Helix and Sheet: Structures With Unique Optimal Foldings in The H-P Model

Gerrit Leder[1]

[1] **Institute for Neuro- and Bioinformatics, University of Lübeck**

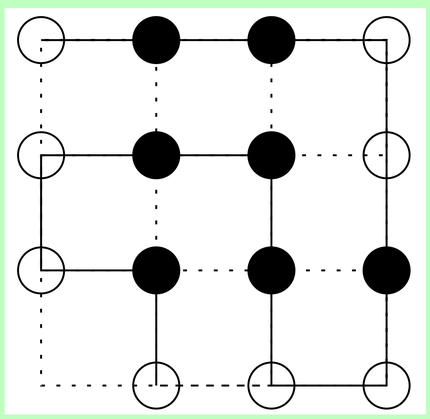
Ratzeburger Allee 160, House 64, D–23538 Lübeck, Germany

Abstract

According to protein structures α-helix and β-sheet we will introduce H-P (hydrophobic-polar) sequences that arrange in the two-dimensional grid as spiral and fold, respectively. We will state two generation rules (according to the form of regular like expressions), that generate such 'helix' and 'sheet' sequences. We proof, that the arrangement in the two-dimensional grid for short lengths is 'unique', except for rotation and symmetry (which is achieved by rotating in a third dimension). I. e. the generated sequences have one and only one optimal folding in the two-dimensional grid (cp. Hayes [Hay98]). In addition we have the strong conjecture that all (long) 'sheet' sequences have a unique optimal folding. These sequences may be relevant for naturally occuring proteins as well as polypeptide chains manufactured in vitro. Furthermore the short spiral and fold of lengths eleven and twelve have the same unique optimal folding, except for the last monomer. This corresponds to Erdmann [Erd04], who states, that "there are even examples of proteins with nearly identical primary sequence in which α-helices have become β-sheets".

1. A mathematical hydrophobic-polar model

According to Crescenzi et al. [CGP+98] we present the following. We model a plane, with horizontal and vertical lines of cavities, in which black and white beads can be placed. The beads are connected through a string, in such a way, that neighbours are always placed horizontally or vertically to one another (diagonal neighbourhood of two beads placed next to one another on the band, as well as more than one bead per cavity are not possible). These threaded beads represent an amino acid sequence with hydrophobic (black beads or '1') and hydrophilic/polar (white beads or '0') monomers. Cp. next figure.



A *two-dimensional grid* is a graph $(\mathbb{Z}^2, L)$ with the set of nodes $\mathbb{Z}^2 = \mathbb{Z} \times \mathbb{Z}$ (points of the grid) and the edges, that run between all horizontally and vertically adjacent points in the grid and belong to the set of edges $L = \{\{(x,y),(x',y')\} : |x-x'|+|y-y'| = 1\}$.
For a set $S = \{s_1, s_2, \ldots, s_m\}$ of 0-1 sequences $s_1, s_2, \ldots, s_m \in \{0,1\}^*$ the following mapping f from S into the two-dimensional grid is a *fold* or *walk*:

$$f : \{(i,j) | 1 \leq i \leq m, 1 \leq j \leq |s_i|\} \to \mathbb{Z}^2$$

and it is necessary $\forall i, 1 \leq i \leq m : \forall j, 1 \leq j \leq |s_i| - 1$:

$$\{f(i,j), f(i,j+1)\} \in L.$$

This is a fold of neighbours in the 0-1 sequence s_i to neighbours in the two-dimensional grid.

We define a *score* called *loss* E of a fold of the 0-1 sequence in S:
Neighbours in the grid $\{(x,y),(x',y')\} \in L$ are a *loss*, if

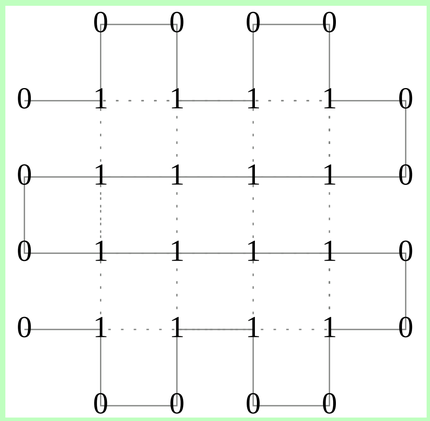
1. they represent no neighbours in the same 0-1 sequence and
2. exactly one of the two points is a 1.

2. Helix and sheet structures

Below we present two examples for sequences in the H-P model, that fold into the two-dimensional grid and form a helix or a sheet. These folds are similar to structures in proteins called α-helix and β-sheet, respectively. We now state a closed formula, similar to regular expressions, for a walk with losses $E = 0$. In this connection, for a string $x \in \{0,1\}^*$, $(x)^i$ means i string(s) x in succession and $(x)^*$ means zero or arbitrarily, but finite, repetitions of string x. The formula of the so called sheet sequence with $i \in \mathbb{N}_0$ is:

$$0(1001)^i 0(0((11)^i 0)^* 0(1001)^i 0 \qquad (1)$$

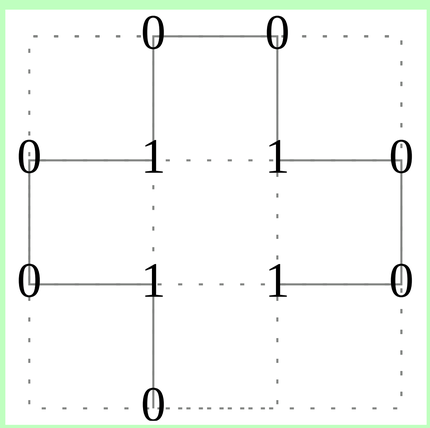
The folding of this sequence (s. formula (1) with $i = 2$) in the two-dimensional grid is shown in the next figure.



We now state the corresponding formula, similar to regular expressions again, for a walk with $E = 1$ loss. The formula is constructed as the one above. The formula of the so called helix sequence with $m \in \mathbb{N}_{\geq 1}$ is:

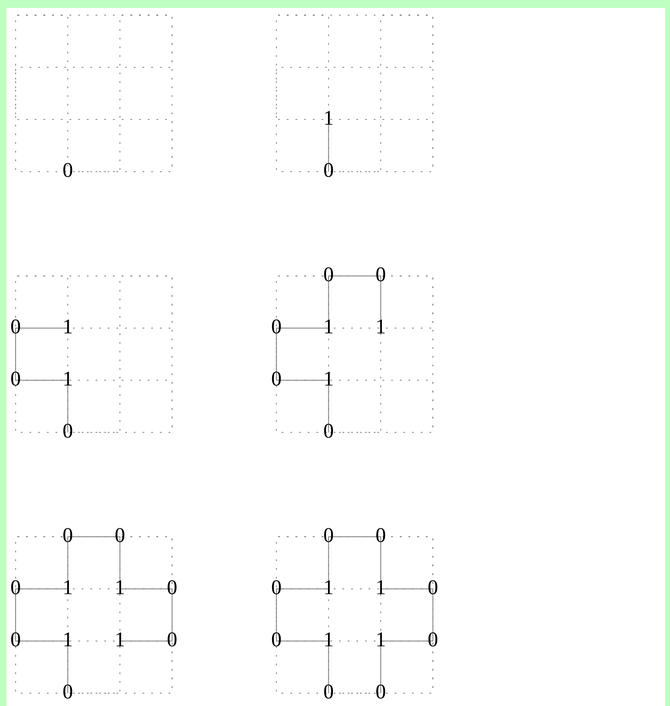
$$(1)^{4m^2-8m+4}(1001)^{(m-1)}(10)(0110)^{(m-1)}(010)(0110)^{(m-1)}(010)(0110)^{(m-1)}(010) \qquad (2)$$

The folding of this sequence (s. formula (2)) in the two-dimensional grid is shown in the next figure for $m = 1$.

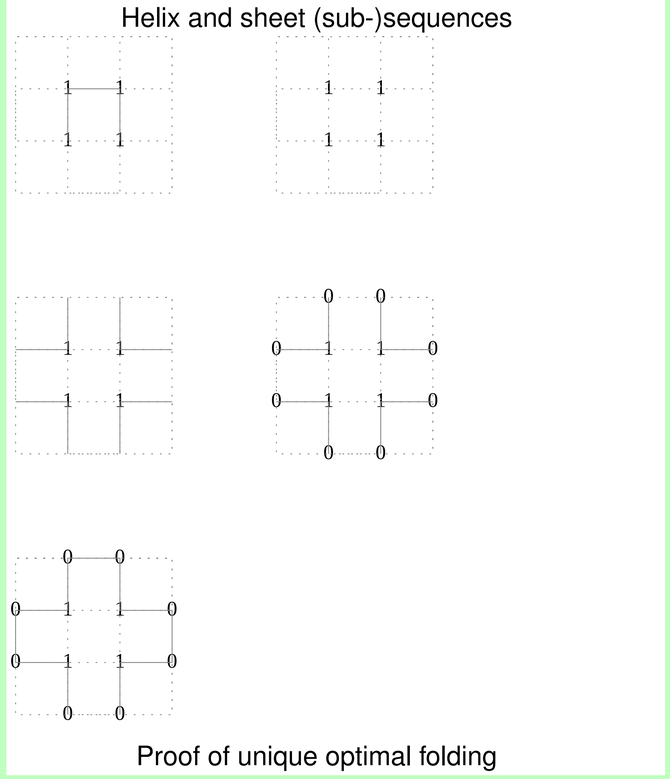


3. Unique optimal foldings

We show unique optimal foldings for short H-P sequences. The sequences are generated by formulas (1) and (2) for the lengths of twelve elements (sheet) and eleven elements (helix), respectively. See the following figures for partial sequences up to the length of twelve and the unique optimal folding of the sheet sequence.



Helix and sheet (sub-)sequences



Proof of unique optimal folding

Future work

- ***All (long) sheet sequences have unique optimal folding***
- ***Validate experimentally the relevance of sequences by generating polymers in vitro***
- ***Compute unique optimal foldings for (sub-)sequences and determine distances by sequence alignment***

References

[CGP+98] P. Crescenzi, D. Goldman, C. H. Papadimitriou, A. Piccolboni, and M. Yannakakis. On the complexity of protein folding. *J. of Comp. Biology*, 5(3):423–465, 1998. NP-completeness of the two-dimensional H-P model.

[Erd04] Michael A. Erdmann. Protein similarity from knot theory and geometric convolution. In *Proceedings of the eighth international conference on computational molecular biology*, pages 195–204. ACM, 2004. Abstract cited.

[Hay98] Brian Hayes. Prototeins. *American Scientist*, 86(3):216, May 1998. Sequences in the two-dimensional H-P model with unique optimal folding.