Lattice models of protein folding permitting disordered native states

Gordon M. Crippen and Mukesh Chhajer
College of Pharmacy, University of Michigan, Ann Arbor, Michigan 48109-1065

(Received 30 August 2001; accepted 15 November 2001)

Self-avoiding lattice walks are often used as minimalist models of proteins. Typically, the polypeptide chain is represented as a lattice walk with each amino acid residue lying on a lattice point, and the Hamiltonian being a sum of interactions between pairs of sequentially nonadjacent residues on adjacent points. Interactions depend on the types of the two residues, and there are usually two or more types. A sequence is said to fold to a particular ''native'' conformation if the ground state is nondegenerate, i.e., that native conformation is the unique global energy minimum conformation. However, real proteins have some flexibility in the folded state. If this is permitted in a lattice model, the most stably and cooperatively folding sequences have very disordered native states unless the Hamiltonian either favors only a few specific interactions or includes a solvation term. The result points the way toward qualitatively more realistic lattice models for protein folding.
 [DOI: 10.1063/1.1433745]

INTRODUCTION

Simple globular proteins are specific linear copolymers of 20 different amino acids that are soluble in water and under biologically appropriate conditions adopt a fairly well defined ''native'' conformation that depends in some complicated way on the sequence of amino acid residues. Sequences that fold to some native state are apparently rare,[1] but the process of evolution has selected at least 10^5 of them. The native or folded state is characterized by a specific native conformation forming a compact globule having axial ratio not much more than 2:1 and a packing density comparable to that seen in crystal structures of small organic molecules.[2] The more hydrophobic amino acid residues tend to be buried and thus less exposed to solvent than the more polar ones.[3] The native state is also stabilized by more specific interactions, such as hydrogen bonds and salt bridges. Varying the temperature, pressure, *p*H, ionic strength, or otherwise the solvent composition can lead to a moderately cooperative transition to a denatured state characterized by increased exposure to solvent, high conformational entropy, and high mean radius of gyration.[4] In particular, increasing temperature leads to the denatured state over a roughly 10 °C transition that can be reasonably reversible.[5]

Given that even small globular proteins consist of thousands of atoms joined in a chain having at least hundreds of degrees of conformational freedom, there has been a long standing interest in simple models that would still possess a well defined and compact native state that makes a cooperative transition to an entropically more favored denatured state with increasing temperature. A favorite model is to simplify each amino acid residue to an isotropic point on some sort of lattice, and the polypeptide chain is then a self-avoiding connected walk over the lattice, just as the real chain is connected and has excluded volume effects.[6] Typically one defines some sort of energy function that is a sum over pairwise interactions between residues located on adjacent lattice points but are not adjacent in sequence. Each interaction term usually depends only on the types of the two residues and the fact that they are in spatial contact. Denote this potential function by $E(c,s)$, where c is the conformation and s is the sequence. In a realistic model, E represents the free energy of a dilute solution of the protein as a function of the conformation, given the particular fixed sequence. If the lowest E is achieved for a unique conformation c^* for some particular sequence s, then that sequence is said to fold to $c^*=c(s)$, in analogy to the low conformational entropy of a real globular protein in its native state. Any other $c \neq c(s)$ is said to belong to the denatured macroscopic state. One commonly used measure of the resemblance between any conformation c and the native $c(s)$ is the order parameter Q, defined[7] to be the fraction of native contacts seen in c relative to the number of contacts in $c(s)$.

It turns out there is a significant problem with these sorts of models. By definition, the native state has zero conformational entropy and a precise set of contacts by which Q is measured, and these contacts are present 100% of the time in the native state. Real proteins, however, have significant conformational flexibility even in the native state. Often the ends of the chain or external loops cannot be resolved in x-ray crystal structures. Slow but large magnitude motions are sometimes required for ligands to approach or leave buried binding sites.[8] Hydrogen exchange experiments in solution under native conditions reveal that labile hydrogens at different positions in the protein vary greatly in the rates at which they exchange with the solvent, which is interpreted as variation in solvent accessibility. Some rapidly exchanging parts, i.e., solvent accessible hydrogens, appear buried in the static picture provided by crystallography.[9]

Others have recognized the need to model the native state as more than a single conformation.[10] For example, any conformation having more than a certain number of the contacts seen in the conformation of globally minimal energy can be included in the native state.[11] Or the native basin of attraction on the energy surface presumably includes all conformations having greater conformational resemblance to the

global minimum structure than the mean resemblance at the thermal transition midpoint.[12] In this work we apply a similar definition to such short chains on simple lattices that all conformations and all sequences may be examined to calculate their folding thermodynamics for a wide variety of assumed energy functions. Such an exhaustive approach avoids search problems over conformations and sequences, and it avoids biasing the results toward native structures, sequences, and interaction energies chosen *a priori*. The price, however, is the limitation in chain length and number of residue types imposed by computational feasibility. Of course our lattice models—and even much longer chain simulations—do not have the same quantitative values of thermodynamic parameters (e.g., absolute temperatures, free energies, and entropies of unfolding) that real proteins have. The point of this work is not to reproduce experimental values, but rather to show how relaxing the assumption of a unique native conformation leads to qualitative changes in the behavior of these models that is more realistic.

METHODS

A variety of different lattices have been used in such studies. Although proteins are obviously three-dimensional, a two-dimensional square lattice makes the calculations easy, the results are easy to depict, and residues can be buried by short chains. Longer chains tend to exhibit more cooperativity,[10] presumably because of more opportunity to bury hydrophobic residues. Here we also include calculations on a three-dimensional cubic lattice, but due to the higher coordination number and greater restrictions on chain length, there is much less opportunity for solvent shielding. Many different choices have been made about residue types, all the way from a different type for every residue in the chain[13] to only two types, as in the hydrophobic/polar (HP) model.[14] Here the emphasis is on an HP model, but more types have been considered. As is customary, E is taken to be a sum over pairwise interactions between residues in close contact, rather than being distance dependent over longer distances. Thus there are three interaction energies in the HP model, e_{PP}, e_{HP}, and e_{HH}. In addition, we consider a solvent exposure factor whenever a residue lattice point is adjacent to an unoccupied lattice point. Denote the residue–water interactions by e_{PW} and e_{HW}. For a chain of n residues in a fairly extended conformation, there are at most $2n+2$ solvation interactions on a square lattice. In all,

$$E(c,s)=\sum_{\text{contacts}}^{n}{}_{i,j}\, e_{t_i,t_j}+\sum_{\text{exposed}}^{n}{}_{i}\, e_{t_i,W}, \tag{1}$$

where t_i is the type of residue i.

Alternatively, one can express such a contact function on a lattice in terms more familiar to polymer theory. Let $q=4$ be the coordination number of the square lattice, φ_{ij} be the number of nonbonded adjacent lattice points occupied by particle types i, $j\in\{\mathrm{H,P,W}\}$, n_i be the total number of lattice points occupied by particles of type i, and $b_{i,j}$ be the number of bonded adjacent lattice points occupied by particle types i,j. Clearly the n_i and $b_{i,j}$ depend on the chain length and HP sequence, but they are independent of conformation. Assuming a large lattice with periodic boundary conditions, there are three equations relating the quantities

$$qn_{\mathrm{H}}=2\varphi_{\mathrm{HH}}+\varphi_{\mathrm{HP}}+\varphi_{\mathrm{HW}}+2b_{\mathrm{HH}}+b_{\mathrm{HP}}, \tag{2}$$

$$qn_{\mathrm{P}}=2\varphi_{\mathrm{PP}}+\varphi_{\mathrm{HP}}+\varphi_{\mathrm{PW}}+2b_{\mathrm{PP}}+b_{\mathrm{HP}}, \tag{3}$$

$$qn_{\mathrm{W}}=2\varphi_{\mathrm{WW}}+\varphi_{\mathrm{PW}}+\varphi_{\mathrm{HW}}. \tag{4}$$

In these terms, Eq. (1) becomes

$$E=\sum_{i,j}\varphi_{ij}e_{ij} \tag{5}$$

but in terms of relative interaction energy χ parameters[15]

$$\chi_{ij}=e_{ij}-(e_{ii}+e_{jj})/2. \tag{6}$$

Equation (5) becomes

$$\begin{aligned}E=\varphi_{\mathrm{HP}}\chi_{\mathrm{HP}}&+([4n_{\mathrm{H}}-2b_{\mathrm{HH}}-b_{\mathrm{HP}}]-2\varphi_{\mathrm{HH}}-\varphi_{\mathrm{HP}})\chi_{\mathrm{HW}}\\&+([4n_{\mathrm{P}}-2b_{\mathrm{PP}}-b_{\mathrm{HP}}]-2\varphi_{\mathrm{PP}}-\varphi_{\mathrm{HP}})\chi_{\mathrm{PW}}+[(4n_{\mathrm{H}}\\&-b_{\mathrm{HP}}-2b_{\mathrm{HH}})e_{\mathrm{HH}}+(4n_{\mathrm{P}}-b_{\mathrm{HP}}-2b_{\mathrm{PP}})e_{\mathrm{PP}}\\&+4n_{\mathrm{W}}e_{\mathrm{WW}}]/2,\end{aligned} \tag{7}$$

where the quantities in square brackets depend on sequence but not on conformation. Thus for the purposes of examining the general behavior of all sequences or the detailed behavior of one given sequence as a function of energy, one need only vary the three χ parameters.

For a given sequence s of n residues there are a finite number of self-avoiding lattice conformations up to translation, rotation, and reflection. There will be one or more of these having the lowest (ground state) energy. Choose one of them having the fewest contacts (favorable or unfavorable) to be the reference microstate in the native macroscopic state. Any other conformation, regardless of its energy, having $Q>0.5$ compared to the reference is included in the native state.[11] All other conformations belong to the denatured state. This cutoff for Q is probably low compared to real proteins under native conditions, but for these small lattice models of short chains, the reference state may have only four or fewer contacts, and the low cutoff permits many sequences to have multiple conformations in their native states. Note that some conformations in the denatured state may have several contacts as long as not many of them involve the same i,j pairs seen in the reference conformation's contacts.

Consider the equilibrium thermodynamics of thermal denaturation in this model. A given sequence is taken to fold and unfold if (1) the reference native conformation has at least one contact, (2) at $T=0$ it is completely in the native state because the lowest energy non-native conformation is strictly above the global minimum of energy, and (3) at $T=\infty$ it is completely denatured because there are fewer native conformations than non-native ones. While there is a real absolute temperature T involved, the scale is arbitrarily determined by the magnitudes of the interaction parameters. Results will be presented with reference to the midpoint of the melting transition, T_m, where the protein is 50% native and 50% denatured. The behavior of different sequences will

be discussed by comparing their T_m's, the higher value indicating greater thermal stability. The cooperativity of the thermal transition is sometimes described by its width ΔT, taken to be the difference in temperatures between 90% native and 10% native, for example.[16] Chan has advocated using as a measure of cooperativity[10,17] the ratio of the calorimetric $\Delta H_{\rm cal}$ over the whole transition to the van't Hoff ΔH_{vH} measured at the midpoint of the transition. Unless the van't Hoff plot is strictly linear, ΔH_{vH} will depend on which definition of midpoint is used: our choice of T_m, or the temperature of the maximum heat capacity, or the halfway point in the enthalpy change.[10] In any case, it has been argued that $\Delta H_{\rm cal} > \Delta H_{vH}$ is to be expected. As will be shown in the Results section, when the two values are very close, sometimes it is the other way around, so here we use

$$h_{\rm ratio} = \max\left[\frac{\Delta H_{\rm cal}}{\Delta H_{vH}}, \frac{\Delta H_{vH}}{\Delta H_{\rm cal}}\right]. \tag{8}$$

Letting $\langle x\rangle_{{\rm state},T}$ denote the Boltzmann weighted mean of property x over macroscopic state=native (nat) or denatured (den) or both at temperature T, it is easy to see that

$$\Delta H_{\rm cal} = \langle E\rangle_{{\rm both},\infty} - \langle E\rangle_{{\rm both},0} \tag{9}$$

or in other words, the unweighted mean E over all native and denatured conformations minus the global minimum E, which by definition is that of the reference native conformation. It is easy to derive ΔH_{vH} starting form the defining equation

$$\left.\frac{d\ln K_{\rm eq}}{d(1/T)}\right|_{T=T_m} = -\frac{\Delta H_{vH}}{R} \tag{10}$$

and the standard relations

$$\Delta G = -RT\ln K_{\rm eq} \tag{11}$$

and

$$G_{\rm state} = -RT\ln\sum_{i\in{\rm state}} \exp(-E_i/RT) \tag{12}$$

to produce

$$\Delta H_{vH} = \langle E\rangle_{{\rm den},T_m} - \langle E\rangle_{{\rm nat},T_m}. \tag{13}$$

Thus, $h_{\rm ratio}$ is simple to calculate for these models, and it is close to unity when above the main native state energy level (having low degeneracy) the lowest denatured state energy level is highly degenerate. Note that $h_{\rm ratio}$ is independent of scaling E by a positive constant, whereas T_m increases when the scaling factor is greater than unity. While much has been made of $\Delta E_{\rm gap}$=the gap between the lowest denatured state energy level and the highest native state level, it has little to do with cooperativity in these and other statistical mechanical models.[16]

2D RESULTS

Consider chains with length $n=9$. It is easy to enumerate all 740 conformations and 512 HP sequences, and it is possible to form a compact 3×3 conformation having an interior completely shielded from solvent. Longer chains would presumably show some sequences having better cooperativity, but the conceptual outcome would be similar. First consider what range of energy parameters permits at least one HP sequence to have a folding transition. Figure 1 shows a quick scan over the three independent parameters $\chi_{\rm HP}$, $\chi_{\rm HW}$, and $\chi_{\rm PW}$ over the range -2 to $+2$ arbitrary energy units. If $\chi_{\rm HW}=\chi_{\rm PW}$ and $\chi_{\rm HP}=0$, then both H and P residues have equal preference for solvent and no preference for forming a hydrophobic (or polar) core, so no sequence exhibits a folding/unfolding transition. Otherwise, as long as H residues are hydrophobic ($\chi_{\rm HW}>0$), at least some sequences will fold. An alternative way to produce folding transitions is to strongly favor HP associations relative to HH or PP contacts ($\chi_{\rm HP}\leqslant 0$) even with mild hydrophilicity of both H and P ($\chi_{\rm HW},\chi_{\rm PW}\leqslant 0$). This last case, however, is not relevant to real proteins folding under the influence of the hydrophobic effect.

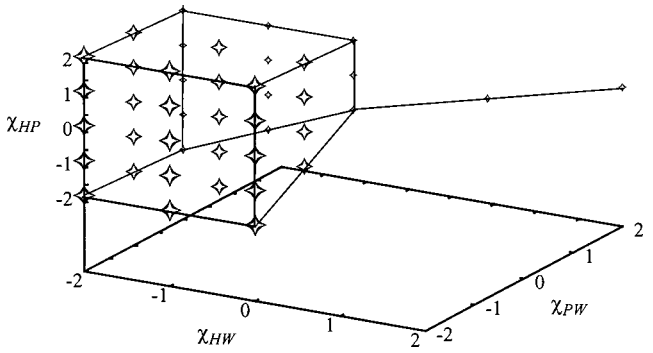

FIG. 1. A coarse scan over the three energetic degrees of freedom $\chi_{\rm HP}$, $\chi_{\rm HW}$, and $\chi_{\rm PW}$, showing the region marked by stars where there are no HP sequences that have a folding transition.

Customarily in these HP lattice models, one assumes that proteinlike behavior can be produced by an energy function that depends only on residue–residue contacts, i.e., $e_{\rm PW}=e_{\rm HW}=e_{\rm WW}=0$. Permitting the native state to include more than one conformation has the curious effect of favoring noncompact native reference conformations for the thermally most stable folding sequences over a wide range of the remaining three energy parameters, $e_{\rm HP}$, $e_{\rm PP}$, and $e_{\rm HH}$. Only

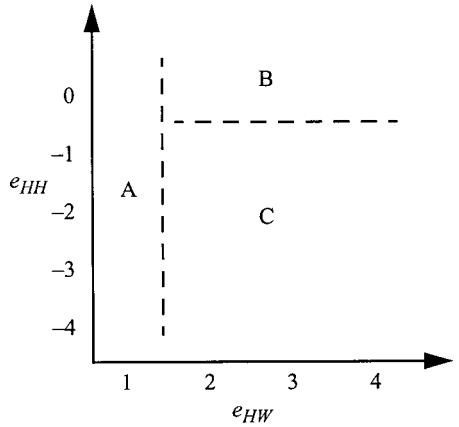

FIG. 2. Phase diagram for square lattice models as a function of hydrophobic–hydrophobic interactions $e_{\rm HH}$ and hydrophobic–solvent $e_{\rm HW}$, holding fixed $e_{\rm PP}=e_{\rm HP}=0$ and $e_{\rm PW}=-1$. See text for a discussion of the three regions A, B, and C.

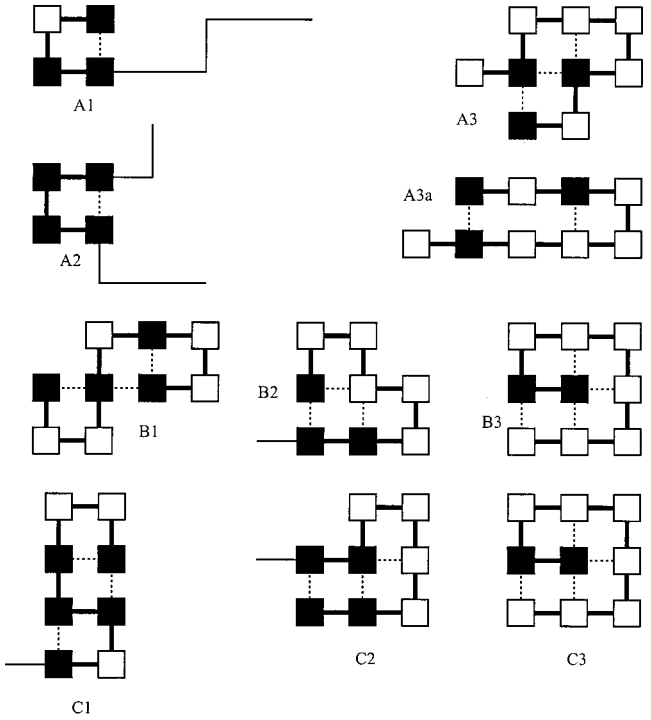

FIG. 3. Reference native conformations and sequence patterns corresponding to different energy functions E in Fig. 2 and tabulated in Table I. Black squares are H residues, white squares are P residues, heavy solid lines mark the fixed chain path, thin solid lines indicate variable portions of the conformation, and dotted lines mark the defining contacts only.

with $e_{HP}<e_{HH}$, $e_{PP}<0$ do sequences like that in Fig. 3 B3 fold to such compact 3×3 conformations and have higher T_m than other sequences. Sequences that fold rather cooperatively ($h_{ratio}\approx1.4$) to compact native reference structures can be found when $e_{HH}=e_{PP}<0$. Yet these are hardly realistic choices of energy parameters where hydrophobic–polar contacts are favored over hydrophobic–hydrophobic ones, or where HH and PP contacts are equally favorable. A more appropriate choice would be e_{HP}, $e_{PP}\geqslant0$ and $e_{HH}<0$, but then the thermally most stable folding sequences have native states with multiple conformations sharing a terminal hairpin loop, as in Fig. 3 A1.

Consider a more realistic model for the hydrophobic effect on protein folding where explicit solvation terms are included in the energy. Suppose $e_{PP}=e_{HP}=0$ and $e_{PW}=1$, i.e., polar–polar and hydrophobic–polar contacts are energetically indifferent, but it is favorable to expose polar residues to solvent. Without loss of generality, we take $e_{WW}=0$ so that e_{PW} and e_{HW} are measured relative to e_{WW}. As the two remaining energy parameters are varied, Fig. 2 shows three qualitatively different outcomes.

Case A. When the penalty for exposing hydrophobic residues is weak, then no matter how favorable HH contacts are, the thermally most stable sequences are those where the native state is represented by a single HH contact at the end of the chain, as illustrated in Fig. 3 item A1. The remainder of the chain is free to move about, resulting in 124 conformations in the native state, which is therefore so entropically favored compared to the total of 740 conformations that even as $T\rightarrow\infty$, the equilibrium mixture is always more than 10% native. Consequently $\Delta T=\infty$, even though h_{ratio} is a reasonably cooperative 1.2 calculated at T_m (see Table I). Since the native state is characterized by only a single contact, trivially the Boltzmann average $Q=1$ for the native state, and it is 0 for the denatured state, even at T_m. In the particular case that $e_{HH}=0$, $e_{PW}=-1$, and $e_{HW}=+1$, the reference native has $E=-8$, but there are other conformations in the native state sharing the defining HH contact while having other PP and HP contacts such that the highest energy native has $E=-6$. This is the same value achieved by the lowest energy denatured conformations that have no contacts and balance unfavorable solvation of the three H residues against favorable solvation of the six P residues. Consequently, we have an example where the sequence with the most thermally stable native state and fairly cooperative unfolding has $\Delta E_{gap}=0$. This peculiar behavior is seen over a wide range of interaction parameters, particularly when there is no solvation energy for H or P, or whenever the solvation penalty for H residues is weak, regardless of how favorable HH contacts are.

Extremely cooperative folding ($h_{ratio}=1.002$) can be achieved at much lower T_m for sequences having a native state characterized by a single local contact in the middle of the chain, Fig. 3 case A2, for example, when $e_{HH}=0$, $e_{PW}=-1$, and $e_{HW}=+1$. Then there are a very few denatured conformations having lower E than the few highest energy native conformations, resulting in $\Delta E_{gap}=-2$. Clearly a

TABLE I. Thermodynamic characteristics of the sequences that are the most stable, most cooperative, and highest number of contacts for various E functions.

E class[a]	T_m	h_{ratio}	No. contacts	ΔE_{gap}	$\langle Q\rangle_{nat,T_m}$	$\langle Q\rangle_{den,T_m}$	Reference native[b]
A	**691**	1.228	1	0	1	0	A1
	443	**1.002**	1	−2	1	0	A2
	320	1.584	**3**	−2	0.939	0.130	A3, A3a
B	**1076**	1.969	3	−4	0.739	0.261	B1
	599	**1.416**	3	−4	0.766	0.196	B2
	243	2.694	**4**	0	0.972	0.303	B3
C	**1357**	1.745	3	−1	0.736	0.236	C1
	982	**1.436**	3	−4	0.769	0.143	C2
	242	2.694	**4**	0	0.972	0.303	C3

[a]See Fig. 2.
[b]See Fig. 3.

large (positive) energy gap is required neither for thermal stability nor for cooperativity in these models that permit multiple conformations in the native state.

Those native sequences having the most contacts defining their native states in the A region of energy parameters do form passably compact native conformations, such as A3, but at half the T_m of the most stable sequences and 50% greater h_{ratio} than the most cooperative sequences (Table I). Obviously in this model, favorable HH contacts alone are insufficient to produce proteinlike, compact, low entropy native states that unfold cooperatively. Figure 3 shows one of the nine alternative native conformations, A3a, that has 2/3 of the reference native conformation's contacts as well as one extra non-native contact. This is just an illustration of the conformational flexibility permitted by the assumptions of the model, even when the reference conformation has three contacts and only nine residues.

Case B. Much more realistic behavior can be achieved even with weak HH interactions as long as the penalty for solvent exposure of H residues is greater than the reward for solvation of P residues. For example, when $e_{\text{HH}}=0$ and $e_{\text{HW}}=+2$, the thermally most stable sequence folds with the complete burial of one H residue and three contacts, as in structure B1. Although there are 36 conformations in the native state, this is not nearly the entropic stabilization of case A, and even at T_m, 43% of the native state consists of only four of those 36. However, the unfolding transition is not as cooperative as for examples A1, A2, or A3. The most cooperatively folding sequences, B2, have only 16 structures in the native state but less complete burial of H residues. Structure B3 is the intuitively preferable case for compactness, since it has the maximum number of contacts, only two native conformations, and deep burial of H residues. Yet it has a substantially lower T_m and higher h_{ratio}.

Case C. When there is simultaneously a substantial penalty for solvation of H residues and a reward for HH contacts, the most stable fold is generally Fig. 3 structure C1. The most cooperative folding is seen with sequences having reference native states like C2 or a few other possibilities, and the native state with the most contacts remains C3, still with a low T_m and high h_{ratio}. Compared to case B, a more favorable e_{HH} does not make a qualitative difference in the folding of all HP sequences, although there are quantitative variations in stability and cooperativity, and variations in exactly which conformations are the most interesting native reference states. If one chooses the definition of native state used in this work, then one must adopt a potential function having a solvation component in order to achieve even modestly realistic folding. A residue–residue contact function alone is apparently insufficient.

It is possible to produce compact 3×3 native conformations without solvation but more specific sorts of interactions. For example, suppose HP contacts are favorable while all other contacts are very unfavorable. This would be a simple model of a chain of hydrogen bond donors and acceptors in a polar solvent that formed no hydrogen bonds. Then the most cooperatively folding sequence has h_{ratio} of only 12.3, but the native reference conformation is maximally compact [Fig. 4(a)]. Permitting four residue types (a), (b), (c), and (d) with favorable (a)–(b) and (c)–(d) contacts produces no more cooperatively folding sequences.

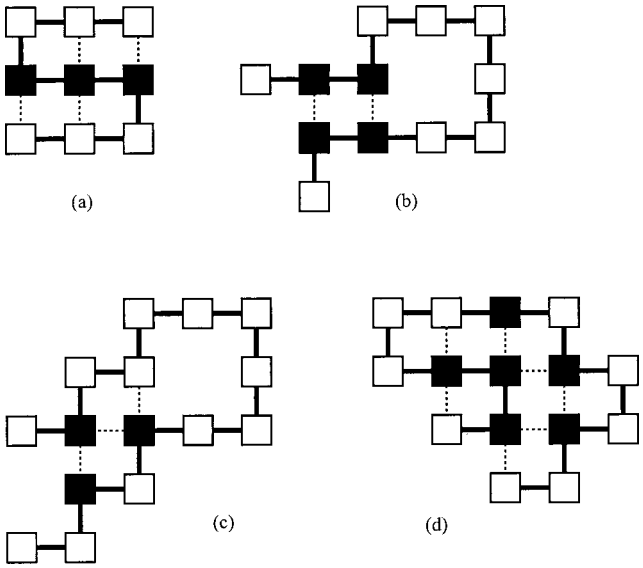

FIG. 4. Reference native conformations and sequence patterns depicted as in Fig. 3. (a) The most cooperatively folding native with no solvation and favorable HP contacts. (b) The most cooperatively folding native for 12 residues and $e_{\text{HH}}=-1$, $e_{\text{PW}}=-1$, $e_{\text{HW}}=+2$, and all other parameters$=0$. (c) The most cooperatively folding native for 15 residues and the same energy parameters. (d) The 15 residue natives having the most contacts, given the same energy parameters.

The outcome is not substantially changed for somewhat longer chains. For 12 residue HP chains in region C ($e_{\text{HH}}=-1$, $e_{\text{PW}}=-1$, $e_{\text{HW}}=+2$, and all other parameters$=0$), the most cooperatively folding sequence has the defining native structure shown in Fig. 4(b) and $h_{\text{ratio}}=1.56$. The most stable sequence has twice the T_m and $h_{\text{ratio}}=2.26$, but still not the expected compact 3×4 native structure. This ideally compact native is achieved by some sequences that are substantially less thermally stable and very uncooperative. For a 15 residue HP chain under the same energy parameters, the most cooperatively folding sequence out of the 32 768 possible sequences has the moderately compact defining native structure seen in Fig. 4(c) and $h_{\text{ratio}}=1.74$. Two of the three hydrophobic residues are completely buried while allowing 538 conformations to fall in the native state, out of 296 806 total conformations, although only 81 of these 538 account for most of the statistical weight at T_m. A more compactly folded sequence having seven contacts in its defining native conformation [Fig. 4(d)] and 322 native conformations folds less cooperatively with $h_{\text{ratio}}=2.05$.

Residual structure. The denatured state of real proteins, even far from the midpoint of the folding transition, often shows substantial residual structure, typically consisting of measurable levels of some features that are seen in the native.[18] One might be concerned that the present model permits so much conformational variation in the native state that the denatured state would have almost no residual native structure. Yet Table I shows that for cases B and C, the Boltzmann averaged Q in the denatured state at T_m is 14–26 % for the most stable and cooperative sequences. This is another realistic feature in favor of the model.

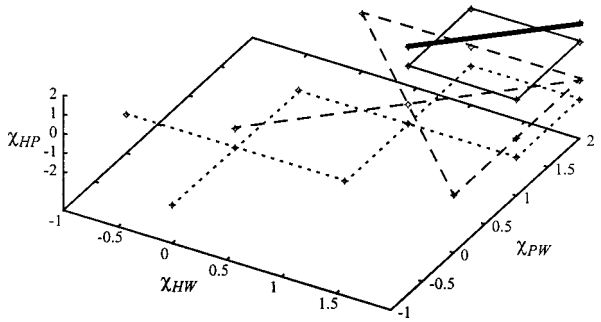

FIG. 5. A coarse scan over the three energetic degrees of freedom χ_{HP}, χ_{HW}, and χ_{PW}, showing the region marked by stars where the sequences having highest T_m fold to a maximally compact $2\times2\times2$ conformation for eight residues. As a visual aid, lines connect stars having the same value of χ_{HP}.

3D RESULTS

Folding vs interaction energies. Consider chains with length $n=8$. It is easy to enumerate all 1832 conformations and 256 HP sequences, and it is possible to form a compact $2\times2\times2$ conformation having five contacts but no residue completely shielded from solvent. As in the two-dimensional case, the values of the three independent parameters χ_{HP}, χ_{HW}, and χ_{PW} that permit at least one HP sequence to have a folding transition are essentially those shown in Fig. 1. Compact native reference conformations for the most stably folding sequences are seen when χ_{HP} is particularly favorable, analogous to stabilization of a compact but solvent exposed peptide conformation that is stabilized by particular interactions between pairs of dissimilar groups, as in salt bridges or hydrogen bonds. The other way that compact native conformations are achieved is when all three parameters are unfavorable, meaning that HH and PP contacts are favored. This is summarized in Fig. 5.

Consider the $n=8$ sequences having highest T_m, lowest h_{ratio}, or most contacts in the reference native structure, given the above sampling of interaction energies, namely $-2\leqslant\chi_{\mathrm{HP}}$, χ_{HW}, $\chi_{\mathrm{PW}}\leqslant2$. Figure 6 shows there is no particular correlation between ΔE_{gap} and T_m. In other words, having a greater energy gap between the native conformations and the denatured conformations does not increase the thermal stability of the native state in these models. Note that negative values of ΔE_{gap} are possible here, because the energy ranges of the native and denatured sets of conformations can overlap, even though the native state includes the conformation having the globally minimal energy. Likewise, Fig. 7 shows there is no special correlation between ΔE_{gap} and the folding cooperativity as measured by h_{ratio}. In fact, the most cooperatively folding sequences generally have reference native conformations with only one or two contacts at the end of the chain, as seen before on the square lattice. The greatest cooperatively seen for sequences having compact reference native conformations is $h_{\mathrm{ratio}}=1.58$. Thermal unfolding is generally a rather broad transition in this sampling of sequences and interaction energies, but Fig. 8 shows a positive correlation between h_{ratio} and $\Delta T/T_m$. The scatter at low values of h_{ratio} is due to sometimes very broad ΔT.

Attempting to model the hydrophobic effect by setting $e_{\mathrm{PP}}=e_{\mathrm{HP}}=0$ and $e_{\mathrm{PW}}=-1$, as in Fig. 2 for the square lattice, results in similar, flat reference native conformations even on the cubic lattice. For each choice of e_{HH} and e_{HW}, the sequence with highest T_m folds to just one of three

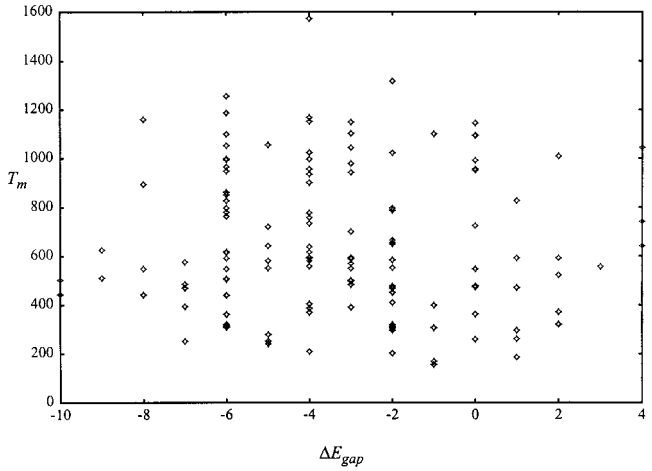

FIG. 6. Correlation between ΔE_{gap} and T_m for the $n=8$ sequences having highest T_m, lowest h_{ratio}, or most contacts in the reference native structure, given $-2\leqslant\chi_{\mathrm{HP}}$, χ_{HW}, $\chi_{\mathrm{PW}}\leqslant2$.

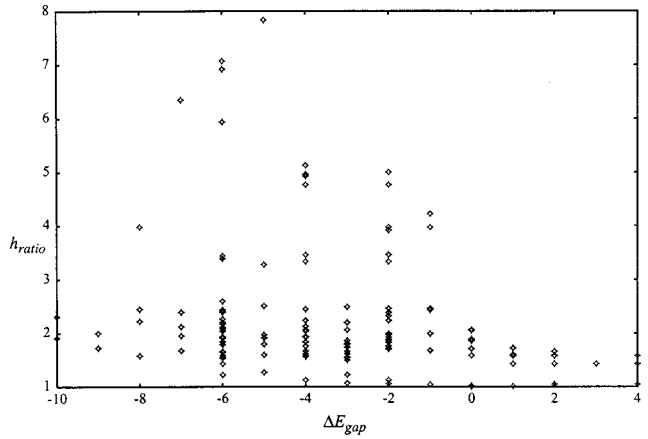

FIG. 7. Correlation between ΔE_{gap} and h_{ratio} for the same sequences as in Fig. 6.

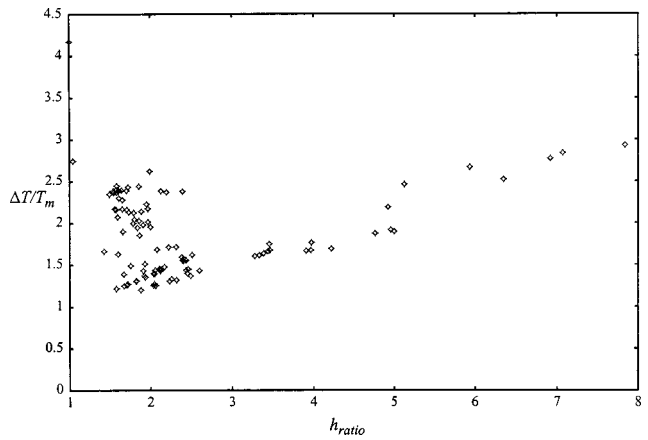

FIG. 8. Correlation between $\Delta T/T_m$ and h_{ratio} for the same sequences as in Fig. 6.

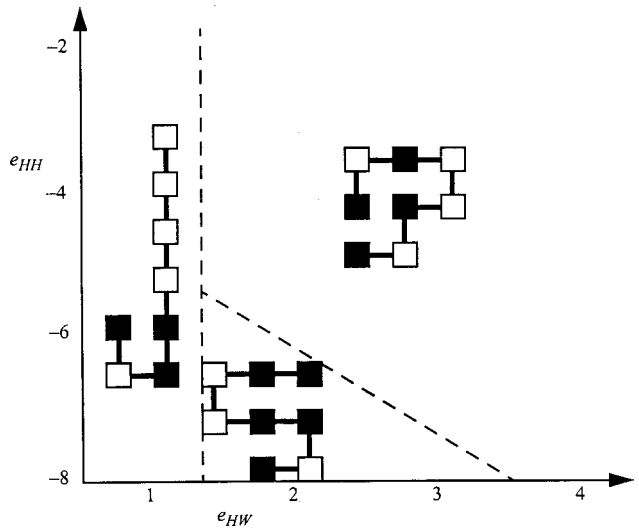

FIG. 9. Phase diagram for cubic lattice models as a function of hydrophobic–hydrophobic interactions e_{HH} and hydrophobic–solvent e_{HW}, holding fixed $e_{\mathrm{PP}}=e_{\mathrm{HP}}=0$ and $e_{\mathrm{PW}}=-1$. Thermally most stable sequences and their native conformations are drawn as in Fig. 3.

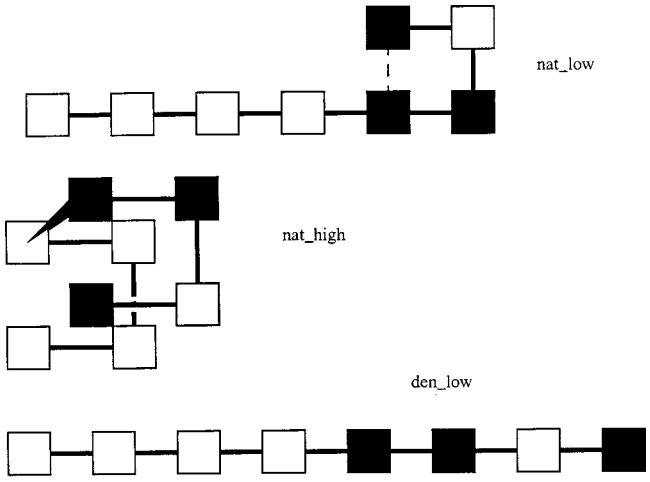

FIG. 10. Reference conformations for the case described in Table II.

choices of native conformation, even when e_{HH} is very favorable, as shown in Fig. 9. In particular, the flexible extended native having a terminal hairpin loop is seen when e_{HW} is only mildly unfavorable.

Examining substantially longer chains on the cubic lattice will have to rely on Monte Carlo sampling techniques. For $n=9$ there are 8453 conformations and 512 HP sequences; for $n=10$ there are 39 640 conformations and 1024 sequences. In the $n=10$ case, a scan of χ_{HP}, χ_{HW}, and χ_{PW} reveals no qualitative difference in the sorts of native conformations and degree of cooperativity of folding compared to $n=8$. There are still many choices of energy parameters whereby the thermally most stable sequences fold to reference native conformations having only one or two contacts, so that the native state is stabilized in large part by entropy.

ΔH vs cooperativity. Occasionally we observe $\Delta H_{vH} > \Delta H_{\mathrm{cal}}$, particularly when the values are close, although it has been argued that this situation is unlikely.[17] A particularly clear case of this occurs when $n=8$, $e_{\mathrm{HH}}=-8$, $e_{\mathrm{HP}}=e_{\mathrm{PP}}=0$, $e_{\mathrm{HW}}=+1$, and $e_{\mathrm{PW}}=-1$. Then the sequence PPPPHHPH folds to a native state characterized by a single contact at the end of the chain, but it also includes fully compact conformations. As shown in Table II (Fig. 10) the native conformations are clearly separated by a substantial energy gap from the highly degenerate lowest denatured energy level, which are filled by conformations having no contacts at all. This is clearly a situation involving only two significant states, and the folding is rather cooperative, judging by the low value of $h_{\mathrm{ratio}}=1.107$. However, this result is obtained due to $\Delta H_{vH}=10.19$ and $\Delta H_{\mathrm{cal}}=9.21$, which come from Eqs. (9) and (13), where $T_m=2796$ K, taking the energies to have units of kcal/mol. Figure 11 shows the enthalpy of the system as a function of temperature near T_m. The slope is monotonically decreasing, so there is no maximum of C_P near the midpoint. Note that the halfway point in calorimetric ΔH is also clearly different from T_m.

Lest one should conclude that these models show none of the usual signs of cooperative unfolding, consider the sequence HHHHPHHP when $e_{\mathrm{HP}}=-1$ and all other interaction parameters are zero. For these short chains, the thermal unfolding transition has a relatively narrow $\Delta T=223$ compared to its $T_m=154$, yet the $h_{\mathrm{ratio}}=2.46$ is not particularly low. Figure 12 shows that the heat capacity of the system peaks near T_m. The reference native conformation is maximally compact, forming three HP contacts, for a unique ground state $E=-3$. Yet there is substantial overlap in the energy distributions of the native and denatured states, namely 1, 11, and 3 native conformations at $E=-3$, -2, and -1, respectively, compared to 23, 305, and 1489 denatured conformations at $E=-2$, -1, and 0, respectively.

TABLE II. Energy levels for an example having $\Delta H_{vH} > \Delta H_{\mathrm{cal}}$.

	E	Degeneracy	Conformation[a]
Denatured	−4	47	
levels	−6	390	
	−8	1144	den_low
Native	−14	2	nat_high
levels	−16	40	
	−18	209	nat_low

[a]See Fig. 10.

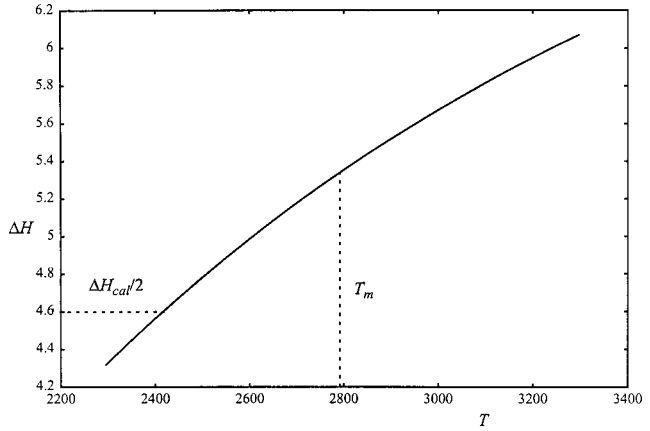

FIG. 11. ΔH_{cal} as a function of T for the case described in Table II.

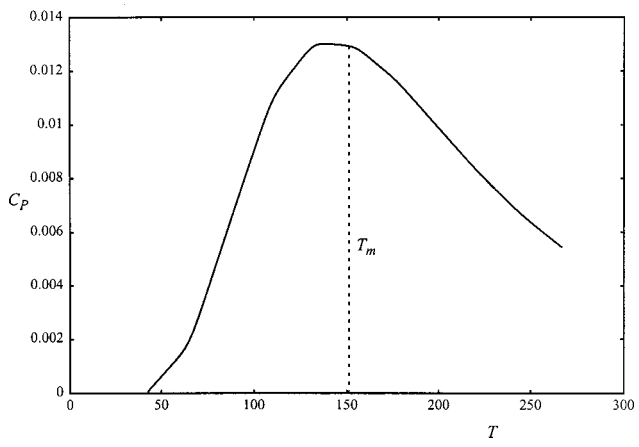

FIG. 12. C_P as a function of T for an example where ΔT is small but $h_{\text{ratio}}=2.46$.

CONCLUSION

Short chains of isotropically interacting points on square and cubic lattices are obviously a grossly simplified model for real proteins, and their thermodynamics of folding cannot be expected to agree even roughly with experimental values. However, they do show some qualitative features in agreement with real proteins, such as some degree of cooperativity in unfolding for some sequences according to some cooperativity measures. Here we have shown that a broad survey over possible energetic models, all sequences, and all conformations for short chains reveals curious disagreement among different cooperativity measures and purported governing factors, apparently due to permitting conformational flexibility in the native state. Globular native conformations are not the rule, and over a substantial range of energy functions, the favored native states for the most thermally stable and cooperatively folding sequences are high entropy ensembles of conformations having little consistent structure.

ACKNOWLEDGMENT

This work was supported by NIH Grant No. GM-59097.

[1] S. Roy, K. J. Helmer, and M. H. Hecht, Folding Des. **2**, 89 (1997).
[2] I. D. Kuntz and G. M. Crippen, Int. J. Pept. Protein Res. **13**, 223 (1979).
[3] P. K. Ponnuswamy, Prog. Biophys. Mol. Biol. **59**, 57 (1993).
[4] T. E. Creighton, *Proteins: Structures and Molecular Properties*, 2nd ed. (Freeman, New York, 1993).
[5] R. W. Hartley, Biochemistry **7**, 2401 (1968).
[6] H. S. Chan and K. A. Dill, J. Chem. Phys. **100**, 9238 (1994).
[7] J. N. Onuchic, Z. Luthey-Schulten, and P. G. Wolynes, Annu. Rev. Phys. Chem. **48**, 545 (1997).
[8] M. Miller, J. Schneider, B. K. Sathyanarayana, M. V. Toth, G. R. Marshall, L. Clawson, L. Selk, S. B. H. Kent, and A. Wlodawer, Science **246**, 1149 (1989).
[9] S. W. Englander, Annu. Rev. Biophys. Biomol. Struct. **29**, 213 (2000).
[10] H. Kaya and H. S. Chan, Proteins **40**, 637 (2000).
[11] A. M. Gutin, V. I. Abkevich, and E. I. Shakhnovich, Folding Des. **3**, 183 (1998).
[12] D. K. Klimov and D. Thirumalai, Folding Des. **3**, 127 (1998).
[13] V. S. Pande, A. Yu. Grosberg, C. Joerg, and T. Tanaka, Phys. Rev. Lett. **76**, 3987 (1996).
[14] H. S. Chan and K. A. Dill, Proteins: Struct., Funct., Genet. **30**, 2 (1998).
[15] P. J. Flory, *Principles of Polymer Chemistry* (Cornell University Press, New York, 1953), pp. 508–509.
[16] G. M. Crippen and Y. Z. Ohkubo, Proteins: Struct., Funct., Genet. **32**, 425 (1998).
[17] H. S. Chan, Proteins **40**, 543 (2000).
[18] P. Hammarström and U. Carlsson, Biochem. Biophys. Res. Commun. **276**, 393 (2000).