

Inverse protein folding in 2D HP model

(Extended abstract)

Arvind Gupta

School of Computing Science
Simon Fraser University, 8888 University Drive
Burnaby, BC, Canada V5A 1S6
arvind@mitacs.ca

Ján Maňuch

School of Computing Science and PIMS
Simon Fraser University, 8888 University Drive
Burnaby, BC, Canada V5A 1S6
jmanuch@sfu.ca

Ladislav Stacho

Department of Mathematics
Simon Fraser University, 8888 University Drive
Burnaby, BC, Canada V5A 1S6
lstacho@sfu.ca

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 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.

Keywords: 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 [7].

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 [6] for details). This led Dill to introduce the *Hydrophobic-Polar Model* [5]. 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 simplify 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 non-neighbor 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 NP-hard, cf. [4] for two- and [3] 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 for this model. A linear time algorithm with approximation factor $3/8$ for 3D square lattice can be found in [9], and linear time algorithms with approximation factors $6/11$ and $3/5$ for 2D and 3D triangular lattices, respectively, have been developed in [1].

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 [11, 2]. However, major forces of

folding are due to hydrophobic and other *non-local* interactions [6]. To compensate for this unbalance, the existing designs work on 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 [13].

The more general *inverse protein folding problem* involves starting with an arbitrary target native fold and designing an amino acid sequence whose native fold is the target. As this problem is more complex the current research concentrates only on a simple HP model. Early work on this problem involved heuristics that bury the H monomers in a central core with the P monomers on the outside [10], find all possible short sequences and put these together [14], perform a sequence evolution, a form of local search [12]. A relationship between symmetries and designability of proteins was observed in [15]. On the other hand, it was shown in [8] that this version of inverse protein folding problem in the 2D HP model is NP-complete.

In this paper we will consider another modification of the inverse protein folding problem in the HP model. We will assume that a target structure (a connected set of lattice sites) is given, and the goal is to find a protein whose native fold occupies all and only the sites of the target structure (a compatible fold). Note that it is easy to choose targets for which there is no compatible fold. We show that it is possible to closely approximate any given structure in 2D and find a protein with a native fold compatible with this approximation. We will work on a refinement of this lattice in which each square is divided into 9 squares (i.e., a 3×3 refinement of the original lattice). Now almost all of these 9 squares must be occupied by the protein. We call our structures approximating any given structure in the original lattice *constructible structures*.

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. An extreme example 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.

For a number of basic structures, we give a *formal* proof that our proteins are stable in the HP model. Note that we are not aware of any other results explicitly showing stability of a infinite class of proteins (in [12] 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.

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., non-polar, 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 non-polar 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 $E = \{\{[a, b], [a + c, b + d]\}; a, b \in \mathbb{Z} \text{ and } (c, d) = (0, 1), (1, 0)\}$. 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 $F_{p,c} : V \rightarrow \{0, 1\}$ 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 H-H 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 $\text{score}(F_{p,c})$, 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. Formally,

Definition 3 (Native conformations). A conformation c of length $|p|$ is *native for protein p* if for any other conformation c' of length $|p|$, $\text{score}(F_{p,c}) \geq \text{score}(F_{p,c'})$. 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.

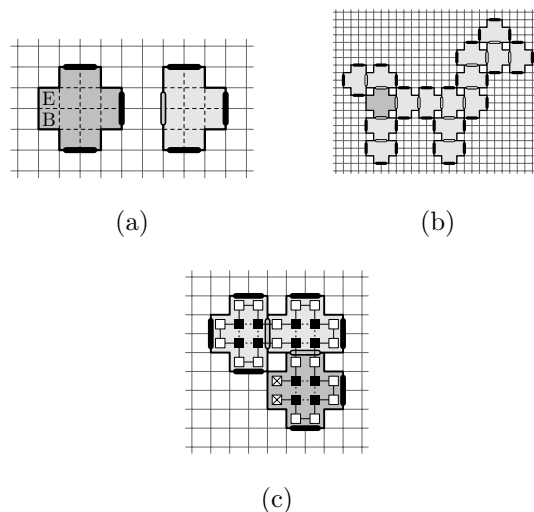


Figure 1. Illustration of: (a) the starting tile (left) and the regular tile (right); (b) a constructible structure; (c) the protein associated with a constructible structure S : $p(S) = 01001001010010010100100100101010$.

Definition 5 (Constructible structures). We have two tiles, depicted in Figure 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 Figure 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 $V(S)$.

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. The second property can be formalized as follows:

Theorem 1. For every constructible structure S , there exists a protein $p(S)$ which has a native fold compatible with S . Furthermore, this fold is saturated.

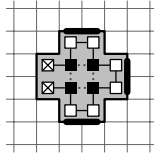


Figure 2. A native conformation of protein $p(S_{\langle \rangle}) = 010010010010$ filling the starting tile.

Due to space limitations we will omit the formal proof of the theorem. However, in the remaining part of the paper we will often refer to the protein $p(S)$ associated with a constructible structure S . Informally, $p(S)$ can be obtained in the following way: Fill four central squares of every tile with 1 and the remaining squares with 0 monomers. There is a unique way of connecting squares of S with a cyclic path. By disconnecting this path between B and E of the starting tile (cf. Figure 1(a)) we obtain a conformation $c(S)$ compatible with S . Reading contents of squares along the path $c(S)$ gives the protein sequence $p(S)$, cf. Figure 1(c).

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:

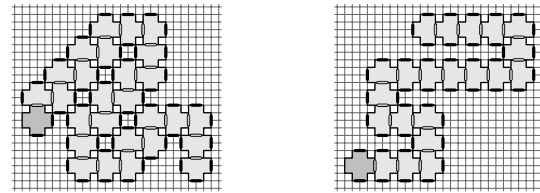
Conjecture 1. For any constructible structure S , the protein $p(S)$ is stable.

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.

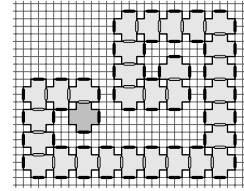
Proof. Since the conformation depicted in Figure 2 is a native conformation for p , by Observation 1, for any native conformation c' for p , the 1-squares form a single 1-cycle of length 4. Now, it is easy to check that there is only one possibility for placing the 0's of the protein in the fold, so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

An extensive computer search shows the conjecture is satisfied for over 15,000 constructible structures including



(a)

(b)



(c)

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

all structures composed of up to 8 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.

Note that a linear constructible structure of length n can be described by a *linear tiling sequence* in $\{1, 2, 3\}^n$ where the number 1 in this tiling sequence means “turn right”, 2 means “continue straight” and 3 means “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 Figure 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:

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

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 $\underbrace{(2, 2, \dots, 2)}_{n-1}$ 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:

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:*

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

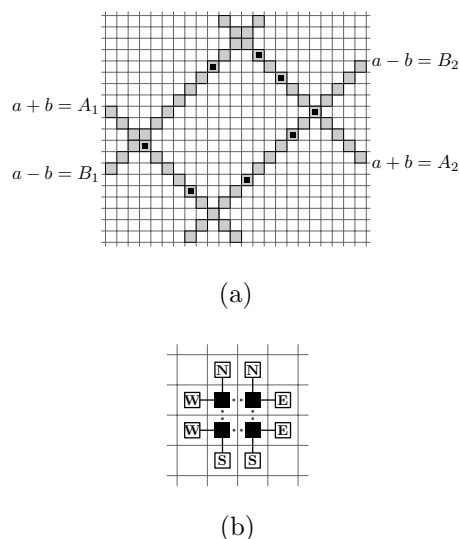


Figure 4. An illustration of (a) a diagonal frame $R(A_1, A_2, B_1, B_2)$, the black squares depict boundary squares; (b) a core.

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. Figure 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 con-

formation 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. Figure 4(b). If the northern (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. \square

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 1. 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 $X \in \{\mathbf{NE}, \mathbf{SE}, \mathbf{SW}, \mathbf{NW}\}$, 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 a 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

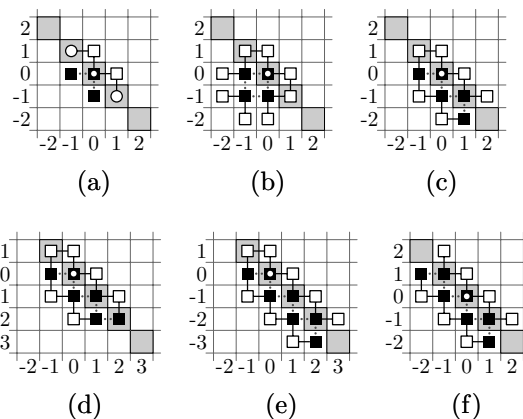


Figure 5. The situation around the boundary square $s = [0, 0]$ (marked with a white dot) which is at distance at least: (a–c) 2 from any terminal; (d–f) 4 from any terminal.

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 non-empty 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 Figure 5(a).

Now we will consider three cases depending on squares $[1, -1]$ and $[-1, 1]$, depicted as empty circles in Figure 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. Figure 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. Figure 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. \square

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 a X -closed core, for every $X \in \{NE, SE, SW, NW\}$.

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. \square

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 1001001 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 1001001001 of the protein, and they are main squares of the same core. Obviously, such a core has to be completely-closed. \square

3.2. \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 $(\underbrace{2, 2, \dots, 2}_{n-1}, \underbrace{3, 2, 2, \dots, 2}_m)$.

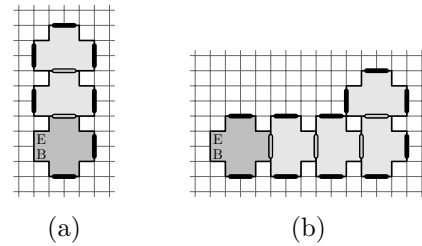


Figure 6. An example of two \mathcal{L}_1 -structures (a) $L_{1,2}$ and (b) $L_{4,1}$.

Observe that if we prove that for all $n, m \geq 1$, 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 $(\underbrace{2, 2, \dots, 2}_{n-1}, \underbrace{1, 2, 2, \dots, 2}_m)$ — 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 6 shows an illustrations of \mathcal{L}_1 -structures $L_{1,2}$ and $L_{4,1}$.

Observe that a constructible structure S is a \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.

Due to space limitations we will omit the rather difficult proof of Theorem 3.

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.

Other interesting problems along these lines are:

- 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;
- find the class of proteins with similar properties for other lattices, namely for the 3D square lattice, and 2D and 3D triangular lattices.

The major obstacle in extending our results to 3D square lattice is the fact that it does not allow saturated folds since the number of neighbors of a site in the lattice 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.

References

- [1] Agarwala, R., Batzoglou, S., Dančák, V., Decatur, S. E., Farach, M., Hannenhalli, S., Skiena, S., Local rules for protein folding on a triangular lattice and generalized hydrophobicity in the HP model, *J. Comp. Biol.* 4:275–296, 1997.
- [2] Andrew, C. D., Penel, S., Jones, G. R., Doig, A. J., Stabilizing nonpolar/polar side-chain interactions in the alpha-helix, *Proteins* 45(4):449–455, 2001.
- [3] Berger, B., Leighton, T., Protein folding in the hydrophobic-hydrophilic (HP) is NP-complete, *J. Comp. Biol.* 5(1):27–40, 1998.
- [4] Crescenzi, P., Goldman, D., Papadimitriou, C., Piccolboni, A., Yannakakis, M., On the complexity of protein folding, *Proc. of STOC'98*, Dallas, 597–603, 1998.
- [5] Dill, K. A., Theory for the folding and stability of globular proteins, *Biochemistry* 24(6):1501–1509, 1985.
- [6] Dill, K. A., Dominant forces in protein folding, *Biochemistry* 29(31):7133–7155, 1990.
- [7] Dill, K. A., Bromberg, S., Yue, K., Fiebig, K. M., Yee, D. P., Thomas, P. D., Chan, H. S., Principles of protein folding: A perspective from simple exact models, *Protein Science* 4:561–602, 1995.
- [8] Hart, W. E., On the computational complexity of sequence design problems, *Proc. of Comp. Molecular Biology*, 128–136, 1997.
- [9] Hart, W. E., Istrail, S., Fast protein folding in the hydrophobic-hydrophilic model within three-eighths of optimal, *Proc. of STOC'95*, 157–168, 1995.
- [10] Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M., Hecht, M.H., Protein design by binary patterning of polar and nonpolar amino acids, *Science* 262:1680–1685, 1993.
- [11] Lyu, P. C., Liff, M. I., Marky, L. A., Kallenbach, N. R., Side chain contribution to the stability of alpha-helical structure in peptides, *Science* 250:669–673, 1990.
- [12] Sun, S., Brem, R., Chan, H. S., Dill, K. A., Designing amino acid sequences to fold with good hydrophobic cores, *Protein Engineering* 8(12):1205–1213, 1995.
- [13] Yu, Y. B., Coiled-coils: stability, specificity, and drug delivery potential, *Advanced Drug Delivery Reviews* 54(8):1113–1129, 2002.
- [14] Yue, K., Dill, K. A., Inverse protein folding problem: Designing polymer sequences, *Proc. Natl. Acad. Sci. USA, Biophysics* 89:4163–4167, 1992.
- [15] Wang, T., Miller, J., Wingreen, N. S., Tang, C., Dill, K. A., Symmetry and designability for lattice protein models, *J. of Chem. Physics* 113(18):8329–8336, 2000.