Sequence-structure relationships in proteins and copolymers

Kaizhi Yue and Ken A. Dill
Department of Pharmaceutical Chemistry, Box 1204, University of California at San Francisco, San Francisco, California 94143
(Received 14 January 1993)

We model proteins as copolymer chains of *H* (hydrophobic) and *P* (polar) monomers configured as self-avoiding flights on three-dimensional simple-cubic lattices. The *HH* interaction is favorable. The folding problem is to find the "native" conformation(s) (lowest free energy) for an *HP* sequence. Using geometric proofs for self-avoiding lattice chains, we develop equations relating a monomer sequence to its native structures. These constraint relations can be used for two purposes: (1) to compute a tight lower bound on the free energy of the native state for *HP* sequences of any length, which is useful for testing conformational search strategies, and (2) to develop a search strategy. In its present implementation, the search strategy finds native states for *HP* lattice chains up to 36 monomers in length, which is a speedup of 5–15 orders of magnitude over existing brute-force exhaustive-search methods.

PACS number(s): 87.10.+e

I. INTRODUCTION

The protein-folding problem is the question of how a linear polymer chain, composed of a specific sequence of amino acids, encodes the unique three dimensional structure to which it folds. The relationship between the amino-acid sequence, on the one hand, and the "native" conformation (i.e., of lowest free energy) of a protein chain, on the other hand, has been explored using simple lattice models [1–4]. In the *HP model* [3,5–7], the 20 different amino acids are assumed to fall into two classes: hydrophobic (*H*) or polar (*P*). Chains are configured as self-avoiding walks on two-dimensional square lattices or three-dimensional cubic lattices. *HH* contacts are favorable, and are assumed to be the dominant interaction [8], so under strong folding conditions the native conformations are those that have the greatest number of *HH* contacts. For chains that are sufficiently short, the globally optimal states have been found by brute-force exhaustive-computer enumeration [5-7]. The *HP* model has the following proteinlike features. When the *HH* sticking energy is small, the chains have an ensemble of open conformations (the "denatured state"), but when sticking is strong, chains with certain sequences of *H* and *P* monomers collapse, through a relatively sharp transition [5,9], to a small ensemble of compact states (often only one or two) [5,6], with cores of *H* monomers, comprised of about the same distribution of helices and sheets as in the known proteins [10,11]. *HP* lattice proteins also resemble real proteins in some mutational [3,7,12] and kinetic [13,14] properties.

A virtue of this simple model is that its partition function can be enumerated exactly, but a major problem is that the global optima cannot be found for longer chains on three-dimensional lattices because the computer time for brute-force enumeration is prohibitive and increases exponentially with chain length [6]. To search for native conformations of longer chains in the *HP* lattice model, O'toole and Panagiotopoulos have developed efficient Monte Carlo search procedures [9], and Unger and Moult have developed a genetic algorithm [15].

One approach to studying three-dimensional chains through exact enumeration involves the use of a somewhat different model. The "perturbed homopolymer" model [1,2] assumes all monomers (*H* and *P*) are sufficiently strongly self-attractive that native states are guaranteed to be among those that are maximally compact. Energetic differences between *H* and *P* monomers are taken to be a small perturbation relative to the strong background attraction of all monomers for each other. The 27-monomer-chain cube has been studied in this model [1,2]. Exhaustive enumeration is computationally prohibitive for longer chains in three dimensions in either the *HP* or perturbed homopolymer models.

Here we explore a different strategy to find native states for longer chains on three-dimensional lattices in the *HP* model. We analyze the geometric packing constraints for models of chains, using the method of discrete geometry [16,17]. We then develop an equation relating an amino-acid sequence to certain features of its compact native confirmations. The search for native states is formulated in terms of a search for conformations that have a core of *H* monomers of minimal surface area. This constrained optimization is treated at three different levels of increasingly detailed accounting for the chain connectivity and sequence. At each level, we can predict the general characteristics of the *H* cores and compute upper bounds on the maximum number of *HH* contact achievable by a given sequence. Such bounds can be used to learn how successful are sampling strategies, such as Monte Carlo, simulated annealing, genetic algorithms, i.e., how closely they come to finding globally optimal conformations. The constraint relations are also useful in guiding a search to find native conformations. A search program is described.

II. THE MODEL AND DEFINITIONS

First, we define some terms. We consider copolymer chains, each consisting of a specific sequence of *H* and *P*

monomers. Let L be the total number of monomers in the chain, and n_H be the number of H-type monomers. Within a sequence, a run of monomers of a single type is called a *segment*. For example, the sequence *PHHHP* contains an H segment of length 3, and *HPPPPPPH* contains a P segment of length 6. A *P singlet* is a P segment of length 1, i.e., the P in ...*HPH*.... A chain is configured as a self-avoiding flight on a lattice. The coordination number, z, is the number of nearest neighbors of a site. Therefore a lattice site or a monomer has z *sides*. For the three-dimensional simple cubic lattices considered here, $z=6$. Sites not occupied by monomers are occupied by solvent molecules. Covalent links are referred to here as *bonds*. A *residue* (i.e., monomer) makes a *contact* with another residue if the two monomers are neighbors on the lattice but are not covalently linked (see Fig. 1). A contact between two H monomers has a favorable free energy, $-\varepsilon$, $\varepsilon>0$, and the contact free energy for all other types of contacts is 0. Conformations are *native* if they have the lowest free energy among all the possible conformations, in a strong folding solvent (i.e., as $\varepsilon\rightarrow\infty$). Thus native conformations have the maximum possible number of *HH* contacts for the given sequence. A sequence may have one or more native structures.

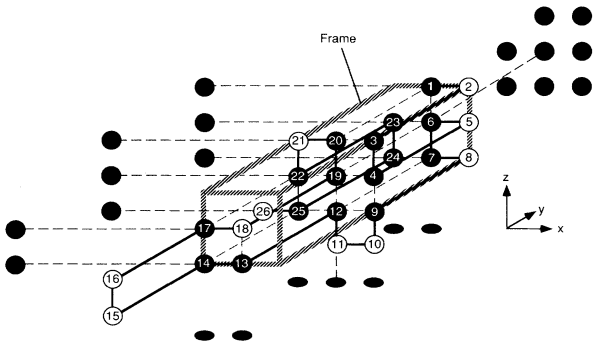

FIG. 1. *HP* sequence and a native conformation. H monomers are dark; P monomers are white. The order of residues in the chain is indicated by numbers. A dark solid line between residues indicates a bond; a light dashed line a contact. For example, there is an *HH* contact between $H6$ and $H19$. The projection areas of *H monomers in the H core* in x, y, and z dimensions (alternatively on y-z, x-z, and x-y planes) are shown as the filled circles or ovals on the sides of the conformation and underneath it. The projection directions are shown as dashed lines. For x dimension, the directions for all the projections of the H residues are shown. For y and z dimensions, only the directions for the center projections are shown. The frame of the H core is indicated by a rectangular solid of thick shaded lines, which has a dimension of $3\times3\times3$. The H core has three *layers* along the z axis. From the top down, the first layer contains $H1$, $H3$, and $H20$; the second $H4$, $H6$, $H17$, $H19$, $H22$, and $H23$; and the third $H7$, $H9$, $H12$, $H13$, $H14$, $H24$, and $H25$. The projection areas in the x dimension are 2, 3, and 3. Projection areas in the y dimension are 2, 3, and 3. The projection area of the H core onto the xy plane is 7.

III. THE FOLDING EQUATION: A RELATIONSHIP BETWEEN SEQUENCE AND STRUCTURE

In this section we develop an equation relating the sequence and the native conformations of an *HP* sequence, for any lattice. This equation describes the *conservation* of sides of H monomers. The top panel of Table I is an accounting of all the sides of H monomers of a given sequence, the total number of which is zn_H. Here b_{HH} and b_{HP} are the numbers of *HH* and *HP* bonds. h_{term} is the number of chain ends that are H residues: i.e., $h_{\text{term}}=0$, 1, or 2. The bottom panel of Table I enumerates sides of H monomers in various contacts and bonds in a given chain conformation. They also sum to zn_H. Here t_{HH}, t_{HP}, and $t_{H\text{-solvent}}$ are the numbers of *HH*, *HP*, and *H*-solvent contacts, respectively. Equating the sequence contributions from the top panel to the conformational contributions from the bottom panel of Table I gives

$$2b_{HH}+b_{HP}+h_{\text{term}}+(z-2)n_H=2t_{HH}+2b_{HH}+b_{HP}+t_{HP}+t_{H\text{-solvent}}\ . \qquad (1)$$

Let G be the number of H segments in a chain. Since the end of an H segment is either a bond between an H monomer and a P monomer or is the end of the chain, the total number of *ends* of H segments is $b_{HP}+h_{\text{term}}$. Since each H segment has two ends, we have

$$G=(b_{HP}+h_{\text{term}})/2\ .$$

For example, in the sequence *PHHPHHHPPH*, $b_{HP}=5$, $h_{\text{term}}=1$, and $G=3$.

We define the *H core(s)* of a conformation as a collection of H monomers connected by contacts or covalent bonds. The *surface area* of the H core, S, equals the number of sides of H monomers that neither contact nor bond to other H monomers. Such sides either adjoin P monomers (through contacts or covalent bonds) or they adjoin solvent sites. Hence $S\equiv b_{HP}+t_{HP}+t_{H\text{-solvent}}$. Thus the surface of the H core involves *all exposed* sides, i.e., that are either bonded to P monomers or just contacting P's or solvent.

A conformation can have one or more H cores. Using the definitions of G and S, Eq. (1) becomes

$$t_{HH}+\frac{S}{2}=G+\frac{(z-2)n_H}{2}\ . \qquad (2)$$

We refer to Eq. (2) as the folding equation. It has the useful property that the right-hand side depends only on the sequence while the left-hand side depends on the conformation. Hence the right-hand side is a constant, for a given sequence. In the *HP* model, the folding problem is that of finding the conformation of maximum t_{HH}. But if the right side is constant for a given sequence, then maximization of t_{HH} is equivalent to minimization of S. Hence Eq. (2) shows that the folding problem can be reduced to the problem of finding the conformations that have the smallest H core surface area S, designated S_{min}, for a given sequence. Since the surface area is a global property, and since mathematical and computational

TABLE I. The sides of H residues in the sequence and in a conformation.

H residue sides in the sequence	
Sides due to covalent bonds between H-H	$2b_{HH}$
Sides due to covalent bonds between H-P	b_{HP}
Nonbonded sides of nonterminal H residue	$(z-2)\times(n_H-h_{\text{term}})$
Nonbonded sides of terminal H residues	$(z-1)\times h_{\text{term}}$
H residue sides in a conformation	
Sides due to H-H covalent bonds and H-H contacts	$2b_{HH}+2t_{HH}$
Sides due to H-P covalent bonds and H-P contacts	$b_{HP}+t_{HP}$
Sides due to H-solvent contacts	$t_{H\text{-solvent}}$

tools are more readily available for surface-area-minimization problems, this change of venue offers advantages for finding native structures.

IV. SURFACE AREAS OF H CORES

We first describe a few properties of H cores. If an H core is a simple solid, with no indentations or cavities (i.e., with no internal sites that contain solvent or P monomers, to be defined more rigorously below), then its surface area will equal the sum of its orthogonally projected areas:

$$S=2(s_x+s_y+s_z)\ , \tag{3}$$

where s_x, s_y, and s_z are the *projected area* of H's in the x, y, and z dimensions, i.e., onto the yz, xz, and xy planes, respectively. The projected area is computed as follows. If at position $(0,y_1,z_1)$ one or more H's occur in a column along the x axis, then it contributes 1 unit to s_x; if there is no H at $(0,y_1,z_1)$ in a column along the x axis; then it contributes 0 to s_x. The quantity s_x is the sum of such contributions over all positions $(0,y,z)$, on the yz plane. The factor of 2 in the equation above arises because solids have 2 surfaces normal to each of the x, y, and z axes. For examples, see Figs. 1 and 2.

More generally, if an H core has H *cavities,* i.e., sites where P residues or solvent are sandwiched between H residues, then S will instead by given by

$$S=2(s_x+s_y+s_z)+Q\ , \tag{4}$$

where Q is the surface area of the H cavities, i.e., the number of sides of H monomers enclosing the cavity; see Fig. 2(b1) and 2(b2). Note that Q is the *excess* surface contributed by buried P residues (or solvent), i.e., relative to an H core that has no buried P residues.

The positions (sites) in an H core conformation can be classified by their numbers of neighbors. For example, *corner, edge, face, or interior* sites have 3, 4, 5, and 6 neighbors, respectively. The location of a P monomer in an H core determines its energy cost (i.e., its loss of HH contacts). If a P residue is surrounded by H residues at an edge position, then $Q=2$ because the P is sandwiched by two H residues (along the edge direction) and shares two sides with H residues (regardless of whether the H residues bond or contact the P residue). At a face position, $Q=4$, because a P residue is surrounded by four H residues. At an interior position, $Q=6$. We call the minimal rectangular solid that completely contains the H core the *frame* (see Fig. 1). A frame is a useful feature of a conformation. It can be determined from knowledge of the HP sequence and serves as a constraint for pruning the search tree.

On a cubic lattice, an H core can be decomposed into a stack of *planar layers* of H monomers along some coordinate direction, as in the stacking of slices of bread; see Fig. 3. Each layer has the thickness (normal to the "plane") of one lattice site. Layers may have different shapes. Each layer is characterized by a cross-sectional area, which is the surface area normal to the direction of the stacking. Also each layer has a lateral surface area [18], which is the surface area in the four directions perpendicular to the direction of stacking. Examples are shown in Fig. 3.

We define the *body of an H core* to be the maximal rectangular solid that contains H residues but contains no P monomers or solvent. Usually, the remaining H monomers (that are not contained within the body of an H core) form groups of layers which are appendages upon the H-core body. That is, in the direction of their layer stacking, the projection of these layers is completely blocked by the H-core body, so their only contribution to the surface area of the H core will be their lateral surface areas. We call these *barnacle layers* [e.g., layers 1, 2, 6, and 7 in the x direction in Fig. 3(b)].

V. SEQUENCE-STRUCTURE RELATIONSHIPS

A. First approximation: The disjoint segment model

To investigate conformations with minimal H-core surface area for a given HP sequence, we begin with a simplest model in which chain connectivity among the H segments is neglected. Since disconnected H segments have greater freedom to configure within a volume than fully connected chains, the disjoint segment model will give a lower bound for the surface area of an H core that could be achieved by an HP sequence.

It is shown in the Appendix that in order for unconnected H segments to configure to have minimal surface area, there can only be one H core and it cannot have any H cavities. It is also shown in the Appendix that for the disjoint segment model, every H core of minimal surface area can be reconfigured to have an H-core body with a single barnacle layer without changing its frame dimen-

sions. Therefore, for a given sequence σ, the problem of finding the lower bound for the minimal H-core surface area and of finding the associated frame dimensions can be formulated as follows. Let $S^1(\sigma)$ represent the surface area of an H core that contains disjoint segments. The superscript 1 refers to the first approximation, the dis-

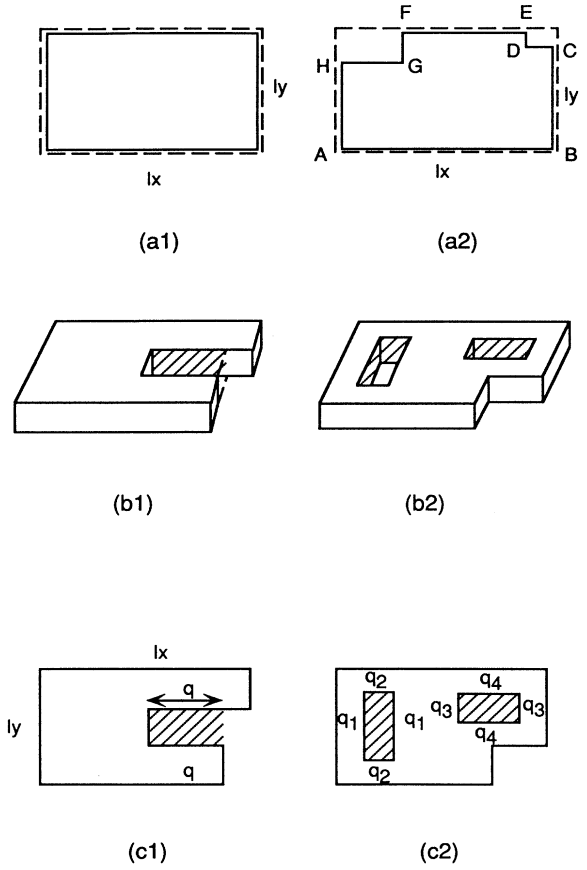

FIG. 2. This group of figures illustrates how projection works in cases with or without cavities. To simplify the problem, we have used two-dimensional (2D) shapes (a1) and (a2) and 2D projections of 3D shapes (c1) and (c2). In these figures, each shape given in solid lines is a "H core" in 2D. The boundaries between individual H residues are omitted. In (a1) and (a2), the projections of the H core in two dimensions are shown as dotted lines, which form a rectangular "frame." There is no cavity in the H core. As a result, the H core has the same perimeter as the rectangular frame that tightly encloses it. That is, the *projection length* of the H core to the two dimensions is exactly half of its perimeter. Let p be the parameter of an H core, $l_x + l_y = p/2$. Applying the same argument, we can show that in 3D, the *projection area* is exactly half of the surface area of an object, when no cavities exist. (b1) and (b2) give examples of H cores with cavities. Their corresponding projections in 2D are shown respectively in (c1) and (c2). Cavities are marked by shaded area. In 2D, cavities add extra length to the total perimeter of non-H contacts, as indicated by the q's. Cavities include indentations (c1) or holes (c2). The added perimeter (added surface area in three dimensions) is not the sum of all lengths enclosing the cavity, but of only the two sides of H residues that sandwich it.

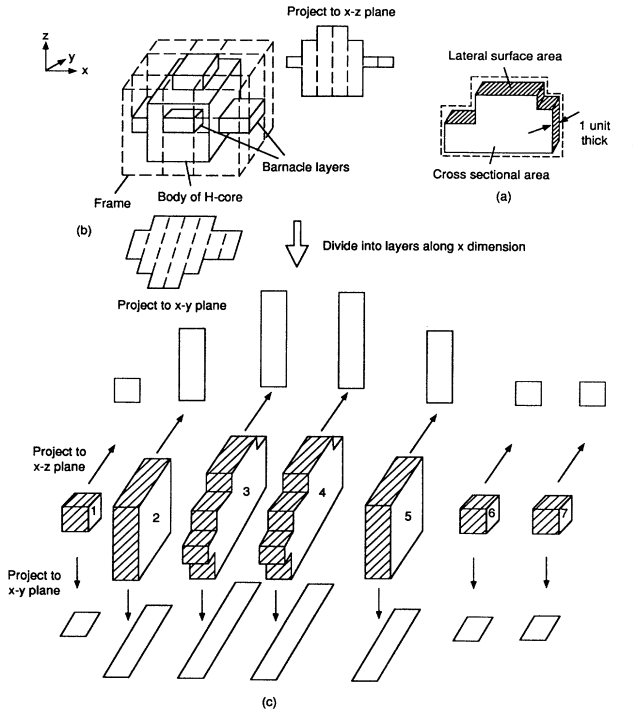

FIG. 3. Illustration of dividing an H core into layers. The H core contains H residues only; it is not a conformation of the HP chain. Here the H core is represented by the collection of unit cells occupied by individual residues. However, the boundaries between individual H residues are not drawn. (a) defines measures of a layer. Because of the unit thickness, the lateral surface area equals the length of the perimeter. (b) shows the division of an H core into a main body and four groups of barnacle layers. Projection areas in two directions are shown. (c) shows division into layers along the x direction. The layers then each have projections in z and y dimensions. It is clear that the sum of their lateral surface areas is the same as the lateral surface area of the H core.

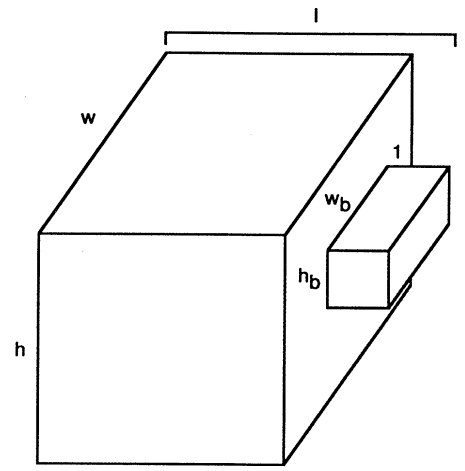

FIG. 4. This figure shows a layout of the H core under the assumption of unconnected H segments. Note that there is a single barnacle layer, indicated by the length being 1.

joint segment model. Let l, w, h be the length, width, and height, of a frame; see Fig. 4. We denote the frame dimensions with brackets: $\langle l,w,h \rangle$. Assume the single barnacle layer is attached to the H-core body along the axis labeled with the length l. Let w_b and h_b be the width and height, respectively, of the barnacle layer in the two directions perpendicular to the l axis. Find the dimensions l, w, h, w_b, h_b which minimize

$$S^1(\sigma)=2[(l-1)(w+h)+wh+w_b+h_b] \tag{5}$$

subject to the constraints

$$lwh \geq n_H \ , \tag{6}$$

$$w_b+h_b \leq w+h \ , \tag{7}$$

$$2(w_b+h_b)=P_m(v_b) \ , \tag{8}$$

where

$$v_b=n_H-(l-1)wh \ . \tag{9}$$

In Eq. (5), the first two terms of the right-hand side give the surface area of the H-core body while w_b and h_b are areas of the barnacle layer (since they are linear dimensions multiplied by the thickness, which is 1). The constraint relations arise from the following considerations. Equation (6) accounts for the requirement that all H segments must appear in either the H-core body or the barnacle layer of the H core, i.e., inside the frame. Equation (7) enforces the definition that the perimeter of the barnacle layer is smaller than the perimeter of any layer in the H-core body; Eqs. (8) and (9) specify the area and volume of the barnacle layer. In Eq. (8), $P_m(N)$ is the minimal perimeter around N residues on a square lattice [19],

$$P_m(N) \equiv 2(a+\lceil N/a \rceil), \quad \text{where } a=\lceil \sqrt{N} \rceil \ . \tag{10}$$

The minimum value of $S^1(\sigma)$ of Eq. (5), denoted $S_m^1(\sigma)$, can be readily found by the Kuhn-Tucker multiplier method using integer mathematical programming [20]. Since disjoint segments are less constrained than fully connected chains, any configuration of H monomers that can be achieved by a fully connected chain can be achieved in the disjoint model. Therefore, the surface area obtained from the disjoint segment model is a lower bound on the quantity we seek, namely, $S_{\min}$; i.e.,

$$S_m^1(\sigma) \leq S_{\min} \ .$$

More than one frame can usually be found from this minimization process. For example, if $n_H=21$, then the minimal surface in the disjoint segment model, $S_m^1=50$, is achieved either by frame $\langle 3,3,3 \rangle$ or by $\langle 4,3,2 \rangle$ [both have $2(w_b+h_b)=8$] [21].

It can be shown (proof omitted) that one frame that is always a solution is that which is closest in shape to a cube [22]. Its dimensions are

$$l_0=\lceil \sqrt[3]{n_H} \rceil \ , \tag{11}$$

$$w_0=\lceil \sqrt{n_H/l_0} \rceil \ , \tag{12}$$

$$h_0=\lceil n_H/l_0w_0 \rceil \ . \tag{13}$$

Substituting Eqs. (11)–(13) into Eq. (5) gives an expression for the minimal surface area in the disjoint segment model:

$$S_m^1=2[(l_0-1)(w_0+h_0)+w_0h_0+w_b+h_b] \ , \tag{14}$$

where w_b and h_b are found using Eqs. (8) and (9).

Hence for a sequence with n_H H residues, Eqs. (11)–(13) give the dimensions of a frame according to the disjoint segment model, and Eq. (14) gives the minimal surface area of an H core that could be achieved upon folding the sequence.

B. Second approximation: Including *P* singlets

1. H equivalents

Model 1 finds minimal-surface-area H cores based on neglecting chain connectivity and sequence information. That approximation predicts the shapes of the H cores and gives an upper bound on the number of HH contacts. But real chains are more constrained, by bond connectivity, and therefore may not be able to achieve so compact an H core or so many HH contacts. In order to predict the shapes of H-cores more accurately and to give a tighter upper bound, we now consider an improvement that takes into account one aspect of chain connectivity and sequence. In model 2, we consider the costs of burying P monomers that occur because of P singlets (i.e., P in . . .HPH. . .). Examples of P singlets are residues 2, 5, 8, 18, and 21 in Fig. 1.

P singlets in a sequence will usually lead to frames that are larger than the size of the original frame of minimal H-core surface area, for the following reason. Since an H core, by definition, contains all monomers of the H type, then the two H monomers that flank a P in a P singlet must be within the H core. Moreover, since the frame of an H core is defined to be a rectangular solid, then it is geometrically impossible for the P in a P singlet to reside outside the frame since the two flanking H monomers must be contained within it [23]. Therefore, to introduce consideration of P singlets into the optimization in Eq. (5), we must now change Eq. (6) to recognize that the frame volume (lwh) must not only exceed the number of H residues, but instead must exceed the sum of the number of H residues plus the number of P singlets. If the number of P singlets is n_{P_1}, then, when computing the frame size using Eqs. (5) and (6), we should now replace n_H with

$$H_{\text{equiv}} \equiv n_H+n_{P_1} \ . \tag{15}$$

H_{equiv} is the number of monomers that are "H equivalents," i.e., H monomers plus P's in P singlets, which are all required, in model 2, to be within the H core.

As an example, when $n_H=8$, according to Eqs. (11)–(13), the frame is $\langle 2,2,2 \rangle$ and the minimal H core surface area is $S_m^1=24$. However, if we have one P singlet, then $H_{\text{equiv}}=9>n_H$, and the frame dimensions must be at least $\langle 3,2,2 \rangle$. By enumerating all possible combinations of arrangements for this example, it can be shown that the minimal surface area of any such se-

quence is at least $S_{\min}=28$. Thus the surface area of the minimal H core for a sequence with eight H's and one P singlet is the same as if there were nine H residues.

2. Polar edges

Model 2, which accounts for P singlets, leads to potentially greater H core surface area than model 1. The amount of surface area increase from model 1 to 2 depends on where the P singlets are placed. If a P singlet is buried in the core, it is energetically costly because it contributes six sides of an H cavity to the surface area [see Eq. (4)]. Lower-cost placements of P singlets are either at corners (residues 18, 21, 2, 8, Fig. 1), or aligned with other P singlets that are at corners of the H core (residue 5, Fig. 1); those configurations introduce no cavities. One configuration in which P singlets add minimal surface area is a line of edge positions occupied by P singlets and/or solvent molecules (but no H monomers; e.g., the line of monomers 2, 5, 8 in Fig. 1). In this case, the surface area is smaller than that which is calculated from Eq. (15). We call this a *polar edge*. To take P singlets and polar edges into account, we optimize a new objective function. Here, we will only give the formulation for sequences in which $n_H \le 27$ and $n_{P_1} \le \lfloor n_H/3 \rfloor \le 9$ [24]. We note that this is a fairly general class of sequences since, when the ratio of the number of H to P monomers is in the range 0.5–1, the value of $H_{\text{equiv}}=27+9=36$ typically corresponds to $70 \ge L \ge 50$.

Let $S^2(\sigma)$ be the surface area of an H core in model 2. Let n_{P_1} be the number of P singlets in sequence σ. To find the minimal surface area, we now minimize

$$S^2(\sigma)=2[(l-1)(w+h)+wh+w_b+h_b]-\Delta s \ , \quad (16)$$

with dimensions as defined in Fig. 5, and subject to the constraints

$$n_H+n_{P_1} \le lwh \ , \quad (17)$$

$$w \ge h \ , \quad (18)$$

$$w \ge w_b \ge h_b \ , \quad (19)$$

$$h \ge h_b \ , \quad (20)$$

$$2(w_b+h_b)=P_m(v_b) \ , \quad (21)$$

where

$$v_b=n_H+n_{P_1}-(l-1)wh \ , \quad (22)$$

$$\Delta s=\text{if } v_b \ge 2n_{P_1} \text{ then } 2(w_b+h_b)-P_m(v_b-n_{P_1}) \ ,$$
$$\text{else } 2e_b+2\lfloor (n_{P_1}-n'_{P_1})/h \rfloor \ , \quad (23)$$

where

$$e_b=\left\lfloor \frac{n_{P_1}}{\left\lfloor \frac{v_b}{w_b} \right\rfloor} \right\rfloor \text{ and if } w>2 \text{ then } e_b \le 2 \text{ else } e_b \le 1 \ , \quad (24)$$

$$n'_{P_1}=v_b-h_b(w_b-e_b) \text{ and } 0 \le n'_{P_1} \le n_{P_1} \ . \quad (25)$$

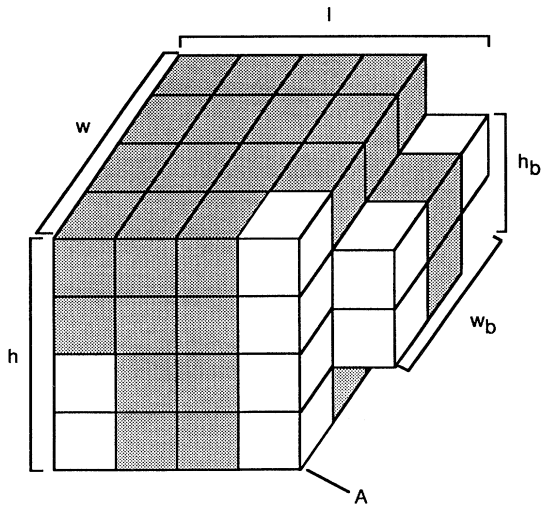

FIG. 5. An H-core body and barnacle, indicating placements of H monomers (dark) and P monomers (light). The barnacle layer has five H equivalents: two are H residues, three indicate possible P singlets. The line of P's indicated by A shows a polar edge.

These equations have the following bases (also see [24]). Equation (16) is identical to Eq. (5), but with a term Δs subtracted to account for the reduction in area due to polar edges. Constraint (17) enforces that all H monomers and P singlets must be within the frame. In the remaining equations, $\lfloor x \rfloor$ is the *floor* function [19]. v_b [Eq. (22)] is the number of H equivalents (i.e., H residues plus P singlets) in the barnacle layer. w_b and h_b are the width and height of the barnacle layer in the dimensions parallel to the width w and height h of the frame, respectively. They are determined by the minimal perimeter for v_b [Eqs. (19)–(21)]. The minimum perimeter of the barnacle layer is determined by Eqs. (8) and (9) in model 1 [Eq. (21)]. e_b is the number of polar edges in the barnacle layer; and n'_{P_1} is the number of P singlets in the barnacle layer. The term $\lfloor v_b/w_b \rfloor$ in Eq. (24) is close to and usually equals h_b. There are two kinds of cases of Δs. The "then" part in Eq. (23) is for cases in which the barnacle layer is sufficient to accommodate the P singlets. In the "else" part, the two terms account for the area reductions from polar edges in the barnacle layer and the H-core body. When $n_{P_1}=0$, $\Delta s=0$, and $2(w_b+h_b)$ in Eq. (21) is the same as $2(w_b+h_b)$ in Eq. (5); hence S^2 reduces to S^1 of Eq. (5). When the number of P singlets, $n_{P_1}>0$, Δs is usually greater than zero.

Since any configuration that can be assumed by a sequence in model 2 can be assumed by the disjoint segment model, model 2 provides a tighter lower bound for the minimal surface area of the H core, $S_{\min}$, than the disjoint segment model, i.e.,

$$S^1_m(\sigma) \le S^2_m(\sigma) \le S_{\min} \ ,$$

and is therefore a more accurate estimate of the best H core for a given HP sequence.

As an example, for the sequence in Fig. 1, we calculate $S^2(\sigma)$. In this case, we have $H_{\text{equiv}}=21$, and the number of P singlets is $n_{P_1}=5$. Therefore, there is a frame in which $l=3$, $w=3$, $h=3$, and the volume of the barnacle layer is

$$v_b=H_{\text{equiv}}-wh\,\lfloor H_{\text{equiv}}/wh\rfloor=3\ .$$

Given this value of v_b, a minimum lateral surface area can only be achieved if h_b is 1 or 2; we arbitrarily choose $h_b=1$. Then $w_b=3$. Thus the number of polar edges is $e_b=2$ and the number of P singlets in the barnacle layer is $n'_{P_1}=2$. This leaves three more P singlets, which make an edge of length 3. Thus $\Delta s=2\times1+2\times2=6$. From Eq. (16), we get $S^2(\sigma)=44$. In this case, this happens to equal $S_m^2(\sigma)$, the minimum possible surface area obtainable within model 2. This lower bound from model 2 is close to the true minimal surface area (found through the constraint-based search explained later) for this sequence, which is $S_{\min}=46$.

In general, for the same number of H equivalents, the more P singlets there are in a sequence, the more polar edges are possible, and the greater will be the reduction of surface area, Δs, provided the H core is sufficiently compact. It can be shown (proof omitted) that because of the tendency to form "polar edges" certain sequence patterns assume amphipathic-helix-like [(25)] configurational patterns. For example, the sequence of $H1$ to $H9$ shown in Fig. 1 most favorably adopts the configurational pattern shown in Fig. 1 [26].

C. Third approximation: *H* segments must reach into the core

In model 3, we incorporate an additional constraint that the sequence imposes on the possible native structures. To achieve an optimal compact core, an HP sequence must have H segments long enough to penetrate to the center of the core then return to the surface (see Fig. 6); otherwise P monomers will be buried, which increases the surface area. For example, a long ($L>54$) sequence . . . *HPPHHPPHHPPHHPPHHP*. . . cannot make a compact (cubelike [27]) core, without introducing H cavities, if there is no H segment longer than two monomers. To account for this effect, we define two quantities here. (i) The *core depth* of a site is the minimal distance (Manhattan distance, in lattice steps) of a path that starts from a position on the surface, passes through the site, and ends at another surface site (see Fig. 7). (ii) The *tether length*, ρ, of an H residue is defined in terms of the distance, d, between the residue and the nearest end of its H segment [28]:

$$\rho\equiv2d+\delta\ ,$$

$\delta=0$ if $d=(\lambda-1)/2$, where λ is the length of the H segment; $\delta=1$, otherwise. The quantity δ defines a sequence symmetry: When an H residue is in the middle of an H segment that has an odd number of H's, then neither direction has a shorter distance and $\delta=0$; for any other

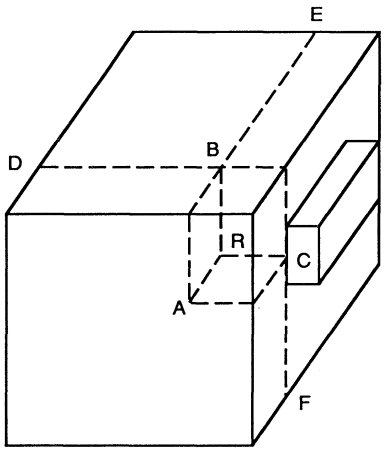

FIG. 6. Illustration of core depth. Here only the H core is shown. The shortest paths from a buried position R to the surface are to A and to B. To reach position R without burying any P monomers requires some string of H monomers in the sequence equal to or longer than $|AR|+|BR|$.

H in any H segment, one direction along the sequence to the nearest P is shorter than the other and $\delta=1$. Table II illustrates these definitions, in which the tether length of residue $H1$ is calculated.

It is straightforward to show that for an H core without a cavity, an H residue having a tether length ρ can occupy a position of core depth r only when

$$\rho\geq r\ . \tag{26}$$

The size and shape of an H core that can be achieved

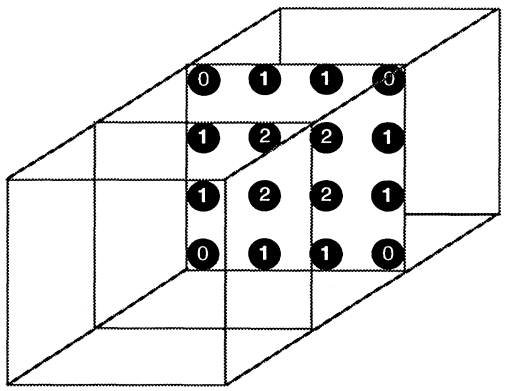

FIG. 7. The depths of the positions in a *middle* layer for a $\langle4,4,4\rangle$ H core. Note that since this H core has layers of the same size on both sides, the positions marked by 0 are edge positions. Those marked by 1 are face positions. Those by 2 interior positions. Typically, a corner or edge position has the depth of 0, since the position has at least two distinct sides to reach the surface. The depth of a face position is 1, because it has only one side to reach the surface directly and need to use the side of its neighboring residue as the other side for reaching the surface. Each of the interior positions in the figure has depth 2, since two residues are required to get to the nearest surface sides.

TABLE II. Examples of tether lengths.

Segment pattern	Minimum distance to the ends	δ	Residue tether length	Segment length
*PH*1*P*	$d=0$	$\delta=0$	$\rho=0$	$\lambda=1$
*PH*1*HP*	$d=0$	$\delta=1$	$\rho=1$	$\lambda=2$
*PHH*1*HP*	$d=1$	$\delta=0$	$\rho=2$	$\lambda=3$
*PH*1*HHP*	$d=0$	$\delta=1$	$\rho=1$	$\lambda=3$
*PHH*1*HHP*	$d=1$	$\delta=1$	$\rho=3$	$\lambda=4$

depends on the distribution of lengths of H segments. The avoidance of cavities in an H core requires that the following relation must hold for each depth r:

$$\sum_{\rho=r}^{\rho_{\max}} [N_\rho] \geq \sum_{i=r}^{r_{\max}} [n_i] , \tag{27}$$

where $\rho_{\max}$ is the maximum tether length of H segments in the given sequence, and $r_{\max}$ is the maximum core depth. N_ρ is the number of H residues in the sequence that have tether length ρ and n_i is the number of positions at depth i. The left-hand side accounts for the number of H residues that can fill positions at depth r. The right-hand side accounts for the number of core positions that are at depth greater than or equal to r. This inequality is the result of simply applying inequality (26) to all depths of the H-core positions. It holds because, for each depth r, only those H residues that have tether lengths equal to or greater than r [29] can occupy that depth *and deeper positions.*

To find the minimal H core surface area for model 3, S_m^3, we introduce inequality (27) as an added constraint to Eq. (16). In practice, we use the inequality (27) to filter the solutions obtained after optimizing Eq. (16). For example, if we have a sequence with $H=24$, $N_\rho=0$ for $\rho \geq 2$ (i.e., there are no H residues of tether length 2 or higher, implying no H segment of length 3 or greater), and if we had found from minimizing Eq. (16) that the frames of minimal surface area included $\langle 3,3,3 \rangle$ and $\langle 4,2,3 \rangle$, then constraint equation (27) would eliminate the frame $\langle 3,3,3 \rangle$ since it contains a position of depth 2 which cannot be accommodated at low-energy cost by this sequence. The example shows that H monomers with long tether lengths are often rare resources. When there are too few long H segments, the H core and the conformation will adopt a flat shape instead of a cubelike compact one. Very long H segments are rare in real proteins; this could be the basis for domains and sandwich-like shapes in the proteins that have them [30].

So far we have considered only cases of a single H core with no H cavities. The cases of multiple H cores and cavities can be treated by reducing them to the former. For example, for a particular *HP* sequence, to see if burying a P singlet can achieve a smaller H core surface area, a P singlet can be treated as if it were an H insofar as the two H segments now become one longer H segment (the length equals one plus the sum of the two H-segment lengths). An optimal core can now be sought using the same methods as above. There are two opposing tendencies. On the one hand, the P seen as H is disadvantageous insofar as it creates an H cavity, which tends to increase the surface area. On the other hand, the P seen as H is advantageous insofar as it provides a longer tether length, which may allow the chain to assemble a more compact (cubelike) H core. When will it be advantageous to bury a P? Suppose σ is the original sequence we attempt to fold, and suppose it has an actual minimal H-core surface area $S_{\min}(\sigma)$. Now suppose σ_1 is a modified sequence constructed by taking a P as if it were an H. The minimal *estimated* surface area for σ_1 is $S^3_{\min}(\sigma_1)$ and a buried P residue will have cavity area Q. Then the native state will not have a cavity area of Q, if there is no σ_1 for which σ_1

$$S^3_{\min}(\sigma_1)+Q \leq S_{\min}(\sigma) . \tag{28}$$

The sum on the left-hand side is the lower bound of the surface area of an H core with cavity area Q for the given sequence σ when a particular P residue acts as an H and is buried in the H core [31]. Inequality (28) is often good enough to determine if we need to search for native conformations with H cavities [32] or whether we need to construct a different sequence and try again. For example, in general if the sequence has many long H segments, then burying P residues only adds surface area and does not make the H core more compact.

Multiple H cores can also be treated, but on the three-dimensional cubic lattice for chain lengths $L < 5^3 = 125$ a single H core is almost always preferred, so we do not consider those cases further here.

VI. THE SEARCH PROGRAM

Based on the foregoing sequence-structure relationships, we have developed a program for fast exhaustive search [33,34] of native conformations. The program

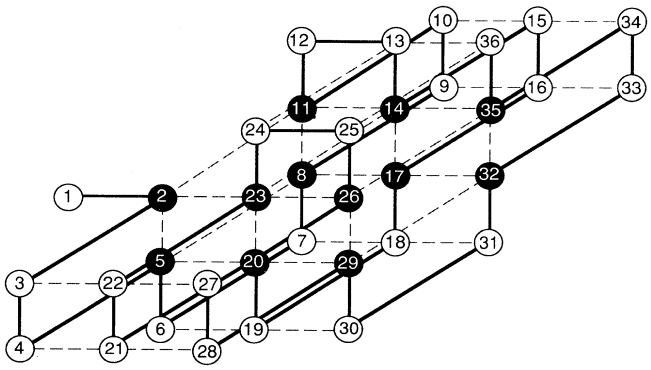

FIG. 8. A native conformation of $L5$ in Table III.

TABLE III. Example sequences.

	Sequence and Conformation	L	Degeneracy
$L1$	*HPPPPHHHHPPHPHPHHHPHPPHHPPH* *RFDDBULLFUBBRRDLLDFFRBBBRF*	27	36691
$L2$	*HPPPHHHHPHPHHPPPHPHHPHPPPHP* *RFDLULDDBRULBLDFFUBUFURBDB*	27	297
$L3$	*HPHHPPHHPPHHHHPPPHPPPHHHPPH* *RFLLFRRDRUUBDBULLURFDFLUBD*	27	25554
$L4$	*HHPHHPHHPHHHHHHPPHHHHHPPHHHHHHH* *RFLDRBDFLBUBURRDLDLLULURFFDBDF*	31	1114
$L5$	*PHPPHPPHPPHPPHPPHPPHPPHPPHPPHPHP* *RFDBDBUBUFURDBDFDFUFUBURDFDBDBUBUFU*	36	3538

first estimates the minimal surface area, the frames for H cores and the allowed cavity area [31]. The program searches the conformational space by constructing a conformation by laying down monomers one at a time. At the first stages of the search we "set up" the frame: Placements of the first few H monomers will not yet specify the frame location. But further assembly makes it clearer where the frame boundaries must be relative to the placed H monomers. When placing a residue, we use the following constraints: (1) an H residue should be placed inside the frame; (2) a P residue should be placed close enough to the frame so that the next H residue can make it into the frame; and (3) the cavity area Q should not exceed the limit determined by the inequality (28). During the search, a branch of the search tree is pruned if any of these constraints is violated. If the search fails to find a solution, the program will increment the estimate for the minimal surface area, recompute the corresponding frames, and start the next round of search. Because this estimate of the minimal surface area is a lower bound, this approach will never miss a solution.

The requirement that H residues must lie within a frame substantially prunes the search tree. For example, the number of conformations of maximally compact homopolymer chains grows only as 1.9^L, compared to 4.7^L for all conformations [35].

The search speed depends on the monomer sequence. For typical sequences with H-residue:P-residue ratio $\geq 50\%$, the speedup is about 10^5. For some particular sequences, the speedup can reach 10^8 or higher. Some chains of $L=36$ can be folded in 20 min running on a Sun 4 workstation. The computer program has been validated using results for short chains, $L \leq 15$.

Five examples of folded HP sequences are shown in Table III. Here, for each sequence a native conformation is shown. A conformation is represented by a sequence of bond directions, R, L, D, U, F, B are right, left, down, up, forward, and backward, respectively. One native conformation for the 36-residue chain $L5$ is shown in Fig. 8. Table III shows the total number of native conformations (degeneracy) for a sequence. The estimated minimal surface areas (S^3), actual minimal surface areas ($S_{\min}$), and the search times on a Sun 4 workstation are shown in Table IV. The search time for brute-force search is extrapolated from the results for shorter chains, also running on a Sun 4 workstation.

VII. CONCLUSIONS

We study the relationship between the monomer sequence and native structures of HP copolymers in the three-dimensional lattice model. We derive a folding equation that serves to reformulate the folding problem to that of finding conformations that have H cores of minimal surface area. The minimal H core surface area that is achievable by an HP sequence is estimated at three different levels of treatment of chain connectivity and sequence. Model 1 neglects connectivity between H segments; model 2 is a refinement that accounts for the placements of P singlets (P in . . . HPH. . .); and model 3 accounts for tether lengths of H segments (the ability of a sequence of consecutive H monomers to reach into, and return from, the center of the core). These models are used to compute the dimensions of a rectangular solid, the frame, within which all the H monomers must reside in order for a conformation to be native. The frame serves as a strong constraint for pruning the decision tree

TABLE IV. Search time for example sequences.

Sequence	n_H	n_{P_1}	S^3	$S_{\min}$	Search time (brute force)	Search time (using constraints)
$L1$	14	3	38	40	9×10^8 h	1 h, 38 min.
$L2$	13	4	36	38	9×10^8 h	1 h, 14 min.
$L3$	14	1	38	38	9×10^8 h	5 h, 19 min.
$L4$	24	3	50	52	4.5×10^{11} h	5 h, 18 min.
$L5$	12	0	16	16	10^{15} h	20 min.

that seeks native structures. These constraints underly a computer algorithm that finds native states for three-dimensional chains up to $L=36$ monomers long in its present implementation.

ACKNOWLEDGMENTS

We thank the Office of Naval Research for financial support and Dr. Sarina Bromberg for helpful comments. Kaizhi Yue thanks Dr. Yuzhong Wang for helpful discussions.

APPENDIX

We show here that under the disjoint H-segment model, the H core of every native conformation can be reconfigured into a body and a single barnacle layer, without changing the size of the frame. This serves as the basis for the optimization equations beginning with Eq. (5). The main idea behind the proofs below is that, in the disjoint H-segment model, H segments in two layers, A and B, can always be reconfigured instead into two different layers, C and D, one of which fills the frame efficiently, and the other of which contains the remaining segments. Before proceeding with this proof, we first explain how we compute the surface areas of H cores. We do it on a layer-by-layer basis. That is, if layers are stacked along the z axis, the total projection in the x and y dimensions is the sum of "lateral surface areas" of each layer "slice" (see Fig. 1). Each slice has a "cross section" (on the yz plane in Fig. 3), the perimeter of which defines its "lateral surface area" (see Fig. 2).

As an example of reconfiguration, for the seven layers shown in Fig. 3, the two rightmost layers can be combined into a single layer. The resulting new layer will make no contributions to surface area in the direction of attachment because its projection is blocked by larger layers. However, since the laterial surface of the new layer is smaller than the sum of those of the original two layers, the combination results in the reduction of surface area. The following three lemmas explore relationships that hold for such transformations. The first lemma states that if the residues of two layers P and Q are reconfigured to form a single layer C, then the minimum cross section perimeter of C will be smaller than the sum of the perimeters of P and Q.

Lemma 1 (combining two layers with no restriction on total volume). Let N_P and N_Q be the numbers of H residues (i.e., volumes) in layers P and Q, respectively. The following relation must hold.

$$P_m(N_P)+P_m(N_Q)>P_m(N_P+N_Q)\ , \tag{A1}$$

where $P_m(\ )$ is the function of the minimal perimeter of a layer defined by Eq. (10).

Proof: $P_m(N)$ is the minimal perimeter of any layer containing N residues. For any integers a and b,

$$\text{if } ab\geq N \text{ then } 2(a+b)\geq P_m(N)\ . \tag{A1'}$$

Any integer N can be expressed as $N=p^2+q$ where p,q are integers and $q<2p+1$. Let

$$N_P=m^2+k,\quad N_Q=n^2+l\ ,$$

where k, l, m, n are integers and

$$0\leq k<2m+1,\quad 0\leq l<2n+1\ .$$

Assume $m\geq n$. From the definition of P_m, it can be verified for cases of $k=0$, $0<k\leq m$, and $m<k\leq 2m$ that

$$P_m(N_P)=2\left[2m+\left\lceil\frac{k}{m}\right\rceil\right]\ .$$

Similarly,

$$P_m(N_Q)=2\left[2n+\left\lceil\frac{l}{n}\right\rceil\right]\ .$$

So we have

$$\begin{aligned}P_m(N_P)+P_m(N_Q)&=2\left[2m+\left\lceil\frac{k}{m}\right\rceil+2n+\left\lceil\frac{l}{n}\right\rceil\right]\\&>2\left[2m+\left\lceil\frac{k}{m}\right\rceil+n+\left\lceil\frac{l}{n}\right\rceil\right]\ .\end{aligned}$$

Now, we separate the right-hand side as in Eq. (A1'). Since $m\lceil k/m\rceil\geq k$, we have

$$(m+\lceil\lceil k/m\rceil)m\geq m^2+k\ .$$

Similarly (since $m\geq n$)

$$\left[m+\left\lceil\frac{k}{m}\right\rceil\right]\left[n+\left\lceil\frac{l}{n}\right\rceil\right]\geq n^2+l\ .$$

Therefore,

$$\left[m+\left\lceil\frac{k}{m}\right\rceil\right]\left[m+n+\left\lceil\frac{l}{n}\right\rceil\right]\geq m^2+k+n^2+l$$

Taking $m+\lfloor k/m\rfloor$ and $m+n+\lfloor l/n\rfloor$, respectively, as a and b in Eq. (A1'), we have

$$\begin{aligned}2\left[2m+\left\lceil\frac{k}{m}\right\rceil+n+\left\lceil\frac{l}{n}\right\rceil\right]&\geq P_m(m^2+k+n^2+l)\\&=P_m(N_P+N_Q)\ .\end{aligned}$$ ■

Thus if there is no constraint on how layers P and Q can reconfigure into a new single layer C, then C will have less lateral surface area. The following is the geometric interpretation of this lemma. Suppose the layers P and Q adopt the shapes that have minimal perimeters. Then by attaching layer A to layer B, some of the sides on the perimeters become shared. Therefore the new layer will always have a smaller perimeter than the separate layers P and Q.

The result above applies when two layers can combine without constraint. Now consider a different situation involving a constraint. Suppose layers P and Q are reconfigured to form two new layers C and D, subject to the constraint that neither C nor D can individually exceed a given number of H monomers. This situation arises when layers are constrained to fit within a given frame. Even subject to this constraint, the

reconfiguration may still lead to a reduced perimeter, under the following conditions. Let the volumes of layers C and D be N_C and N_D.

Lemma 2 (combining two layers with restriction on the resultant layer volumes). Assuming $N_C \geq N_P \geq N_Q$ (condition 1) and $N_C + N_D = H_P + N_Q$ (condition 2). Let

$$N_P = m^2 + k, \quad N_Q = n^2 + l \ ,$$

where k, l, m, n are integers, and

$$0 \leq k < 2m + 1, \quad 0 \leq l < 2n + 1 \ .$$

If N_C can be expressed in the form of mm' where m and m' are integers and $m' - m$ is 0 or 1 (condition 3), then

$$P_m(H_P) + P_m(N_Q) \geq P_m(N_C) + P_m(N_D) \ . \qquad \text{(A2)}$$

Condition 1 states that the volume of one of the resulting layers will be larger than either original layer. Condition 2 conserves volume. Layer C can be constructed by moving residues from layer Q to layer P. Condition 3 allows only transformations that cause Layer C to be a nearly square rectangle. The proof (omitted) follows the same reasoning as for inequality (A1).

Next, consider reconfigurations of more than two layers.

Lemma 3 (combining multiple layers subject to a maximum volume per layer). If the maximum volume of a layer is A and the total volume (number of residues) V satisfies

$$NA \leq V \ ,$$

where N is an integer, then if there are $N' > N$ layers in the frame, the surface area of the H core is not minimal for the given maximal volume A.

This lemma states that unless the number of layers has reached the minimum, we can always reduce the surface area by combining layers. Again, the proof (omitted) follows the same reasoning as for lemma 1. Here we give an intuitive argument for the lemma. According to lemmas 1 and 2, to minimize the surface area of the layers, two smaller layers can combine into a larger layer, without a residual layer (if there are no constraints on volume per layer), or with a residual layer (if layer sizes are constrained). Under these conditions the lateral surface areas (perimeters) will not increase. As for the *total* surface area, which also includes the area in the dimension perpendicular to the layer, it does not increase either. This is because the volume of any resulting layer is less than A and the largest layer (with area A) can block the projections of other layers and determine the projection area in the direction perpendicular to the layers. We can repeatedly reconfigure layers so that the layers have the largest possible volumes. Each step of reconfiguration may result in either 1 or 2 layers. If it results in two different layers, then the surface area at least will not increase (lemma 2). However, since $N' > N$, at least for one step of the reconfiguration, two layers will combine into one layer. According to lemma 1, at this step, the surface area will be reduced. This shows the original N' layer arrangement is not optimal.

Finally, assuming a model of disjoint H segments, we show that the H core of a native conformation can be reconfigured into a body with a single barnacle layer. On a cubic lattice, a layer achieves its minimal lateral surface area by configuring to have a rectangular cross section. By lemma 3, if an H core has multiple layers, then we should combine multiple small layers into a smaller number of layers, each one as large as possible, until the remaining H residues do not completely fill the largest layer. Taken together, these arguments imply that the optimal configuration is a stack of rectangular layers of the same volume and a remainder layer that contains the rest of the H monomers. The volume of the remainder layer will be

$$v_b = V - A \lfloor V/A \rfloor \ ,$$

where V is the total number of residues and A is the area of the large rectangular layer.

[1] E. Shakhnovich, G. Farztdinov, A. M. Gutin, and M. Karplus, Phys. Rev. Lett. **67**, 1665 (1991).

[2] P. Leopold, M. Montal, and J. Onuchic, Proc. Natl. Acad. Sci. **89**, 8721 (1992).

[3] K. F. Lau and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 638 (1990).

[4] H. Taketomi, F. Kano, N. Go, Biopolymers **27**, 527 (1988).

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

[6] H. S. Chan and K. A. Dill, J. Chem. Phys. **95**, 3775 (1991).

[7] D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (1992).

[8] K. A. Dill, Biochemistry **29**, 7133 (1990).

[9] E. A. O'toole and A. Z. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).

[10] H. S. Chan and K. A. Dill, Macromolecules **22**, 4559 (1989).

[11] H. S. Chan and K. A. Dill, Proc. Natl. Acad. Sci. USA **87**, 6388 (1990).

[12] D. J. Lipman and W. J. Wilbur, Proc. R. Soc. Lon. Ser. B **245**, 7 (1991).

[13] R. Miller, C. A. Danko, M. J. Fasolka, A. C. Balazs, H. S. Chan, and K. A. Dill, J. Chem. Phys. **96**, 768 (1992).

[14] J. D. Honeycutt and D. Thirumalai, Proc. Natl. Acad. Sci. USA **87**, 3526 (1990).

[15] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).

[16] P. M. Gruber and C. G. Lekkerkerker, *Geometry of Numbers* (Elsevier, New York, 1987).

[17] *Discrete Geometry and Convexity,* edited by J. E. Goodman (New York Academy of Sciences, New York, 1985).

[18] H. R. Jacobs, *Geometry* (Freeman, San Francisco, 1974).

[19] Here, $\lceil x \rceil$ is the *ceiling* function, it is defined as the smallest integer that is greater than or equal to x. Similarly, $\lfloor x \rfloor$ is the *floor* function, it is the largest integer that is

smaller than or equal to x.

[20] G. Mitra, *Theory and Applications of Mathematical Programming* (Academic, London, 1976).

[21] In the first case, $v_b=21-3\times3\times2=3$, $w_b+h_b=4$, and $(l-1)(w+h)+wh=2\times6+9=21$. In the second case, for $l=4$, we also have $v_b=21-3\times3\times2=3$, $w_b+h_b=4$, and $(l-1)(w+h)+wh=3\times5+6=21$. In both cases, $S^1=2\times(21+4)$, which we know equals S_m^1 in this case [see below Eqs. (11)–(13)].

[22] The physical basis for this is simple. Just as a hydrocarbon droplet will adopt a spherical shape in water to minimize its interfacial tension, the lattice solid that has the smallest surface area is a rectangular solid with a small appended remainder piece.

[23] While the exact form of this constraint is dependent on the lattice geometry, the constraint itself is physical. For an off-lattice chain to place a P singlet outside a hydrophobic core requires energetically unfavorable tight bending [see A. V. Finkelstein and O. B. Ptitsyn, Prog. Biophys. Mol. Bio. **50**, 171 (1987), and references therein].

[24] The proof of Eq. (16) follows the same line of reasoning as the proof for Eq. (5), i.e., by reconfiguring layers. Here we only give a brief explanation of the condition of $n_H \leq 27$ and $n_{P_1} \leq \lfloor n_H/3 \rfloor$. Equation (16) calculates a compact H core in terms of H equivalents, not in terms of H residues. Thus, it requires all the layers in the frame except one be completely filled with H residues or P singlets. It implicitly assumes that there is only one barnacle layer *in terms of H equivalents* in the maximally compact H core in model 2. However, this is true only for simple cases. In general many more combinatorial factors, e.g., several barnacle layers and their respective widths and heights, need to be considered than in Eqs. (16)–(25). Only when $n_H \leq 27$ and $n_{P_1} \leq \lfloor n_H/3 \rfloor$, we have a simple case: The minimal surface area can be calculated from Eqs. (16)–(25). As an example of other cases, if the number of P singlets is too large, the most compact H core could not accommodate them all along polar edges (since the more compact, i.e., cubelike, an H core is, the fewer edges it has), and more than one barnacle layer is required.

[25] D. Eisenberg, R. M. Weiss, and T. C. Terwillinger, Proc. Natl. Acad. Sci. USA **81**, 140 (1984).

[26] Because of the limitations of the geometry of a cubic lattice, a pattern in the shape of helices in real proteins is not attainable. The pattern in Fig. 1 is one of the closest to real helices.

[27] A cubelike shape is the most compact shape on the cubic lattice; this corresponds to a spherical shape in continuous space.

[28] The distance along the chain is the difference between the residue numbers. There are two minor details about the tether length that we have omitted in the main text. First, when an H segment is at the terminal of a sequence, only the end neighboring a P residue of the H segment, i.e., the nonterminal end, is considered for computing the distance. This is because the terminal H, the other end of the H segment, is not connected to a P residue and therefore is free to be anywhere inside the H core. Second, when P singlets are considered, we usually consider an H core in terms of H equivalents (instead of only H residues), this complicates the matter a little bit. When no cavity is allowed, P singlets will always be positioned at the surface. Therefore, the neighboring H segment will have one "additional H residue." Correspondingly, the tether lengths of the H residues on the segment may increase by 1.

[29] Equation (26) implies the following equation: $\sum_{\rho=r}^{\rho_{\max}} N_\rho \geq n_r$. That is, for a depth r, only those H residues that have tether length equal to or greater than r can fill positions at depth r. This relationship, however, can be made stronger because if there are positions of depths $>r$ then some of the H residues of tether lengths $>r$ have to be used for filling those positions. For example, suppose we have a sequence of *PPHHHHPPHHHHPPHPPHPPH-PPHPPHPPHPPHPPHPP*. It has two H segments of length 4 and eight H segments of length 1. Thus, it has eight residues of tether length 0, four residues of tether length 3, and four residues of tether length 1. These H residues can fill the positions in Fig. 6 according to the above equation, because there are four $r=2$ positions, eight $r=1$ positions, and four $r=0$ positions. But this will not actually work, because the $\rho=2$ residues are *counted twice* in the equation.

[30] J. Richardson and D. C. Richardson, in *Prediction of Protein Structure and the Principles and Protein Conformation,* edited by Gerald Fasman (Plenum, New York, 1989).

[31] If for some σ_1 the inequality does not hold, then it is possible that an H core of a native conformation has a cavity area of Q, which we call the *allowed cavity area.*

[32] How is Eq. (28) useful for long chains, for which we cannot know the actual minimal surface area? The main point is that we do not need to know the exact actual minimal surface area to use the relationship. Whenever we can find a conformation with a surface area, say S_1, which is less than that of the left-hand side, there cannot be a cavity. S_1 can be the result of any search, in particular a nonexhaustive one.

[33] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms* (Addison-Wesley, Reading, MH, 1974).

[34] N. Nilsson, *Principle of Artifical Intelligence* (Tioga, Palo Alto, CA, 1980).

[35] W. J. C. Orr, Trans. Faraday Soc. **43**, (1947).