Long Proteins with Unique Optimal Foldings in the H-P Model

Oswin Aichholzer* David Bremner† Erik D. Demaine‡ Henk Meijer§

Vera Sacristán¶ Michael Soss‖

Abstract

We explore a problem suggested by Brian Hayes in 1998: what proteins in the two-dimensional hydrophilic-hydrophobic (H-P) model have *unique* optimal foldings? In particular, we prove that there are closed chains of monomers (amino acids) with this property for all (even) lengths; and that there are open monomer chains with this property for all lengths divisible by four. Along the way, we prove and conjecture several structural results about bonds in the H-P model.

1 Introduction

Protein folding is a central problem in bioinformatics with the potential to reveal an understanding of the function and behavior of proteins, the building blocks of life. Such an understanding would greatly influence many areas in biology and medicine such as drug design.

One of the most popular models of protein folding is the hydrophilic-hydrophobic (H-P) model [2, 4, 6], which defines both a geometry and a quality metric of foldings. This combinatorial model is attractive in its simplicity, and already seems to capture several essential features of protein folding such as the tendency for the hydrophobic components to fold to the center of a globular protein [2]. While the H-P model is most intuitively defined in 3-D to match the physical world, in fact it is more realistic as a 2-D model for computationally feasible sizes. The basic reason for this is that the perimeter-to-area ratio of a short 2-D chain is a close approximation to the surface-to-volume ratio of a long 3-D chain [2, 6].

Much work has been done on the H-P model. Recently, on the computational side, Berger and Leighton [1] proved NP-completeness of finding the optimal folding in 3-D, and Crescenzi et al. [3] proved NP-completeness in 2-D. Hart and Istrail [5] have developed a 3/8-approximation for protein folding in 3-D and a 1/4-approximation for protein folding in 2-D.

In this paper we study several structural aspects of the 2-D H-P model. In particular we explore a problem suggested by Brian Hayes [6] about the existence of stable protein foldings

*Institut für Grundlagen der Informationsverarbeitung, Technische Universität Graz, Inffeldgasse 16b, A-8010 Graz, Austria, email: `oaich@igi.tu-graz.ac.at`. Supported by the Austrian Programme for Advanced Research and Technology (APART).

†Faculty of Computer Science, University of New Brunswick, P. O. Box 4400, Fredericton, N. B. E3B 5A3, Canada, email: `bremner@unb.ca`

‡Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada, email: `eddemaine@uwaterloo.ca`.

§Department of Computing and Information Science, Queen's University, Kingston, Ontario K7L 3N6, Canada, email: `henk@cs.queensu.ca`.

¶Departament de Matemàtica Aplicada II, Universitat Politècnica de Catalunya, Pau Gargallo 5, 08028 Barcelona, Spain, email: `vera@ma2.upc.es`. Supported by DURSI Gen. Cat. 1999SGR00356 and Proyecto DGES-MEC PB98-0933.

‖School of Computer Science, McGill University, 3480 University Street, Montréal, Québec H3A 2A7, Canada, email: `soss@cs.mcgill.ca`.

of all lengths. We solve this problem in a positive sense for circular protein strands. We also nearly solve the problem for open strands by exhibiting an infinite class of proteins with unique optimal foldings.

More precisely, we prove the following main results, in a sense establishing the existence of stable protein foldings:

1. We exhibit a simple family of closed chains of monomers, one for every possible (even) length, and prove that each chain has a unique optimal folding according to the H-P model.

2. We exhibit a related family of open chains of monomers, one for every length divisible by 4, with the same uniquely-foldable property. Note that a result as strong as (1) cannot be obtained for open chains, because there are some lengths for which no uniquely foldable open chains exist.

We also make an interesting conjecture about the structure of optimal protein foldings in the H-P model; see Section 3.1. Finally, we present experimental results supporting another conjecture, strengthing the second result described above to all even lengths.

2 H-P Model

In this section we review the H-P model and the very basics of the biology behind it.

Proteins are chains of monomers, each monomer one of the 20 naturally occurring amino acids. In the H-P model, only two types of monomers are distinguished: hydrophobic (H), which tend to bundle together to avoid surrouding water, and hydrophilic (P), which are attracted to water and are frequently found on the surface of a folding [2]. In our figures we use small gray disks to denote H monomers and black disks to denote P monomers. These monomers are strung together in some combination to form an *H-P chain*, either an open chain (path or arc) or a closed chain (cycle or polygon).

Proteins are folded onto the regular square lattice. More formally, a *lattice embedding* of a graph is a placement of vertices on distinct points of the (regular square) lattice such that each edge of the graph maps to two adjacent (unit-distance) points on the lattice. In the H-P model, proteins must fold according to lattice embeddings, so we also call such embeddings *foldings*.

The quality of a folding in the H-P model is simply given by the number of hydrophobic monomers (light-gray H nodes) that are not adjacent in the protein but adjacent in the folding. More formally, the *bond graph* of a folding has the same vertex set as the chain, and there is an edge between every two H vertices that are adjacent in the folding onto the lattice, but not adjacent along the chain. The edges of the bond graph are called *bonds*; in our figures, bonds are drawn as light-gray edges.

An *optimal* folding maximizes the number of bonds over all foldings. Intuitively, if a protein is folded to bond together many hydrophobic monomers (H nodes), then those monomers are hidden from the surrounding water as much as possible.

There is a natural bijection between strings in $\{\mathsf{H}, \mathsf{P}\}^*$ and protein chains. We consider the nodes in a chain as labeled by their order in the string. We sometimes use a limited form of regular expressions to describe chains where e.g. H^k indicates k H nodes in sequence. Similarly, if we walk along an embedded chain in the order given and read off the direction of each edge, we can encode foldings as strings in $\{\mathsf{E}, \mathsf{W}, \mathsf{N}, \mathsf{S}\}^*$.

3 General Observations and Ambiguous Foldings

In this section we describe some basic structural and combinatorial results about bonds in the H-P model.

Fact 3.1 *A folding of an open (closed) chain with h* H *nodes has at most $h+1$ (h) bonds.*

Fact 3.2 *Any lattice-embeddable graph is bipartite.*

Corollary 3.3 *If a folding of a closed chain with h* H *nodes has h bonds, then its bond graph is a union of vertex-disjoint even cycles.*

Corollary 3.4 *There can be a bond between two* H *nodes only if they have opposite parity (i.e., there is an even number of nodes between them) in the chain.*

Fact 3.5 *Any optimal folding of the closed chain* $(\mathsf{PHP})^{4k}$ *has a bond graph consisting of k 4-cycles. (See Figure 1.)*

Fact 3.6 *For any n, there exists an n-node closed chain with at least $2^{\Omega(n)}$ optimal foldings, all with isomorphic bond graphs. In addition to the trivial example of P^n, the chain* $(\mathsf{PHP})^{4k}$ *is an example(see Figure 1).*

3.1 Internal and External Bonds

For any closed chain, we can consider the area of a folding, and whether a bond is internal or external to the folding (polygon).

Conjecture 3.7 *If there is a lattice point strictly interior to a folding of a closed chain, then there is another folding of that chain with smaller area and no fewer bonds.*

A proof of this conjecture would support the experimental evidence that most proteins fold into a tightly packed "globule" [2].

Fact 3.8 *For any folding of an n-node closed chain there are at most $\frac{n-4}{2}$ internal bonds.*

Corollary 3.9 *There exists a closed chain whose optimal folding requires at least one external bond. Namely,* H^{12} *is an example. (See Figure 3.)*

Fact 3.10 *Any folding of an n-node closed chain has at most $\frac{n-p}{2}$ external bonds, where p is the perimeter of the bounding box of the folding.*

Let $\rho(k)$ denote the minimum perimeter of a rectangle containing k lattice points.

Corollary 3.11 *In any folding of an n-node closed chain, there are at most $\frac{n-\rho(n)}{2}$ external bonds and at most $\frac{2n-\rho(n)-4}{2}$ total bonds.*

Fact 3.12 $\rho(k) = 4(\lfloor\sqrt{k}\rfloor - 1) + 2(\lceil k/\lfloor\sqrt{k}\rfloor^2\rceil - 1)$.

4 Uniquely Foldable Closed Chains

In this section we are concerned with closed H-P chains whose optimal foldings are unique (modulo isometries).

For each $k \geq 1$, we define a closed chain S_k as follows; see Figure 2. Let A_m denote the sequence $(\mathsf{HP})^m$. Define $u = \lceil k/2 \rceil$ and $d = \lfloor k/2 \rfloor$. Then define S_k as $\mathsf{P}\,A_u\,\mathsf{P}\,A_d$.

We also define a folding F_k of S_k as follows; see Figure 2. Let D_m (a "down staircase") denote the alternating path $(\mathsf{ES})^m$. Let U_m (an "up staircase") denote the alternating path $(\mathsf{WN})^m$. If k is even, define F_k as $\mathsf{E}\,D_d\,\mathsf{W}\,U_u$. If k is odd, define F_k as $\mathsf{E}\,D_d\,\mathsf{S}\,U_u$.

Theorem 4.1 *For each $k \geq 1$, F_k is the unique optimal folding of S_k.*

5 Uniquely Foldable Open Chains

Hayes [6] has established experimentally that for each $1 \leq n \leq 14$ except 3 and 5 there is an open chain with a unique optimal folding. We have extended Hayes's results as shown in Figure 4. Figure 5 shows the unique foldings for these chains.

A natural question is for what values of n there is an n-node open chain with a unique optimal folding. Based on our results about closed chains, one approach is to consider the open version of S_k with the first and last nodes removed. That is, define $Z_k = (\mathsf{HP})^u(\mathsf{PH})^d$ where $u = \lceil k/2 \rceil$ and $d = \lfloor k/2 \rfloor$. It turns out that this chain has multiple optimal folding for odd k, but only one optimal folding for even k:

Theorem 5.1 *The open chain $Z_{2j} = (\mathsf{HP})^j(\mathsf{PH})^j$ has a unique optimal embedding for each positive j. (See Figure 6 for an example.)*

Combining this theorem with results from [6] and Figure 4, we know that there are open chains with unique optimal foldings for $n = 2$, $n = 4$, and $6 \leq n \leq 20$.

Acknowledgments. The authors would like to thank Godfried Toussaint for pointing us to this topic and for organizing the workshop at which the research was initiated. The authors would also like to thank Vida Dujmović, Jeff Erickson, Ferran Hurtado, and Suneeta Ramaswami for stimulating conversations on this and other topics.

References

[1] 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.

[2] H. S. Chan and K. A. Dill. The protein folding problem. *Physics Today*, pages 24–32, February 1993.

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

[4] K. A. Dill. Dominant forces in protein folding. *Biochemistry*, 29(31):7133–7155, August 1990.

[5] W. E. Hart and S. Istrail. Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal. In *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*, pages 157–168, Las Vegas, Nevada, May–June 1995.

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

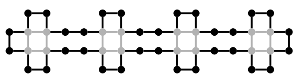


Figure 1: Example of an optimal folding of $(\mathsf{PHP})^{4k}$.

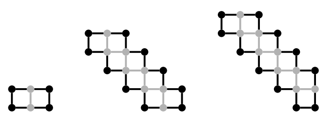


Figure 2: Examples of S_k folded according to F_k for $k \in \{2, 8, 9\}$.

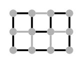
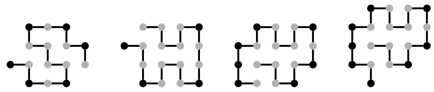


Figure 3: A folding of H^{12} with 5 bonds.

Len.	Chain	Max. Bonds
15	$(\mathsf{PH})^3\mathsf{H}^3(\mathsf{PH})^3$	7
17	$\mathsf{H}^5(\mathsf{PHHP})\mathsf{H}^4(\mathsf{PHHP})$	9
18	$\mathsf{H}^3(\mathsf{PH})^3\mathsf{H}^3\mathsf{PHPPPH}$	9
19	$\mathsf{H}^3(\mathsf{PH})^3\mathsf{H}^3\mathsf{PHPPPHP}$	9

Figure 4: Experimentally computed open chains with unique optimal foldings.



15 17 18 19

Figure 5: Unique foldings for the chains in Table 4.

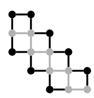


Figure 6: Unique optimal folding of Z_8.