Designability and cooperative folding in a four-letter hydrophobic-polar model of proteins

Hai-guang Liu[1,2] and Lei-Han Tang[1]

[1]*Department of Physics, Hong Kong Baptist University, Kowloon Tong, Kowloon, Hong Kong*

[2]*UC Davis Genome Center and Department of Applied Science, University of California, Davis, California 95616-8254, USA*

(Received 27 February 2006; revised manuscript received 25 July 2006; published 29 November 2006)

The two-letter hydrophobic-polar (HP) model of Lau and Dill [Macromolecules **22**, 3986 (1989)] has been widely used in theoretical studies of protein folding due to its conceptual and computational simplicity. Despite its success in elucidating various aspects of the sequence-structure relationship, thermodynamic behavior of the model is not in agreement with a sharp two-state folding transition of many single-domain proteins. To gain a better understanding of this discrepancy, we consider an extension of the HP model by including an "antiferromagnetic" (AF) interaction in the contact potential that favors amino acid residues with complementary attributes. With an enlarged four-letter alphabet, the density of states on the low energy side can be significantly decreased. Computational studies of the four-letter HP model are performed on 36-mer sequences on a square lattice. It is found that the designability of folded structures in the extended model exhibits strong correlation with that of the two-letter HP model, while the AF interaction alone selects a very different class of structures that resembles the Greek key motif for beta sheets. A procedure is introduced to select sequences which have the largest energy gap to the native state. Based on density of states and specific heat calculations in the full configuration space, we show that the optimized sequence is able to fold nearly as cooperatively as a corresponding Gō model.

DOI: 10.1103/PhysRevE.74.051918 PACS number(s): 87.15.Cc, 05.50.+q, 82.35.Rs, 87.15.Aa

I. INTRODUCTION

Most proteins in normal physiological conditions have a definitive three-dimensional structure that supports their specific biological function. Denaturation (or unfolding) of this native state can be induced by a rise in temperature, variation in *p*H or salt concentration, or exposure to a variety of denaturants such as urea [1]. Experimental studies of the folding transition of short globular proteins all indicate a two-state transition. The rate of the transition, however, varies greatly among different proteins [2–5].

From a theoretical point of view, the sharpness and often two-state kinetics of the thermodynamic transition is quite a mystery considering the short chain length and the heterogeneous composition [6–8]. It is generally believed that, out of the exponentially large number of possible amino acid sequences, real proteins gained their *atypical* cooperativity in folding through evolution and natural selection [9]. Microscopically, attempts have been made to attribute cooperativity to side-chain packing, hydrogen bonding, formation of a nucleus, etc. [10].

The frequent occurrence of alpha helices and beta sheets in real proteins suggests that organization at the intermediate level may be Nature's solution to the folding complexity of a heteropolymer [11,12]. Indeed, when amino acid residues are grouped into these structural elements, the available conformation space is greatly reduced. The resulting free energy landscape is much simpler, allowing folding to proceed in a more orderly fashion. While this picture is conceptually appealing, interaction between the twenty amino acid residues in an aqueous environment is sufficiently rich and complex so that many different possibilities still exist for the organization to take place in a sequence-specific manner [13,14]. Existence of a limited number of protein structure "families" has reinforced the belief that such organization may be governed by certain generic principles. Recent attempts at uncovering such principles based on mutational stability [15] or geometry of peptide chains [16] have met with some success, though many details remain controversial [17].

The well-known computational difficulty associated with "all-atom" simulation of even the smallest proteins underlies the continuing interest in heteropolymers on a lattice that contain certain key features of the protein folding problem [16,18–21]. The much reduced degree of freedom of such systems allow one to carry out a thorough statistical mechanical study of the folding transition, and to explore how interactions of different physical origin work together to produce "protein-like" behavior, one of which being the sharp two-state transition mentioned above. Studies of the two-letter HP (hydrophobic and polar residues) model on a lattice [22] have shown that cooperativity of the folding transition can be enhanced in structures (and corresponding amino acid sequences) that are highly "designable" [23]. Designability is a concept used by Li *et al.* [15,24] to capture the mutational stability of a folded structure under generic interactions. Therefore, the observation that the most designable structures selected by hydrophobicity also fold more cooperatively provides extra appeal to the idea. It is, however, clear that a truly two-state transition requires presence of other interactions [25]. For example, amino acid side chains need to be properly packed in the hydrophobic core of a protein to maintain a sufficiently compact structure [26]. This imposes certain selections on the size of side chains in order to optimize the van der Waals interactions, as demonstrated in the stability study of T4 lysozyme and its mutants [27]. Hydrogen bonding is another significant factor, responsible for the formation and stability of alpha helices and beta sheets [28,29]. The additional specificity acquired through these interactions is needed in order to suppress the population of collapsed but non-native protein chain conformations which form the folding intermediates in the HP model, so that a highly cooperative folding transition can be achieved.

To bring these ideas to quantitative verification, we report in this paper results from numerical investigations of a two-dimensional lattice protein model with four-letter sequences. The expanded alphabet allows us to code additional, sequence-specific contact interactions which are absent in the two-letter HP model. Physically, such interactions originate from the steric repulsion of amino acids of varying size, or the donor/acceptor parity of hydrogen bonding. We, therefore, assume the contact interaction to be of an antiferromagnetic (AF) type, so that amino acids with complementary properties can gain energy when they are nearest neighbors on the lattice.

It has been suggested that real proteins are minimally frustrated so as to achieve a "funnel landscape" for fast folding [11,12]. Yet minimal frustration, as in a homopolymer, does not guarantee cooperative folding. With the enlarged sequence space, one may explore emergence of organization through "fine-tuning" of the sequence, both at the structural level and at the density of states level. In particular, one may examine whether the additional interaction serves to strengthen certain local motifs as in alpha helices, or the matching of distant segments along the chain as in beta sheets. Although the lattice model considered here offers only a very limited set of structural alternatives as compared to real proteins, it may nevertheless give us a glimpse of the origin of specificity in the contact map of monomers without having to assume it from the outset as in the Gō-like models [30].

The paper is organized as follows. In Sec. II we introduce the four-letter lattice protein model and discuss its basic symmetries. Section III presents exact and Monte Carlo calculations of the designability of all compact structures on a 6×6 lattice for three representative interactions: HP-only, AF-only, and the mixed case where HP and AF interactions assume equal strength. A comparison of top designable structures in the three cases is given. Section IV contains simulation results on the thermodynamics of the folding transition for several representative sequences that share the same native state. The cooperativity of the transition is analyzed and compared with two previously studied cases, the HP-only case and the Gō model. Discussion of our findings and conclusions are given in Sec. V.

II. THE MODEL

The four-letter lattice protein model considered in this paper is an extension of the HP model first introduced by Lau and Dill for theoretical analysis of the sequence-structure relationship of proteins [22]. The 20 amino acid residues found in natural proteins can be divided into two groups according to their affinity to water: H-type if they are hydrophobic and P-type if they are hydrophilic or polar. Due to the large surface to volume ratio, the hydrophobicity of amino acid residues provides the main thermodynamic driving force for protein folding, which enables shielding of a hydrophobic core by polar residues on the surface. In the HP model, a polypeptide chain is represented by a self-avoiding lattice polymer of hydrophobic (H) and polar (P) beads. Two monomers on the chain are said to form a contact if they occupy neighboring sites on the lattice but are not adjacent on the chain. The energy of a given polymer configuration is expressed in terms of the number of HH, PP, and HP contacts $n_{\rm HH}$, $n_{\rm PP}$, and $n_{\rm HP}$, respectively,

$$E_{\rm HP}=n_{\rm HH}\epsilon_{\rm HH}+n_{\rm PP}\epsilon_{\rm PP}+n_{\rm HP}\epsilon_{\rm HP}. \quad (1)$$

Here $\epsilon_{\rm HH},\epsilon_{\rm PP}$, and $\epsilon_{\rm HP}$ are the energy parameters for the respective contacts. In the special case

$$\epsilon_{\rm HP}=(\epsilon_{\rm HH}+\epsilon_{\rm PP})/2, \quad (2)$$

one arrives at the solvation model where the energy (1) is obtained alternatively from the number of exposed faces of hydrophobic beads in the lattice configuration plus an overall energy for condensation [31].

The HP interaction alone has limited ability to differentiate structures with different core configurations. Hence, to base the study of protein folding entirely on hydrophobic properties of amino acids neglects certain key aspects of the problem [13]. Real proteins are organized through the formation of secondary structures that distinguish them from a randomly collapsed polypeptide. The importance of additional interactions in stabilizing secondary structures has been demonstrated in a statistical analysis of environment-specific amino acid contacts [32]. One of the challenges for modelers is to understand how such organizational units emerge under simple yet realistic interactions.

Here we extend the HP model by introducing a second attribute, the "spin" $s_i=0,1$, to each monomer on the chain, expanding the alphabet to four: (H,0), (H,1), (P,0), and (P,1). Two monomers in contact acquire an energy u_0 if they carry the same spin, and $u_1(<u_0)$ if they carry different spins. The total energy of a given self-avoiding polymer configuration $\{\mathbf{r}_i\}$ can then be written as

$$E_{\rm mix}=E_{\rm HP}+E_{\rm AF}, \quad (3)$$

where

$$E_{\rm AF}=\sum_{i\neq j,j\pm1}[u_0+(u_1-u_0)(s_i+s_j-2s_is_j)]\Delta(|\mathbf{r}_i-\mathbf{r}_j|) \quad (4)$$

is the energy due to an "antiferromagnetic" interaction. Here, $\Delta(r)=1$ for $r=a$ (lattice constant) and $\Delta(r)=0$ otherwise, restricting the interaction to nearest neighbor monomers only. In all numerical calculations reported below, we take $\epsilon_{\rm HH}=-2$, $\epsilon_{\rm HP}=-1$, $\epsilon_{\rm PP}=0$, and $u_1=-1$. The interaction energy of spins of the same type is set to $u_0=0$, unless specified otherwise.

The AF interaction favors matching of monomers with complementary attributes in the folded structure. Both steric repulsion and hydrogen bonding between amino acid residues have the complementary character as in our model and, hence, can be considered as the physical interactions we try to capture in the simplified lattice model. Naturally, there are also apparent differences in each case when examined in detail. For example, intramolecular hydrogen bonding is of the donor and acceptor type, hence each residue seldom has more than one partner at a time. The steric repulsion is less selective, yet part of the effect can be absorbed by reorientation of the side chains and by elastic deformations. These complications are likely to be important to the folding of real proteins. Nevertheless, statistics performed on structures kept at the Protein Data Bank (PDB) have shown that amino

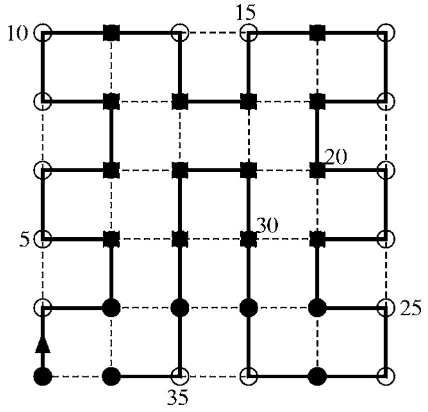

FIG. 1. Contacting monomers (linked by dashed lines) on a lattice walk (solid line) can be grouped into clusters. The two largest contact clusters are indicated by filled circles and squares, respectively. Arrow at the lower-left corner indicates the direction along which monomers are numbered along the chain.

acids have strong preference for their neighbors [33,34]. Therefore, study of the simple lattice model may still furnish useful insights.

The energy E_{AF} has a global symmetry with respect to the parity transformation $s_i \to 1-s_i$, all i. In fact, in a given lattice polymer conformation, contacting monomers form several linked clusters as shown in Fig. 1. Each such cluster is a chain of monomers connected by dashed lines. For monomers in a given cluster, the AF bonding occurs exclusively within the cluster. Therefore, the parity transformation can be applied individually to any given cluster without affecting the AF energy of the system.

III. COMPACT LATTICE WALKS

A. Designability under three different energy functions

In an attempt to understand the evolutionary stability of native protein structures, Li, Tang, and Wingreen [15,24] introduced the concept of designability for any given chain conformation. In the case of compact lattice walks, the designability of a structure is defined to be the number of sequences that choose the structure as their unique ground state. Conformations ranked high in designability are more robust against mutations once selected as the ground state, and the corresponding optimal sequences fold more cooperatively [23]. Recent work that extends the designability concept to the more general Miyazawa-Jernigan interactions [35] concluded that the main features of the HP model survive [36], although a completely different set of structures may emerge when the HP interaction is replaced by a random contact interaction [37,38].

In the following we present results of our numerical study of the designability of all self-avoiding compact walks on a 6×6 lattice under the energy function (3). For comparison, we have performed the same calculation using the energy functions (1) and (4) for the HP-only and AF-only interactions, respectively. The designability of structures under the HP energy function has been analyzed previously by Li *et al.* [15,24].

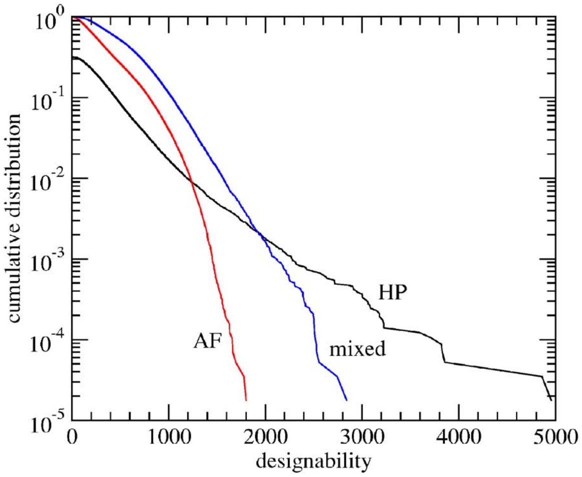

FIG. 2. (Color online) Cumulative distribution of designability of the 57 337 structures obtained from random sampling of 84 000 000 sequences using three different energy functions.

On a 6×6 lattice, there are 57 337 possible self-avoiding 36-mer chain conformations not related by rotation or reflection. They can be constructed using an exact enumeration algorithm as reported earlier by Li *et al.* Since the number of possible four-letter sequences increases as 4^N with the chain length N, it is impractical to compute the designability by going through all possible sequences at $N=36$. Instead, the designability can be calculated from uniform sampling of the sequence space. For each four-letter sequence generated, we compute E_{HP}, E_{AF}, and E_{mix} for all the compact structures and record, for each energy function, the ground state energy, and its degeneracy. Let M be the total number of sequences generated. The designability of a given structure, s, is estimated by the number of times, D_s, that s is selected as the unique ground state among all possible structures. Elementary calculations show that D_s follows the Poisson distribution,

$$\mathrm{Prob}(D_s = D) = \frac{\lambda_s^D}{D!} e^{-\lambda_s}, \qquad (5)$$

where $\lambda_s = M\rho_s$, and ρ_s is the density of sequences that choose s as the unique ground state in sequence space. It then follows that the average $\langle D_s \rangle = \lambda_s$ and the variance $\langle D_s^2 \rangle - \langle D_s \rangle^2 = \lambda_s$. Hence, the statistical error of D_s measured using uniform sampling is of the order of $D_s^{1/2}$.

The results presented below are obtained from a simulation with $M=84\,000\,000$ random sequences. For the HP-only case, the fraction of sequences with a unique ground state is 8.82%. This increases to 25.2% for the AF-only case, and 39% in the mixed case. The cumulative distributions of the designability of all structures are shown in Fig. 2. In all three cases, the distribution decays roughly exponentially with increasing D. The HP-only case has the largest spread of D, and also only 18 213 (32%) structures have a nonzero D. In comparison, the range of D under the AF interaction is much narrower. In particular, the change in D for the top 1% structures is within 1.5 fold, compared to the fourfold change under the HP energy function.

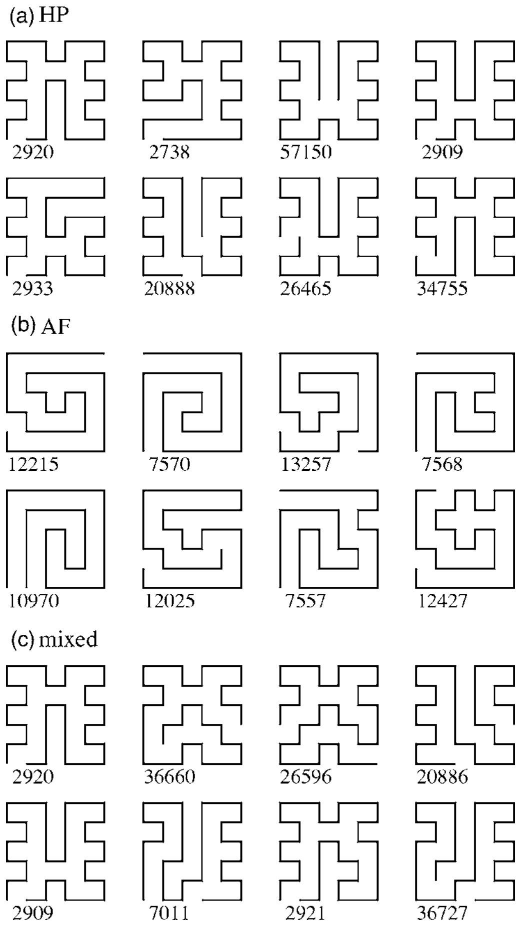

FIG. 3. Top eight, nonsymmetry-related structures with the highest designability, arranged from left to right and continued in the second row in each case: (a) HP interaction; (b) AF interaction; (c) mixed interaction.

Figure 3 shows the top eight, nonsymmetry-related structures in designability under the three energy functions. The structures are arranged from left to right and continued in the second row in decreasing order. Each structure is identified by a serial number used in our exact enumeration, which appears at the lower left corner of each diagram. (They will be referred to with a prefix "S" in the text.) In each of the three cases, one or more distinct local features are selected in the top group of structures. For example, the HP interaction favors "in-and-out" patterns on the surface [39]. The presence of a "thick" surface layer leaves little room for alternative core configurations on the 6×6 lattice. In contrast, the AF interaction singles out structures that contain long, antiparallel running edges that resemble suspiciously the Greek key motif of beta sheets [1]. As is well known, beta sheets are stabilized by alternating hydrogen bonds along neighboring peptide chains. Our result suggests that this arrangement indeed has a high designability under the AF interaction. In the mixed case, constraining the core configuration with a thick in-and-out surface pattern on the 6×6 lattice remains an effective strategy to achieve high designability as in the HP model.

The overall relationship among the three designabilities for each of the 57 337 structures can be visualized in a scatter plot as shown in Fig. 4. The top designable structures under the HP interaction have only a medium level of designability under the AF interaction (cf., Fig. 2 for the designability scales). Conversely, the top designable structures under the AF interaction belong almost exclusively to the set of degenerate structures under the HP interaction. The most designable structures under the mixed interaction, on the other hand, shows strong correlation with HP designability, but a much weaker correlation with AF designability (data not shown).

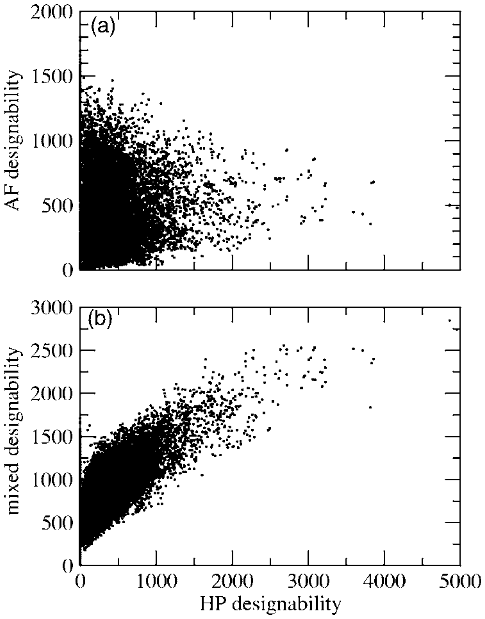

FIG. 4. Comparison of designability under different interactions for all 57 337 structures. (a) HP against AF; (b) HP against mixed.

B. Density of states and the energy gap

As noted previously, the sequence that minimizes E_{HP} of a given structure is unique [24]. For example, the sequence that minimizes the HP energy on S2920 (as shown in Fig. 1) is given by

$$\mathrm{PPHHPPHHPPPPHHPPPPHHPPHHPPPPHHHHHHPP}. \tag{6}$$

Although we are not aware of a general proof, it is plausible that this is also the (or one of the) sequence that has the largest energy gap between its ground state, which is the structure in question, and other compact structures.

A different situation arises in the mixed energy model. The parity transformation ensures that there is a large number of sequences that share the same lowest AF energy on any given structure. For S2920, there are 11 contact clusters as shown in Fig. 1, and, hence, from any given AF sequence, one can construct a total of 2048 AF sequences, all having the same AF interaction energy. An AF sequence that minimizes the AF interaction energy can be obtained by minimizing the energy of all AF bonds as indicated by the dashed lines in Fig. 1. The symmetry mentioned above allows us to generate all 2048 sequences which choose S2920 as their ground state and have the lowest possible AF energy. Fixing the global parity (say, setting $s_1=0$), we are still left with 1024 sequences to consider.

Even though the 1024 AF sequences share a common ground state and the same ground state energy, their energies on structures other than S2920 are different. To gain some intuition on which property of the sequence is important in this discussion, we have compared the excitation spectrum of these sequences [with the HP part given by (6)] under the mixed energy model (3) in the compact walk space. We observe that the relative "phase" of the two largest contact clusters (filled circles and squares) in Fig. 1 is the most important quantity that determines the gap size ΔE to the ground state. When the two clusters are "in-phase" (i.e., monomers 3 and 4 have different spin), ΔE lies between 2 and 6. On the other hand, when the two clusters are "out-of-phase", ΔE is between 6 and 8. This suggests that, in the out-of-phase situation, it is much more difficult to form alternative pairing of monomers with complementary properties. The energy gap with $E_{\rm HP}$ alone is 2.

At the largest energy gap $\Delta E=8$, there are 25 AF sequences. One of these sequences is given by

AF857: 0000101010111111000011000111110100101101.

The remaining 24 sequences differ from this one only in the parity of five small contact clusters: monomers 6 and 9, monomer 10, monomers 12 and 15, monomer 17, and monomers 18 and 21. Among the $2^5=32$ possible choices, monomers 6, 10, and 12 are not allowed to all take the same value. In addition, when $s_{15}=0$, we must have $s_{17}=s_{21}=1$. Violation of any of these rules reduces ΔE.

Among the sequences which have the smallest energy gap $\Delta E=2$, the sequence

AF1024: 0101010101010101010101010101010101010101

has the largest number of low energy excited states. In fact, since two contacting monomers on the square lattice is always connected by an odd number of links on the chain, all possible contacts are of the (0,1) type for this sequence. Therefore, this case is identical to the two-letter HP model with $\epsilon_{\rm HH}=-3, \epsilon_{\rm HP}=-2$, and $\epsilon_{\rm PP}=-1$.

In addition to these two sequences, we have selected a few other representative cases for detailed study. Two of these sequences are

AF241: 0001100110000111100110011100010101111,

AF289: 0000101010101010101010101011110000101111.

Sequence AF241 is in the "in-phase" family but has long stretches of 0's and 1's to give it the largest energy gap $\Delta E=6$ within the set. On the other hand, AF289 is in the "out-of-phase" family but has a long stretch of alternating 0 and 1 and the smallest energy gap $\Delta E=6$ in the set.

Figure 5 shows the density of states of the four sequences in the compact walk space. Although they share the same native state S2920 and the same ground state energy $E_0=-57$, there is a big difference in the excitation energy spectrum. This property allows for fine-tuning of the sequence to achieve a more stable native state and a more cooperative folding transition, as we demonstrate below. We have inspected the density of states of a number of other sequences of the 1024 set whose behavior generally lies between two extreme cases AF857 and AF1024.

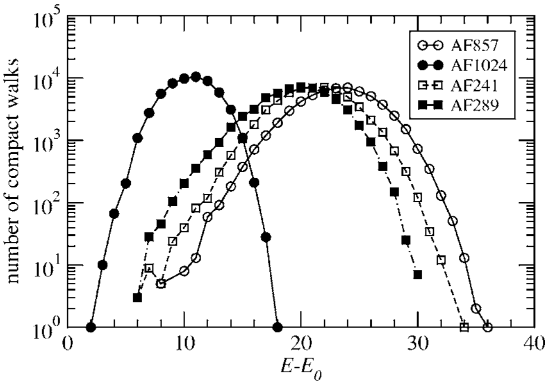

FIG. 5. Density of states in the compact walk space for four sequences all having S2920 as the ground state.

IV. THE FOLDING TRANSITION

The study of compact lattice walks in the previous section shows that certain protein folds (such as S2920) have an exceptionally large mutational stability as a native state. The energy gap between such a structure and other structures can be extended with carefully chosen sequences. To show that the selected sequence indeed exhibits cooperative folding, we performed thermodynamic studies in the full configuration space using a multi canonical Monte Carlo (MCMC) scheme. Since the method has been described in great detail elsewhere [40–43], we shall only give a brief summary of the simulation procedure.

The MCMC (also known as entropic sampling) scheme defers from the usual canonical Monte Carlo (MC) scheme in the sampling weight one assigns to the allowed configurations. Instead of the Boltzmann factor, $\exp(-E/T)$, for MC simulation at a given temperature, T, the sampling weight is chosen to be proportional to $1/\Omega(E)$, where $\Omega(E)$ is the total number of configurations at energy, E, or equivalently the density of states function. Since $\Omega(E)$ is unknown *a priori*, one constructs a series of trial sampling functions $P_k(E)$ that converge to $1/\Omega(E)$. In each iterative step, simulation is performed using $P_k(E)$ to generate an energy histogram, $h_k(E)$. An estimate of $\Omega(E)$ is obtained from $\Omega_k(E)\sim h_k(E)/P_k(E)$. From this estimate one constructs an updated sampling function $P_{k+1}=P_k(E)/h_k(E)$ and repeats the process again. Due to the intrinsically noisy nature of the Monte Carlo data and the lack of information when a certain energy range (usually on the low energy side) is not sampled in the simulation, care must be taken in the implementation of the iterative process [43]. Fortunately for our problem, the number of degrees of freedom is relatively small, and the system is unfrustrated despite the heterogeneous nature of interactions defined by the sequence. Therefore convergence is not an obstacle. To allow the lattice polymer to move more efficiently in the configuration space within the collapsed state, we have implemented a multidimensional MCMC scheme as described in Ref. [44]. As demonstrated by the results pre-

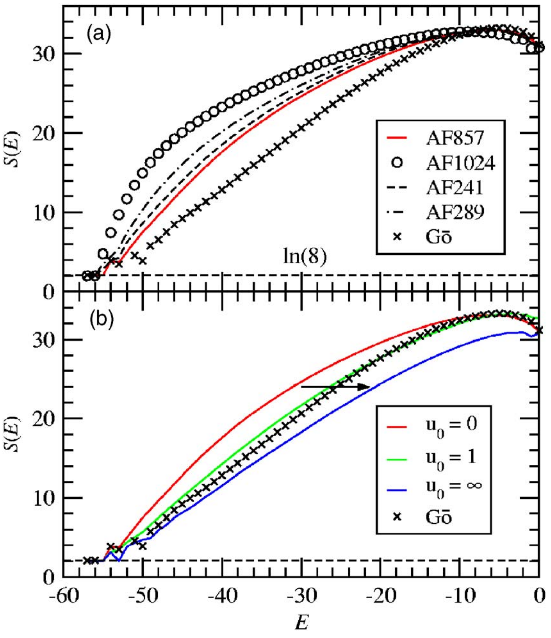

FIG. 6. (Color online) Entropy against energy for the four AF sequences and the Gō model. The exact value of the ground state entropy is indicated by the horizontal line. (a) $u_0=0$; (b) the sequence AF857 at other values of u_0, in increasing order along the arrow.

sented below, the density of states calculated in this scheme is extremely accurate.

In addition to the four AF sequences introduced above, we have also simulated a Gō model defined with reference to S2920. Here E_{AF} in Eq. (3) is replaced by

$$E_{\mathrm{G\bar{o}}} = - \sum_{i\neq j, j\pm 1} \delta_{i,j}^{(n)} \Delta(r_{ij}), \tag{7}$$

where $\delta_{i,j}^{(n)}=1$ if monomers i and j are in contact in S2920, and $\delta_{i,j}^{(n)}=0$ otherwise. The HP interaction is kept as in the mixed model.

Figure 6 shows the entropy function $S(E)=\ln \Omega(E)$ obtained from our simulations. From an independent study of self-avoiding walks (SAW) on the square lattice, we obtained an estimate of the total number of self-avoiding 36-mer configurations with a fixed end, $\Omega_0=\exp(S_0)$, where $S_0 = 35.351\pm 0.002$. This is used to calibrate the absolute value of entropy. From the figure we see that the ground state entropies obtained from the simulations agree very well with the exact result ln 8, where 8 is the ground state degeneracy from the fourfold rotation and twofold mirror reflection of S2920.

As shown in Fig. 6(a), the trend seen in Fig. 5 from the study of compact walks is well-preserved in the full configuration space. In the case of the optimized sequence AF857, the density of states on the low energy side is significantly reduced as compared to AF1024, which is equivalent to the HP model. The unfolding temperature T_f in each case can be estimated graphically from the $S(E)$ curve using a standard procedure [19]. Starting from the point representing the ground state [$E=-57$ and $S=\ln(8)$ in all five cases shown in Fig. 6(a)], we draw a straight line tangential to the $S(E)$ curve. The slope of this line is given by $1/T_f$. It is evident from the figure that AF1024 has the lowest T_f and also the smallest energy difference (or heat of denaturation) between the ground state and the denatured state just above the transition, while the Gō model has the highest transition temperature and the largest heat of denaturation. The transition temperature and the heat of denaturation for other AF sequences lie between these two limits.

We have also investigated the effect of a repulsive contact energy $u_0>0$ for monomers of the same type on the density of states. As seen in Fig. 6(b), even at a moderate $u_0=1$, a behavior similar to the Gō model is obtained. The improvement in cooperativity is easy to understand. Since the ground state is unfrustrated, increasing u_0 does not affect its energy. However, a large number of conformations with noncomplementary contacts have increased their energy. Consequently, the pathway(s) leading to the native state becomes more pronounced and specific. On the high energy side, since the number of monomer contacts in the extended conformations is small, increasing u_0 does not affect these states significantly.

Once the density of states $\Omega(E)$ is known, one may proceed to calculate the partition function

$$Z(T) = \sum_{E} \Omega(E) \exp(-E/T), \tag{8}$$

and the free energy $F(T)=-T\ln Z(T)$ at temperature T. The internal energy $E(T)$ and the specific heat $c(T)$ can be calculated from F using standard thermodynamic relations. Figure 7 shows $E(T)$ and $c(T)$ for AF857, AF1024, and the Gō model. As expected, the Gō model has the highest transition temperature and the sharpest specific heat peak, while AF1024 has the lowest unfolding temperature and the smallest specific heat peak. From the specific heat data, we see that the unfolding transition of sequence AF857 is much more cooperative than sequence AF1024, with a twofold increase in the peak temperature and a nearly similar amount of increase in peak value. Thus a significant improvement in cooperativity of the transition is achieved through the introduction of the AF interaction. We should mention that, in a perfect two-state transition, the energy (or enthalpy for proteins in an aqueous environment) on the denatured side of the transition should reach its completely unfolded (i.e., a free self-avoiding walk) value [25]. This is not quite the case for the lattice Gō model and much less so for sequence AF857. We suspect that part of this unsatisfactory state of affairs has to do with the two-dimensional geometry used in the present study.

Finally, we present data on the average number of native contacts as a function of energy and temperature of the system, respectively. The number of native contacts B_n is defined by the number of contacts (i,j) in a given configuration which is also present in the native state, which in our case is S2920. It has a maximum value of 25 for a 36-mer lattice protein. In our MCMC calculation, we keep a histogram of the number of configurations visited in the sampling process that have a given value of B_n and E. The average value of B_n

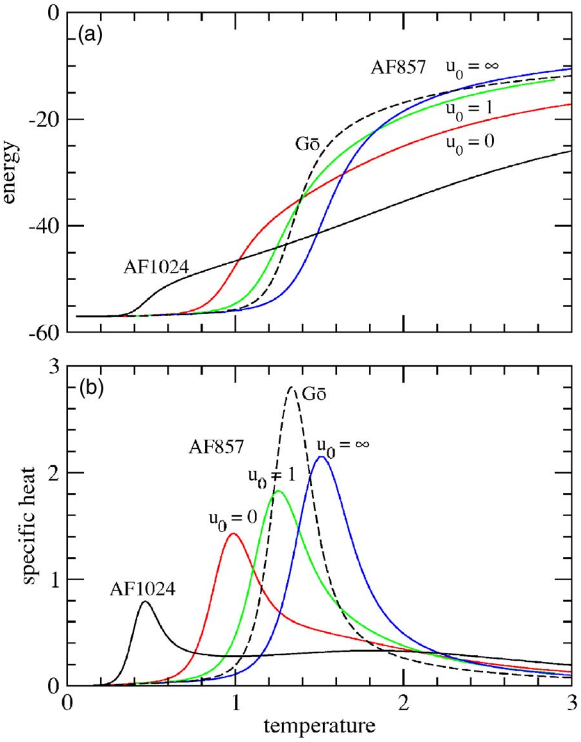

FIG. 7. (Color online) (a) Energy and (b) specific heat against temperature for sequences AF1024 and AF857 (at $u_0=0,1,$ and ∞), as well as the Gō model.

at a given energy is plotted in Fig. 8(a) against E. For the Gō mode, the data fall into two approximately linear regimes joined at the unfolding energy. This is quite reasonable considering the fact that $E_{\mathrm{G\bar{o}}}$ is simply given by $-B_n$. For sequence AF857, this description is still approximately valid, though as u_0 decreases, the number of native contacts drops at a given energy. For AF1024, this trend is accelerated, indicating the presence of many low energy configurations that are different from the native structure.

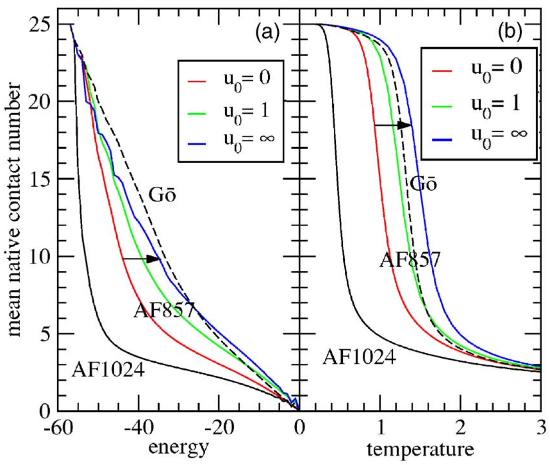

FIG. 8. (Color online) The average number of native pairwise monomer contacts against (a) energy and (b) temperature for sequences AF1024 and AF857 (at $u_0=0,1,$ and ∞, following the arrow direction), as well as the Gō model (dashed line).

Figure 8(b) shows the temperature dependence of average B_n. In all cases, the number of native contacts drops sharply at the unfolding transition.

V. DISCUSSION AND CONCLUSIONS

It has been appreciated for nearly a decade that proteins are minimally frustrated heteropolymers and as such occupy a distinct place in the statistical mechanics of chain molecules [6–13]. Minimal frustration in a given native state can often be achieved through sequence optimization, as in models considered in the present paper. This property alone, however, is not sufficient to guarantee cooperative folding. Real proteins often possess an additional "order" that makes them "protein-like," but the precise meaning of these terms, as well as their relation to the thermodynamics of folding, remain to be clarified.

In the present paper, we examined an extension of the two-letter HP model by including a (bulk) term in the total energy that describes matching of neighboring amino acid residues in a folded structure. We have argued that such interactions arise naturally from hydrogen bonding and steric repulsion between amino acid side chains. They provide additional specificity required to achieve a higher degree of cooperativity in folding transitions as exhibited by real proteins.

With an energy function which gives equal weight to HP and matching (or antiferromagnetic) energies, we have computed the designability of all 57 337 compact lattice walks on a 6×6 lattice. Designability of the mixed model is strongly correlated to that of the two-letter HP model. In particular, structure S2920, which is the most designable in the HP model, also ranks topmost in the mixed case. Under a given energy function, the top structures all share certain structural characteristics. In this respect one may draw analogies with the common secondary structure motifs of real proteins. However, the particular type of structural features selected depends on the type of interactions considered, and is also expected to be sensitive to the geometry and chain length. Interestingly, the highly designable structures under the AF interaction bear striking resemblance to the Greek key motif of beta sheets. This is compared with the alpha helix like "in-and-out" patterns selected by the HP interaction.

Unlike the two-letter HP model, many different four-letter sequences are able to minimize the energy of a given structure in our mixed model. This property has to do with the existence of many subdomains defined by the structure which we call "contact clusters." To attain the minimum energy, the complementary requirement must be fulfilled within each subdomain, but not between subdomains. This extra degrees of freedom allows one to engineer the four-letter sequence to achieve higher cooperativity, as we have demonstrated in Sec. IV. The best sequence in the family indeed exhibits a much improved cooperative folding behavior than the two-letter HP model.

We have also observed a strong correlation between the excitation energy spectra of a given sequence in the compact walk space and in the full configuration space. This property can be very useful in the selection of sequences that fold most cooperatively onto a given structure by computational methods, as the number of compact structures is much smaller than the number of all possible conformations.

Within the premise that solvent environment has the strongest effect on the formation of a compact structure, cooperative folding does require participation of intramolecular interactions. Introduction of matching energies in the extended HP model is a way to incorporate such interactions in the thermodynamic study of the folding transition. Even with a moderate four-letter alphabet, sharpness of the transition can be much improved. We believe similar ideas can be applied in the study of models that employ more realistic geometries [16] for better agreement with the folding energetics and kinetics of real proteins.

ACKNOWLEDGMENTS

L.H.T. would like to thank Chao Tang and Ned Wingreen for many enlightening discussions on protein designability during the early part of the work. Research is supported in part by the Research Grants Council of the Hong Kong SAR under Grant No. 2076/02P, and by the Hong Kong Baptist University (HKBU) under Grant No. FRG/02-03/II-44. Computations were carried out at HKBU's High Performance Cluster Computing Centre supported by Dell and Intel. Support of the Hong Kong University Computer Centre which provided bioinformatics resources through the BIO-SUPPORT project is gratefully acknowledged.

[1] T. E. Creighton, *Proteins*, 2nd edition (W. H. Freeman, New York, 1993).

[2] S. E. Jackson, Folding Des. **3**, R81 (1998).

[3] W. A. Eaton, V. Munoz, S. J. Hagen, G. S. Jas, L. J. Lapidus, E. R. Henry, and J. Hofrichter, Annu. Rev. Biophys. Biomol. Struct. **29**, 327 (2000).

[4] K. L. Maxwell *et al.* Protein Sci. **14**, 602 (2005).

[5] A. Akmal and V. Munoz, Proteins **57**, 142 (2004).

[6] A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules* (AIP, New York, 1994).

[7] T. Garel, H. Orland, and E. Pitard, in *Spin Glasses and Random Fields*, edited by A. P. Young (World Scientific, Singapore, 1998).

[8] H. S. Chan, S. Shimizu, and H. Kaya, Methods Enzymol. **380**, 350 (2004).

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

[10] L. Mirny and E. Shakhnovich, Annu. Rev. Biophys. Biomol. Struct. **30**, 361 (2001).

[11] P. G. Wolynes, J. N. Onuchic, and D. Thirumalai, Science **267**, 1619 (1995); J. N. Onuchic, P. G. Wolynes, Z. Luthey-Schulten, and N. D. Socci, Proc. Natl. Acad. Sci. U.S.A. **92**, 3626 (1995).

[12] K. A. Dill and H. S. Chan, Nat. Struct. Biol. **4**, 10 (1997).

[13] O. Schueler-Furman, C. Wang, P. Bradley, K. Misura, and D. Baker, Science **310**, 638 (2005).

[14] W. W. Chen and E. I. Shakhnovich, Protein Sci. **14**, 1741 (2005).

[15] H. Li, C. Tang, and N. S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).

[16] J. R. Banavar and A. Maritan, Rev. Mod. Phys. **75**, 23 (2003); T. X. Hoang, A. Trovato, F. Seno, J. R. Banavar, and A. Maritan, Proc. Natl. Acad. Sci. U.S.A. **101**, 7960 (2004); J. R. Banavar, M. Cieplak, and A. Maritan, Phys. Rev. Lett. **93**, 238101 (2004); J. R. Banavar, M. Cieplak, A. Flammini, T. X. Hoang, R. D. Kamien, T. Lezon, D. Marenduzzo, A. Maritan, F. Seno, Y. Snir, and A. Trovato, Phys. Rev. E **73**, 031921 (2006).

[17] I. A. Hubner and E. I. Shakhnovich, Phys. Rev. E **72**, 022901 (2005).

[18] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, Rev. Mod. Phys. **72**, 259 (2000).

[19] E. I. Shakhnovich and A. M. Gutin, Proc. Natl. Acad. Sci. U.S.A. **90**, 7195 (1993).

[20] H. Kaya and H. S. Chan, Phys. Rev. Lett. **85**, 4823 (2000).

[21] H. Abe and H. Wako, Phys. Rev. E **74**, 011913 (2006).

[22] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989).

[23] R. Mélin, H. Li, N. S. Wingreen, and C. Tang, J. Chem. Phys. **110**, 1252 (1999).

[24] H. Li, C. Tang, and N. S. Wingreen, Proc. Natl. Acad. Sci. U.S.A. **95**, 4987 (1998).

[25] H. S. Chan, Proteins **40**, 543 (2000).

[26] M. Levitt, M. Gerstein, E. Huang, S. Subbiah, and J. Tsai, Annu. Rev. Biochem. **66**, 549 (1997).

[27] B. W. Matthews, Adv. Protein Chem. **46**, 249 (1995).

[28] S. Marqusee and R. T. Sauer, Protein Sci. **3**, 2217 (1994).

[29] J. K. Myers and C. N. Pace, Biophys. Chem. **71**, 2033 (1996).

[30] H. Taketomi, Y. Ueda, and N. Gō, Int. J. Pept. Protein Res. **7**, 445 (1975).

[31] For a recent review, see T. Lazaridis and M. Karplus, Biophys. Chem. **100**, 367 (2003).

[32] C. Zhang and S.-H. Kim, Proc. Natl. Acad. Sci. U.S.A. **97**, 2550 (2000).

[33] R. P. Saha, R. P. Bahadur, and P. Chakrabarti, J. Proteome. Res. **4**, 1600 (2005).

[34] F. Glaser, D. M. Steinberg, I. A. Vakser, and N. Ben-Tal, Proteins **43**, 89 (2001).

[35] S. Miyazawa and R. Jernigan, Macromolecules **18**, 534 (1985); J. Mol. Biol. **256**, 623 (1996); Proteins **36**, 357 (1999).

[36] H. Li, C. Tang, and N. S. Wingreen, Proteins **49**, 403 (2002).

[37] N. E. G. Buchler and R. A. Goldstein, Proteins **34**, 113 (1999).

[38] R. C. Ball and T. M. A. Fink, Phys. Rev. E **66**, 031902 (2002).

[39] C. T. Shih, Z. Y. Su, J. F. Gwan, B. L. Hao, C. H. Hsieh, J. L. Lo, and H. C. Lee, Phys. Rev. E **65**, 041923 (2002).

[40] B. A. Berg, Fields Inst. Commun. **26**, 1 (2000).

[41] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).

[42] U. H. E. Hansmann and Y. Okamoto, Curr. Opin. Struct. Biol. **9**, 177 (1999), and references therein.

[43] L.-H. Tang, in *Computational Physics: Proceedings of the Joint Conference of ICCP6 and CCP2003*, edited by X.-G. Zhao, S. Jiang, and X.-J. Yu (Rinton Press, New Jersey, 2006).

[44] Y. Iba, G. Chikenji, and M. Kikuchi, J. Phys. Soc. Jpn. **67**, 3327 (1998).