Few Optimal Foldings of HP Protein Chains on Various Lattices*

Sheung-Hung Poon[†] Shripad Thite[†]

Abstract

We consider whether or not protein chains in the HP model have unique or few optimal foldings. We solve the conjecture proposed by Aichholzer et al. that the open chain $\mathcal{L}_{2k-1} = (HP)^k(PH)^{k-1}$ for $k \geq 3$ has exactly two optimal foldings on the square lattice. We show that some closed and open chains have unique optimal foldings on the hexagonal and triangular lattices, respectively.

1 Introduction

Protein folding is a central and long-standing problem in molecular and computational biology. Due to the complexity of the problem, a variety of simplified models have been proposed to simulate how real proteins fold. In the Hydrophobic-Polar (HP) model, the amino acids in proteins are grouped into two types: *hydrophobic* (H) monomers and *hydrophilic* or *polar* (P) monomers. H monomers tend to attract each other while P monomers are neutral. Proteins are modeled as chains of H and P nodes, or equivalently, strings from $\{H, P\}^+$. The chains are embedded in some lattice in two or three dimensions such that monomers which are adjacent in the given chain must be placed at adjacent points in the lattice. Two non-adjacent nodes on the chain are in *contact* if they occupy a pair of neighboring lattice points. An *optimal folding* of a chain is an embedding in the lattice which maximizes the number of HH contacts.

Much research has been done on the HP model. In particular, Berger and Lieghton [2] showed the NP-completeness of finding the optimal folding on the cubic lattice in 3D, and Crescenzi et al. [3] proved the NP-completeness on the square lattice in 2D. Constant-factor approximation algorithms were also developed for various lattices in both 2D and 3D. We consider the question of whether or not chains in HP model have unique or few optimal foldings. The problem is related to the folding stability of protein chains, and was first suggested by Hayes [4]. Aichholzer et al. [1] exhibited families of closed and open chains in the square lattice, each of which has a unique optimal folding. In this paper, we obtain several results for the square, hexagonal and triangular lattices in two dimensions.

*S.-H.P. was supported by the Netherlands' Organisation for Scientific Research (NWO) under project no. 612.065.307. S.T. was supported by the Netherlands' Organisation for Scientific Research (NWO) under project no. 639.023.301.

[†]Department of Mathematics and Computer Science, TU Eindhoven, 5600 MB, Eindhoven, the Netherlands. {spoon,sthite}@win.tue.nl

2 Open Chain in Square Lattice

Consider the open chain $\mathcal{L}_{2k-1} = (HP)^k(PH)^{k-1}$. In this section, we solve a conjecture proposed by Aichholzer et al. [1] by showing the theorem below.

Theorem 1 *The open chain $\mathcal{L}_{2k-1}$ for $k \geq 3$ has exactly two optimal foldings on the square lattice.*

First, we need the theorem from [1] about unique optimal folding of the closed chain as stated below. See Figure 1 for examples. Note that, in our figures, we use small circles to denote H nodes and small black disks to denote P nodes; we use solid segments to denote chain edges and dashed ones to denote HH contacts.

Theorem 2 *[1] The closed chain $\mathcal{S}_k = P(HP)^{\lceil k/2 \rceil} P(HP)^{\lfloor k/2 \rfloor}$ for $k \geq 1$ has a unique optimal folding on the square lattice.*

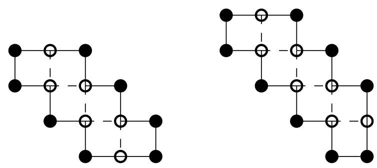

Figure 1: Optimal foldings of $\mathcal{S}_6$ and $\mathcal{S}_7$.

Aichholzer et al. [1] show that Fact 18 to Lemma 29 in their paper hold for the open chain $\mathcal{L}_{2k} = (HP)^k(PH)^k$ for $k \geq 1$. We can verify that these properties are also true for $\mathcal{L}_{2k-1}$. However, for the later lemmas and theorems in their paper, adjustments need to be made to be suitable for the chain $\mathcal{L}_{2k-1}$. The two lemmas below simulate Lemmas 30 and 31 in [1], and their proofs can be adapted with slight modifications. A *straight* node is a node collinear with both its preceding and following nodes on the chain. A *solitary straight H node* v is a straight

H node on the bounding box B of the chain such that both its preceding and following H nodes are not on the same side of B as v.

Lemma 3 *In an optimal folding of $\mathcal{L}_{2k-1}$, there are either one or two solitary straight H nodes on its bounding box B. In particular, if there are exactly two solitary straight H nodes on B, then (see Figure 2(a))*

(i) They lie on opposite sides of B.

(ii) One of them is adjacent to the PP edge, and the other is adjacent to an end edge uv and in contact with an endpoint.

(iii) The PP edge and the end edge uv lie on opposite sides of B.

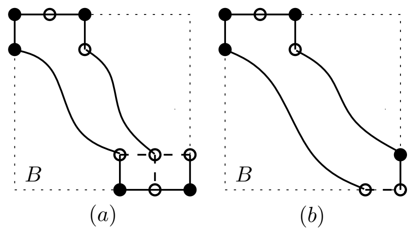

Figure 2: Optimal foldings: (a) when there are two solitary straight H nodes; (b) when there are only one.

Lemma 4 *In an optimal folding of $\mathcal{L}_{2k-1}$, if there is exactly one solitary straight H node on its bounding box B, then (see Figure 2(b))*

(i) The solitary H node is adjacent to the PP edge.

(ii) The solitary H node and the contact of the two endpoints of the chain lie on opposite sides of B.

(iii) The PP edge and an end edge of the chain lie on opposite sides of B.

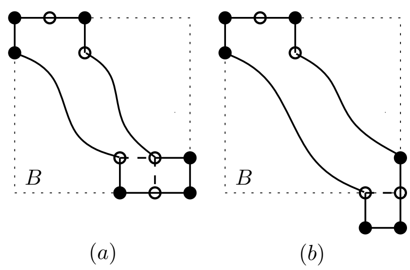

Figure 3: Modify cases (a) and (b) in Figure 2 to closed chains $\mathcal{S}_{2k-2}$ and $\mathcal{S}_{2k-1}$ respectively.

Now we are ready to prove our main theorem.

Proof of Theorem 1. Case (a): If there are exactly two solitary H nodes, by Lemma 3 we can modify the optimal folding of $\mathcal{L}_{2k-1}$ to an optimal folding of $\mathcal{S}_{2k-2}$ by adding a chain edge between the contact of the two end nodes and replacing the end H node on the chain bounding box to a P node. See Figure 3 (a). Thus in this case, the number of optimal folding(s) of $\mathcal{L}_{2k-1}$ is equal to that of $\mathcal{S}_{2k-2}$, which is one by Theorem 2.

Case (b): If there is exactly one solitary H node, by Lemma 4 we can modify the optimal folding of $\mathcal{L}_{2k-1}$ to an optimal folding of $\mathcal{S}_{2k-1}$ by connecting the two end H nodes by a short chain $HPPH$. See Figure 3 (b). Thus in this case, the number of optimal folding(s) of $\mathcal{L}_{2k-1}$ is equal to that of $\mathcal{S}_{2k-1}$, which is one by Theorem 2. □

3 Hexagonal Lattice

3.1 Closed chain

Consider the closed chain $\mathcal{H}_k = (HP)^k PPP(HP)^k PPP$ for $k \geq 1$. We call the two subchains $PPPP$ the two *ends* of $\mathcal{H}_k$. In the above expression of $\mathcal{H}_k$, we denote the ith H node by H_i for $1 \leq i \leq 2k$. We consider the folding $\mathcal{F}_k$, in which each H_i for $1 \leq i \leq k$ is in contact with H_{2k-i+1}. See Figure 4 for an example of folding $\mathcal{F}_k$. We call a contact between an H node and a non-H node a *missing contact*.

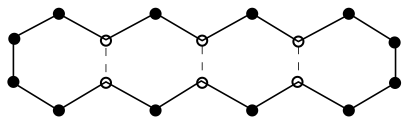

Figure 4: Folding $\mathcal{F}_3$ for $\mathcal{H}_3$.

As in folding $\mathcal{F}_k$, all H nodes are in contact with other H nodes. As there is no missing contact in $\mathcal{F}_k$, there is also none in the optimal folding. Now suppose each H_i for $1 \leq i \leq k$ is in contact with H_{c_i} in the optimal folding. Due to the parity of the positions of H nodes, we have $c_i > k$. We claim that c_i decreases as i increases in the lemma below. After we have the claim, our theorem is immediate.

Lemma 5 *Suppose each H_i for $1 \leq i \leq k$ is in contact with H_{c_i} in the optimal folding. Then c_i decreases as i increases.*

Proof. Suppose to the contrary that there exist $i, i'(i < i')$ such that $c_i < c_{i'}$. Note that H_i (resp. $H_{i'}$) is in contact with H_{c_i} (resp. $H_{c_{i'}}$). Denote the subchain from H_i to $H_{i'}$ (resp. from H_{c_i} to $H_{c_{i'}}$) not containing any end of $\mathcal{H}_k$ by C_1 (resp. C_2). Denote the subchain from H_i to $H_{c_{i'}}$ containing one end of $\mathcal{H}_k$ by E_1. And also denote the subchain from $H_{i'}$ to

H_{c_i} containing another end of $\mathcal{H}_k$ by E_2. See Figure 5 for illustration.

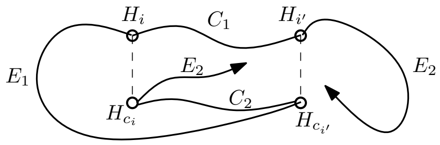

Figure 5: Illustration for the proof of Lemma 5.

Note that there are no chain edges or contacts that can intersect the contact $H_{i'}H_{c_{i'}}$. Consider the cycle $D = C \cup E_1 \cup H_{i'}H_{c_{i'}}$. As H_i is in contact with H_{c_i}, it is not hard to see that H_{c_i} must be in the interior of cycle D. Also it is clear that E_2 lies in the exterior of cycle D. As E_2 connects $H_{i'}$ and H_{c_i}, E_2 must intersect the contact $H_{i'}H_{c_{i'}}$. This is a contradiction. □

Theorem 6 *The closed chain $\mathcal{H}_k$ for $k \geq 1$ has the unique optimal folding $\mathcal{F}_k$ on the hexagonal lattice.*

Proof. By Lemma 5, c_i decreases as i increases from 1 to k in any optimal folding. As all c_i are different and $c_i \in \{k+1, \ldots, 2k\}$, c_i must be $2k - i + 1$. Thus $\mathcal{F}_k$ is the unique optimal folding. □

3.2 Open chain

Consider the open chain $\mathcal{H}'_k = P(HP)^k PPP(HP)^k$ for $k \geq 1$. In the above expression, we denote the ith H node by H_i for $1 \leq i \leq 2k$. We consider the folding $\mathcal{F}'_k$, in which each H_i for $1 \leq i \leq k$ is in contact with H_{2k-i+1}. Notice that $\mathcal{F}'_k$ simulates $\mathcal{F}_k$. See Figure 6 for an example of $\mathcal{F}'_k$.

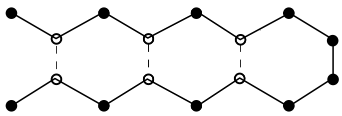

Figure 6: Folding $\mathcal{F}'_3$ for $\mathcal{H}'_3$.

The uniqueness of the optimal folding for $\mathcal{H}'_k$ can be shown by following the similar proof skeleton as Theorem 6, but with slightly more involved arguments.

Theorem 7 *The open chain $\mathcal{H}'_k$ for $k \geq 1$ has the unique optimal folding $\mathcal{F}'_k$ on the hexagonal lattice.*

4 Triangular Lattice

4.1 Closed chain

Consider the closed chain $\mathcal{T}_k = (HP)^k$. We consider its folding $\mathcal{G}_k$ defined as shown in Figure 7.

In this section, we show the following uniqueness theorem. Note that the theorem is not true for $k = 6$.

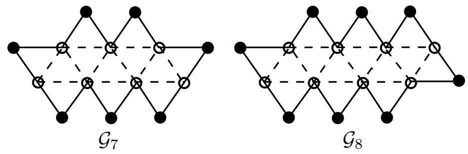

Figure 7: Foldings $\mathcal{G}_7$ & $\mathcal{G}_8$ for $\mathcal{T}_7$ & $\mathcal{T}_8$ respectively.

Theorem 8 *The closed triangular chain $\mathcal{T}_k$ for $k \geq 2$ and $k \neq 6$ has the unique optimal folding $\mathcal{G}_k$ on the triangular lattice.*

When k is small, we can show the uniqueness of the optimal folding by enumerating the configurations of the HH-contact graph with maximum number of contacts.

Lemma 9 *The chain $\mathcal{T}_k$ for $2 \leq k \leq 5$ or $k = 7$ has the unique optimal folding $\mathcal{G}_k$. The chain $\mathcal{T}_6$ has two optimal foldings including $\mathcal{G}_k$ as shown in Figure 8.*

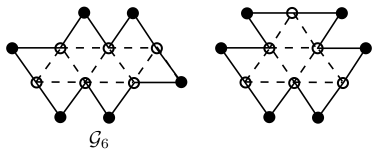

Figure 8: Two optimal foldings of $\mathcal{T}_6$.

It remains to show the uniqueness of the optimal folding of long chains as stated in the following main lemma.

Lemma 10 *The chain $\mathcal{T}_k$ for $k \geq 8$ has the unique optimal folding $\mathcal{G}_k$.*

As there are six missing contacts in $\mathcal{G}_k$, we observe that an optimal folding has at most six missing contacts.

We call an H node *fully-contacted* if there is no missing contact from it. The optimal folding of $\mathcal{T}_k$ for $k \geq 8$ contains at least two fully-contacted H nodes due to the above observation. By careful examination of the neighborhoods of the two H nodes, we can show that there must be a pair of contacting H nodes that are both fully-contacted and non-straight.

Lemma 11 *An optimal folding of $\mathcal{T}_k$ for $k \geq 8$ contains two fully-contacted non-straight H nodes in contact with each other.*

Using the above lemma, we can divide the whole chain at a pair of contacting H nodes into two "quite-long" paths.

Lemma 12 *An optimal folding of $\mathcal{T}_k$ for $k \geq 8$ contains two non-straight contacting H nodes such that they divide $\mathcal{T}_k$ into two paths, each of which contains at least two internal H nodes.*

We define a *U-line* (resp. *D-line*) as a line of slope $\sqrt{3}$ (resp. $-\sqrt{3}$). We define a *canonical line* of the triangular lattice as a horizontal line, a U-line, or a D-line. A *canonical strip* of a lattice edge e in the triangular lattice is a strip between the two parallel canonical lines, each of which passes through exactly one endpoint of e. Note that each lattice edge has exactly two canonical strips.

Lemma 13 *Suppose $\mathcal{C}$ is a path along $\mathcal{T}_k$ connecting a pair of contacting H nodes such that $\mathcal{C}$ contains either a non-straight internal H node or two internal H nodes. Then there are at least three missing contacts from internal H nodes of $\mathcal{C}$.*

Proof. (*Sketch*) Suppose X is a canonical strip of the contacting edge e between the pair of ending H nodes such that the two end edges of $\mathcal{C}$ are separated by X. Without loss of generality, we assume that X runs horizontally, the contact edge e between the two end H nodes of $\mathcal{C}$ lies on a U-line, and $\mathcal{C}$ crosses X to the right of e in an odd number of times. See Figure 9 for illustration. Let H_a, H_b be the upper and lower ends of e respectively.

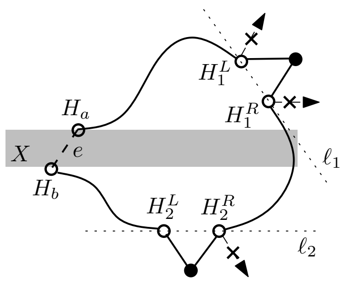

Figure 9: Illustration for the proof of Lemma 13.

Sweep a D-line to the right until it reaches some extremal H node of $\mathcal{C}$. We call the D-line at current position ℓ_1. Let H_1^L and H_1^R be the leftmost and rightmost H nodes on ℓ_1 respectively. We define ℓ_2, H_2^L, H_2^R similarly by sweeping a horizontal line downwards.

It is clear that the right-contact of H_1^R and the bottom-right-contact of H_2^R are both missing. With the given conditions, it is easy to show that $H_1^L = H_a$ and $H_2^L = H_b$ cannot both be true. Without loss of generality, we assume that the former is not true. Then we have that the top-right-contact of H_1^L is also missing. $\square$

Now by an involved analysis, we can show that in order for each of these two paths to contain exactly three missing contacts, it must possess the pattern as shown in Figure 10 (a) or (b). The details are omitted in this abstract. With this property, it is immediate to claim our main lemma, Lemma 10, and we finish the proof of Theorem 8.

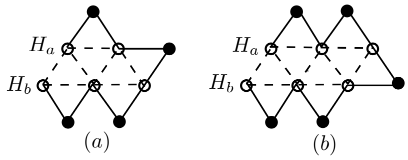

Figure 10: Patterns in an optimal folding.

4.2 Open chain

However, the open chain $\mathcal{T}'_k = (HP)^{k-1}H$ can have several optimal foldings on the triangular lattice. Instead, we show the following theorem for the open chain $\mathcal{T}''_k = (HP)^k(PHP)^2(PH)^k$ for $k \geq 3$ by using the similar technique we use for the closed chain $\mathcal{T}_k$, but with a more involved analysis. See Figure 11 for an example of the unique optimal folding.

Theorem 14 *The open chain $\mathcal{T}''_k$ for $k \geq 3$ has a unique optimal folding on the triangular lattice.*

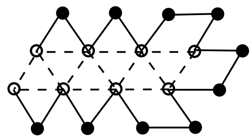

Figure 11: The unique optimal folding of $\mathcal{T}''_3$.

5 Conclusion & Discussion

We solve a conjecture about an open chain in the square lattice. We obtain unique optimal foldings for chains in the hexagonal and triangular lattices, respectively. All of our results are in two dimensions. Is there any family of chains that have unique optimal foldings on some lattice in three dimensions?

References

[1] O. Aichholzer, D. Bremner, E. Demaine, H. Meijer, V. Sacristan, and M. Soss. Long proteins with unique optimal foldings in the H-P model. *Computational Geometry: Theory and Applications*, 25(1-2), 139–159, 2003.

[2] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5(1), 27–40, 1998.

[3] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5(3), 423–466, 1998.

[4] B. Hayes. Prototeins. *American Scientist*, 86, 216–221, 1998.