JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 12, Number 10, 2005

Pp. 1328–1345

Structure-Approximating Inverse Protein Folding Problem in the 2D HP Model

ARVIND GUPTA,[1] JÁN MAŇUCH,[1] and LADISLAV STACHO[2]

ABSTRACT

The inverse protein folding problem is that of designing an amino acid sequence which has a particular native protein fold. This problem arises in drug design where a particular structure is necessary to ensure proper protein–protein interactions. In this paper, we show that in the 2D HP model of Dill it is possible to solve this problem for a broad class of structures. These structures can be used to closely approximate any given structure. One of the most important properties of a good protein (in drug design) is its stability—the aptitude not to fold simultaneously into other structures. We show that for a number of basic structures, our sequences have a unique fold.

Key words: inverse protein folding, HP model, protein stability, protein design.

1. INTRODUCTION

IT HAS LONG BEEN KNOWN THAT PROTEIN INTERACTIONS depend on their native three-dimensional fold, and understanding the processes and determining these folds is a long standing problem in molecular biology. Naturally occurring proteins fold so as to minimize total free energy. However, it is not known how a protein can choose the minimum energy fold amongst all possible folds (Dill *et al.*, 1995).

Many forces act on the protein which contribute to changes in free energy, including hydrogen bonding, van der Waals interactions, intrinsic propensities, ion pairing, and hydrophobic interaction. Of these, the most significant is hydrophobic interaction (see Dill [1990] for details). This led Dill to introduce the *Hydrophobic-Polar Model* (Dill, 1985). Here, the 20 amino acids from which proteins are formed are replaced by two monomers, hydrophobic (H) or polar (P), depending on their affinity to water. To simply the problem, the protein is laid out on a 2D spatial lattice with each monomer occupying exactly one square and neighboring monomers occupy neighboring squares. The free energy is minimized when the maximum number of nonneighbor hydrophobic monomers are adjacent in the lattice. Therefore, the "native" conformations are those with the maximum number of such HH contacts, also called *bonds*.

Even though the hydrophobic-polar model is the simplest model of the protein folding process, computationally it is an NP-hard problem (cf. Crescenzi *et al.* [1998] for two- and Berger and Leighton [1998] for three-dimensional square lattices). Interestingly, in the first case, the result is deduced from the NP-completeness of the Hamilton cycle problem for special planar graphs, while in the second case, the result follows from the NP-completeness of the modified bin-packing problem. The problem is still open for other types of lattices, namely, triangular and diamond lattices. Research has focused on approximations

[1]School of Computing Science, Simon Fraser University, Burnaby BC V5A 1S6, Canada.
[2]Department of Mathematics, Simon Fraser University, Burnaby BC V5A 1S6, Canada.

for this model. A linear time algorithm with approximation factor 3/8 for 3D square lattice can be found in Hart and Istrail (1995), and linear time algorithms with approximation factors 6/11 and 3/5 for 2D and 3D triangular lattices, respectively, have been developed by Agarwala *et al.* (1997).

In many applications, such as drug design, we are actually interested in the complement problem to protein folding: *protein design*. Current protein designs often focus on local interactions, such as intrinsic propensities of amino acids to form helices and turns (Lyu *et al.*, 1990; Andrew *et al.*, 2001). However, major forces of folding are due to hydrophobic and other *nonlocal* interactions (Dill, 1990). To compensate for this unbalance, the existing designs work on a selected small group of very stable protein motifs altering only some parts of the sequence appearing at the surface of the fold. For instance, due to its simplicity and regularity, the most extensively studied protein motif is the "coiled coil": alpha-helices wrapping around each other (Yu, 2002).

A major challenge in designing proteins that attain a specific native fold is to avoid proteins that have multiple native folds. We say that a protein is *stable* if the minimum free energy fold is unique. It is generally believed that all naturally occurring proteins are stable; however, this is usually not true for arbitrary protein sequences. Extreme examples are proteins containing only polar monomers in the HP model. In this case, every fold achieves lowest free energy. We note that the proteins used to prove NP-hardness of the protein folding problem are not stable.

The more general *inverse protein folding problem* involves starting with an arbitrary target fold and designing an amino acid sequence whose native fold is the target (positive design) and which is stable (negative design). As this problem is more complex, the current research concentrates only on a simple HP model. Even in the HP model, the complexity of this problem is unknown but conjectured to be NP-hard. Early work on this problem involved heuristics that bury the H monomers in a central core with the P monomers on the outside (Kamtekar *et al.*, 1993), find all possible short sequences and put these together (Yue and Dill, 1992), or perform a sequence evolution, a form of local search (Sun *et al.*, 1995). A relationship between symmetries and designability of proteins was observed by Wang *et al.* (2000).

Another approach to this difficult problem is a heuristic sequence design, i.e., design of a sequence fulfilling easier alternative criteria which is likely to solve the original inverse protein folding problem. There are currently two sets of criteria studied, called canonical and grand canonical models, introduced by Shakhnovich and Gutin (1993) and Sun *et al.* (1995), respectively. It has been shown that the protein sequence design problem can be solved in polynomial time in the grand canonical model for both 2D and 3D square lattices (cf. Hart [1997]), and in polynomial time for 2D lattices while the problem is NP-hard for 3D square lattice in the canonical model (cf. Berman *et al.* [2004]). Note, however, that design of heuristic sequences does not guarantee that the generated sequence satisfies the two criteria (positive and negative design) of the inverse protein folding problem.

In this paper, we consider a completely new version of the inverse protein folding problem: instead of a target fold we are given a target structure. This structure is given on a lattice by specifying (marking) a connected collection of lattice squares which the amino acids must occupy. We show that it is possible to design a protein whose native fold closely approximates any given 2D structure. We will work on a refinement of this lattice in which each square is divided into nine squares (i.e., a 3×3 refinement of the original lattice). Now, for a marked square, almost all (eight) of its nine subsquares must be occupied by the monomers of the protein, and for an unmarked square at most $2n$ of its subsquares are occupied where n is the number of marked neighboring squares. We call our structures *constructible structures*.

For a number of basic structures, we give a *formal* proof that our proteins are stable in the 2D HP model. Note that we are not aware of any other results explicitly showing stability of an infinite class of proteins (in Sun *et al.* [1995], a heuristic method generating stable proteins was proposed; however, this is only supported by computer testing). Based on our results and on an extensive computer search, we conjecture that our proteins for all constructible structures are stable.

An extended abstract of this paper appeared in Gupta *et al.* (2004).

2. PRELIMINARIES

2.1. Hydrophobic-polar model

In this section, we will formally define the hydrophobic-polar model. We will restrict our attention to the two-dimensional square lattice.

Proteins are chains of monomers where each monomer is either hydrophobic, i.e., nonpolar, or hydrophilic, i.e., polar. We can represent a protein chain as a binary string $p = p_1 p_2 \dots p_{|p|}$ in $\{0,1\}^*$, where "0" represents a polar monomer and "1" a nonpolar monomer. In our figures, "0" will be depicted as "□" and "1" as "■."

Let us consider a tiling of $\mathbb{R}^2$ with unit squares. Obviously, such a tiling can be represented by a two-dimensional square lattice L where the vertices (squares of the tiling) are represented as ordered pairs and two vertices are adjacent if and only if the corresponding squares share a side. More formally, L is a graph with vertex set $V = \{[a,b];\ a,b \in \mathbb{Z}\}$ and edge set $E = \{\{[a,b],[a+c,b+d]\}; a,b \in \mathbb{Z}$ and $(c,d) = (0,1),(1,0)\}$. The squares adjacent to $[a,b] \in V$ are called *neighbors* of $[a,b]$. In particular, $[a,b+1]$ is the *northern*, $[a+1,b]$ is the *eastern*, $[a,b-1]$ is the *southern*, and $[a-1,b]$ is the *western* neighbor of $[a,b]$.

Next we define a conformation of a protein as a self-avoiding walk in the lattice and a fold as a placement of monomers into the lattice. More formally:

Definition 1 (Conformations and folds). *For every $n \geq 2$, a path $c = (c_1, c_2, \dots, c_n)$ in L is called a* conformation of length *n. An edge $e = \{s_1, s_2\}$ of c, i.e., $e \in E(c)$, is called a* c-edge*, and we say that the squares s_1 and s_2 are* c-connected*, or that they are* c-neighbors*. A* fold $F_{p,c}$ of a protein $p \in \{0,1\}^n$ with respect to a conformation c of length n *is a partial mapping $F_{p,c} : V \to \{0,1\}$ such that for every $k = 1, \dots, n$, $F_{p,c}(c_k) = p_k$. If no confusion can arise, we will retain the phrase "$u \in V$* is an *a*-square*" for the fact that $F_{p,c}(u) = a$. The squares c_1 and c_n are called* terminals*; in pictures these are marked with a cross. Denote the set of all* 1*-squares as $1_{p,c}$ and the graph induced by these vertices by $L[1_{p,c}]$.*

A protein will fold into a conformation with minimum free energy. In the HP model, only hydrophobic interactions between adjacent hydrophobic monomers (which are not consecutive in the protein) contribute to the score. Hence, a conformation with the lowest free energy corresponds to a conformation with the highest score, that is, the conformation with the largest number of HH bonds.

Definition 2 (Bonds and score). *For every fold $F_{p,c}$, a* bond of $F_{p,c}$ *is an edge $\{u,v\}$ of L such that u and v are* 1*-squares and they are not consecutive in c; i.e., a bond is an edge in $L[1_{p,c}] - E(c)$.* The score *of a fold $F_{p,c}$, denoted by* score$(F_{p,c})$*, is the number of bonds in $F_{p,c}$.*

The conformations with the highest score (corresponding to the lowest free energy) are called native conformations, which is stated more formally as follows.

Definition 3 (Native conformations). *A conformation c of length $|p|$ is* native for protein *p if for any other conformation c' of length $|p|$,* score$(F_{p,c}) \geq$ score$(F_{p,c'})$*. The fold of p with respect to a native conformation is called a* native fold*.*

Note that there might be several native conformations for p. The set of all native conformations is denoted by $C(p)$. From a biological point of view, the proteins having a single native conformation are more likely to stay in the same state without changing their structure.

Definition 4 (Stable proteins). *A protein p is* stable *if it has exactly one native conformation, i.e., if $|C(p)| = 1$.*

The proteins we are going to describe have a special property. The score of their native conformations is the maximal possible score with respect to the number of hydrophobic "1" monomers contained in the protein. The following useful observation characterizes native conformations of such proteins.

Observation 1 (Saturated folds). *Let $p \in 0\{0,1\}^*0$ be a protein, and F be the fold of p with respect to a conformation c. If for every* 1*-square s, two out of four edges incident with s are bonds, then c is a native conformation for p. We will call the fold F a* saturated *fold.*

Furthermore, if there exists a conformation c such that the fold of p with respect to c is saturated, then for any native conformation c' of p, its fold is also saturated.

Note that the fold F of p with respect to c is saturated if and only if the graph $L[1_{p,c}] - E(c)$ is a 2-factor of L, i.e., every connected component is a cycle, called a 1-cycle. *All edges of such a* 1*-cycle are bonds.*

The proof of the observation follows by a simple argument that any 1-square s has at most two bonds. Note that not every protein has a saturated fold. For instance the necessary condition for protein p to have a saturated fold is that p contains an even number of hydrophobic "1" monomers.

2.2. Constructible structures and their proteins

In this section, we define a wide class of structures which can be used to approximate any given shape, called *constructible structures*. Next, to each constructible structure we assign a protein which has a native conformation exactly filling the constructible structure. We conjecture that such proteins are stable; cf. the next section.

Definition 5 (Constructible structures). *We have two tiles, depicted in Fig. 1(a), a starting tile in the shape of "+" and a regular tile in the shape of "⊢." Both tiles have three* ligands*, depicted with black lines, and in addition, the regular tile has one* receptor*, depicted with a gray line. A* constructible structure *is a partial tiling of the two-dimensional grid L obtained by the following procedure:*

1. *Place the starting tile into the grid.*
2. *Place a regular tile into the grid so that its receptor is attached to a ligand of a tile already in the grid and it does not overlap with any other tile.*
3. *Continue with step 2 or end the procedure.*

An example of a constructible structure is shown in Fig. 1(c). Let $V(S) \subset V$ be the set of squares covered by tiles of S. A conformation c is compatible with *S if $V(c) = V(S)$. Similarly, a fold F is* compatible *with S if its current domain (the set of squares containing monomers) is equal to $V(S)$.*

Note that at each step of the above construction we remove one ligand from the set of available ligands (the one into which the new tile is attached) and add three ligands (the ligands of the new tile). Therefore, if we give distinct numbers to ligands, it is possible to describe the process of tiling by a sequence of ligand labelings called the *tiling sequence*

$$T = \langle t_1, \ldots, t_k \rangle$$

such that for every $\ell = 1, \ldots, k$, $t_\ell \leq 3\ell$ and $t_1, \ldots, t_k$ is an increasing sequence of positive integers. Ligands of the starting tile are numbered counterclockwise 1, 2, 3 as depicted in Fig. 1(b). The ℓ-th regular tile is attached by its receptor to ligand t_ℓ, and its three new ligands are numbered counterclockwise $3\ell+1$, $3\ell+2$, $3\ell+3$. To make this coding unique, all numbers are placed on tiles in such a way that the receptor of a tile is between ligands $3\ell+1$ and $3\ell+3$. The numbering of tiles is depicted in Fig. 1(b). The tiling sequence of the constructible structure depicted in Fig. 1(c) is $\langle 1, 2, 3, 5, 8, 12, 17, 22, 23, 26, 30, 35, 37, 41, 42 \rangle$.

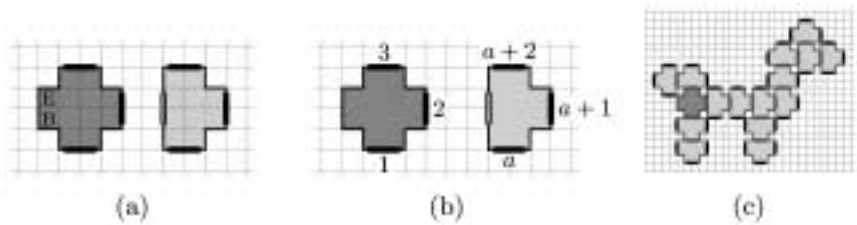

FIG. 1. Illustration of: (**a**) the starting tile (left) and the regular tile (right); (**b**) numbering of the ligands of tiles; (**c**) a constructible structure described by a tiling sequence $\langle 1, 2, 3, 5, 8, 12, 17, 22, 23, 26, 30, 35, 37, 41, 42 \rangle$.

Note that not every tiling sequence describes a constructible structure, since its definition does not guarantee that the placed tiles do not overlap. However, if the tiling sequence T describes a constructible structure, then this structure is denoted by S_T. The following lemma provides a construction of a conformation compatible with a constructible structure S.

Lemma 1. *For every constructible structure S, there exists a conformation compatible with S.*

Proof. We will prove a stronger claim by induction on the number k of regular tiles used: For every constructible structure S which uses k regular tiles, there exists a conformation $c(S)$ of length $12 + 10k$ such that

- $c(S)$ is compatible with S;
- the terminals $c(S)_1$ and $c(S)_{12+10k}$ are the squares B and E of the starting tile, respectively; cf. Fig. 1(a); and
- for every ligand $t_\ell \in \{1, \ldots, 3k+3\} - \{t_1, \ldots, t_k\}$ of S, the two squares of S adjacent to t_ℓ form a $c(S)$-edge (the other two squares adjacent to t_ℓ lie outside of S).

Consider a constructible structure S described by a tiling sequence $T = \langle t_1, \ldots, t_k \rangle$. If $k = 0$, then S contains only the starting tile, and the conformation

$$c(S) = ([0,0], [1,0], [1,-1], [2,-1], [2,0], [3,0], [3,1], [2,1], [2,2], [1,2], [1,1], [0,1])$$

has the desired property; cf. Fig. 2(a).

Otherwise, let S' be a constructible structure described by $\langle t_1, \ldots, t_{k-1} \rangle$. By the induction hypothesis, there exists a conformation $c' = c(S')$ of length $2 + 10k$ compatible with S'. Moreover, c'_1 and c'_{2+10k} are the squares covered by B and by E of the starting tile, respectively. Since, $t_k \in \{1, \ldots, 3k\} - \{t_1, \ldots, t_{k-1}\}$, the squares u and v of S' adjacent to the ligand t_k form a c'-edge. Without loss of generality, we can assume that

$$u = [a, b] = c'_i \quad \text{and} \quad v = [a, b+1] = c'_{i+1}.$$

Now, we can define a conformation c compatible with S:

$$c(S) = (c'_1, \ldots, c'_i, [a+1, b], [a+1, b-1], [a+2, b-1], [a+2, b], [a+3, b],$$
$$[a+3, b+1], [a+2, b+1], [a+2, b+2], [a+1, b+2], [a+1, b+1], c'_{i+1}, \ldots, c'_{2+10k}).$$

As can easily be seen in Fig. 2(b), the vertices added to $c(S)$ will exactly fill the squares covered by the last tile of S, and the squares adjacent to new ligands $3k+1$, $3k+2$, and $3k+3$ are consecutive in c, respectively. ■

Observe that it can easily be proved that the conformation constructed in Lemma 1 is the only conformation compatible with S starting at the square B and ending at the square E of the starting tile. In what follows, we will denote this conformation by $c(S)$.

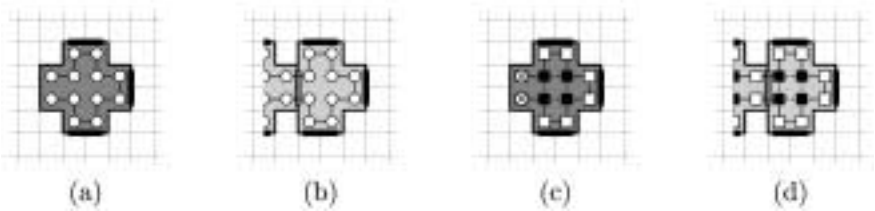

FIG. 2. How to place the conformation $c(S)$ in the starting tile (**a**) and a regular tile (**b**) and how to fill in the squares of the conformation $c(S)$ with the symbols of the protein $p(S)$ in the starting tile (**c**) and a regular tile (**d**).

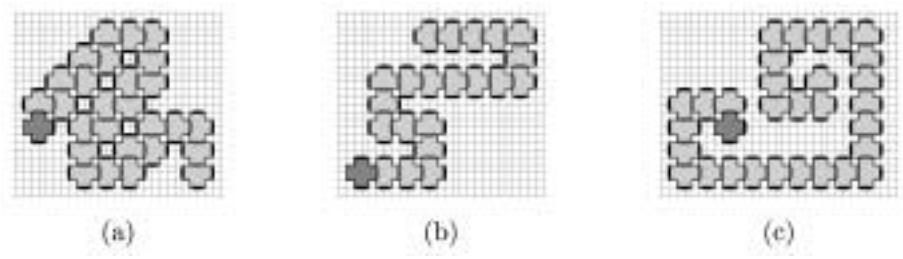

FIG. 3. Illustration of (**a**) a linear constructible structure; (**b**) a slowly bending constructible structure; (**c**) a spiral constructible structure.

Now, for each constructible structure S, we define a protein $p(S)$ such that the conformation $c(S)$ is a native conformations of $p(S)$. Our goal is to show that $p(S)$ is stable. However, this seems extremely difficult; here we show the result for particular special cases.

The constructible structures are specially designed to have the following two properties: (a) they can approximate any given shape; and (b) it is easy to construct the proteins with native folds compatible with the structures (see Fig. 3). The second property can be formalized as follows:

Theorem 1. *For every constructible structure S, there exists a protein $p(S)$ whose fold with respect to $c(S)$ is saturated. Hence, $c(S)$ is a native conformation of $p(S)$ and any native fold of $p(S)$ is saturated.*

Proof. Let $T = \langle t_1, \ldots, t_k \rangle$ be the tiling sequence of S. We will define $p(S)$ inductively. If $k = 0$, then set

$$p(S) = 010010010010\,.$$

The fold of $p(S)$ with respect to $c(S)$, depicted in Fig. 2(c), is saturated. Note that the squares of S adjacent to all three ligands are 0-squares.

If $k > 0$, let S' be the constructible structure described by $\langle t_1, \ldots, t_{k-1} \rangle$. Consider a protein $p(S') = p_1 \ldots p_{i-1}00p_{i+2} \ldots p_{2+10k}$, where the 0-squares c_i and c_{i+1} of the conformation $c(S')$ are adjacent to the ligand t_k. Set

$$p(S) = p_1 \ldots p_{i-1}010010010010p_{i+2} \ldots p_{2+10k}\,.$$

From the construction of the protein $p(S)$, it is easy to check that the fold of the protein $p(S)$ with respect to $c(S)$ is saturated. Indeed, in every inductive step, we add four "1" and six "0" monomers to the protein which completely fill one regular tile in such a way that the added 1-squares form a 1-cycle of length 4; cf. Fig. 2(d). The second part of the theorem follows by Observation 1. ■

For every constructible structure S, consider all proteins having saturated fold with respect to conformation $c(S)$. Note that the protein defined in the above proof is the one with the maximum number of hydrophobic "1" monomers among those proteins. In what follows, we will denote this protein by $p(S)$.

Our goal is to show that $p(S)$ is stable. However, this seems extremely difficult; here we show the result for particular special cases. The following local properties of the proteins $p(S)$ can easily be seen.

Observation 2. *For any constructible structure S, the protein $p(S)$ satisfies the following properties:*

- $p(S) \in 0\{0,1\}^*0$, *and*
- $p(S)$ *does not contain any of* 11, 000, 1010101, *and* $100100100100 = (100)^4$ *as a substring.*

2.3. Our results

We believe that for all constructible structures S, $p(S)$ is stable:

Conjecture 1. *For any constructible structure S, the protein $p(S)$ is stable.*

It is easy to prove that the conjecture is true for the constructible structure with the empty tiling sequence $\langle\rangle$.

Claim 1. *The protein* $p = p(S_{\langle\rangle}) = 010010010010$ *is stable.*

Proof. Since the conformation depicted in Fig. 2(c) is a native conformation for p, by Observation 1, for any native conformation c' for p, the 1-squares form a single 1-cycle of length 4. Now, it is easy to check that there is only one possibility for placing the 0's of the protein in the fold, so p is stable. ■

An extensive computer search shows the conjecture is satisfied for over $20,000$ constructible structures including all structures composed of up to eight tiles. To tackle this conjecture, we first consider a broad subclass, the linear constructible structures.

Definition 6 (Linear structures). *We say that a constructible structure S is* linear *if it is constructed such that every regular tile is attached to the ligand of the last placed tile, i.e., the tiling sequence $\langle t_1, \ldots, t_k\rangle$ of S satisfies the following condition: $t_\ell \in \{3\ell - 2, 3\ell - 1, 3\ell\}$, for every $\ell = 1, \ldots, k$.*

Note that a linear constructible structure of length n can be described by a *linear tiling sequence* in $\{1, 2, 3\}^n$ where $i \in \{1, 2, 3\}$ in position k denotes that the k-th regular tile is attached to the ligand $3(k-1) + i$ of the $(k-1)$-th tile. We can interpret the number 1 in this tiling sequence to mean "turn right," 2 to mean "continue straight," and 3 to mean "turn left" when traveling along the linear chain of tiles. Note that 1, 1 (resp., 3, 3) can be a subsequence of a linear tiling sequence describing a constructible structure only if it is the prefix of the sequence. An example of a linear constructible structure with the linear tiling sequence

$$\langle 3, 1, 3, 1, 3, 1, 3, 1, 2, 1, 2, 1, 3, 1, 3, 1, 3, 2, 3, 2, 3, 1, 3, 1, 2, 1, 2\rangle$$

is depicted in Fig. 3(a).

Since, for any linear constructible structure S, the protein $p(S)$ contains exactly one substring 1001001001 corresponding to "the turning point," i.e., the last added regular tile, we believe that it should be easier to identify the last tile in the fold of $p(S)$ and continue backwards showing that the conformation $c(S)$ is the only possibility for $p(S)$.

Clearly if Conjecture 1 holds, then it also holds for all linear constructible structures. Let us factorize the class of linear constructible structures by the number of "bends," i.e., the number of 1's and 3's in the sequence:

$$\mathcal{L}_n = \{S_T : T = 2^{i_0}, t_1, 2^{i_1}, \ldots, t_{n-1}, 2^{i_{n-1}}, t_n, \text{ where } t_1, \ldots, t_{n-1} \in \{1, 3\} \text{ and } S_T \text{ is constructible}\}.$$

A structure in $\mathcal{L}_n$ is called $\mathcal{L}_n$*-structure*. Our main result is that the conjecture holds for $\mathcal{L}_0$ and $\mathcal{L}_1$ (proved in the next section). We believe that our proof techniques form the basis for proving the conjecture for all linear constructible structures.

3. CLASSES $\mathcal{L}_0$ AND $\mathcal{L}_1$

In this section, we will first prove the conjecture for all $\mathcal{L}_0$-structures, and then we extend this result to all $\mathcal{L}_1$-structures.

3.1. $\mathcal{L}_0$-structures

Let us first characterize all $\mathcal{L}_0$-structures. For any integer $n \geq 1$, a constructible structure with the linear tiling sequence $\langle \underbrace{2, 2, \ldots, 2}_{n-1}\rangle$ is a $\mathcal{L}_0$-structure and is denoted by S_n. Observe that a constructible structure

S is a $\mathcal{L}_0$-structure if and only if $p(S)$ does not contain 10101 as a substring and if and only if $p(S)$ contains exactly two occurrences of the substring 1001001. This observation will help us to prove stability of $\mathcal{L}_0$-structures.

The main result of this subsection is the following.

Theorem 2. *For every $n \geq 1$, the protein $p(S_n) = 0(10010)^n(01001)^n0$ is stable. Consequently, for every structure S in $\mathcal{L}_0$, the protein $p(S)$ is stable.*

Consider a native fold $\hat{F}$ of $p(S_n)$. By Theorem 1, it is saturated. To prove Theorem 2, it is enough to show that $\hat{F}$ must be the fold of $p(S_n)$ with respect to $c(S_n)$. Let us start by observing simple properties of $\hat{F}$.

Observation 3. *Let $p \in 0\{0,1\}^*0$ be a protein not containing 11 and 000 as a substring. Then every saturated fold of p has the following properties:*

(a) every 1-square has two 1-squares and two 0-squares as neighbors;
(b) every 0-square has at least one adjacent 1-square;
(c) an adjacent 1-square and 0-square are c-connected where c is a conformation of the fold; and
(d) adjacent 1-squares are connected by a bond.

In particular, the above properties are satisfied for any protein $p(S)$ where S is a constructible structure.

Next, we will enclose all 1-squares of the fold in a rectangular region.

Definition 7 (Diagonal frame). *Let $A_1 \leq A_2, B_1 \leq B_2$ be integer constants. The* diagonal rectangle *$R(A_1, A_2, B_1, B_2)$ is the set of squares $[a, b]$ which satisfy the inequalities $A_1 \leq a + b \leq A_2$ and $B_1 \leq a - b \leq B_2$. The* **SW**-border line, **NE**-border line, **NW**-border line, *and* **SE**-border line *of the rectangle $R(A_1, A_2, B_1, B_2)$ are the sets of squares $\{[a, b];\ a+b = A_1\}$, $\{[a, b];\ a+b = A_2\}$, $\{[a, b];\ a-b = B_1\}$, and $\{[a, b];\ a - b = B_2\}$, respectively. Let F be a fold containing at least one 1-square. The* diagonal frame *of the fold F is the smallest diagonal rectangle $R(A_1, A_2, B_1, B_2)$ containing all 1-squares of the fold F; cf. Fig. 4(a).*

Consider one border line of the diagonal frame of the fold $\hat{F}$, say $a + b = A_2$. This divides the grid into two parts, the *inner* part $a + b \leq A_2$ and the *outer* part $a + b > A_2$. The squares of the outer part are either empty or 0-squares. Since, by Observation 3(b), at least one neighbor of a 0-square must be a 1-square, among the squares of the outer part, the 0-squares can appear only on the diagonal line next to the border line of the frame, i.e., on $a + b = A_2 + 1$.

A 1-square lying on a border line is called a *boundary square*. We will show that each boundary square lies on a 1-cycle of length 4. Such 1-cycles will be called *cores*. More formally:

Definition 8 (Cores). *Consider a fold with respect to a conformation c. A* core *is a 1-cycle of length 4 such that every 1-square of the 1-cycle is c-connected to two 0-squares; cf. Fig. 4(b). If the northern*

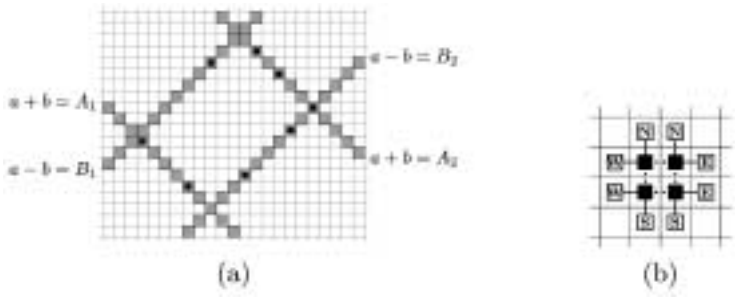

FIG. 4. An illustration of (**a**) a diagonal frame $R(A_1, A_2, B_1, B_2)$ (the black squares depict boundary squares), (**b**) a core.

(resp., eastern, southern, western) 0*-squares of a core, marked with* ***N****(resp.,* ***E****,* ***S****,* ***W****), are c-connected, we say that the core is* **N**-closed *(resp.,* **E**-closed, **S**-closed, **W**-closed*). If, for instance, a core is* ***N****-closed and* ***E****-closed, we say that it is* **NE**-closed*. The* main square *of a* ***NE****-closed (resp.,* ***SE****-closed,* ***SW****-closed,* ***NW****-closed) core is the northeast (resp., southeast, southwest, northwest)* 1*-square of the core. A core closed from two adjacent sides is called a* corner-closed *core, and a core closed from three sides is called a* completely-closed *core.*

Let $\mathcal{B}$ be the set of all boundary squares. In general, the cardinality of $\mathcal{B}$ is at least three if the fold contains at least three 1-squares. However, for the folds which we are interested in, we have a slightly better bound.

Observation 4. *Let* $p \in 0\{0,1\}^*0$ *be a protein not containing* 11 *and* 000 *as a substring. For any saturated fold of p, we have that each boundary square lies on exactly one border line. Hence, the number of boundary squares is at least 4, and there are at least two boundary squares which are not adjacent to a terminal.*

Proof. Assume, that a boundary square s lies on two border lines. For instance, on the **NE**-border line and the **SE**-border line. Since it lies on **NE**-border line, its northern and eastern neighbors cannot be 1-squares, and since it also lies on the **SE**-border line, its southern neighbor cannot be a 1-square as well. Then, at most one neighbor of s can be a 1-square which contradicts Observation 3(a). The first part of the observation follows.

Since each border line contains at least one boundary square, the cardinality of the set $\mathcal{B}$ is at least 4. Since both terminals are 0-squares, by Observation 3(c), they can be adjacent to at most one 1-square. Hence, there are at most two of boundary squares adjacent to terminals. ■

The above observation guarantees the existence of boundary squares not adjacent to either of the two terminals. The following lemma shows why such squares are very useful for our purposes.

Lemma 2. *Let* $p \in 0\{0,1\}^*0$ *be a protein not containing* 11, 000 *and* 10101 *as a substring. For every saturated fold of p and every* $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$*, each boundary square s lying on the X-border line not adjacent to a terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core.*

Proof. Let $\hat{c}$ be the conformation of the fold. Without loss of generality, assume that $s = [0, 0]$ and that it lies on the **NE**-border line. By Observation 3(a), out of four neighbors of s, two are 1-squares and two are 0-squares. But since the outer part $a + b > A_2$ cannot contain any 1-square, the eastern neighbor $[1, 0]$ and northern neighbor $[0, 1]$ are 0-squares, while the neighbors $[-1, 0]$ and $[0, -1]$ are 1-squares. By the assumption of the lemma, none of the neighbors of s is a terminal; hence, the squares $[1, 0]$ and $[0, 1]$ are both $\hat{c}$-connected to nonempty squares other than s. Those must be the squares $[1, -1]$ and $[-1, 1]$, respectively, since all other adjacent squares are too far from the border line. We have the situation depicted in Fig. 5(a).

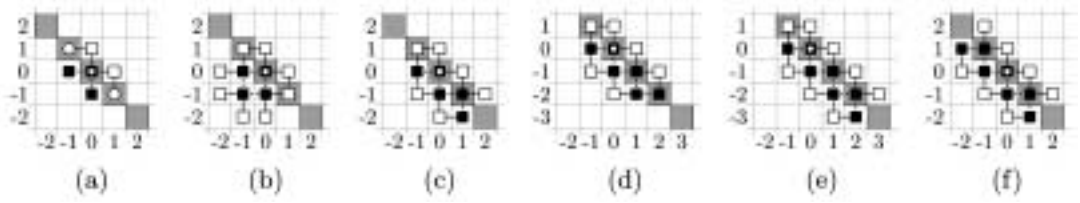

FIG. 5. The situation around the boundary square $s = [0, 0]$ (marked with a white dot) which is at distance at least (**a**–**c**) 2 from any terminal, (**d**–**f**) 4 from any terminal.

Now we will consider three cases depending on squares $[1, -1]$ and $[-1, 1]$, depicted as empty circles in Fig. 5(a).

Case 1. They both are 0-squares. If the square $[-1, -1]$ is also a 0-square then, by Observation 3(c), the fold $\hat{F}$ would contain a closed $\hat{c}$-path which is not possible. Hence, assume that $[-1, -1]$ is a 1-square; cf. Fig. 5(b). By Observation 3(c), the squares $[-2, 0]$, $[-2, -1]$, $[-1, -2]$, and $[0, -2]$ are all 0-squares. We are done: the square s is the main square of a **NE**-closed core.

Case 2. One of the squares is a 1-square and the other is a 0-square. Without loss of generality assume that $[1, -1]$ is a 1-square and $[-1, 1]$ is a 0-square. Since two neighbors of $[0, -1]$ are 1-squares, by Observation 3(a), the remaining two neighbors, $[-1, -1]$ and $[0, -2]$, are 0-squares. Similarly, by Observation 3(a) applied on the 1-square $[1, -1]$ and the fact that the outer part $a + b > A_2$ cannot contain 1-squares, we have that $[2, -1]$ is a 0-square and $[1, -2]$ is a 1-square; cf. Fig. 5(c). This yields a contradiction, as

$$([-1, 0], [-1, -1], [0, -1], [0, -2], [1, -2])$$

conforms to 10101.

Case 3. They both are 1-squares. We have again an occurrence of the substring 10101 starting at $[-1, 1]$ and ending at $[1, -1]$ in the fold, a contradiction. ■

Corollary 1. *Consider a native fold of the protein $p(S_n)$, $n \geq 1$. A boundary square s lying on the X-border line not adjacent to a terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core, for every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$.*

Proof. By Theorem 1, $p(S_n)$ has a saturated fold, and hence, by Observation 1, any native fold is also saturated. Furthermore, by Observation 2, it is enough to notice that the string $p(S_n) = 0(10010)^n(01001)^n0$ does not contain 10101 as a substring. ■

Now, we are ready to prove Theorem 2.

Proof of Theorem 2. We will prove, by induction on n, that every saturated fold of $p(S_n)$ is the fold $F_{p(S_n),c(S_n)}$ (recall Definition 1). The base case $n = 1$ of the induction follows by Claim 1. Hence, take an $n > 1$. Let $\hat{c}$ be a native conformation for $p(S_n)$ and $\hat{F}$ be the saturated fold of $p(S_n)$ with respect to $\hat{c}$. Our goal is to identify the substring 1001001001 of the protein $p(S_n)$ in $\hat{F}$ and show that it folds as a completely-closed core. Then, if we cut out this completely-closed core from $\hat{F}$, we obtain a saturated fold F' of the protein $p(S_{n-1})$. By the induction hypothesis, $F' = F_{p(S_{n-1}),c(S_{n-1})}$. If we attach the completely-closed core to the fold F' back to its original place, we can easily observe that $\hat{F} = F_{p(S_n),c(S_n)}$.

Hence, it suffices to find a completely-closed core in $\hat{F}$. Take two boundary squares not adjacent to any terminal (their existence is guaranteed by Observation 4). By Corollary 1, such squares are main squares of their corner-closed cores. Hence, each of the two boundary squares is a 1-square corresponding to the underlined 1 in a substring 100$\underline{1}$001 of the protein. There are only two occurrences of this substring in $p(S_n)$. Therefore, the two boundary 1-squares correspond to the underlined 1's in the substring 100$\underline{1}$00$\underline{1}$001 of the protein, and they are main squares of the same core. Obviously, such a core has to be completely closed. ■

3.2. Modifications of Lemma 2 and its corollary

Before we show that Conjecture 1 holds for $\mathcal{L}_1$-structures, we will prove two modifications of Lemma 2 and its corollary. Both modifications might be useful for proving similar results for other special constructible structures, whereas the latter will be used in the proof of stability of L-structures. By the *distance* of two squares $[a, b]$ and $[c, d]$, we mean the Manhattan distance, i.e., $d([a, b], [c, d]) = |a - c| + |b - d|$.

Lemma 3. *Let $p \in 0\{0,1\}^*0$ be a protein not containing* 11, 000, *and* 1010101 *as a substring. For every saturated fold of p and every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$, each boundary square s lying on the X-border line at distance at least* 4 *from any terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core.*

Proof. The assumptions of the lemma satisfy all assumptions of Lemma 2 except the condition that the string p does not contain 10101 as a substring. This condition is used only in the last step of the second and third cases of the proof to derive a contradiction. Reconsider these two cases:

Case 2. Let us consider the situation as depicted in Fig. 5(c). By the additional assumption of Lemma 3, the square $[2, -1]$ at distance 3 from $s = [0, 0]$ cannot be a terminal. Hence, it has a nonempty neighbor besides $[1, -1]$, which has to be the square $[2, -2]$. If this neighbor is a 0-square, then the fold contains a closed $\hat{c}$-path, a contradiction. Thus, the square $[2, -2]$ is a 1-square; cf. Fig. 5(d). Now, using Observation 3(a) and the fact that the outer part $a + b > A_2$ does not contain any 1-squares, we deduce that $[1, -3]$ and $[3, -2]$ are 0-squares and $[2, -3]$ is a 1-square. As is easily seen in Fig. 5(e), the protein p contains a substring 1010101 which is a contradiction.

Case 3. Assume that both $[1, -1]$ and $[-1, 1]$ are 1-squares. The eastern neighbor of the 1-square $[1, -1]$ lies in the outer part; hence, it is a 0-square. Applying Observation 3(a) on 1-squares $[0, -1]$ and $[1, -1]$, we have that the squares $[-1, -1]$ and $[0, -2]$ are 0-squares and the square $[1, -2]$ is a 1-square. By symmetry, we can apply the similar argument around squares $[-1, 0]$ and $[-1, 1]$, thus, obtaining the situation depicted in Fig. 5(f). In this situation, we have an occurrence of the substring 1010101 starting at $[-2, 1]$ and ending at $[1, -2]$ in the fold, a contradiction. ■

Lemma 3 together with Observation 2 yield:

Corollary 2. *Consider a native fold of the protein $p(S)$, where S is a constructible structure. For every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$, a boundary square s lying on the X-border line at distance at least* 4 *from any terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core.*

The above claims work for any constructible structure; however, they require the boundary square to be at distance at least 4 from any terminal. In the proofs of stability of the proteins $p(S)$ for particular constructible structures S, it might be useful to have a similar result for all boundary squares nonadjacent to a terminal. This was achieved in Lemma 2 and its corollary, but the price was too high: they work only for the simplest linear constructible structures, namely, $\mathcal{L}_0$-structures. The next lemma generalizes Lemma 2 for a much richer variety of constructible structures.

Lemma 4. *Let $p \in 0\{0,1\}^*0$ be a protein which does not contain the string* 01010010101 *as a prefix or* 10101001010 *as a suffix and does not contain* 11, 000 *and* 1010101 *as substrings. For every saturated fold of p and every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$, we have that each boundary square s lying on the X-border line not adjacent to a terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core.*

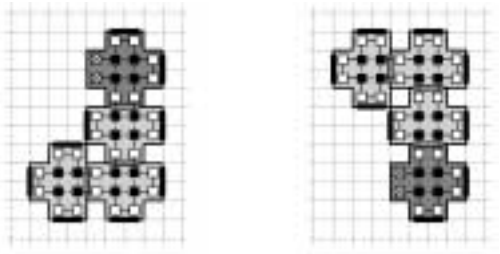

FIG. 6. Forbidden substructures for Corollary 3.

Proof. We can apply the proof of Lemma 3, as the square $[2, -1]$ cannot be any of the two terminals, unless either 01010010101 is a prefix or 10101001010 is a suffix of p; cf. Fig. 5(c). ■

Corollary 3. *Consider a native fold of the protein $p(S)$, where S is a constructible structure which does not contain either of the structures depicted in Fig. 6 as a substructure; i.e., at least one of three regular tiles in either of the pictures is not a part of S. For every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$, a boundary square lying on the X-border line not adjacent to a terminal lying outside of the diagonal frame of the fold is the main square of an X-closed core.*

3.3. $\mathcal{L}_1$-structures

In this subsection, we further extend the result of Theorem 2 for the second class of linear constructible structures. Consider the following set of constructible structures:

> For any pair of integers $n \geq 1$ and $m \geq 0$, let $L_{n,m}$ be a linear constructible structure with the linear tiling sequence
>
> $$\langle \underbrace{2, 2, \ldots, 2}_{n-1}, \underbrace{3, 2, 2, \ldots, 2}_{m} \rangle .$$

Observe that if we prove that for all $n, m \geq 1$ that the proteins $p(L_{n,m})$ are stable, then by symmetry—the reverse image of a protein $p(L_{n,m})$ is a protein $p(S)$ where S is a constructible structure with the linear tiling sequence $\langle \underbrace{2, 2, \ldots, 2}_{n-1}, \underbrace{1, 2, 2, \ldots, 2}_{m} \rangle$— we have that Conjecture 1 holds for all $\mathcal{L}_1$-structures.

Note also that a degenerated structure $L_{n,0}$ is actually the $\mathcal{L}_0$-structure S_n. It will be used as a base case of the induction in the proof of stability of proteins $p(L_{n,m})$. Figure 7(a) shows an illustrations of $\mathcal{L}_1$-structures $L_{1,1}$ and $L_{2,2}$.

Observe that a constructible structure S is an $\mathcal{L}_1$-structure if and only if $p(S)$ contains exactly one occurrence of the substring 10101, and if and only if $p(S)$ contains exactly three occurrences of the substring 1001001. This observation will help us to prove stability of $\mathcal{L}_0$-structures. As for $\mathcal{L}_0$-structures, this observation will help us to show that Conjecture 1 also holds for $\mathcal{L}_1$-structures. The proof involves a lengthy case analysis.

Theorem 3. *For every $n \geq 1$ and $m \geq 0$, the protein*

$$p(L_{n,m}) = 0(10010)^n 010(10010)^m (01001)^m 01(01001)^{n-1} 0$$

is stable. Consequently, for every constructible structure S in $\mathcal{L}_1$, the protein $p(S)$ compatible with S is stable.

First, let us prove an auxiliary lemma. The distance of square s from a line (set of squares) is the minimum Manhattan distance between s and any square lying on the line.

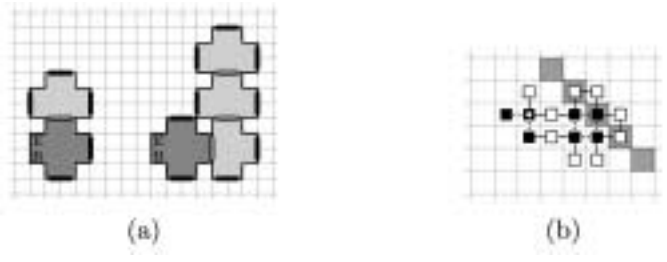

FIG. 7. An example (**a**) of two $\mathcal{L}_1$-structures $L_{1,1}$ (left) and $L_{2,2}$ (right); (**b**) why the proof of Lemma 3 cannot be directly applied on 1-squares at distance greater than 2 from the border line.

Lemma 5. *Let $p \in 0\{0,1\}^*0$ be a protein not containing* 11, 000 *and* 1010101 *as a substring. Consider a saturated fold of p such that both terminals lie outside of its diagonal frame. For every $X \in \{\boldsymbol{NE}, \boldsymbol{SE}, \boldsymbol{SW}, \boldsymbol{NW}\}$, if the X-border line contains only* 1*-squares which are main squares of their X-closed cores, then each* 1*-square at distance at most* 2 *from the X-border line is a square of an X-closed core.*

Proof. Let c be a conformation of the fold. Without loss of generality, consider the **NE**-border line $a + b = A_2$.

First, we prove the claim for all 1-squares at distance 1 from the border line, and then using this fact, we extend the claim to 1-squares at distance 2. Assume that we have a 1-square s lying on the line $a + b = A_2 - 1$ which is not a square of any **NE**-closed core. We can apply the proof of Lemma 3 on s taking $a + b = A_2 - 1$ as the new border line. The proof starts with the 1-square s and tries to extend the 1-cycle of s in both directions. The 1-neighbors of a square t are the first 1-squares different from t encountered when traveling along c-edges from t towards the terminals.

The same proof applies because the following conditions hold:

1. All 0-squares lying inside the new diagonal frame (hence also those at distance at most 3 from s) lie inside the original frame; hence by the assumption, they cannot be terminals.
2. Both 1-neighbors of each 0-square lying on the line $a+b = A_2+1$ are parts of a **NE**-closed core. Since s is not a part of any **NE**-closed core, no 0-square considered in the proof of Lemma 3 can c-connect to any 0-square on the line $a + b = A_2 + 1$.
3. All 1-squares lying on the line $a+b = A_2$ are parts of **NE**-closed cores. Since all 1-squares considered in the proof of Lemma 3 are part of the same 1-cycle which is not **NE**-closed core, all their neighbors lying outside of the new frame are 0-squares.

Now, assuming that all 1-squares at distance 1 from the border line are parts of cores, we can apply the proof of Lemma 3 to 1-squares at distance 2 taking $a + b = A_2 - 2$ as a new border line. Conditions 2 and 3 hold for squares on line $a + b = A_2$ and $a + b = A_2 - 1$, respectively. To ensure that 0-squares lying in the outer part of the new border line have to connect to squares in the inner part, we need another condition:

4. Both 1-neighbors of each 1-square lying on the line $a + b = A_2$ are parts of **NE**-closed cores. Hence, no 0-square considered in the proof of Lemma 3 can c-connect to any 1-square on the line $a+b = A_2$.

■

Note that the above lemma cannot be recursively applied. For example, condition 4 and later also condition 2 will not necessary be satisfied when we get further from the original border line. Figure 7(b) shows an example of when condition 4 fails. The 1-square s (marked with a white dot) is at distance 3 from the **NE**-border line. However, the eastern neighbor of s is not necessarily c-connected to its southern neighbor.

Now, we are ready for the proof of the theorem.

Proof of Theorem 3. We will prove the theorem by induction on m. The base case $m = 0$ follows by Theorem 2. Now, take integers $n, m \geq 1$, and let us assume that for all $k < m$, the protein $p(L_{n,k})$ is stable. Let $\hat{c}$ be a native conformation of $p(L_{n,m})$ and $\hat{F}$ be the fold with respect to $\hat{c}$. Recall that, by Theorem 1, $\hat{F}$ is saturated.

Note that the protein $p(L_{n,m})$ contains exactly three occurrences of the substring 1001001. We will identify these occurrences in the fold $\hat{F}$. Let $R(A_1, A_2, B_1, B_2)$ be the diagonal frame of $\hat{F}$ and $\mathcal{B}$ the set of all boundary squares; cf. Fig. 4(a). Note that $L_{n,m}$ does not contain any of two forbidden structures depicted in Fig. 6. Hence, if at least one of the following conditions hold:

(a) the cardinality of $\mathcal{B}$ is at least 5, or
(b) there is at most one element in $\mathcal{B}$ adjacent to a terminal lying outside of the diagonal frame,

then, by Observation 4 and Corollary 3, there are three boundary squares in $\mathcal{B}$ which are main squares of corner-closed cores. Hence, in this case, we have identified all three occurrences of 1001001, and we can apply a similar argument as in the proof of Theorem 2. Since two of the occurrences of the substring 1001001 intersect, two of the boundary squares are main squares of the same core. Hence, the core must be complete. Cutting the core out, we get a saturated fold for the protein $L_{n,m-1}$ which, by the induction hypothesis, folds in a unique way as $c(L_{n,m-1})$. Hence, the only native conformation for $L_{n,m}$ is indeed $c(L_{n,m})$.

Now suppose that both conditions (a) and (b) are false. By falsity of (a), the cardinality of $\mathcal{B}$ is 4; i.e., each border line of the diagonal frame contains exactly one 1-square. By falsity of (b), both terminals are located outside of the diagonal frame. By Observation 4 and Corollary 3, we still have at least two boundary squares s and t which are main squares of their corner-closed cores. They are 1-squares corresponding to two out of three underlined 1's in the substrings 100$\underline{1}$00$\underline{1}$001 and 10100$\underline{1}$00101 of the protein $p(L_{n,m})$. If they both correspond to the underlined 1's of the first substring, then as above, they are main squares of the same completely closed core and we can apply induction. Therefore, assume that s corresponds to one of the underlined 1's of the first substring and that t corresponds to the underlined 1 of the second substring.

Our proof relies heavily on properties of Observation 3, and we use these without explicit reference to Observation 3.

Without loss of generality, assume that $s = [0, 0]$ and that it lies on the **NE**-border line. The square $t = [e, f]$ cannot lie on the **NE**-border line, and therefore there are three possibilities on which border line $[e, f]$ lies. We will assume that $[e, f]$ lies on the **NW**-border line. This is without loss of generality if we consider that the determined part of the fold around $[e, f]$ can appear in $\hat{F}$ correspondingly rotated.

First, we will determine the part of the fold around the square $[e, f]$. By Corollary 3, $[e, f]$ is the main square of a **NW**-closed core $C_{[e,f]}$; cf. Fig. 8(a). Since $[e, f]$ is a 1-square corresponding to the underlined 1 in the substring 10100$\underline{1}$00101 of the protein, the square $[e + 2, f]$ is $\hat{c}$-connected to another 1-square. Since, $[e + 2, f - 1]$ is already a 0-square, there are only two candidates: $[e + 2, f + 1]$ and $[e + 3, f]$. But the square $[e + 2, f + 1]$ cannot be a 1-square; otherwise, it would have to be $\hat{c}$-connected to the 0-square $[e + 1, f + 1]$ which is already adjacent to two $\hat{c}$-edges. Hence, we can assume that $[e + 2, f]$ is $\hat{c}$-connected to its eastern neighbor $[e + 3, f]$, which in turn has to be a 1-square.

Claim 3.1. *The northern neighbor $u = [e + 3, f + 1]$ of the 1-square $[e + 3, f]$ is a 0-square.*

Proof. Assume that u is a 1-square; cf. Fig. 8(b). By falsity of (a), the **NW**-border line contains no other 1-square. Hence, by Lemma 5, u is a part of a **NW**-closed core. Since, the 1-square $[e + 3, f]$ is adjacent to u and $[e + 2, f]$ is a 0-square, the main square of this core is u; cf. Fig. 8(c). Since the core is **W**-closed, the 0-square $[e + 2, f]$ should be $\hat{c}$-connected to $[e + 2, f + 1]$, a contradiction. ■

The two remaining neighbors of the 1-square $[e + 3, f]$, squares $[e + 4, f]$ and $[e + 3, f - 1]$, are 1-squares. By symmetry, we have the same configuration on the southern side of the core $C_{[e,f]}$. The situation is depicted in Fig. 8(d). Observe that the part of the fold in the figure contains an occurrence of the substring 10101 starting at square $[e + 1, f - 3]$ and ending at square $[e + 3, f - 1]$. Since the protein

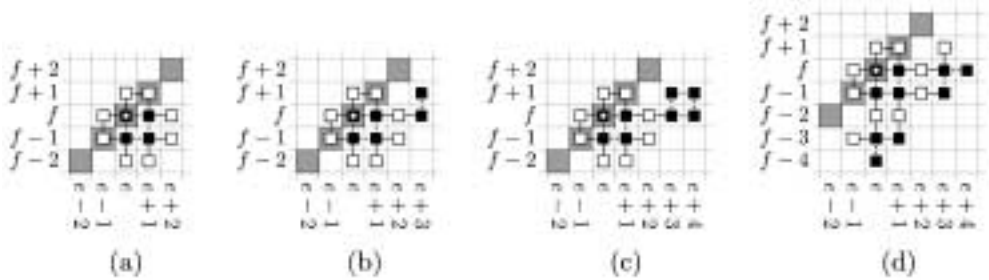

FIG. 8. The situation around the boundary 1-square $t = [e, f]$ (marked with a white circle) corresponding to the underline 1 in the substring 10100$\underline{1}$00101 of the protein $p(L_{n,m})$.

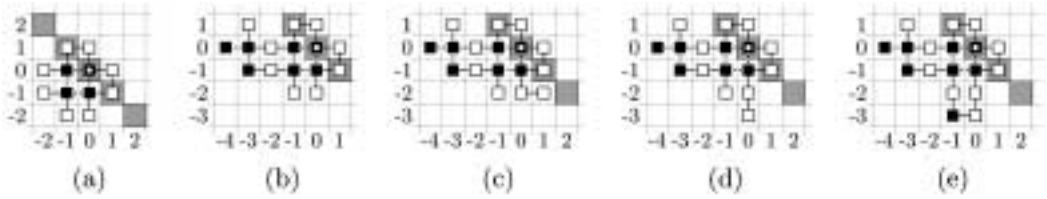

FIG. 9. The situation around the boundary 1-square $[0, 0]$ (marked with a white circle) corresponding to one of the underlined 1's in the substring $100\underline{1}00\underline{1}001$ of the protein $p(L_{n,m})$.

$p(L_{n,m})$ contains only one occurrence of this substring, in the following considerations, *if we find the part of the conformation containing the substring* 10101, *then it has to match the part of the conformation in Fig. 8(d) (up to rotations)*. This observation will turn out to be very useful.

Now, consider the part of the fold around the square $[0, 0]$. Assume that the square $[0, -2]$ is $\hat{c}$-connected to another 0-square. Consequently, since the protein does not contain $(100)^4$ as a substring, $[-2, 0]$ is $\hat{c}$-connected to a 1-square.

By Corollary 3, the square $[0, 0]$ is the main square of a **NE**-closed core $C_{[0,0]}$; cf. Fig. 9(a). Further, by falsity of (a), the **NE**-border line does not contain any other 1-square. Hence, we can apply Lemma 5: if there is a 1-square at distance at most 2 from the **NE**-border line, it lies in a **NE**-closed core. However, this is possible only for 1-squares of the core $C_{[0,0]}$. Indeed, each **NE**-closed core contains an occurrence of the substring 1001001, but we have only three occurrences of the substring: one in the core $C_{[e,f]}$ which is not **NE**-closed, one in $C_{[0,0]}$, and the last one starting in $[0, 0]$ which is partially contained in $C_{[0,0]}$ and cannot form a **NE**-closed core. Hence, we have the following claim.

Claim 3.2. *All* 1*-squares at distance at most* 2 *from the* ***NE****-border line are the* 1*-squares of the core* $C_{[0,0]}$.

The square $[-2, 0]$ must be $\hat{c}$-connected to another 1-square. We can apply the same proof as when considering the neighborhood of the square $[e, f]$ and derive that the squares $[-3, 0]$, $[-3, -1]$, and $[-4, 0]$ are 1-squares and the square $[-3, 1]$ is a 0-square; cf. Fig. 9(b).

On the southern side of the core $C_{[0,0]}$ we have:

Claim 3.3. *The second* $\hat{c}$*-neighbor of* $[0, -2]$ *is either* $[-1, -2]$ *or* $[0, -3]$.

Proof. As we assumed, the second $\hat{c}$-neighbor of $[0, -2]$ is a 0-square. It is enough to prove that it is not the square $[1, -2]$. Assume it is; cf. Fig. 9(c). The another $\hat{c}$-neighbor of the 0-square $[1, -2]$ must be a 1-square. All neighbors of $[1, -2]$ lies at distance at most 2 from the **NE**-border line; hence, by Claim 3.2, they cannot be 1-squares, a contradiction. ■

Since in the first case we obtain a completely-closed core, we can apply the induction. Thus, we consider the second case, the square $[0, -2]$ is $\hat{c}$-connected to the 0-square $[0, -3]$, as depicted in Fig. 9(d).

Claim 3.4. *The second* $\hat{c}$*-neighbor of* $[0, -3]$ *is the* 1*-square* $[0, -4]$.

Proof. Obviously, the second $\hat{c}$-neighbor of $[0, -3]$ is a 1-square. It can be one of the following three candidates: $[1, -3]$, $[-1, -3]$, or $[0, -4]$. By Claim 3.2, the square $[1, -3]$ cannot be a 1-square. Assume that $[-1, -3]$ is a 1-square. We have an occurrence of the substring 10101 starting at $[-1, -3]$ and ending at $[-3, -1]$; cf. Fig. 9(e). Since there is only one occurrence of this substring in the protein $p(L_{n,m})$, it should match the occurrence depicted in Fig. 8(d) up to rotation. If we align these two parts of the fold, we see that the square $[0, -3]$ should be at the same time a 0-square and a 1-square, a contradiction. Note also that the orientation of the piece of conformation containing the substring 10101 in Fig. 9(e) is contradictory as well: it would imply that the boundary square $[e, f]$ lies on the same border line as $[0, 0]$.

Since, out of three candidates, only $[0, -4]$ can be a 1-square, the 0-square $[0, -3]$ is $\hat{c}$-connected to it. ■

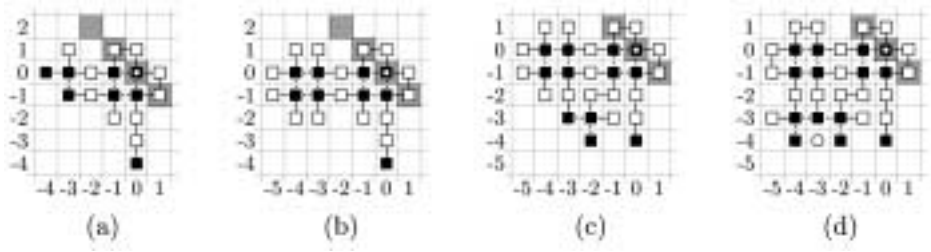

FIG. 10. The situation around the boundary 1-square $[0, 0]$ (marked with a white circle) corresponding to one of the underlined 1's in the substring 100$\underline{1}$00$\underline{1}$001 of the protein $p(L_{n,m})$. The case when the 0-square $[0, -3]$ is $\hat{c}$-connected to 1-square $[0, -4]$.

We have a situation depicted in Fig. 10(a). The square $[-4, -1]$ has to be a 1-square. Indeed, if $[-4, -1]$ is a 0-square, we have an occurrence of the substring 10101 starting at $[-4, 0]$ and ending at $[-1, -1]$ in the fold not matching the occurrence in Fig. 8(d), a contradiction. The 1-square $[-4, -1]$ has completed a core; hence, we can add the necessary 0-squares around the core; cf. Fig. 10(b).

Claim 3.5. *The square* $[-1, -2]$ *is* $\hat{c}$*-connected to the* 0*-square* $[-1, -3]$.

Proof. Obviously, $[-1, -3]$ cannot be a 1-square. If it is $\hat{c}$-connected to $[-1, -2]$, it has to be a 0-square, and we are done. Assume that it is not $\hat{c}$-connected to $[-1, -2]$. Since the square $[-1, -2]$ cannot be a terminal, it is $\hat{c}$-connected to its western neighbor $[-2, -2]$, which can only be a 0-square. The another $\hat{c}$-neighbor of $[-2, -2]$ has to be a 1-square, and the square $[-2, -3]$ is the only possibility. Now, the eastern neighbor $[-1, -3]$ of the 1-square $[-2, -3]$ has to be a 0-square, and the remaining two neighbors, $[-3, -3]$ and $[-2, -4]$, have to be 1-squares; cf. Fig. 10(c). Note that we have an occurrence of the substring 10101 starting at $[-1, -1]$ and ending at $[-3, -3]$ in the fold, and it has to match the occurrence in Fig. 8(d). Hence, we have a situation depicted in Fig. 10(d). The square $[-3, -4]$, depicted as a circle, is adjacent either to three $\hat{c}$-edges or three bonds, depending whether it is a 0-square or a 1-square, a contradiction. ■

The second $\hat{c}$-neighbor of $[-1, -3]$ has to be a 1-square. It is either the western neighbor $[-2, -3]$ or the southern neighbor $[-1, -4]$. In the first case, the northern neighbor of 1-square $[-2, -3]$ has to be a 0-square and has to be $\hat{c}$-connected to $[-3, -2]$. We have a closed $\hat{c}$-path; cf. Fig. 11(a), which is a contradiction. Hence, assume that $[-1, -3]$ is $\hat{c}$-connected to 1-square $[-1, -4]$. We have the situation depicted in Fig. 11(b).

Claim 3.6. *The second* $\hat{c}$*-neighbor of* $[0, -4]$ *is the* 0*-square* $[1, -4]$.

Proof. Obviously, the next $\hat{c}$-neighbor of $[0, -4]$ is a 0-square. It can be either square $[0, -5]$ or $[1, -4]$. Consider the first possibility. The remaining neighbor $[1, -4]$ of the 1-square $[0, -4]$ has to be a 1-square. The northern neighbor $[1, -3]$ of the 1-square $[1, -4]$ can only be a 0-square; cf. Fig. 11(c). Consider its

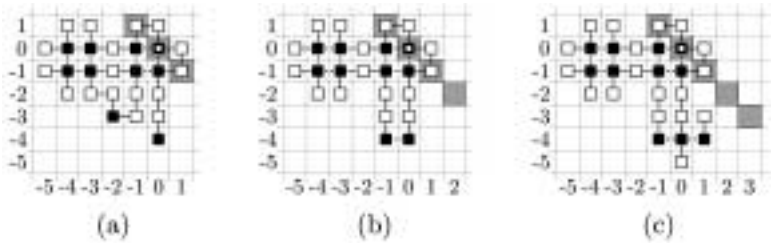

FIG. 11. The situation around the boundary 1-square $[0, 0]$ (marked with a white circle) corresponding to one of the underlined 1's in the substring 100$\underline{1}$00$\underline{1}$001 of the protein $p(L_{n,m})$ (continued).

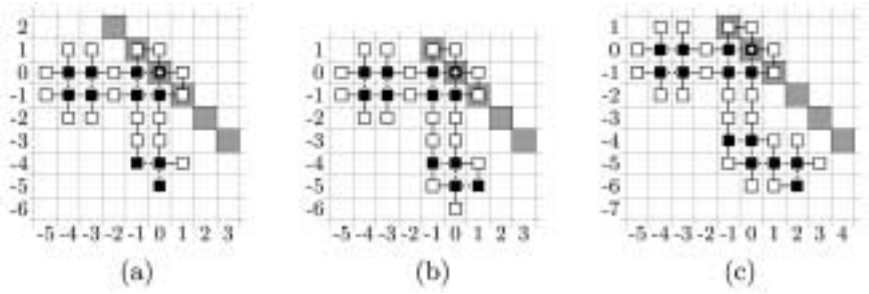

FIG. 12. The situation around the boundary 1-square $[0, 0]$ (marked with a white circle) corresponding to one of the underlined 1's in the substring 100$\underline{1}$00$\underline{1}$001 of the protein $p(L_{n,m})$ (continued).

another $\hat{c}$-neighbor u. It can be either $[1, -2]$ or $[2, -3]$. By Claim 3.2, neither of two candidates for u can be a 1-square. The other $\hat{c}$-neighbor of u must be a 0-square, but all feasible candidates are at distance at most 2 from the **NE**-border line, a contradiction with Claim 3.2. ■

The remaining neighbor $[0, -5]$ of the 1-square $[0, -4]$ has to be a 1-square; cf. Fig. 12(a). Since the protein does not contain $(100)^4$ as a substring, the next $\hat{c}$-neighbor of 0-square $[1, -4]$ has to be a 1-square. By Claim 3.2, the northern and eastern neighbors of $[1, -4]$ cannot be a 1-square. Hence, $[1, -4]$ is $\hat{c}$-connected to its southern neighbor $[1, -5]$. The remaining neighbors of the 1-square $[0, -5]$, i.e., $[-1, -5]$ and $[0, -6]$, are 0-squares, cf. Fig. 12(b).

Consider the southern neighbor of the 1-square $[1, -5]$. If it is a 1-square, then the conformation would contain the substring 10101 starting at $[-1, -4]$ and ending at $[1, -6]$ in the fold which does not match the one in Fig. 8(d), a contradiction. Hence, assume that $[1, -6]$ is a 0-square and the remaining neighbor $[2, -5]$ of the 1-square $[1, -5]$ is also a 1-square. By Claim 3.2, the northern and the eastern neighbors of the 1-square $[2, -5]$ cannot be a 1-square. The remaining neighbor $[2, -6]$ of 1-square $[2, -5]$ is necessarily a 1-square; cf. Fig. 12(c). But this is a contradiction, since we have an occurrence of the substring 10101 starting at $[0, -4]$ and ending at $[2, -6]$ in the fold not matching the one in Fig. 8(d). ■

CONCLUSIONS

We have proven Conjecture 1 for a number of basic constructible structures. We believe that our results can be generalized to prove at least the following relaxation of Conjecture 1:

Conjecture 2 (Linear structures). *For any linear constructible structure S, the protein $p(S)$ is stable.*

Proving Conjecture 2 may not be so easy, so we suggest a number of relaxations which restrict the bends in the linear constructible structure. Below are two possible relaxations.

Conjecture 3 (Slowly bending structures). *We say that a linear constructible structure S is* slowly bending *if its linear tiling sequence*

- *starts with* 2, *and*
- *does not contain any of* $1, 3$ *and* $3, 1$ *as a subsequence, that is after each turn the chain of tiles continues straight for at least one tile.*

We conjecture that for any slowly bending constructible structure S, the protein $p(S)$ is stable.

Conjecture 4 (Spiral structures). *A linear constructible structure $S \in \{1, 2\}^* \cup \{2, 3\}^*$ is called* spiral, *i.e., the chain of tiles can turn only in one direction. We conjecture that for any spiral constructible structure S, the protein $p(S)$ is stable.*

Examples of a slowly bending constructible structure and a spiral constructible structure are depicted in Fig. 3(b) and Fig. 3(c), respectively.

Other interesting problems along these lines are

- to find the class of proteins with similar expressible properties which are strongly stable—there is a big gap between the score of the native conformation and the score of any other conformation of the particular protein from the class;
- to find the class of proteins with similar properties for other lattices, namely, for the 3D square lattice and the 2D and 3D triangular lattices.

The major obstacle in extending our results to a 3D square lattice is the fact that it does not allow saturated folds since the degree of the lattice, i.e., the number of neighbors of a site, is six. Perhaps, the most optimal design would be placing tiles on top of each other, creating 2×2 columns of hydrophobic monomers, resembling helices. Another promising direction is to use other 3D lattices with degree 4 or 5.

REFERENCES

Agarwala, R., Batzoglou, S., Dančík, V., Decatur, S.E., Farach, M., Hannenhalli, S., and Skiena, S. 1997. Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model. *J. Comp. Biol.* 4, 275–296.

Andrew, C.D., Penel, S., Jones, G.R., and Doig, A.J. 2001. Stabilizing nonpolar/polar side-chain interactions in the alpha-helix. *Proteins* 45, 449–455.

Berger, B., and Leighton, T. 1998. Protein folding in the hydrophobic–hydrophilic (HP) model is NP-complete. *J. Comp. Biol.* 5, 27–40.

Berman, P., DasGupta, B., Mubayi, D., Sloan, R., Turán, G., and Zhang, Y. 2004. The protein sequence design problem in canonical model on 2D and 3D lattices. *Proc. CPM '04*, 244–253.

Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., and Yannakakis, M. 1998. On the complexity of protein folding. *Proc. of STOC '98*, 597–603.

Dill, K.A. 1985. Theory for the folding and stability of globular proteins. *Biochemistry* 24, 1501–1509.

Dill, K.A. 1990. Dominant forces in protein folding. *Biochemistry* 29, 7133–7155.

Dill, K.A., Bromberg, S., Yue, K., Fiebig, K.M., Yee, D.P., Thomas, P.D., and Chan, H.S. 1995. Principles of protein folding: A perspective from simple exact models. *Protein Sci.* 4, 561–602.

Gupta, A., Maňuch, J., and Stacho, L. 2004. Inverse protein folding in 2D HP model. *In Proc. of CSB '04*. To appear.

Hart, W.E. 1997. On the computational complexity of sequence design problems. *Proc. of Comp. Mol. Biol.*, 128–136.

Hart, W.E., and Istrail, S. 1995. Fast protein folding in the hydrophobic–hydrophilic model within three-eights of optimal. *Proc. of STOC '95*, 157–168.

Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., and Hecht, M.H. 1993. Protein design by binary patterning of polar and nonpolar amino acids. *Science* 262, 1680–1685.

Lyu, P.C., Liff, M.I., Marky, L.A., and Kallenbach, N.R. 1990. Side chain contribution to the stability of alpha-helical structure in peptides. *Science* 250, 669–673.

Shakhnovich, E.I., and Gutin, A.M. 1993. Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl. Acad. Sci.* 90, 7195–7199.

Sun, S., Brem, R., Chan, H.S., and Dill, K.A. 1995. Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.* 8, 1205–1213.

Wang, T., Miller, J., Wingreen, N.S., Tang, C., and Dill, K.A. 2000. Symmetry and designability for lattice protein models. *J. Chem. Phys.* 113, 8329–8336.

Yu, Y.B. 2002. Coiled-coils: Stability, specificity, and drug delivery potential. *Advanced Drug Delivery Reviews* 54, 1113–1129.

Yue, K., and Dill, K.A. 1992. Inverse protein folding problem: Designing polymer sequences. *Proc. Natl. Acad. Sci. USA, Biophysics* 89, 4163–4167.

Address correspondence to:
Ján Maňuch
School of Computing Science
Simon Fraser University
8888 University Drive
Burnaby BC V5A 1S6, Canada

E-mail: jmanuch@sfu.ca

Author
Right running head okay as shown (short title)?

AU1
Figure 3 citation okay as added? If not, please cite elsewhere.

AU2
What does “(continued)” refer to? Continued from a previous figure? Please see Figures 11 and 12.