Geometric and Statistical Properties of the Mean-Field HP Model, the LS Model and Real Protein Sequences

C.T. Shih[1], Z.Y. Su[1], J.F. Gwan[2], B.L. Hao[3], C.H. Hsieh[1] and H.C. Lee[4,5,6]
[1] *National Center for High-Performance Computing, Hsinchu, Taiwan, ROC*
[2] *Forum Modellierung, Forschungszentrum Jülich, D 52425 Jülich, Germany*
[3] *Inst. of Theoretical Physics, Academia Sinica, Beijing, China*
[4] *Dept. of Physics and Dept. of Life Science, National Central University, Chungli, Taiwan, ROC*
[5] *National Center for Theoretical Science, Hsinchu, Taiwan, ROC*
[6] *Department of Physics, Stanford University, Palo Alto, CA 94305, USA*
(March 29, 2001)

Geometric and statistical properties of the mean-field HP model, which takes the hydrophobicity of the residues to be the main driving force for early protein folding, and the LS model, which takes steric repulsion on the residues to be the most important folding interaction, are studied. It is shown that the two models have almost identical mathematical structures. The mechanism leading to the previously reported result that the most foldable peptides in the mean-field HP model have a high similarity to binarized peptide sequences that code for α-helices in real proteins is examined in detail and the result is confirmed. This study supports the view that hydrophobicity plays a major role in the early formation of α-helices in proteins, but does not support the view that steric effects play a similar role.

PACS number: 87.10.+e, 87.15.-v, 87.15.By

I. INTRODUCTION

The unraveling of the complex structure of proteins is a great challenge in science. It is believed that the dynamical folding process and stable structure, or native conformation, of a protein are determined by its primary structure, namely its amino acid sequence [1]. A number of coarse-grained models have been advanced to provide insight to these very complicated issues. One is the HP model proposed by Dill *et al.* [2], in which the 20 kinds of amino acids are divided into two types, hydrophobic (H) and polar (P). This model has been studied extensively by several groups in the last decade [2–7]. One of the important results from numerical studies of the model on lattices is that there are a small number of structures with exceptionally high designability. A highly designable structure is one that attracts a large number of protein sequences to choose it as the ground state. These highly designable structures are found to have protein-like secondary structures [3,5,6].

Another one is the LS model proposed by Micheletti *et al.* [8]. Similar to the HP model, it divides the 20 kinds of amino acids into two types, large (L) and small (S), but unlike the HP model, it assumes the deciding factor in folding to be the steric hindrance effect which depends on the size of the amino acids [8]. It was shown in ref. [8] that on a lattice, structures in the LS model too have uneven designability (there called encodability score); only a small portion of structures, also claimed to have protein-like secondary structures, are selected by large numbers of peptide sequences as unique ground states. It will be shown here that the LS model is mathematically equivalent to the mean-field HP model.

In [6] the designability problem of the mean-field HP model was reduced to a geometric problem. Here the connection between geometrical constraints and designability on the lattice is studied in detail. Highly foldable (binary) peptides in the two models, namely those peptides that choose highly designable structures as ground states, are then compared with protein sequences in data banks. It is found that, as expected, when the highly foldable peptides are thought of as pure mathematical objects, namely binary sequences pure and simple, then such peptides in the HP and LS models have a high similarity with each other. However, when the highly foldable peptides are thought of as representations of protein sequences, which in the HP (LS) model maps the digit 0 to P-type (L-type) residues and the digit 1 to H-type (S-type) residues, then the highly foldable peptides in the HP model match well with real protein sequences in general and with segments of sequences that fold to α-helices in particular (but not well with segments of sequences that fold to β-sheets), whereas those in the LS model match poorly with real protein sequences.

While in this paper attention is focused on the HP and LS models on two-dimensional lattices, the method used can be applied to other models in more general settings.

II. THE HP MODEL

The Hamiltonian of the HP model is:

$$H = \sum_{i<j} E_{p_i p_j} \Delta(\vec{r_i} - \vec{r_j}) \tag{1}$$

where p_i is the type, H for hydrophobic and P for polar, of the ith residue, or amino acid, in the peptide chain [2]; $\Delta(\vec{r_i} - \vec{r_j}) = 1$ if $\vec{r_i}$ and $\vec{r_j}$ are nearest neighbors in

the lattice but not adjacent along the peptide sequence, and $\Delta(\vec{r_i} - \vec{r_j}) = 0$ otherwise; $E_{p_i p_j}$ specifies the residue contact energies that depend on the types of residues in contact.

Several sets of contact energies (E_{HH}, E_{HP}, E_{PP}) have been used: $(-1, 0, 0)$ for the original HP model [2], $(-2.3, -1, 0)$ by Li *et al.* [5], and $(-\pi, -1, 0)$ by Buchler and Goldstein [9]. Li *et al.* suggested that the contact energies should satisfy the following constraints: 1) compact shapes have lower energies than non-compact shapes; 2) $E_{PP} > E_{HP} > E_{HH}$ so that hydrophobic residues are buried as much as possible; and 3) different types of residues tend to segregate, which is a condition induced by having $2E_{HP} > E_{PP} + E_{HH}$ [5,10]. In this work these will be adopted with the modification that 3) is replaced by the additive relation $2E_{HP} = E_{PP} + E_{HH}$. Then the potential simplifies to:

$$E_{p_i p_j} = -(p_i + p_j) \tag{2}$$

where $p_i = 1$ for H and $p_i = 0$ for P residue [11]. Henceforth only structures that correspond to self-avoiding compact paths on a lattice will be considered.

In an $N \times N$ two-dimensional square lattices, there are four corner sites with coordination number $N_n = 2$, $4(N-2)$ side sites with $N_n = 3$ and $(N-2)^2$ core sites with $N_n = 4$. With the exception of the two ends of the peptide chain, which we ignore, each lattice point has $N_n - 2$ contacts. So the Hamiltonian Eq.(1) becomes:

$$\begin{aligned} H &= -(0 \times \sum_{i \in corner} + 1 \times \sum_{i \in side} + 2 \times \sum_{i \in core}) p_i \\ &= -\sum_i p_i - \sum_{i \in core} p_i + \sum_{i \in corner} p_i \end{aligned} \tag{3}$$

The first term on the right-hand side of Eq.(3) is a constant for a given peptide sequence. It is independent of whatever conformation the peptide resides in and, since Eq.(3) will only be used here to determine the native structure of a particular peptide sequence, it will be omitted. The third term means that it is costly to put H residues in the corner sites. Since it is of order $1/N^2$ it too will be omitted. The Hamiltonian then simplifies to what is known as the mean-field HP model [6]:

$$H(\mathbf{p}, \mathbf{s}) = -\mathbf{p} \cdot \mathbf{s} = \frac{1}{2}(|\mathbf{s} - \mathbf{p}|^2 - \mathbf{p}^2 - \mathbf{s}^2) \tag{4}$$

where $\mathbf{p} = (p_1, p_2, \ldots, p_n)$, $n = N^2$, is the binary peptide sequence and $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ is a binary structural sequence converted from a self-avoiding compact path on the lattice with the assignment: $s_i = 1$ (0) if the ith site of the structure is a core (surface) site. In this new form the Hamiltonian has an interpretation quite different from its original meaning. There it was an expression of inter-residual interaction. Here in Eq. (4) it is no longer inter-residual, rather it has the form of a site-dependent potential. With $\mathbf{s}^2$ fixed for a given lattice and $\mathbf{p}^2$ a constant for a given peptide sequence, both are irrelevant to the determination of the ground state structure of the peptide. They will be ignored in the ensuing calculation. The Hamiltonian now reduces to one-half of $|\mathbf{s}-\mathbf{p}|^2$ and a neat geometric interpretation for it emerges [6]. When $\mathbf{p}$ and $\mathbf{s}$ are viewed as n-component vectors, this quantity is just the Hamming distance between two corner points in a unit n-dimensional hypercube.

When the energy matrix elements are not additive, that is, when $E_{HH} = -2 - \gamma$ with $\gamma > 0$ as was used in [2,5,9], the model cannot be reduced to the simple site-dependent form of Eq.(4). The effect of γ is to stabilize the low-lying states in the mean-field model further by increasing the number of H-H contacts.

III. THE LS MODEL

It was shown by Micheletti *et al.* that in the LS model the designability (called encodability score by the authors) distribution of structures is similar to that in the mean-field HP model [8]. The Hamiltonian of this model is

$$H = -\sum_i z_i(\Gamma) \cdot A(z(\sigma_i) - z_i(\Gamma)) \tag{5}$$

where $\sigma_i \in \{L, S\}$; $z(\sigma_i)$ is the maximal number of nearest contacts without steric repulsion belonging to residue i; on a square lattice, $z(\sigma_i)$ is equal to 1 (2) for L (S) residues inside the chain, and to 2 (3) for L (S) residues at chain ends; $z_i(\Gamma)$ is the number of contacts of the ith residue in a conformation Γ; and $A(x)$ equals to 1 if $x \geq 0$ and $-a < 0$ otherwise. The Hamiltonian implies that if the number of contacts of the ith residue is larger than $z(\sigma_i)$, then the contact energy will be increased by a owing to steric effects.

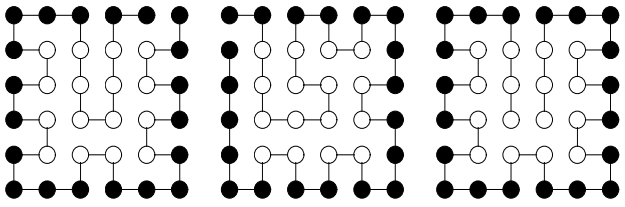

FIG. 1. (a) The most (third most) designable, (b) the second most (most) designable and (c) the third (second) most designable structures in the mean-field HP (LS) model, respectively, on a 6×6 lattice.

The results in Ref. [8], where a was set equal to ∞, show that the distribution of designability of structures in LS model is very similar to that in the HP model. In fact most of the highly designable structures in one model are likewise in the other model (see Appendix). The highly designable structures in the LS model also have

protein-like secondary substructure and tertiary symmetries. Three among the most designable structures in the two models are shown in Fig. 1.

Just as practiced in the last section, we consider only compact structures and neglect the effect of the two end points on a peptide chain. Table I gives the values of x, $A(x)$ and Hamiltonian for the two types of residues at corner, side and core sites on a square lattice. Let o, s and c denote the number of corner, side and core sites, respectively; $n = o + s + c = N^2$ the total number of sites; and the subscripts L and S denote residue type, then

$$\begin{aligned} H &= -s_L + 2ac_L - s_S - 2c_S \\ &= 2an_L - (1+2a)s_L - 2ao_L - n_S - c_S + o_S \end{aligned} \quad (6)$$

TABLE I. Action of the Hamiltonian for the LS model on a square lattice; end points of chains are ignored and $x = z(\sigma) - z(\Gamma)$.

type		corner	side	core
	$z(\Gamma)$	0	1	2
	x	2	1	0
S	$A(x)$	1	1	1
	H	0	-1	-2
	x	1	0	-1
L	$A(x)$	1	1	-a
	H	0	-1	2a

For a given peptide sequence, n_L and n_S are fixed. First consider the case when the steric repulsion is strong but finite, namely, $a \gg 1$. Dropping the corner term o_S one gets for a given peptide sequence,

$$H = -(2a+1)c_S + \text{const.} \approx -2a\mathbf{p}\cdot\mathbf{s} + \text{const.} \quad (7)$$

where $\mathbf{p}$ and $\mathbf{s}$ are the peptide and structure binary vectors defined before, with the exception that in $\mathbf{p}$ the digit 0 (1) now stands for L (S). Comparison of this equation with Eq. (4) reveals that, at least on a square lattice, the mathematical form of the two models are essentially identical, provided that here the pair H and P in the HP model is replaced by S and L, respectively. Since there is only one scale in either model, the size of a does not matter so far as it is much greater than unity but finite.

When $a \to \infty$, as was the case in [8], the term $2ac_L$ in the first line of Eq. (6) becomes a constraint that L residues are prohibited from core sites, namely $c_L = 0$ strictly, and the rest of the Hamiltonian becomes

$$H = -c_S + o_L - n_L + o_S - n_S \approx -\mathbf{p}\cdot\mathbf{s} + \text{const.} \quad (8)$$

which again coincides with Eq. (4).

IV. GEOMETRICAL PROPERTIES OF THE 2D SQUARE LATTICE

Since Eqs. (4), (7) and (8) reduce the Hamiltonians of the mean-field HP and LS models to the same problem in geometry, namely one of the Hamming distance between the two vectors $\mathbf{s}$ and $\mathbf{p}$, we now study the space of these vectors (in the HP model). Consider an $N\times N$ square lattice with $n = N^2$ sites. Recall that every structure is a self-avoiding compact path on the lattice. The set $\mathcal{P}$ of all binary peptides $\mathbf{p}$ is then just the set of 2^n binary sequences. Because of geometric constraints, the set $\mathcal{S}\subset\mathcal{P}$ of binary structure sequences $\mathbf{s}$ is far smaller than $\mathcal{P}$. For a very rough estimate for the upper limit of the size of $\mathcal{S}$, consider the construction of compact paths by random walk on the lattice. At any given point during the walk after the first step, the maximum number of allowed next steps is the coordination number minus one, which is between 2 and 3. As the number of steps taken increases, the average number of allowed next steps will decrease. We take the average number to be 2 up to the point when the lattice is half full. For a randomly chosen path, after the lattice is half full, chances are that the number of allowed next steps will be either one or zero most of the time. So the number of allowed $\mathbf{s}$' should be much less than $2^{n/2}$. On a 6×6 lattice this last number is 262144, whereas the size of $\mathcal{S}$ is 30408, and the size of $\mathcal{P}$ is $2^{36} = 68,719,476,736$. An example of an allowed $\mathbf{s}$ on the 6×6 lattice is shown in Fig. 2 (a). If we think of $\mathcal{P}$ as the set of all the corner points in the n-dimensional unit hypercube, then the set $\mathcal{S}$ is composed of a tiny subset of corner points. It was shown earlier that the designability of an $\mathbf{s}\in\mathcal{S}$ is the Voronoi polytope of $\mathbf{s}$ in $\mathcal{P}$; it is clear what characterizes the designability problem is the distribution of the contents of $\mathcal{S}$ in the unit hypercube.

We now examine how geometric constraints reduce $\mathcal{P}$ down to $\mathcal{S}$. A sequence in $\mathcal{P}$ may be viewed as a chain of 0's and 1's connected by $n-1$ links of three types, those connecting 0 and 0 sites, 0 and 1 or 1 and 0 sites, and 1 and 1 sites, respectively. Let the numbers of such links be n_{00}, n_{10} and n_{11}, respectively. The sequence is partitioned by the 1-0 links into $n_{10}+1$ segments of contiguous 1's or 0's. Whereas the link numbers for a $\mathbf{p}$ are devoid of geometric meaning, that for $\mathbf{s}$ are the consequences of geometric constraints. To illustrate this, consider the case $N > 4$ (the surface to core ratio in smaller lattices are too lop-sided to be of interest). Some of the simplest constraints that must be satisfied by an allowed $\mathbf{s}$ are:

1. An isolated single 0 may only occur at an end of a path;
2. An isolated single 1 may only either occur at or be one 0-segment away from an end of a path;
3. Each of the four corners on the lattice belongs to a 0-segment at least 4 sites long, except when the corner is an end of a path;
4. For a path having the pattern $\mathbf{s} = (1\cdots1)$ (both the ends of the path are 1-sites), $2n_{00} + n_{10} = 8N - 8$ and $2 \le n_{10} \le 4N - 12$;

5. For $\mathbf{s} = (0010011\cdots1)$, $2n_{00} + n_{10} = 8N - 9$ and $5 \le n_{10} \le 4N - 11$;
6. For $\mathbf{s} = (0010011\cdots1100100)$, $2n_{00}+n_{10} = 8N-10$ and $10 \le n_{10} \le 4N - 10$ if $N > 6$, the last relation is replaced by $8 \le n_{10} \le 4N - 10$ if $N \le 6$;
7. For $\mathbf{s} = (0010011\cdots0) \neq (0010011\cdots1100100)$, $2n_{00} + n_{10} = 8N - 10$ and $4 \le n_{10} \le 4N - 12$;
8. For $\mathbf{s} = (0\cdots0) \neq (0010011\cdots0)$ and $\neq (0010011\cdots1100100)$, $2n_{00} + n_{10} = 8N - 10$ and $2 \le n_{10} \le 4N - 12$;
9. For $\mathbf{s} = (0\cdots1) \neq (0010011\cdots1)$, $2n_{00} + n_{10} = 8N - 9$ and $1 \le n_{10} \le 4N - 13$.

The first two rules are obvious on a square lattice. The third rule implies that the polar residues tend to accumulate around corners. This fortuitously reflects a property of real proteins: the relative abundance of polar residues on surface areas with large curvatures. Figs. 2 (b) and (c) illustrate the origin of the fourth rule on a 6×6 lattice. The two structures are both of the type $(1\cdots1)$, that is, they begin and end both on core sites. The dark solid links in the figures define "templates" for constructing $\mathbf{s}$' that respectively have the maximum (twelve) and minimum (two) values for n_{10}. Rules (5)-(8) can be shown in a similar way. By explicitly applying the above rules in the selection of $\mathbf{s}$ (as opposed to requiring an $\mathbf{s}$ to be a compact self-avoiding path), the total number of $2^{36} = 68,719,476,736$ binary sequences in $\mathcal{P}$ is reduced to a set of 537549 candidate paths which, relatively speaking, is now only slightly greater than the exact number (30408) of $\mathbf{s}$' in $\mathcal{S}$. This implies that the set of rules given above embodies the essence of the geometric requirement that guarantees elements in $\mathcal{S}$ to be compact self-avoiding paths.

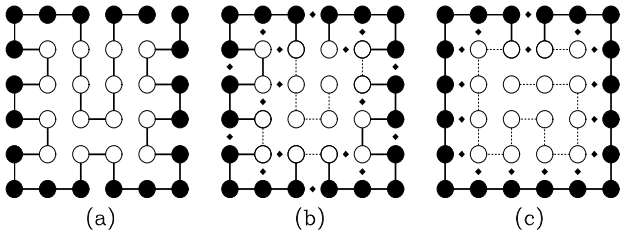

FIG. 2. (a) A structure defined by a compact, self-avoiding path, which is in turn represented by the binary sequence (001100 110000 110000 110011 000011 111100). Black (white) discs represent surface (core) sites coded by the digit 0 (1). In (b) and (c), the dark, solid links define "templates" for constructing structures of the type $(1\cdots1)$ whose n_{10} values are 12 and 2, respectively.

V. DISTRIBUTION OF THE ALLOWED STRUCTURES IN THE HYPERCUBE

Here we show that only a small portion of the structures in $\mathcal{S}$ have large n_{10}. On an $N\times N$ square lattice, there is a total of $2N^2 - 2N$ links and $N^2 - 1$ among them need to be chosen to form a structure. For the 6×6 case these numbers are 60 and 35, respectively. For the structure shown in Fig. 2 (b), of the total number of 60 links on the lattice, 28 links are used to define the template (that has n_{10}=12) and 17 links, marked by filled diamonds in the figure, are forbidden because they would form close loops or connect sites which already have two links. This means that to complete an $\mathbf{s}$ from the template, one needs to select $35 - 28 = 7$ links from among $60 - 28 - 17 = 15$ links on the lattice. Hence at most $\binom{15}{7} = 6435$ $\mathbf{s}$' with $n_{10} = 12$ can be constructed from the template. A similar argument shows that $\binom{23}{14} = 817190$ $\mathbf{s}$' with $n_{10} = 2$ can be constructed from the template shown in Fig. 2 (c), which has 21 predetermined links. The ratio 817190 : 6435 illustrates the point that the number of $\mathbf{s}$' with high n_{10} values is much smaller than the number of $\mathbf{s}$' with low n_{10} values.

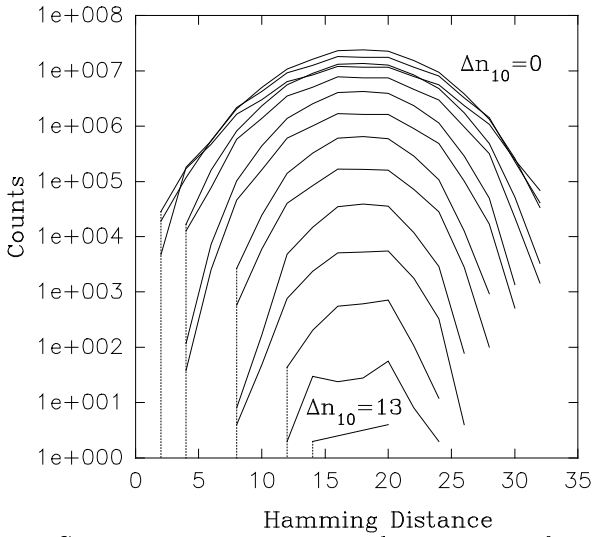

FIG. 3. The Hamming distances between pairs of all the 30408 structural sequences on a 6×6 lattice. The vertical dashed lines indicate the minimal Hamming distances for different Δn_{10}.

We now give a heuristic argument showing that there is an approximate relation between the smallest possible Hamming distance $d_{min}(\mathbf{s}_1, \mathbf{s}_2)$ between two structures $\mathbf{s}_1$ and $\mathbf{s}_2$ and the difference in the n_{10} values of the two structures, Δn_{10}=$n_{10}(\mathbf{s}_1) - n_{10}(\mathbf{s}_2)$; for simplicity we assume that $n_{10}(\mathbf{s}_1) > n_{10}(\mathbf{s}_2)$. For this discussion we ignore the two end points of the structures, so that (on a square lattice) all the segments on an $\mathbf{s}$ partitioned by 0-1 links have at least two 0 or two 1 digits. We begin by considering the case when $\mathbf{s}_2$=$\mathbf{s}_1$. Then both $d(\mathbf{s}_1, \mathbf{s}_2)$ and Δn_{10} are zero. Suppose we can generate $\mathbf{s}_2$ by swapping the positions of a pair of 0's and a pair of 1's in $\mathbf{s}_1$ (while keeping in mind that in most cases such an operation would not give an $\mathbf{s}$; it would give a $\mathbf{p}$ that is not in $\mathcal{S}$). Then $d(\mathbf{s}_1, \mathbf{s}_2) = 2$ and, depending on the position of the replaced pair of 0's in $\mathbf{s}_1$, $\Delta n_{10} = 0$ or 2. Any other

pair of $\mathbf{s}_2$ and $\mathbf{s}_1$ having $\Delta n_{10} = 2$ will have $d(\mathbf{s}_1, \mathbf{s}_2) > 2$. Thus $d_{min}(\mathbf{s}_1, \mathbf{s}_2)$ is 2 for $\Delta n_{10} = 2$. Similarly, if we generate $\mathbf{s}_2$ by exchanging the positions of a pair of 0's and a pair 1's in $\mathbf{s}_1$, for example:

$$\begin{aligned} &(\cdots 0111111110 \cdots 1000000001 \cdots) \\ \rightarrow &(\cdots 0111111000 \cdots 1001100001 \cdots) \end{aligned} \quad (9)$$

$$\begin{aligned} \text{or} \quad &(\cdots 0111111110 \cdots 1000000001 \cdots) \\ \rightarrow &(\cdots 0111100110 \cdots 1001100001 \cdots) \end{aligned} \quad (10)$$

then $d(\mathbf{s}_1, \mathbf{s}_2) = 4$ and $\Delta n_{10} = 2$ (Eq.(9)) or 4 (Eq.(10)). Again any other $\mathbf{s}_2$ and $\mathbf{s}_1$ having $\Delta n_{10} = 2$ or 4 will have $d(\mathbf{s}_1, \mathbf{s}_2) > 4$. Thus $d_{min}(\mathbf{s}_1, \mathbf{s}_2)$ is 4 for $\Delta n_{10} = 4$. Arguing along this line it can be shown that $d_{min}(\mathbf{s}_1, \mathbf{s}_2) \approx \Delta n_{10}$. In Fig. 3, the logarithmic distributions of the Hamming distances between pairs of $\mathbf{s}$' with fixed values of Δn_{10} are plotted for a 6×6 lattice. The relation between $d_{min}(\mathbf{s}_1, \mathbf{s}_2)$ and Δn_{10} is clearly displayed. Notice that all distributions peak at a Hamming distance of 15-20, with the width of the distribution decreasing monotonically with Δn_{10}.

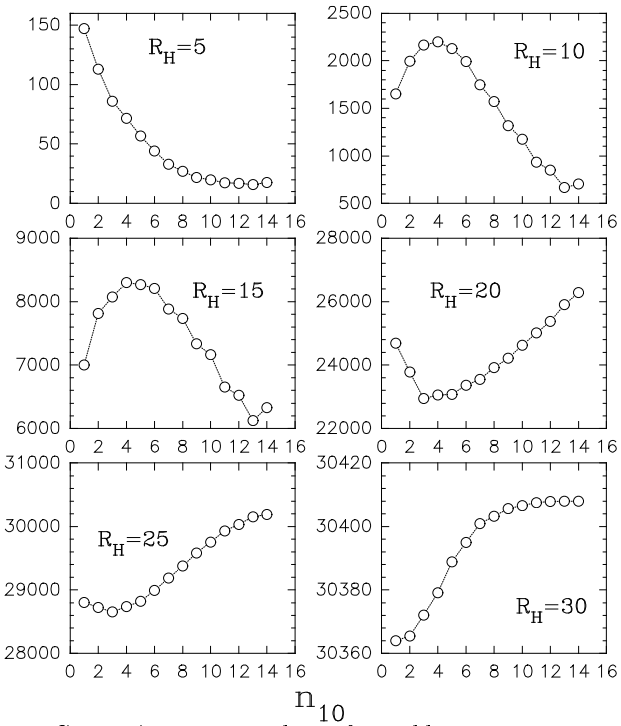

FIG. 4. Average number of neighboring structures within different Hamming distances R_H for a 6×6 lattice.

It has already been shown that the number of $\mathbf{s}$' with large n_{10} is much smaller than the number of $\mathbf{s}$' with small n_{10}. Hence the former kinds of $\mathbf{s}$' will be even more sparsely distributed in $\mathcal{P}$ than the latter kinds. Thus given an arbitrary $\mathbf{s}$ the chances are that most of its nearest neighbors will have relatively small n_{10}'s. An $\mathbf{s}$ with large n_{10} will be farther away from its nearest neighbors than if it has a smaller n_{10}. This is indeed brought out in Fig. 4, where each curve plots as a function of n_{10} the number of neighboring $\mathbf{s}$' in $\mathcal{S}$ within a Hamming distance R_H, averaged over those $\mathbf{s}$' specified by n_{10}. It is seen that so long as $R_H \leq 15$, $\mathbf{s}$' with large n_{10} has far fewer nearby neighbors (in $\mathcal{S}$) than $\mathbf{s}$' with smaller n_{10}. It follows that $\mathbf{s}$' with large n_{10} will on average have large Voronoi polytopes, hence high designabilities. Note that the approximate proportional relation between Δn_{10} and $d_{min}(\mathbf{s}_1, \mathbf{s}_2)$ is not expected to be limited to square lattices although the proportional constant is expected to be dependent on lattice type.

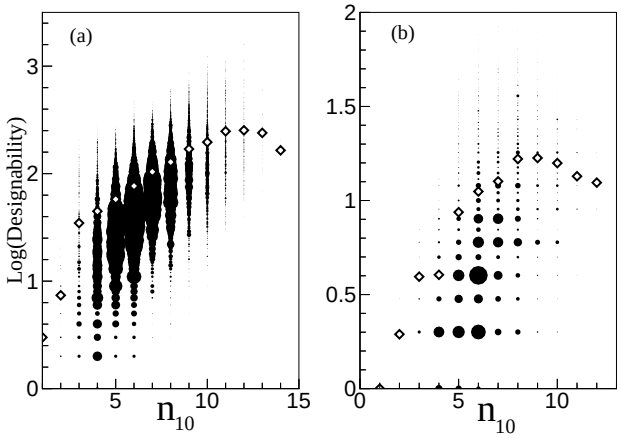

FIG. 5. Designability distributions for (a) 6×6 square lattice and (b) 21-site triangular lattice. See the text for detail.

TABLE II. n_{10}^{max} and n_{10}^{peak} for several lattices

lattice	n_{10}^{max}	n_{10}^{peak}
4 × 4	6	4
4 × 6	9	8
5 × 5	10	7
4 × 7	11	10
5 × 6	12	9
6×6	14	12
21-site triangle	12	9

In Fig. 5 (a) and (b) the logarithmic designability is plotted as a function of n_{10} for a 6×6 square lattice and a 21-site triangular lattice, respectively. The size of each disc indicates the number of $\mathbf{s}$' having the specific n_{10} and designability and an open diamond indicates the average designability of all $\mathbf{s}$' having the specified n_{10}. On the whole the average designability increases with n_{10} up to near the maximum n_{10}. For n_{10} near the maximum value it appears that the heuristic argument given above breaks down, probably partly for boundary effects, and partly because the number of structures with the largest

values of n_{10} is very small (3 for $n_{10} = 14$ and 24 for $n_{10} = 13$ among the 30408 $\mathbf{s} \in \mathcal{S}$ on a 6×6 square lattice) so that statistical fluctuations become important. The designability distributions on several other lattices were studied and the pattern shown in Fig. 5 persisted. The result is summarized in Table II, where n_{10}^{max}, the maximum n_{10} and n_{10}^{peak}, the n_{10} where the largest average designability occurs, are given for each lattice. In all the cases $n_{10}^{peak} = n_{10}^{max} - 2 \pm 1$. Results for three-dimensional lattices will be shown elsewhere.

VI. COMPARISON WITH REAL PROTEINS

It has been shown that the mathematical contents of the mean-field HP model and the LS model are essentially identical. The physical (or biological) interpretations given to the two models are however entirely different. The mean-field HP model is based on the assumption that hydrophobic residues would congregate in the core as much as possible. The LS model is based on the assumption that large residues would be excluded from the core as much as possible. To see which model is based on a more correct premise we compare the results of the two models with real proteins by matching model peptide sequences against protein sequences culled from data banks. For either model, the model sequences are the two sets of sequences among a total $26,000,000$ randomly sampled 36-word binary sequences that select the most highly designable and least designable structures, respectively, on a 6×6 lattice.

We consider the frequency distributions of the set of sequences $\{\mathcal{P}_\lambda | \lambda = h, l, S, \phi, \alpha, \beta, \phi', \alpha', \beta'\}$, where the subscript h denotes the concatenated 27006 peptides mapped to the 15 most highly designable structures in the mean-field HP model; l, the concatenated 24134 peptide sequences mapped to the 1545 least designable structures in the mean-field HP model; S, the concatenated 22789 peptides mapped to the 364 most highly encodable structures in the LS model [12]; ϕ, the concatenated protein sequences in PDB [13], converted to a binary sequences based on the hydrophobicity of the peptides; α, same as ϕ, but includes only segments of protein sequences that fold to α-helices; β, same as ϕ, but includes only segments of protein sequences that fold to β-sheets; ϕ', α' and β', same as ϕ, α and β, respectively, except that protein sequences are converted to binary ones based on the volume of residues. The ten residues designated polar (P) are: Lys, Arg, His, Glu, Asp, Gln, Asn, Ser, Thr, Cys [14] and the ten residues designated as L-type residues are, in descending order of volume, Trp, Tyr, Phe, Arg, Lys, Leu, Ile, Met, His and Gln [15]. That the HP and LS models differ in physical and biological contents is predicated by the fact that the two lists overlap poorly. This predication will not change if the cut-off points of either or both lists are varied slightly. The sequences $\mathcal{P}_h$ and $\mathcal{P}_S$ will be referred to as the most foldable peptides in the HP and LS models, respectively.

To compare the sequences, we employ a Cartesian coordinate representation for symbolic sequences [16], here applied to binary sequences. Let $\mathcal{S}$ denote the set of 2^l binary strings σ of length l. Given a binary sequence $\mathcal{P}_\lambda$ of length L and a string length l (we are interested only in cases when $L >> l$), there is the set $\{f_\lambda^{(l)}(\sigma) | \sigma \in \mathcal{S}\}$ of frequencies of occurrence of the string σ in λ. The frequencies may be obtained, say, by counting while sliding a window l digits wide along λ. The frequency depends on the ratio of 0 to 1 digits in the sequence. This ratio, r_λ, is 0.983, 1.039, 0.553, 0.960, 0.993, 0.720, 0.734, 0.917 and 0.934, respectively, for the sequences $\mathcal{P}_\lambda$, $\lambda = h, l, S, \phi, \alpha, \beta, \phi', \alpha', \beta'$. In order to make a fair comparison of the sequences adjustments need to be made to compensate for the disparity in the 0 to 1 ratios. For this purpose we define a normalized frequency f' by

$$f'^{(l)}_\lambda(\sigma) = (r_\lambda)^{n_\sigma} f^{(l)}_\lambda(\sigma) \tag{11}$$

where n_σ is the number of 0's in σ. Sequences in the normalized frequency set $\{f'^{(l)}_\lambda(\sigma)\}$ now have 0 to 1 ratios equal to unity.

In what follows we consider only cases when l is even, $l = 2k$. Let $\mathcal{L}$ be a $2^k \times 2^k$ lattice with spacing 2^{-k}, and π be a one-to-one mapping from $\mathcal{S}$ to $\mathcal{L}$, $\pi : \mathcal{S} \to \mathcal{L}$ by:

$$\pi(\sigma) = (x, y) \equiv \left(\sum_{i=1}^{k} \sigma_{k+i} \cdot 2^{-i},\ \sum_{i=1}^{k} \sigma_i \cdot 2^{-(k-i+1)} \right) \tag{12}$$

where $\sigma = [\sigma_1, \sigma_2, \cdots, \sigma_{2k}]$ is a string in $\mathcal{S}$ and (x, y) is a site on $\mathcal{L}$. From the set $\{f'^{(l)}_\lambda(\sigma)\}$ we define a normalized relative frequency distribution of λ on the lattice $\mathcal{L}$:

$$F^{(l)}_\lambda(x, y) \equiv F^{(l)}_\lambda(\pi(\sigma)) = \left(f'^{(l)}_\lambda(\sigma) - \bar{f}^{(l)}_\lambda \right) / Z_\lambda \tag{13}$$

where $\bar{f}^{(l)}_\lambda$ is the mean frequency and

$$Z_\lambda = \left(\sum_{\sigma \in \mathcal{S}} f'^{(l)}_\lambda(\sigma) - \bar{f}^{(l)}_\lambda \right)^{1/2} \tag{14}$$

Figs. 6 and 7 show the distributions $F^{(6)}_\lambda$, $\lambda = \phi$, α, β and h, and $\lambda = \phi'$, α', β', and S, respectively. In the figures, the magnitude of the distribution is coded into the gray scale shown at the top of the figures. From the fact that (b) and (d) in Fig. 6 have their brightest and darkest regions, respectively, at generally the same locations, it is evident that $\mathcal{P}_h$ ((d)), the most foldable peptides in the HP-model, is closest to $\mathcal{P}_\alpha$ ((b)), the sequence that represents α-helix segments in real protein sequences. In comparison, although (a) looks similar to (b), it is not so similar to (d). In particular, some of the

brightest regions in (a) are dark in (d), and vice versa. In sharp contrast (c), which represents β-sheet segments in real protein sequences, is entirely different from all the other distributions in Fig. 6.

Turning to Fig. 7, it is noticed that (d), representing the most foldable peptides in the LS model, is very similar to its counterpart in the HP model, Fig. 6 (d). This is as expected because the mathematical contents of the two models are essentially identical. On the other hand, (d) is very dissimilar to (a), which represents all protein sequences in PDB, but with the residues partitioned according to the LS model. This shows that size of the residue is not the most dominant factor in protein structure.

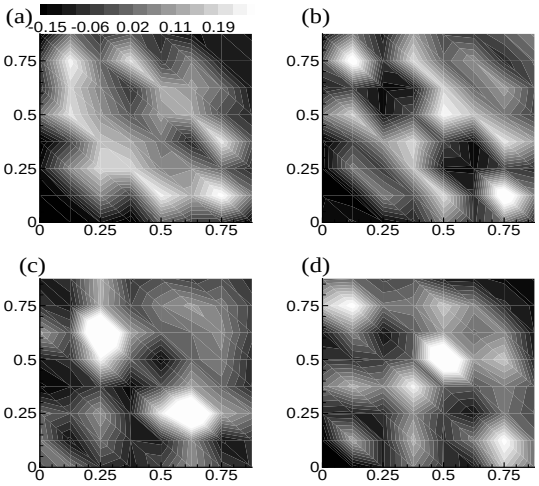

FIG. 6. Frequency distributions of strings of length 6 in the sequences (a) $\mathcal{P}_\phi$, (b) $\mathcal{P}_\alpha$, (c) $\mathcal{P}_\beta$, and (d) $\mathcal{P}_h$; see text for description.

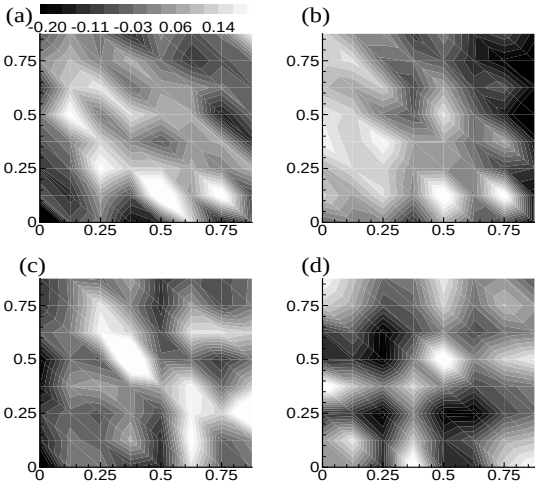

FIG. 7. Frequency distributions of strings of length 6 in the sequences (a) $\mathcal{P}'_\phi$, (b) $\mathcal{P}'_\alpha$, (c) $\mathcal{P}'_\beta$, and (d) $\mathcal{P}_S$; see text for description.

The frequency distributions shown in Figs. 6 and 7 are repeated in Figs. 8 and 9, except that the word length l is now eight instead of six. This implies that the sequences $\mathcal{P}_\lambda$ are now examined with a finer resolution. The result is similar to the $l = 6$ case: the most foldable peptides in the HP model closely resemble the α-helix segments of real protein, while the foldable peptides in the LS model do not resemble real proteins.

The sequences $\mathcal{P}_\lambda$ may be compared in a more quantitative manner through the overlap of frequency distributions:

$$O^{(l)}_{\lambda\lambda'} = \sum_{\sigma\in\mathcal{S}} F^{(l)}_\lambda(\pi(\sigma))F^{(l)}_{\lambda'}(\pi(\sigma)). \tag{15}$$

The overlaps $O^{(l)}_{\lambda\lambda'}$, for a number of pairs (λ, λ') selected from the set $\{h, l, S, \phi, \alpha, \beta, \phi', \alpha', \beta'\}$, and for $l = 4 \sim 14$ are given in Fig. 10.

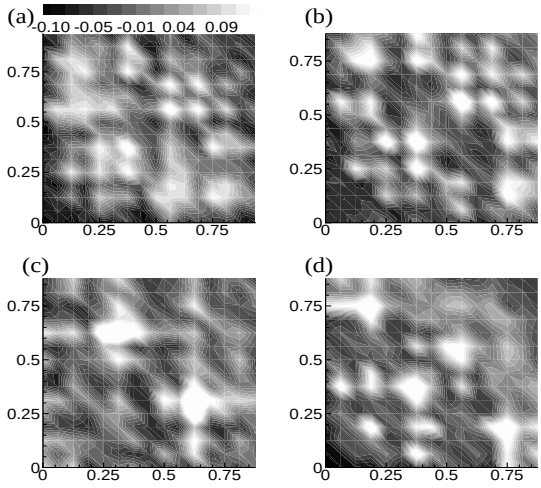

FIG. 8. Frequency distributions of strings of length 8 in the sequences (a) $\mathcal{P}_\phi$, (b) $\mathcal{P}_\alpha$, (c) $\mathcal{P}_\beta$, and (d) $\mathcal{P}_h$.

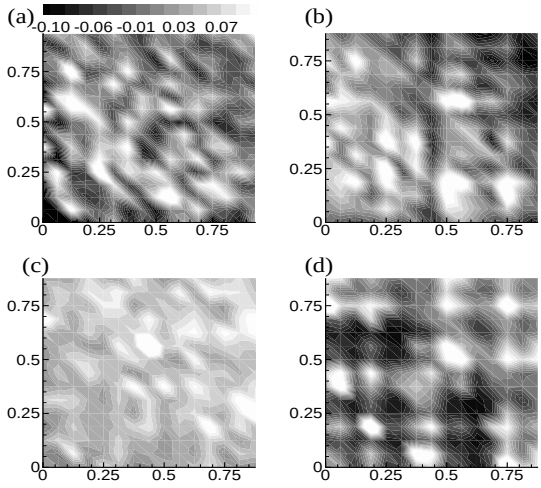

FIG. 9. Frequency distributions of strings of length 8 in the sequences (a) $\mathcal{P}'_\phi$, (b) $\mathcal{P}'_\alpha$, (c) $\mathcal{P}'_\beta$, and (d) $\mathcal{P}_S$.

One first notices that, with the exception of $O^{(l)}_{hS}$ (■ in Fig. 10), all the overlaps approach zero as the word length

l increases. This is so because the resolving power of the method increases with l; for sufficiently large l, the resolution becomes so large that any two sequence that does not have substantial and extended sequence identity will have zero overlap. That $O_{hS}^{(l)}$ has large positive correlation throughout the whole range of l studied is expected from the mathematical equivalence of the HP and LS models. In Ref. [8], the parameter a in Eq.(5) was taken to be infinity to emphasize the steric constraint on the residues. Here we had done the same just to conform to Ref. [8]. On the other hand, since in the present study all the structures are self-avoiding paths on a discrete lattice, the steric constraint caused by the existence of the backbone is automatically satisfied. Therefore, so far as the intention of the LS model is concerned, a small and positive, but not infinite, value for a would have sufficed.

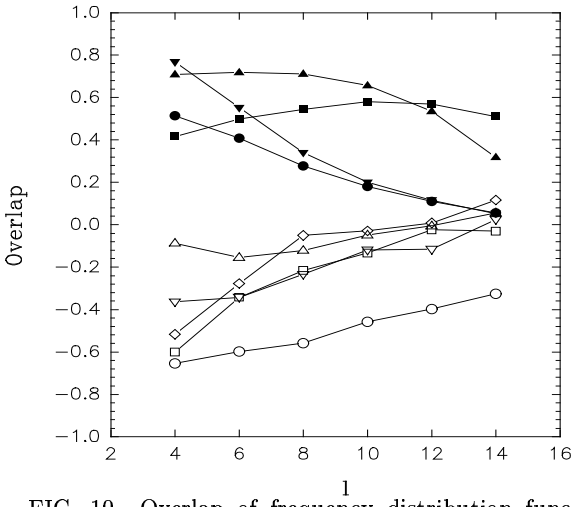

FIG. 10. Overlap of frequency distribution functions versus word length l: $O_{\phi\alpha}^{(l)}$ (filled $\triangle$), $O_{\alpha h}^{(l)}$ (filled $\triangledown$), $O_{\phi h}^{(l)}$ ($\bullet$), $O_{hS}^{(l)}$ ($\blacksquare$), $O_{\alpha' S}^{(l)}$ ($\triangle$), $O_{\beta h}^{(l)}$ ($\triangledown$), $O_{\beta' S}^{(l)}$ ($\diamond$), $O_{\phi' S}^{(l)}$ ($\square$) and $O_{hl}^{(l)}$ ($\bigcirc$). See text for the description of the subscripts $h, l, S, \phi, \alpha, \beta, \phi', \alpha'$ and β'.

The overlap $O_{\phi\alpha}^{(l)}$ (filled $\triangle$) is larger than most other overlaps for much of l's shown in the figure. This is connected to a basic fact of proteins: α-helices account for almost half of the total amount of protein sequences in PDB. The overlap drops sharply when $l \geq 12$ because most α-helix segments are shorter than 15 residues long [19].

Next in order of magnitude are the overlaps $O_{\alpha h}^{(l)}$ and $O_{\phi h}^{(l)}$ (filled $\triangledown$ and $\bullet$); these have large positive values for the smaller l's. This reveals that the mean-field HP model provides a coarse-grained description of some features of the real proteins and suggests that the basic assumption of the model - that local residue-water interaction is the dominant cause for protein folding - is consistent with the mechanism for the formation of α-helices. The overlaps decrease with increasing l for the general reason given above. On the other hand, the negative correlation shown by the negative value of the overlap $O_{\beta h}^{(l)}$ ($\triangledown$) shows that the same assumption is inconsistent with what causes the formation of β-sheets. Two of the obvious reasons are: whereas most β-sheets are buried in the interior of proteins, the mean-field HP model differentiates only surface from core sites but has no means of influencing the interior structure of proteins; the stability of most β-sheets depends on long-range interactions that are absent in the model.

The negative value of the overlaps between $\mathcal{P}_S$ and $\mathcal{P}_{\phi',\alpha',\beta'}$ ($\square$, $\triangle$ and $\diamond$, respectively) indicates that the highly foldable peptide sequences in the LS model are anti-correlated with the real protein sequences for $l \leq 6$ and uncorrelated for larger l. This confirms what is already seen in Figs. 7 and 9: that size effect is not the dominant factor determining the formation of a stable protein conformation. Finally, the large negative values of the overlap $O_{hl}^{(l)}$ ($\bigcirc$) for all values of l tested simply verify that the most and least foldable peptides in the HP model are highly dissimilar however they are compared.

VII. CONCLUSION

The mean-field HP model and the LS model were examined and compared. Both are binary models. The mean-field HP model favors having hydrophobic (H) residues congregating in the core of the protein, while the LS model discourages (when a is finite) or bars (when $a=\infty$) large (L) residues from the core. It was demonstrated that although the two models are expressed in different languages, their mathematical contents are essentially identical when H and P in the HP model are identified with S and L, respectively, in the LS model. However, because these identifications are not realized in nature - a half of the ten largest residues are hydrophobic - the two models describe different physics.

In a two-dimensional lattice model, where potential native conformations are restricted to the set of compact self-avoiding paths on the lattice, the search for the most designable structures in both models can be mathematically reduced to the search for structures with the largest Voronoi volumes in the n-dimensional unit hypercube. In a previous paper it was asserted that structures with the highest designabilities are essentially structures with the largest number of surface-core switch-backs [3]. Here it was demonstrated that indeed there is a close connection between a structure having a large Voronoi volume and a large number of surface-core switch-backs. The connection is a property of binary strings and not a property of the dimensionality of the lattice.

The most designable structures from both models were compared with binary sequences converted from protein sequences culled from data banks. The comparison were made by representing long binary sequences by the set

of frequencies of occurrence of binary strings of a fixed length. The set corresponding to each sequence was presented graphically as distribution in a Cartesian representation (see Eq.(12)) for visual inspection and comparison. Overlaps of sets were also computed to get a quantitative measure of the similarity of the sequences. It was observed that, as expected, the most designable structures from the two models are very similar. However, when the 0 and 1 digits are given their designated meanings - polar and hydrophobic residues, respectively, in the HP model and large and small residues respectively, in the LS model, a different story emerges. The most foldable binary peptide sequences in the mean-field HP model matched well with the α-helix segments of protein sequences, but poorly with the β-sheets segments of protein sequences. But those in the LS model do not match well with any protein sequences.

Because the mean-field HP model is very coarse-grained, it can be expected to model only early protein folding. This study suggests that the rough formation of α-helices and the collapse of proteins into globular shapes are primarily determined by hydrophobicity, and that most α-helices are located near the surface of proteins. The size of the residues does not seem to be an important deciding factor. That the most designable structures in the HP model do not look like β-sheets could be the result of several factors. One is the constraint on the structures imposed by a square lattice; β-sheets have a preponderance of HP repeats [19] which, contrary to HHPP repeats that characterize α-helices, are not energetically favored on a square lattice. Another could be that hydrophobicity is not the dominant factor that causes the formation of β-sheets; it is known that many motifs of secondary structure involve hydrogen bonds between residues on β-sheets that are distant from each other on the primary structure. A third factor, related to the first two, is that β-sheets are mostly buried in the protein. Since the mean-field HP model treats all core sites the same, it does not have any mechanism to promote the formation of β-sheets. If hydrophobicity is indeed the main driving force in early folding, then we can better understand why the formation of α-helices and the collapse would happen on a similar time scale, of the order 10^{-7}s [20,21], and why it takes ten times longer for the formation of β-sheets, which involves interactions between residues distantly separated on the primary structure. This scenario is consistent with the finding in a recent statistical analysis of experimental data: local contacts play the key role in fast processes during folding [22].

We thank the National Center for High-Performance Computing (NCHC) for providing support in computation and accesses to PDB. This work is partly supported by grants NSC89-2213-E-321-004 to ZYS, NSC89-M-2112-008-0022 to HCL and NSC87-M-2112-007-004 to BLH from the National Science Council. HCL thanks the Physics Department of Stanford University where this work was partly written.

[1] C. Anfinsen, Science **181**, 223 (1973).
[2] K.A. Dill, Biochemistry **24**, 1501 (1985); H.S. Chan and K.A. Dill, Macromolecules **22** 4559 (1989).
[3] C.T. Shih, Z.Y. Su, J.F. Gwan, B.L. Hao, and C.H. Hsieh, and H.C. Lee, Phys. Rev. Lett. **84**, 386 (2000).
[4] E.I. Shakhnovich, Phys. Rev. Lett. **72**, 3907 (1994); P.G. Wolynes, J.N. Onuchic and D. Thirumalai, Science **267**, 1619 (1995); H.S. Chan and K.A. Dill, Proteins **24**, 335 (1996); C. Micheletti, F. Seno, A. Maritan, and J.R. Banavar, Phys. Rev. Lett **80**, 2237 (1998); F. Seno, C. Micheletti, A. Maritan, and J.R. Banavar, Phys. Rev. Lett **81**, 2172 (1998).
[5] H. Li, R. Helling, C. Tang and N.S. Wingreen, Science **273**, 666 (1996).
[6] H. Li, C. Tang, and N.S. Wingreen, Natl. Acad. Sci. USA **95**, 4987 (1998).
[7] E.I. Shakhnovich, Curr. Biol. **8**, R478 (1998).
[8] C. Micheletti, J.R. Banavar, A. Maritan, and F. Seno, Phys. Rev. Lett. **80**, 5683 (1998).
[9] N.E.G. Buchler, and R.A. Goldstein, Proteins **34**, 113 (1999).
[10] H. Li, C. Tang, and N.S. Wingreen, Phys. Rev. Lett. **79**, 765 (1997).
[11] M.R. Ejtehadi, N. Hamedani, H. Seyed-Allaei, V. Shahrezaei, and M. Yahyanejad, Phys. Rev. **E 57**, 3298 (1998).
[12] (It turns out that in the LS model, because an L (*i.e.*, large) residue is strictly forbidden - when $a = \infty$ - to occupy a core site, for a same set of sample peptides, the encodabilities of highly encodable structures are generally much lower than the designabilities of highly designable structures in HP model.)
[13] Protein Data Bank ver.91, released Jan. 2000; H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Research, **28**, 235-242, 2000.
[14] A. Radzicka et al., Biochemistry **27**, 1664 (1988).
[15] A.A. Zamyatin, Prog. Biophys. Mol. Biol. **24**, 107 (1972).
[16] Bai-Lin Hao, and Wei-Mou Zheng, *Applied Symbolic Dynamics and Chaos*, World Scientific, 1998.
[17] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee and M.H. Zehfus, Science **229**, 834 (1985).
[18] Wen-Hsiung Li, *Molecular Evolution*, (Sinauer Associates, 1997) p. 14.
[19] J.L. Lo, private communication.
[20] V. Munõz, P.A. Thomson, J. Hofrichter and W.A. Eaton, Nature **390**, 196 (1997).
[21] S. Williams *et al.*, Biochemistry **35**, 691 (1996).
[22] K.W. Plaxco, K.T. Simons and D. Baker, J. Mol. Biol. **277**, 985 (1998); H.S. Chan, Nature **392**, 761 (1998).

APPENDIX

Here we show how the two lattice models differ by comparing strings of several lengths that have the high-

est and lowest frequencies of occurrence, called the most and least favored strings, respectively, in the sequences $\mathcal{P}_h$ and $\mathcal{P}_S$, which are the concatenated sequences of peptides mapped to the group of most designable structures in the mean-field HP and LS models, respectively. In Table III, the first and sixth columns list such strings. Strings of different lengths are ranked separately by their normalized relative frequency of occurrence (Eq. (14)); the string with the highest (lowest) frequency is ranked 1 (2^l). By definition, an unfavored string has negative frequency. Table III shows that the most favored strings are quite well correlated in the two models but the least favored strings are not so. It is seen that the repeats (0011) are the most favored pattern in both models, long repeats of 1's and 0's are the least favored string patterns in the HP model and (01) are the the least favored string repeats in the LS model. The reason for this is clear: (0011) repeats are the favored pattern in most foldable structures in both models and each of the (peptide) strings (0000), (1111) and (0101) is separated from it by the greatest *frame independent* Hamming distance. There is an additional disincentive for a peptide to have (01) repeats in the LS model. On a square lattice such repeats do not appear in a structure sequence, hence, with L-type residues (represented by 0 digits) strictly forbidden on core sites (represented by 1 digits), a peptide string with 01 repeats can only occupy a structure sequence composed entirely of surface sites. This gives the peptide zero binding energy in the LS model. The situation is different in the HP model. There a peptide string with 01 repeats can occupy a structure sequence with 0011 repeats and non-zero binding energy.

TABLE III. Strings most and least favored in the mean-field HP and LS models. Strings of different lengths are ranked separately; e.g., the least favored string of length 4 is ranked 2^4=16.

Strings most/least	HP model		LS model		Strings most/least	LS model		HP model	
favored in HP model	freq.	rank	freq.	rank	favored in LS model	freq.	rank	freq.	rank
(0110)	0.4459	1	-0.0468	10	(0011)	0.3834	1	0.4272	2
(0011)	0.4272	2	0.3834	1	(1100)	0.3693	2	0.4224	3
(0000)	-0.3883	15	0.2732	3	(1010)	-0.3815	15	-0.1572	11
(1111)	-0.3903	16	0.0109	9	(0101)	-0.3892	16	-0.1594	12
(001100)	0.4605	1	0.2694	1	(001100)	0.2694	1	0.4605	1
(011001)	0.2746	2	0.0656	20	(000011)	0.2694	2	0.0515	18
(100110)	0.2698	3	0.0672	19	(110000)	0.2680	3	0.0369	23
(000001)	-0.1725	62	0.0379	22	(101010)	-0.2186	62	-0.1253	58
(100000)	-0.1741	63	0.0385	21	(010101)	-0.2222	63	-0.1234	57
(000000)	-0.2694	64	0.0274	25	(001010)	-0.2224	64	-0.0589	39
(00110011)	0.2101	1	0.1016	19	(11000011)	0.2318	1	0.1875	4
(01100110)	0.2089	2	0.0541	51	(00001100)	0.2141	2	0.1332	15
(11001100)	0.1977	3	0.1001	20	(00110000)	0.2110	3	0.1191	23
(11000011)	0.1875	4	0.2318	1	(00111100)	0.1684	4	-0.0466	200
(00000011)	-0.0927	253	0.0293	74	(01010100)	-0.0989	253	-0.0401	180
(00000001)	-0.1015	254	0.0301	72	(01010010)	-0.1008	254	-0.0418	188
(10000000)	-0.1023	255	0.0334	63	(01001010)	-0.1013	255	-0.0436	194
(00000000)	-0.1060	256	0.0088	94	(00101010)	-0.1017	256	-0.0379	172
(0011001100)	0.1682	1	0.902	14	(0011000011)	0.1837	1	0.1400	4
(1100001100)	0.1574	2	0.1830	2	(1100001100)	0.1830	2	0.1574	2
(0110000110)	0.1548	3	0.1335	3	(0110000110)	0.1335	3	0.1548	3
(0011000011)	0.1400	4	0.1837	1	(1001100001)	0.1230	4	0.1211	8
(1111000000)	-0.0408	1021	0.0220	214	(0101001010)	-0.0441	1021	-0.0173	693
(1110000000)	-0.0414	1022	0.0508	58	(0100001010)	-0.440	1022	-0.0102	528
(0000000000)	-0.0426	1023	-0.0219	773	(0101010101)	-0.0444	1023	0.0268	893
(1111111111)	-0.0427	1024	-0.0358	914	(1010101010)	-0.0446	1024	0.0250	869