum point $Q = Q_0$ decreases and corresponds to the energy, $E(T) = -\epsilon M Q_0(T)$. The gap between the free energies at $Q = 0$ in the second column is caused by differences in $S(Q = 0)$ due to excluded volume. The middle temperature in each case corresponds to the peak in C_V .

be stabilized or destabilized upon the addition of excluded volume.

The stability of the peptide is measured by the energy

$$\tilde{E}(S(T), Q) = -\epsilon M Q(T), \quad (45)$$

derived from Eq. (38). Then, from Eqs. (38), (28), and (32), a relationship between the T dependence and the Q dependence can be shown. $Q(T)$ is obtained by solving

$$\left(\frac{\partial S(Q)}{\partial Q} \right) = -\frac{\epsilon M}{T}, \quad (46)$$

that is, $Q(T)$, and therefore $E(T)$, is implicitly dependent on the gradient of $S(Q)$. Differentiating Eq. (46) with respect to T ,

$$\frac{dQ(T)}{dT} \left(\frac{d^2 S(Q)}{dQ^2} \right) = \frac{\epsilon M}{T^2}. \quad (47)$$

Noting that $\epsilon, T > 0$ and $dQ(T)/dT < 0$ gives $S'(Q), S''(Q) < 0$. This means $S(Q)$ is an always decreasing concave down function. Remember that the cooperativity is proportional to $dQ(T)/dT$. At a fixed T , a small second derivative $|S''(Q)|$ gives a large gradient $dQ(T)/dT$. The cooperativity is inversely proportional to $S''(Q)$.

Consider the quantity $\delta S(Q) \equiv S(Q; \sigma_R = 0 \text{ \AA}) - S(Q; \sigma_R = 4 \text{ \AA})$. If $\delta S'(Q)$ is positive, $S'(Q; \sigma_R = 0 \text{ \AA}) > S'(Q; \sigma_R = 4 \text{ \AA})$. Using Eq. (46) this implies that, at a fixed $T = \tilde{T}$, $Q(\tilde{T}, \sigma_R = 0 \text{ \AA}) > Q(\tilde{T}, \sigma_R = 4 \text{ \AA})$ and therefore $E(\tilde{T}, \sigma_R = 0 \text{ \AA}) < E(\tilde{T}, \sigma_R = 4 \text{ \AA})$. Since lower energy means greater stability, $\delta S'(Q) > 0$ means the protein is destabilized upon the addition of excluded volume and $\delta S'(Q) < 0$ means the protein is stabilized.

Figure 14 shows $S(Q)$ and $\delta S(Q)$ for contact maps $\mathcal{M}_S$ and $\mathcal{M}_D$. $\delta S'(Q)$ is always positive for $\mathcal{M}_S$ which means

FIG. 14. Entropy dependence on contact map. (a) Entropy dependence on Q for sparse $\mathcal{M}_S$ and dense $\mathcal{M}_D$ contact maps. The solid and dotted lines corresponds to the case with and without excluded volume. As expected, the entropy decreases as more native contacts are formed. (b) Plots the difference between the entropy with and without excluded volume, $\delta S(Q) = S_{\sigma=0}(Q) - S_{\sigma=4}(Q)$. The difference between $\mathcal{M}_S$ and $\mathcal{M}_D$ arises from the excluded volume affecting D_{ns} relatively less for $\mathcal{M}_D$. A dense contact map is highly constrained and therefore already prevented from visiting many possible configurations in D_{ex} .

that addition of excluded volume into the sparse map destabilizes the protein for all Q . In $\mathcal{M}_D$, $\delta S(Q)$ has a maximum at $Q = 0.7$ which means the derivative changes sign. The native ensemble ($Q > 0.7$) is stabilized upon addition of excluded volume since $\delta S'(Q) < 0$. For $Q < 0.7$, $\delta S'(Q) > 0$, and $\mathcal{M}_D$ becomes destabilized with excluded volume as was $\mathcal{M}_S$. The change in sign means that the curves $E(T, \sigma_R = 0 \text{ \AA})$ and $E(T, \sigma_R = 4 \text{ \AA})$ must intersect at a temperature T_1 that satisfies $Q(T_1) = 0.7$ (see Fig. 6).

An explanation for the change in sign of $\delta S'(Q)$ is obtained by considering the meaning of $\delta S(Q)$. $\delta S(Q)$ measures the overlap between D_{config} and D_{ex} . When $\delta S(Q)$ is large the excluded volume has a relatively larger effect than if $\delta S(Q)$ is small. As Q increases the protein becomes more compact, the radius of gyration decreases. A more compact protein has more opportunity for steric clashes and therefore $\delta S(Q)$ increases. In the sparse map $\mathcal{M}_S$, $\delta S(Q)$ increases all the way to $Q = 1$. In the dense map $\mathcal{M}_D$ though, the excluded volume does not have as large an effect near the native state since most of the residues are constrained to be interacting through a native contact. The small overlap between D_{ex} and D_{ns} gives $\delta S'(Q) < 0$ near the native state. The switch over point is at $Q = 0.7$, excluded volume stabilizing configurations with $Q > 0.7$ and destabilizing those with $Q < 0.7$.

IV. CONCLUSIONS

Much effort has been spent on studying the interplay between geometric and energetic effects during protein folding. While geometrical information is usually explicit in numerical simulations, analytical approaches tend to treat protein geometry in an average way, focusing instead on energetic terms. Here, we characterize a protein model that accounts for protein geometry without the need for phenomenological descriptions. The peptide is modeled by a Gaussian chain and a structure-based Hamiltonian. The resulting partition function can be factorized into an energetic part and a geometrical part and it is shown that the folding mechanism is completely determined by the geometrical part. The energetic part is modulated by a many-body interaction in the contact potential. The geometrical part is explored with different contact maps and by the introduction of excluded volume through a

non-specific repulsive core. Excluded volume is capable of changing the folding mechanism, i.e., change the order with which the contacts form during folding, but only slightly. The property of factorability exposes the origin of robust geometric features in the current model and connects with results from larger protein folding simulations.

The cooperativity of the folding transition is increased by both a many-body interaction and an excluded volume interaction. Interestingly, in order to capture the same level of cooperativity imparted by a typical excluded volume, the many-body effect needs to contain $> 50\%$ of the stabilizing energy (Figure 5). In our earlier model that neglected excluded volume, a small many-body effect ($\sim 10\%$ stabilizing energy) was needed to fit with the simulation data. These results show that an excluded volume effect more than compensates for the cooperativity gap in the previous model. This allows native specificity to be relaxed, while not losing cooperativity. Here, the width δ of the native contacts is twice the size of the previous model. Also, it was shown that removing a subset of native contacts reduced cooperativity, but that the cooperativity can be restored by including a non-specific excluded volume interaction. Excluded volume is seen to be an important component of a protein model.

The techniques developed in this study will allow closer comparison with other analytical approaches to protein folding. Through a Legendre transformation, we are able to study the dependence of the thermodynamical functions on Q , an order parameter often used in free energy functional models. Since the presented model contains the full density of states, the effects of approximations introduced through reference Hamiltonians and mean field theories can be directly probed. Applying this model to larger protein systems will undoubtedly require some approximations, and a fuller understanding of the consequences of the approximations should prove useful going forward.

ACKNOWLEDGMENTS

We would like to thank Swarnendu Tripathi for useful discussion of the variational model. This work was supported by the Center for Theoretical Biological Physics sponsored by the NSF (Grant PHY-0822283) and by NSF-MCB-1051438. Y.S. has been supported by Japan Society for the Promotion of Science, Grant-in-Aid for Scientific Research (C) 20540398. J.K.N. acknowledges support from National Institutes of Health (NIH) Molecular Biophysics Training Program, Grant Number: T32GM08326.

¹J. N. Onuchic and P. G. Wolynes, *Curr. Opin. Struct. Biol.* **14**, 70 (2004).

²J. N. Onuchic, H. Nymeyer, A. E. Garcia, J. Chahine, and N. D. Socci, *Adv. Protein Chem.* **53**, 87 (2000).

³J. D. Bryngelson and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **84**, 7524 (1987).

- ⁴P. E. Leopold, M. Montal, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **89**, 8721 (1992).
- ⁵J. E. Shea and C. L. Brooks III, *Annu. Rev. Phys. Chem.* **52**, 499 (2001).
- ⁶H. Taketomi, Y. Ueda, and N. Go, *Int. J. Pept. Protein Res.* **7**, 445 (1975).
- ⁷C. Clementi, H. Nymeyer, and J. N. Onuchic, *J. Mol. Biol.* **298**, 937 (2000).
- ⁸P. C. Whitford, J. K. Noel, S. Gosavi, A. Schug, K. Y. Sanbonmatsu, and J. N. Onuchic, *Proteins* **75**, 430 (2009).
- ⁹L. L. Chavez, J. N. Onuchic, and C. Clementi, *J. Am. Chem. Soc.* **126**, 8426 (2004).
- ¹⁰M. Kouza, M. S. Li, E. P. O'Brien, Jr., C. K. Hu, and D. Thirumalai, *J. Phys. Chem. A* **110**, 671 (2006).
- ¹¹C. Clementi, A. E. García, and J. N. Onuchic, *J. Mol. Biol.* **326**, 933 (2003).
- ¹²S. Gosavi, P. C. Whitford, P. A. Jennings, and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **105**, 10384 (2008).
- ¹³F. Liu, D. Du, A. A. Fuller, J. E. Davoren, P. Wipf, J. W. Kelly, and M. Gruebele, *Proc. Natl. Acad. Sci. USA* **105**, 2369 (2008).
- ¹⁴B. A. Shoemaker, J. Wang, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **94**, 777 (1997).
- ¹⁵B. A. Shoemaker, J. Wang, and P. G. Wolynes, *J. Mol. Biol.* **287**, 675 (1999).
- ¹⁶S. S. Plotkin and J. N. Onuchic, *Q. Rev. Biophys.* **35**, 111 (2002).
- ¹⁷S. S. Plotkin and J. N. Onuchic, *Q. Rev. Biophys.* **35**, 205 (2002).
- ¹⁸S. S. Plotkin and J. N. Onuchic, *Proc. Natl. Acad. Sci. USA* **97**, 6509 (2000).
- ¹⁹S. S. Plotkin and J. N. Onuchic, *J. Chem. Phys.* **116**, 5263 (2002).
- ²⁰Y. Suzuki and J. N. Onuchic, *J. Phys. Chem. B* **109**, 16503 (2005).
- ²¹J. D. Bryngelson and D. Thirumalai, *Phys. Rev. Lett.* **76**, 542 (1996).
- ²²S. Takada and P. G. Wolynes, *Phys. Rev. E* **55**, 4562 (1997).
- ²³M. Sasai and P. G. Wolynes, *Phys. Rev. Lett.* **65**, 2740 (1990).
- ²⁴M. Sasai and P. G. Wolynes, *Phys. Rev. A* **46**, 7979 (1992).
- ²⁵J. J. Portman, S. Takada, and P. G. Wolynes, *Phys. Rev. Lett.* **81**, 5237 (1998).
- ²⁶J. J. Portman, S. Takada, and P. G. Wolynes, *J. Chem. Phys.* **114**, 5069 (2001).
- ²⁷J. J. Portman, S. Takada, and P. G. Wolynes, *J. Chem. Phys.* **114**, 5082 (2001).
- ²⁸X. Qi and J. J. Portman, *Proc. Natl. Acad. Sci. USA* **104**, 10841 (2007).
- ²⁹T. Shen, C. Zong, J. J. Portman, and P. G. Wolynes, *J. Phys. Chem. B* **112**, 6074 (2008).
- ³⁰S. Tripathi and J. J. Portman, *J. Chem. Phys.* **128**, 205104 (2009).
- ³¹S. Tripathi and J. J. Portman, *Proc. Natl. Acad. Sci. USA* **106**, 2104 (2009).
- ³²Y. Suzuki, J. K. Noel, and J. N. Onuchic, *J. Chem. Phys.* **128**, 025101 (2008).
- ³³P. J. Flory, *Statistical Mechanics of Chain Molecules* (Wiley, New York, 1969).
- ³⁴V. Munoz and W. A. Eaton, *Proc. Natl. Acad. Sci. USA* **96**, 11311 (1999).
- ³⁵E. Alm, A. V. Morozov, T. Kortemme, and D. Baker, *J. Mol. Biol.* **322**, 463 (2002).
- ³⁶R. K. Pathria, *Statistical Mechanics* 2nd ed. (Butterworth-Heinemann, Oxford, 1996).
- ³⁷H. X. Zhou, *Proc. Natl. Acad. Sci. USA* **99**, 3569 (2002).
- ³⁸H. Lammert, A. Schug, and J. N. Onuchic, *Proteins* **77**, 881 (2009).
- ³⁹G. A. Papoian, J. Ulander, M. P. Eastwood, Z. Luthey-Schulten and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **101**, 3352 (2004).
- ⁴⁰L. C. Oliveira, A. Schug, and J. N. Onuchic, *J. Phys. Chem. B* **112**, 6131 (2008).
- ⁴¹M. R. Ejtehadi, S. P. Avall, and S. S. Plotkin, *Proc. Natl. Acad. Sci. USA* **101**, 15088 (2004).
- ⁴²M. Eastwood, P. G. Wolynes, *J. Chem. Phys.* **114**, 4702 (2001).
- ⁴³J. K. Noel, P. C. Whitford, K. Y. Sanbonmatsu, and J. N. Onuchic, *Nucleic Acids Res.* **38**, 657 (2010).