RAIRO-Theor. Inf. Appl. **41** (2007) 375–402
DOI: 10.1051/ita:2007023

A WEIGHTED HP MODEL FOR PROTEIN FOLDING WITH DIAGONAL CONTACTS*

Hans-Joachim Böckenhauer[1] and Dirk Bongartz[2]

Abstract. The HP model is one of the most popular discretized models for attacking the protein folding problem, *i.e.*, for the computational prediction of the tertiary structure of a protein from its amino acid sequence. It is based on the assumption that interactions between hydrophobic amino acids are the main force in the folding process. Therefore, it distinguishes between polar and hydrophobic amino acids only and tries to embed the amino acid sequence into a two- or three-dimensional grid lattice such as to maximize the number of contacts, *i.e.*, of pairs of hydrophobic amino acids that are embedded into neighboring positions of the grid.

In this paper, we propose a new generalization of the HP model which overcomes one of the major drawbacks of the original HP model, namely the bipartiteness of the underlying grid structure which severely restricts the set of possible contacts. Moreover, we introduce the (biologically well-motivated) concept of weighted contacts, where each contact gets assigned a weight depending on the spatial distance between the embedded amino acids. We analyze the applicability of existing approximation algorithms for the original HP model to our new setting and design a new approximation algorithm for this generalized model.

Mathematics Subject Classification. 68W25, 92C40.

Keywords and phrases. Protein folding, HP model, approximation algorithms.

* *This work was partially supported by SNF grant* 200021-109252/1.

[1] Department of Computer Science, ETH Zurich, Switzerland; hjb@inf.ethz.ch

[2] Gymnasium St. Wolfhelm, Schwalmtal, Germany; bongartz@gym-st-wolfhelm.de

1. Introduction

To determine the three-dimensional structure of proteins is one of the most important and challenging problems in computational biology with many applications, *e.g.* in the area of pharmaceutics and drug design.

Proteins are chains of smaller molecular entities, the so-called amino acids, they are expected to fold in space according to the chemical characteristics of these amino acids [2, 3]. In particular, one distinguishes between hydrophobic and hydrophilic (polar) amino acids, and interactions between hydrophobic amino acids are assumed to be the driving force in the folding of proteins.

To study the process of protein folding incorporating these hydrophobic forces, Dill *et al.* [7, 8] introduced the HP model. In this model, one searches for an embedding of a protein, given in terms of a string over 0 and 1 (encoding for polar and hydrophobic amino acids, respectively), into a grid. Counting the number of ones placed on adjacent positions (that are not already adjacent in the string) in the grid, gives a measure of the stability of the folding. This setting was algorithmically analyzed by Hart and Istrail [9], who proposed a 4-approximation for this problem, which was further improved by Newman [10] obtaining a 3-approximation for two-dimensional grids. Further results on this topic include other lattice types, heuristic algorithms, or extended models. An overview of the existing literature is for instance given in the survey paper [5].

A major drawback of the original HP model is the bipartiteness of the underlying grid. One possible approach for overcoming this drawback was proposed in [1], where the embedding of protein sequences into a triangular grid lattice was studied.

Another approach was taken in [4], where the authors studied the HP problem on a rectangular grid with additional diagonals. While this removes the bipartiteness of the underlying grid, it introduces a new weakness to the model. In particular, 45°, 90°, 135°, and 180° become feasible angles between two consecutive amino acids in this model. Even though the angles in every previously considered underlying lattice for the HP problem do not perfectly fit the chemical reality, especially the 45° angles seem to be problematic.

On the other hand, binding forces are not restricted to possible folding angles of the molecule. Therefore, to avoid these sharp angles and the bipartiteness as well, we will now introduce a kind of *two-level lattice*, where the folding is restricted to the grid of the original HP model, but the binding forces are also effective along the diagonals in this grid structure.

Moreover, we want to account for the different distance between two embedded amino acids along a diagonal edge or along a horizontal/vertical edge of the grid. More formally, for a given folding, we will evaluate this folding by weighting each adjacency of two hydrophobic amino acids along a horizontal or vertical edge of the underlying grid lattice by 1 and each adjacency of two hydrophobic amino acids along a diagonal edge by a parameter α, where $0 \leq \alpha \leq 1$. We refer to the resulting model as the α-DC-HP$_{2d}$ model in the following. The resulting maximization problem is studied in this paper.

In the sequel we will study the α-DC-HP$_{2d}$ problem in more detail. We will first give some basic notions and observations in Section 2 and present some upper bounds on the overall contact weight that might be achieved by any conformation in Section 3. After that, we turn our attention to approximation algorithms and investigate the performance of a classical algorithm for the HP model in the context of this model and of a specially designed algorithm in Sections 4.2 and 4.3, respectively. We conclude the paper by a discussion of the proposed algorithms and the presentation of some open problems in Section 5.

2. Preliminary notions and observations

Before we start with the formal description of the problem, we first define the two underlying lattice types.

Definition 2.1. The *two-dimensional grid lattice* is the infinite graph $\mathcal{L}^{\square} = (V, E)$ with vertex set $V = \mathbb{Z}^2$ and edge set $E = \{\{x, x'\} \mid x, x' \in \mathbb{Z}^2, |x - x'| = 1\}$.

The *two-dimensional grid lattice with diagonals* is the infinite graph $\mathcal{L}^{\boxtimes} = (V, E)$ with vertex set $V = \mathbb{Z}^2$ and edge set $E = \{\{x, x'\} \mid x, x' \in \mathbb{Z}^2, |x-x'| \leq \sqrt{2}\}$.

We consider the input of our problem, namely a chain of amino acids, as a string over $\{0, 1\}$ encoding the hydrophobicity of each amino acid. We will use a 1 to denote a hydrophobic amino acid and a 0 to denote a polar (hydrophilic) amino acid.

Definition 2.2. Let $p = p_1 \dots p_m$ be a string of length m over the alphabet $\{0, 1\}$, and let $\mathcal{L} = (V, E) \in \{\mathcal{L}^{\boxtimes}, \mathcal{L}^{\square}\}$ be a lattice. A *conformation* of p into $\mathcal{L}$ is an injective function $\varphi : \{1, \dots, m\} \to V$ from the positions of the string to the vertices of the lattice that assigns adjacent positions in p to adjacent vertices in $\mathcal{L}$, *i.e.*, $\{\varphi(i), \varphi(i + 1)\} \in E$ for all $1 \leq i \leq m - 1$. These edges $\{\varphi(i), \varphi(i + 1)\} \in E$ for $1 \leq i \leq m - 1$ are called *binding edges*.

An edge $\{x, x'\}$ of $\mathcal{L}$ is called a *contact edge*, if it is no binding edge, but there exist $i, j \in \{1, \dots, m\}$ such that $\varphi(i) = x$, $\varphi(j) = x'$, and $p_i = p_j = 1$.

Whenever a conformation of a string is given, we will call a vertex of the lattice to which there was assigned a one [zero] by this conformation a *one-vertex* [*zero-vertex*] or simply a *one* [*zero*]. The vertices of the lattice which are not occupied by a one or a zero are called *unused*. A binding edge connecting a one with a zero will be called an *alternation edge* and a non-binding edge adjacent to a one that is no contact edge is called a *loss edge*.

A further refinement of our model concerns the intensity of the chemical forces along the different types of edges in $\mathcal{L}^{\boxtimes}$. As the adjacency of two amino acids *via* a diagonal edge implies a greater distance and thus a smaller chemical binding force as between two amino acids which are connected by a horizontal/vertical edge, we introduce an additional parameter α ($0 \leq \alpha \leq 1$) to measure this loss of binding power relatively to the binding power given by horizontal/vertical edges. In particular, we will count a contact weight of 1 for each horizontal/vertical contact edge and a contact weight of α for each diagonal contact edge.

Now, the resulting problem is defined as follows.

Definition 2.3. The protein folding problem in the 2-dimensional α-DC-HP$_{2d}$ model, denoted as *α-DC-HP$_{2d}$ problem*, is the following optimization problem:

Input: A string $p = p_1 \dots p_m$ over the alphabet $\{0, 1\}$ and a parameter α with $0 \leq \alpha \leq 1$.

Feasible solutions: All conformations of p in $\mathcal{L}^{\square}$.

Costs: The cost of a conformation φ is the overall contact weight of all contact edges of φ in $\mathcal{L}^{\boxtimes}$, *i.e.*,

$$\text{cost}(\varphi) = \sum_{\substack{\text{contact edges} \\ \text{in } \mathcal{L}^{\square}}} 1 \quad + \sum_{\substack{\text{contact edges} \\ \text{in } \mathcal{L}^{\boxtimes} \setminus \mathcal{L}^{\square}}} \alpha.$$

Goal: Maximization.

Note that the conformation is restricted to be an embedding in $\mathcal{L}^{\square}$, but the overall contact weight is computed with respect to $\mathcal{L}^{\boxtimes}$, *i.e.*, we also allow for diagonal contacts.

Using this definition, we have removed the possibility of sharp 45° angles by restricting the conformation to $\mathcal{L}^{\square}$ which is the lattice also used in the original HP model, and furthermore we got around the weakness of bipartiteness of $\mathcal{L}^{\square}$ by counting contacts in the $\mathcal{L}^{\boxtimes}$ lattice.

Please note that, although we defined the α-DC-HP$_{2d}$ problem with respect to computing the overall contact weight of *contact edges*, we will for convenience locally count for every one-vertex of the lattice the number of incident contact edges, and we will call these incident contact edges the *contacts* of this one[1]. By summing up the number of contacts over all ones, we will count every contact edge exactly twice. Since we will use this way of counting both for the contacts achieved by our algorithm and for the number of hypothetically possible contacts, this will not affect the approximation ratio.

For an input string p over the alphabet $\{0, 1\}$, we denote a maximal block of a single one as a *singleton*, a maximal block of two ones as a *pair*, and a maximal block of three ones as a *triple* in p.

We will investigate our algorithms with respect to asymptotic approximation only. Thus, we can in particular assume without loss of generality that input strings for the α-DC-HP$_{2d}$ problem will start and end with zeros.

Moreover, as we are looking for a folding in $\mathcal{L}^{\square}$, also the bipartiteness of this grid structure will play a role. Therefore, for a string p over $\{0, 1\}$, let $odds(p)$ and $evens(p)$ denote the number of ones on odd and even positions in p, respectively.

Let $\mu = 2 \cdot \min\{odds(p), evens(p)\}$. Thus, μ denotes the maximal number of horizontal/vertical contacts that could be established for p. Note that, due to the bipartiteness of $\mathcal{L}^{\square}$, horizontal/vertical contacts can occur between ones with different parity only. Hence, at most $2 \cdot \min\{odds(p), evens(p)\}$ contact edges

[1]Please note that the contact edges according to our definition are sometimes called contacts in the existing literature on the HP model.

are possible, as each one may have two incident contact edges. (This is exactly the upper bound also provided by [9, 10].) Therefore, the number of possible horizontal/vertical contacts is bounded by 2μ.

Clearly, as the α-DC-HP$_{2d}$ problem for $\alpha = 0$ is exactly the original HP problem, the α-DC-HP$_{2d}$ problem inherits its hardness from the hardness result for the two-dimensional HP problem [6].

Theorem 2.4. *The α-DC-HP$_{2d}$ problem is NP-hard.* □

We would like to point out that, concerning the analysis of loss edges, we do not need to distinguish between vertices of the grid which are labeled by zero and those which are completely unlabeled. Thus, for convenience, we can assume that unlabeled vertices are labeled with zero for our considerations.

3. Upper bounds on the overall contact weight

In this section we will establish some upper bounds on the overall contact weight that could be achieved for the α-DC-HP$_{2d}$ problem.

The first bound easily results from counting the number of possible neighbors.

Lemma 3.1. *Let $p \in \{0,1\}$ be a string starting and ending with a zero, let φ be a conformation of p in $\mathcal{L}^{\square}$, and let n denote the number of ones in p. Then the overall contact weight is at most $2\mu + 4\alpha n$.*

Proof. Each one has at most 4 horizontal/vertical neighbors in the grid. Two of these neighboring positions are occupied with the adjacent positions in the input string and thus cannot serve as contacts. Therefore, we achieve the same bound on the number of horizontal/vertical contacts as in the original HP model, which is 2μ as discussed above. Moreover, there are at most 4 possible diagonal contacts incident to any 1-vertex that each contribute α to the overall contact weight. □

A more sophisticated upper bound is based on the observation that a zero, which is adjacent to a one in the input string, will prevent the ones in its neighborhood from being incident to the maximal number of six contact edges.

To establish this result, we consider the number (and the weight) of loss edges that are unavoidable in the neighborhood of an alternation edge. Here, we follow the similar argument from [1,4]. Informally, the neighborhood of an edge e consists of the intersection of the sets of vertices that are adjacent to both of the endpoints of e and of those edges that are adjacent to e as shown in Figure 1.

Definition 3.2. Let $e = \{x, y\}$ be any edge in $\mathcal{L}^{\boxtimes} = (V(\mathcal{L}^{\boxtimes}), E(\mathcal{L}^{\boxtimes}))$. We define the *neighborhood* of e as the subgraph $N(e) = (V, E)$ of $\mathcal{L}^{\boxtimes}$, where

$$\begin{aligned} V &= \{v \in V(\mathcal{L}^{\boxtimes}) \mid \{v, x\}, \{v, y\} \in E(\mathcal{L}^{\boxtimes})\} \cup \{x, y\} \text{ and} \\ E &= \{\{x, v\}, \{y, v\} \in E(\mathcal{L}^{\boxtimes}) \mid v \in V\}. \end{aligned}$$

(a) (b)

FIGURE 1. The neighborhood $N(e)$ of the edge $e = \{x, y\}$, if e is a horizontal/vertical edge (a), or if e is a diagonal edge (b). The edges belonging to the neighborhood are shown as dotted lines.

With the following lemma, we will analyze the neighborhood of a loss edge.

Lemma 3.3. *Let p be an input string for the α-DC-HP$_{2d}$ problem, let φ be any conformation of p, and let $l = \{x, y\}$ be any loss edge of φ, where y denotes the embedded* 1*-vertex.*

(i) *If l is a horizontal/vertical edge and y is a singleton, there are at most four alternation edges inside $N(l)$.*
(ii) *If l is a horizontal/vertical edge and y is not a singleton, there are at most three alternation edges inside $N(l)$.*
(iii) *If l is a diagonal edge, there are at most two alternation edges inside $N(l)$.*

Proof. Since both vertices x and y can be adjacent to at most two binding edges, it follows immediately that there are at most four alternation edges adjacent to l. But in this case, y is forced to be a singleton. Otherwise, y could be incident to at most one alternation edge. This immediately gives the proof for (i) and (ii). For (iii) consider Figure 1b, there can be at most two alternation edges inside $N(l)$ here, since, if there are two alternation edges, the remaining vertices in $N(l)$ are labeled zero and one, respectively. Thus, none of the remaining edges in $N(l)$ can be alternation edges. Please note that also the diagonal edge crossing l cannot be an alternation edge, since the embedding is restricted to the rectangular grid lattice only. □

Lemma 3.4. *If there exist two orthogonal alternation edges which are adjacent to a horizontal/vertical loss edge l such as shown in Figure 2, then there are*

(i) *at most three alternation edges inside $N(l)$, if the vertex z is a one, and*
(ii) *at most two alternation edges inside $N(l)$, if the vertex z is a zero.*

Proof. The edge f_1 cannot be an alternation edge since otherwise the vertex y would be incident to three binding edges, which proves (i). If z is a zero, then both endpoints of f_2 are zeros, hence f_2 is no alternation edge, which proves (ii). □

Lemma 3.5. *Let $p = 0^+b_10^+b_20^+ \dots 0^+b_k0^+$ be an input string for the α-DC-HP$_{2d}$ problem for some $0 \leq \alpha \leq 1$, where $b_i \in \{1\}^+$ for $1 \leq i \leq k$, let $n = \sum_{i=1}^{k} |b_i|$ be the number of ones in p. Then the overall contact weight in any conformation is at most $2n + 4\alpha \cdot n - 2k \cdot \min\{\frac{2}{3}, \alpha\}$.*

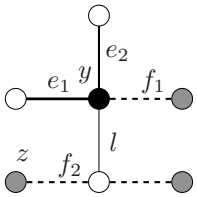

FIGURE 2. A vertical loss edge l in the neighborhood of at most three alternation edges.

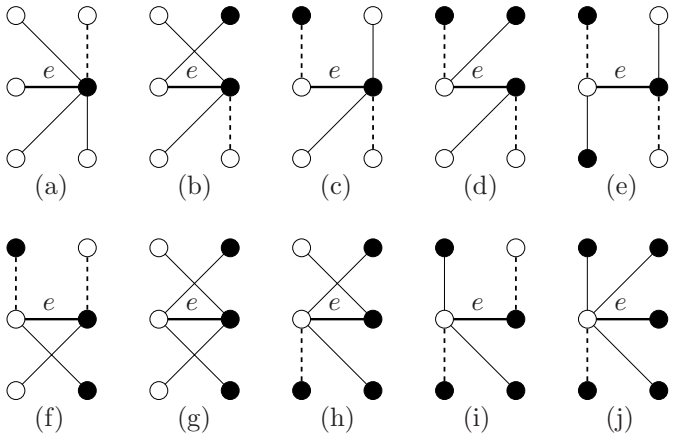

FIGURE 3. The neighborhood of an alternation edge. The alternation edge e under consideration is depicted by a bold line, the possible further alternation edges in its neighborhood are depicted by bold dashed lines, and the loss edges in its neighborhood are shown as thin lines.

Proof. Since every one can have at most two horizontal/vertical contacts and four diagonal contacts, we know that the overall contact weight is at most $2n+4\alpha\cdot n$. Let φ be any conformation of p. We will now analyze the weight of the indespensible losses around the alternation edges of p. For this we will investigate the loss edges in the neighborhood of any alternation edge e. We will distinguish ten cases according to the labeling of the vertices in this neighborhood $N(e)$. These cases are depicted in Figure 3, all cases not shown in this figure are obviously symmetric to one of the shown cases.

Every subfigure shows all edges between zero-vertices and one-vertices within $N(e)$. The more of these edges are also alternation edges, the smaller the weight of the loss edges in $N(e)$ becomes. Thus, the maximal number of alternation edges

TABLE 1. The maximal total weights of loss edges in the neighborhood of an alternation edge.

case	(a), (j)	(b), (h)	(c), (i)	(d), (f)	(e)	(g)
weight of loss edges	$\frac{1}{2} + \frac{3}{2} \cdot \alpha$	$\frac{3}{2} \cdot \alpha$	$\frac{1}{3} + \frac{1}{2} \cdot \alpha$	α	$\frac{2}{3}$	$2 \cdot \alpha$

inside $N(e)$ is also shown in the figure. The weight of the remaining loss edges can be calculated as follows.

Case (a): In this case, there are one vertical and two diagonal loss edges. The upper diagonal loss edge is adjacent to two alternation edges and thus can be counted with weight $\frac{\alpha}{2}$ according to Lemma 3.3 (iii), while the lower diagonal loss edge can be counted with weight α, since there cannot be another alternation edge besides e in its neighborhood. The vertical loss edge can be inside the neighborhoods of at most two alternation edges, as it satisfies the preconditions of Lemma 3.4 (ii), thus its weight can be counted as $\frac{1}{2}$. Altogether, this gives a weight of $\frac{1}{2} + \frac{3}{2} \cdot \alpha$ in this case.

Case (b): In this case, $N(e)$ contains three diagonal loss edges. All of these edges could be adjacent to two alternation edges, thus their weight can be counted as $\frac{3}{2} \cdot \alpha$ altogether.

Case (c): Here, the neighborhood of e contains one vertical and one diagonal loss edge. The vertical loss edge is adjacent to at most three alternation edges, according to Lemma 3.4 (i). Hence it contributes a weight of $\frac{1}{3}$. The diagonal loss edge is adjacent to two alternation edges and thus contributes a weight of $\frac{\alpha}{2}$. This adds up to an overall weight of $\frac{1}{3} + \frac{1}{2} \cdot \alpha$.

Case (d): In this case, there are two diagonal loss edges, both of which are adjacent to two alternation edges. The overall weight thus sums up to α.

Case (e): The neighborhood of e contains two vertical loss edges in this case. Lemma 3.4 (i) is applicable to both of these edges, thus the total weight is $\frac{2}{3}$ in this case.

Case (f): Here, there are two diagonal loss edges, which both can be adjacent to two alternation edges, this results in an overall weight of α.

Case (g): In this case, there are four diagonal loss edges. All of them could be adjacent to two alternation edges, hence the total weight can be estimated as $2 \cdot \alpha$.

Case (h): This case is symmetric to case (b).

Case (i): This case is symmetric to case (c).

Case (j): This case is symmetric to case (a).

We summarize the results of the particular cases in Table 1.

An easy calculation shows that the weight associated with one alternation edge takes its minimum value in case (e), if $\alpha > \frac{2}{3}$, and in cases (d) or (f), if $\alpha \leq \frac{2}{3}$. This proves the claim of the lemma. □

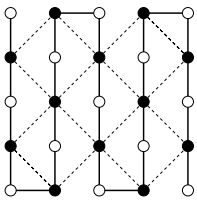

FIGURE 4. A conformation of the string p_2 from Lemma 3.7. Binding edges are shown by bold lines, contact edges are depicted as dashed lines.

Theorem 3.6. *Let $p = 0^+b_10^+b_20^+\dots0^+b_k0^+$ be an input string for the α-DC-HP$_{2d}$ problem for some $0 \leq \alpha \leq 1$, where $b_i \in \{1\}^+$ for $1 \leq i \leq k$, let $n = \sum_{i=1}^{k} |b_i|$ be the number of ones in p. Then the overall contact weight in any conformation is at most*

$$4\alpha \cdot n + \min\{2\mu, 2n - 2k \cdot \min\{\tfrac{2}{3}, \alpha\}\}.$$

Proof. The claim follows immediately from Lemmas 3.1 and 3.5. □

The next lemma shows that the bound from Theorem 3.6 cannot be improved for the case that μ is small. To prove this, we construct a string that will nearly fit the proved upper bound. The idea is to fold this string into the shape of a square, where at each edge side m ones interspaced by single zeros occur and also the corners are labeled by zeros. This will result in a chessboard like pattern. The assumed string thus consists of $(2m+1)^2 = 4m^2 + 4m + 1 = 1 + 2(2(m^2 + m))$ symbols, ones and zeros alternating with each other.

Lemma 3.7. *For any $m \in \mathbb{N}$, the string $p_m = 0(10)^{2(m^2+m)}$ can be embedded into $\mathcal{L}^{\square}$ such that it achieves a overall contact weight of $4\alpha \cdot 2(m^2+m) - 2\alpha \cdot 4m$ in $\mathcal{L}_{\alpha}^{\boxtimes}$.*

Proof. A conformation of p_2 is shown in Figure 4, the generalization to other values of m is straightforward. In this conformation, every inner one-vertex achieves four contacts of weight α, and every one-vertex at the border of the conformation achieves two such contacts, which adds up to the claimed weight. □

Please note that, for the strings p_m from Lemma 3.7, the parameter μ is zero since all ones are on even positions in the string. Thus, the conformation as described above meets the upper bound from Theorem 3.6 up to an additive second-order term of $\Omega(\alpha \cdot \sqrt{n})$, where n denotes as usual the number of ones.

4. Approximation algorithms

4.1. HP algorithms in the α-DC-HP$_{2d}$ model

Our goal is now to compute embeddings of 0/1 strings into $\mathcal{L}^{\square}$ that achieve as many contacts as possible. In principle, all algorithms for the original HP problem, where we do not consider diagonal contact edges, yield feasible solutions for the α-DC-HP$_{2d}$ problem. So it seems to be meaningful to analyze for instance the algorithms given by Hart and Istrail [9] and Newman [10] with respect to this model.

However, these algorithms cannot guarantee any approximation ratio in general. This is due to the fact that the structure of the embedding computed by these algorithms heavily depends on the upper bound derived from the bipartiteness of the underlying grid and therefore partitions the ones in the input string rigorously into an odd and an even part according to their index in the string. If, in the extreme, the input string contains only ones at even positions, the so far proposed HP-algorithms guarantee no contact at all. Anyway, we will examine the algorithm proposed by Newman with respect to our approach of weighted diagonals, to investigate for which cases it might nevertheless be reasonable.

Clearly, if μ is about n, the number of ones in the input, Newman's algorithm will achieve a constant approximation ratio (at least 9), since for at least $\frac{1}{3}$ of all possible contacts (we have 4 potential diagonal contacts and 2 potential horizontal/vertical contacts) it achieves at least one third of the possible overall contact weight.

4.2. Newman's algorithm in the α-DC-HP$_{2d}$ model

To be selfcontained, we first rephrase the algorithm proposed by Newman in [10] and essentially perform the same analysis with additionally counting diagonal edges.

Since Newman's algorithm is originally designed for the classical HP model, where we ignore diagonal edges, it was compared in [10] to the upper bound of 2μ only. Therefore, it was quite reasonable and sufficient for proving a 3-approximation to assume that all inputs have the same number of even and odd ones. We will include this assumption also here, although it does not remain meaningful in our context. However, at the end of this section we will discuss the possibility of relaxations of this assumption.

So, let p be a 0/1 string with $evens(p) = odds(p)$ that starts and ends with a zero.[2] Furthermore, we may also consider the loop $p_\circ$ instead of p, where the end positions of p are joined to form a cycle. If there exists a folding guaranteeing a certain overall contact weight for $p_\circ$, then the same folding obviously guarantees the same overall contact weight for p.

[2]As we deal with asymptotic approximation only, this restriction concerning the end positions of p does not matter.

The basic idea of Newman's algorithm is to compute a folding point and to construct an advantageous folding of the substrings on the left- and right-hand side of this point subsequently. This will eventually result in a staircase-like arrangement of odd ones on one side and even ones on the other side.

Newman established the following result to guarantee the existence of an appropriate folding point.

Lemma 4.1 (Newman [10], Lem. 2.2)**.** *Let $p = p_0 \dots p_{m-1}$ be a string over $\{0,1\}$, where $p_0 = p_{m-1} = 0$ and $evens(p) = odds(p)$, and let $p_\circ$ denote the corresponding loop. Then there exists a point p_x such that if we go around $p_\circ$ in one direction (i.e., either clockwise or counter-clockwise) starting at p_x to any point p_j, then*

$$odds(p_x p_{x+1} \dots p_j) \geq evens(p_x p_{x+1} \dots p_j),$$

and if we go around $p_\circ$ in the other direction from p_{x-1} to any point p_k, then

$$evens(p_{x-1} p_{x-2} \dots p_k) \geq odds(p_{x-1} p_{x-2} \dots p_k).$$

(Here, the indices j and k are considered to be modulo m.)

Informally, we may simply say that there exists a point such that going into one direction starting from that particular point, we will always meet at least as many odd ones as even ones, and going into the other direction, we will always meet at least as many even ones as odd ones at any point.

Having determined such a point as described in the previous lemma, we may use it as a folding point for the algorithm. Before we actually start with the presentation of the algorithm we introduce some notation that will allow us to specify the folding.

Let $p = p_0 \dots p_{m-1}$ be our input string over $\{0,1\}$ and $p_\circ$ the corresponding loop.

Firstly, without loss of generality, we can assume that there is a point p_x as in Lemma 4.1, such that going in clockwise direction we have $odds(p_x p_{x+1} \dots p_j) \geq evens(p_x p_{x+1} \dots p_j)$, for any point p_j, and *vice versa* for the counter-clockwise direction.

By the ith odd one we denote the i-th odd one starting from p_{x+1} going in clockwise direction. Similarly, by the ith even one we denote the i-th even one starting from p_{x-2} going in counter-clockwise direction[3]. We denote the substring starting at the element of $p_\circ$ directly following the $(i-1)$th odd one up to and including the ith odd one by $S_{\text{odd}}(i)$. Its corresponding length is denoted as $l_{\text{odd}}(i)+1$. Similarly, we denote the substring starting at the element of $p_\circ$ directly following the $(i-1)$th even one up to and including the ith even one by $S_{\text{even}}(i)$. Its corresponding length is denoted as $l_{\text{even}}(i)+1$. Note that $l_{\text{odd}}(i)$ and $l_{\text{even}}(i)$ thus denote the number of intermediate positions between two consecutive odd (respectively even) ones, and are always odd integers. Furthermore, by oddpos(i) we map the ith odd one to its actual position in the string, and in the same way we denote by evenpos(i) the mapping of the ith even one to its actual position. Figure 5 illustrates these notations.

[3]We use this particular definition for the ith ones here, since p_{x-2} and p_{x+1} will be paired in Newman's algorithm.

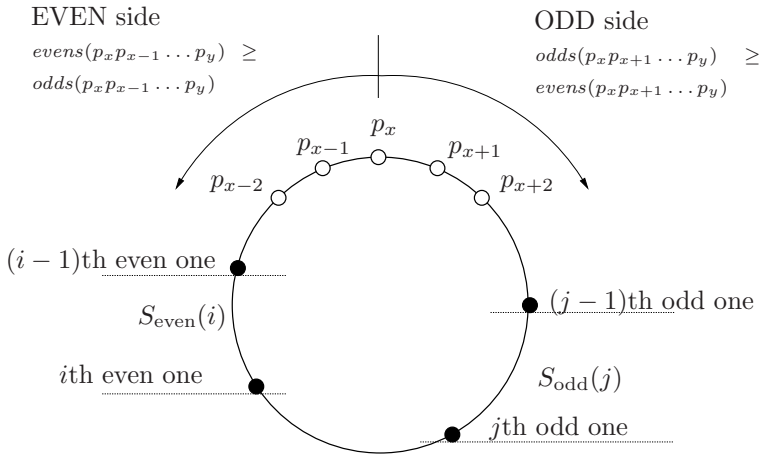

FIGURE 5. Illustration of the notions required in the presentation of Newman's algorithm.

Algorithm 1 NEWMAN'S ALGORITHM

Input: A loop $p_\circ$ over $\{0,1\}$.

(1) For $p_\circ$ compute a folding point p_x as considered in Lemma 4.1.
(2) Arrange the symbols arround p_x (*i.e.*, $p_{x-2}, p_{x-1}, p_x, p_{x+1}$) according to Figure 6. Set $i := 1$ and $j := 1$ (these will be the counters for walking in counter-clockwise and clockwise direction, respectively).
(3) Distinguish four cases according to the lengths of $S_{\text{even}}(i)$ and $S_{\text{odd}}(j)$.
   a If $l_{\text{even}}(i) = l_{\text{odd}}(j) = 1$, then fold $S_{\text{even}}(i), S_{\text{even}}(i+1)$ and $S_{\text{odd}}(j), S_{\text{odd}}(j+1)$ according to Figure 7.
   Set $i := i+2$ and $j := j+2$.
   b If $l_{\text{even}}(i) \geq 3$ and $l_{\text{odd}}(j) \geq 3$, then perform essentially the same folding of $S_{\text{even}}(i), S_{\text{even}}(i+1)$ and $S_{\text{odd}}(j), S_{\text{odd}}(j+1)$ as in the previous case, additionally arranging the intermediate symbols in appropriate side arms (see Fig. 8).
   Set $i := i+2$ and $j := j+2$.
   c If $l_{\text{even}}(i) = 1$ and $l_{\text{odd}}(j) \geq 3$, then fold $S_{\text{even}}(i), S_{\text{even}}(i+1)$ and $S_{\text{odd}}(j)$ according to Figure 9.
   Set $i := i+2$ and $j := j+1$.
   d If $l_{\text{even}}(i) \geq 3$ and $l_{\text{odd}}(j) = 1$, then fold $S_{\text{even}}(i)$ and $S_{\text{odd}}(j), S_{\text{odd}}(j+1)$ according to Figure 10.
   Set $i := i+1$ and $j := j+2$.
(4) Iterate the folding process described in Step 2 until $S_{\text{even}}(i)$ and $S_{\text{odd}}(j)$ overlap at some point.

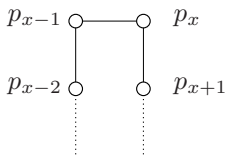

FIGURE 6. Arrangement of the symbols around the folding point p_x.

We are now ready to present Newman's algorithm. As usual, we describe the algorithm in terms of folding patterns to prevent an irksome and confusing index-based notation. Although Newman's algorithm was originally not designed to account for diagonals, we include also diagonal contacts in the drawings of the folding patterns.

There is no special reason for pairing p_{x-2} and p_{x+1} instead of p_{x-1} and p_{x+2} in Step 2. However, a pairing of p_{x-1} and p_{x+1} is impossible due to the parity constraints in $\mathcal{L}^{\square}$, so without loss of generality, we have to decide for one of both.

The situation described in the cases 3(a) and 3(b) is quite advantageous. Unfortunately, it is not clear how to come up with an equally favorable folding for the case, where the number of intermediate zeros is one in one of the considered substrings and greater or equal 3 in the other one (cases 3(c) and 3(d)).

Next, we follow the same lines in the analysis of Newman's algorithm as in [10], but additionally account for diagonal edges.

Lemma 4.2. *Newman's algorithm asymptotically guarantees an overall contact weight of at least* $\frac{2}{3}\mu + \frac{1}{3}\alpha\mu$ *for the* α*-DC-HP*$_{2d}$ *problem.*

Proof. Let $p = p_0 \dots p_{m-1}$ be a string over $\{0,1\}$ and $p_\circ$ its corresponding loop. Denote by i^* and j^* the values of i and j after the last iteration of Newman's algorithm on the input $p_\circ$, *i.e.*, i^* and j^* are the first values for i and j such that $p_{x-2}p_{x-3}\cdots p_{\mathrm{evenpos}(i^*)}$ and $p_{x+1}p_{x+2}\cdots p_{\mathrm{oddpos}(j^*)}$ overlap each other.

Before we go into the details of the proof, we will first give a informal outline and describe some necessary considerations. Our plan is to estimate the number of odd ones participating in the folding and thus contributing to the overall contact weight. To do so, we have to look at all odd ones that are considered by the folding performed by Newman's algorithm, these are roughly the odd ones in $p_{x+1}p_{x+2}\cdots p_{\mathrm{oddpos}(j^*)}$. But for a rigorous estimation we have to take into account that, due to the overlapping of $p_{x+1}p_{x+2}\cdots p_{\mathrm{oddpos}(j^*)}$ and $p_{x-2}p_{x-3}\cdots p_{\mathrm{evenpos}(i^*)}$ some odd ones "at the end" of $p_{x+1}p_{x+2}\cdots p_{\mathrm{oddpos}(j^*)}$ might not contribute to the folding. In particular, $p_{\mathrm{oddpos}(j^*)}$ cannot be paired by the algorithm, since it occurs inside $p_{x-2}p_{x-3}\cdots p_{\mathrm{evenpos}(i^*)}$. Moreover, also $p_{\mathrm{oddpos}(j^*-1)}$ might be unpaired if the situation of Case (a) or Case (b) occurs but $S_{\mathrm{even}}(i+1)$ and $S_{\mathrm{odd}}(j+1)$ already overlap in some point. Then $i^* = i+1$ and $j^* = j+1$ and we do not pair $j = j^*-1$.

Furthermore, to estimate the number of odd ones in $p_{x+1}p_{x+2}\cdots p_{\mathrm{oddpos}(j^*)}$ with respect to the overall number of odd ones or to μ (which will eventually help us to establish an approximation ratio), we have to account for the situation around the folding point, too. Here, Newman's algorithm folds the positions p_{x-2}, p_{x-1}, p_x,

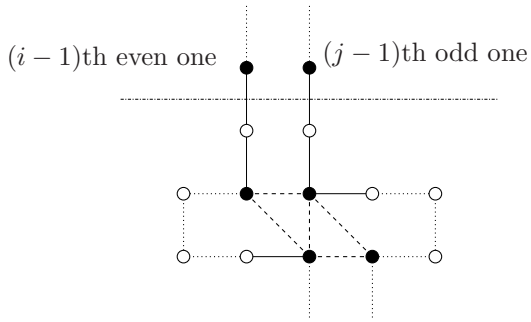

FIGURE 7. Folding pattern if both $l_{\text{even}}(i) = l_{\text{odd}}(j) = 1$.

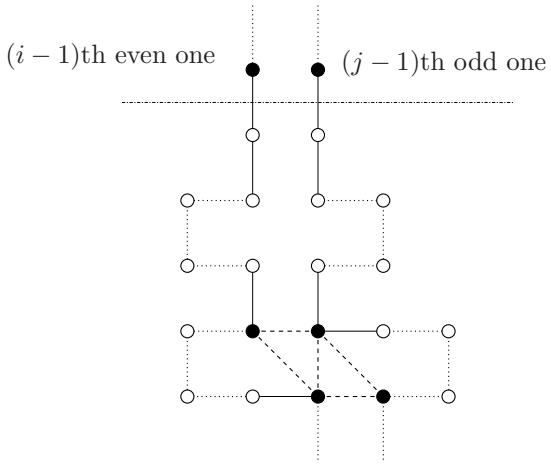

FIGURE 8. Folding pattern if both $l_{\text{even}}(i) \geq 3$ and $l_{\text{odd}}(j) \geq 3$.

and p_{x+1} according to Figure 6. At most two of these positions may be odd ones and therefore we have to consider them in our estimation.

Now, after estimating the number of odd ones participating in the folding, we will have a closer look on the types of the foldings and determine the corresponding contact weight for each type and sum up. We will finally express the estimated contact weight in terms of $odds(p_\circ)$ and ignore all additive constants, since we are interested in the asymptotic amount of contact weight only.

Now, let us continue with the proof. From the setting of i^* and j^*, we can conclude that at least $odds(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) - 2$ odd ones participate in some contacts. We have to subtract 2, since $p_{\text{oddpos}(j^*-2)}$ might be the last odd one (in clockwise direction) that is paired by the algorithm. On the other hand, at most $odds(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}) + 1$ odd ones potentially do not participate

in any contacts. Here, we added 1 to account for the fact that either p_{x-1} or p_x might be an odd one, too. The overlapping of $p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}$ and $p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}$ guarantees that we do not underestimate the number of unpaired odd ones.

By Lemma 4.1 we know

$$\mathit{odds}(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}) \le \mathit{evens}(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}). \tag{1}$$

Furthermore, since $p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}$ and $p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}$ are overlapping and either p_{x-1} or p_x might be an odd one, too, the following holds

$$\mathit{odds}(p_\circ) \le \mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) + \mathit{odds}(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}) + 1. \tag{2}$$

Combining equations (1) and (2), we obtain

$$\mathit{odds}(p_\circ) \le \mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) + \mathit{evens}(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}) + 1. \tag{3}$$

In the following analysis, we will pair together as many of the foldings of type (c) and (d) as possible and denote these as type (c-d) folds. We assume without loss of generality that u type (c) folds remain unpaired. (The case of unpaired type (d) folds is symmetric.) Then, the number of odd ones participating in type (a), (b), or (c-d) folds is

$$\mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) - 2 - 2u, \tag{4}$$

since there occur 2 odd ones in each type (c) fold.

In these folds of type (a), (b), or (c-d), the number of odd ones matches the number of even ones. Moreover, there are u additional even ones involved in type (c) folds. Again, we have to take into account that two even ones may remain unconsidered in $p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}$ due to its overlapping with $p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}$. Hence,

$$\mathit{evens}(p_{x-2}p_{x-3}\dots p_{\text{evenpos}(i^*)}) \le \mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) - u + 2. \tag{5}$$

Combining equations (3) and (5) yields

$$\mathit{odds}(p_\circ) \le \mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) + \mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) - u + 3. \tag{6}$$

This is equivalent to

$$\mathit{odds}(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) \ge \frac{\mathit{odds}(p_\circ)}{2} + \frac{u}{2} - \frac{3}{2}. \tag{7}$$

Next, we consider how much contact weight is contributed by the odd ones in $p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}$. For the two odd ones participating in a type (a) or type (b) fold, we obtain a contact weight of $6 + 4\alpha$, since the folding establishes three horizontal/vertical contact edges and two diagonal contact edges. For the three odd ones participating in a type (c-d) fold, we obtain a contact weight of $8 + 4\alpha$.

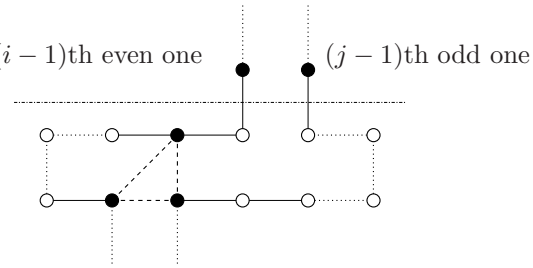

FIGURE 9. Folding pattern if $l_{\text{even}}(i) = 1$ and $l_{\text{odd}}(j) \geq 3$.

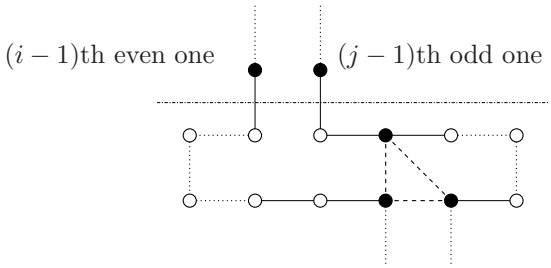

FIGURE 10. Folding pattern if both $l_{\text{even}}(i) \geq 3$ and $l_{\text{odd}}(j) = 1$.

In these cases we can thus guarantee a contact weight of $\frac{6+4\alpha}{2} = 3+2\alpha$ or $\frac{8+4\alpha}{3} = \frac{8}{3} + \frac{4}{3}\alpha$ for each odd one on average.

For the u remaining type (c) folds, we achieve a contact weight of $2+\alpha$ for each odd one. Furthermore, in each of the u remaining folds of type (c), two odd ones occur.

This implies that we can guarantee at least a contact weight of

$$\left(\frac{8}{3} + \frac{4}{3}\alpha\right) \cdot \big(odds(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)}) - 2 - 2u\big) + 2 \cdot (2+\alpha)u. \qquad (8)$$

Estimating $odds(p_{x+1}p_{x+2}\dots p_{\text{oddpos}(j^*)})$ according to equation (7), we can bound this value from below as follows:

$$\begin{aligned} & \left(\frac{8}{3} + \frac{4}{3}\alpha\right) \cdot (odds(p_{x+1}p_{x+2}\dots p_{j^*}) - 2 - 2u) + 2 \cdot (2+\alpha)u \\ \geq\ & \left(\frac{8}{3} + \frac{4}{3}\alpha\right) \cdot \left(\frac{odds(p_\circ)}{2} + \frac{u}{2} - 2u - \frac{7}{2}\right) + 2 \cdot (2+\alpha)u \\ =\ & \frac{4}{3}odds(p_\circ) - 4u + \frac{2}{3}\alpha odds(p_\circ) - 2\alpha u + 4u + 2\alpha u - \frac{7}{2}\left(\frac{8}{3} + \frac{4}{3}\alpha\right) \cdot \end{aligned}$$

As we consider the amount of contact weight only asymptotically, we can skip the additive constant $-\frac{7}{2}\left(\frac{8}{3}+\frac{4}{3}\alpha\right)$ and obtain an asymptotical contact weight of at least

$$\frac{4}{3}\mathit{odds}(p_\circ)+\frac{2}{3}\alpha\,\mathit{odds}(p_\circ).$$

Since we assumed $\mathit{odds}(p_\circ)=\mathit{evens}(p_\circ)$ we can set $\mathit{odds}(p_\circ)=\frac{\mu}{2}$, leading to an overall contact weight achieved by Newman's algorithm of

$$\frac{2}{3}\mu+\frac{1}{3}\alpha\mu$$

which completes the proof. □

Theorem 4.3. *Newman's algorithm is an asymptotic approximation algorithm for the α-DC-HP$_{2d}$ problem with a ratio of*

$$\left(\frac{4\alpha\cdot n+\min\{2\mu,2n-2k\cdot\min\{\frac{2}{3},\alpha\}\}}{\frac{2}{3}\mu+\frac{1}{3}\alpha\mu}\right).$$

Proof. To compute the approximation ratio of Newman's algorithm, we compute the fraction of the upper bound from Theorem 3.6 and the overall contact weight guaranteed by the algorithm according to Lemma 4.2 and obtain a ratio of

$$\frac{4\alpha\cdot n+\min\{2\mu,2n-2k\cdot\min\{\frac{2}{3},\alpha\}\}}{\frac{2}{3}\mu+\frac{1}{3}\alpha\mu}.$$ □

To give an idea of the ratio established above, we now discuss the corresponding ratios for specific values of α and μ.

- Clearly, if $\mu=0$, *i.e.*, if there cannot be any horizontal/vertical contacts at all, the approximation ratio is infinite.
- For $\alpha=0$, the α-DC-HP$_{2d}$ problem corresponds to the original HP problem and the approximation ratio is 3 (as already shown in [10]).
- For $\mu=n$, we can guarantee a ratio of $\frac{6+12\alpha}{2+\alpha}$ which is worst for $\alpha=1$ and yields a ratio of 6 in this case.

Extending the analysis of Newman's algorithm to concern all ones included in the input and not to assume $\mathit{odds}(p_\circ)=\mathit{evens}(p_\circ)$, appears to be quite problematic due to the following reasons.

- Applying the folding strategy of Newman's algorithm inevitably implies that only $\min\{\mathit{odds}(p_\circ),\ \mathit{evens}(p_\circ)\}$ ones are considered.
- Refining the analysis with respect to the block length seems to be a reasonable idea, since longer blocks of consecutive ones that participate in the staircase-like folding, will imply even more contact weight. Nevertheless, it remains unclear, how we can guarantee that certain blocks really participate in the folding. However, it appears to be a suitable heuristic to obtain the condition $\min\{\mathit{odds}(p_\circ),\mathit{evens}(p_\circ)\}$ of the algorithm by ignoring singletons first and longer blocks after that. Because in each block

that is not a singleton both, odd and even ones will occur, we will obtain an improved contact weight in this case independent whether the block belongs to the "odd side" or to the "even side".

4.3. Algorithm Meander

In this section, we present an alternative approach for computing a good folding for the α-DC-HP$_{2d}$ problem. This algorithm is not based on the computation of a folding point, but computes a folding while traversing the input string and analyzing the particular situation locally.

Algorithm 2 Meander

Input: A string p over $\{0,1\}$.

Execute: Walk along p and embed each long block (of length at least 4) in a meander-like way into two or three consecutive rows of the grid such that there are either two or three consecutive ones in the leftmost and rightmost column of the embedding of this block as shown in Figure 11m and n. Embed each shorter block into a single column of the grid. Place consecutive blocks next to each other as shown in Figure 11a to l, arranging the zeros in appropriate side-arms.

Lemma 4.4. *For a given input p, algorithm* Meander *guarantees an overall contact weight of at least $l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}$, where l_1 denotes the number of singletons, and n denotes the overall number of ones in p.*

Proof. Let in the following a *transition* denote a pair of embedded consecutive blocks. To prove this lemma, we count the average contact weight contributed by a block participating in the transitions shown in Figure 11a–l. Here, we always assume a worst-case scenario.

Let us consider the different types of blocks separately.

(1) Let x denote a singleton. It is quite obvious in this case that, with respect to the achieved contact weight, the worst-case in the folding performed by algorithm Meander occurs, if x is framed by two other singletons which are both connected to x by an odd number of intermediate zeros. In this case x contributes a contact weight of 2α (see Fig. 11b).

(2) Let y denote a pair of ones. In this case, we have to consider different kinds of transitions, namely all types of transitions where a pair may participate in, *i.e.*, those shown in Figure 11c, d, g–j. Actually, we would have to consider all possible combinations of these cases for the transitions of the right- and left-hand side of y. But, since we are only interested in the worst-case situation, we can restrict ourselves to the consideration of the same transitions on both sides. The worst case there is clearly also the worst case concerning all possible combinations.

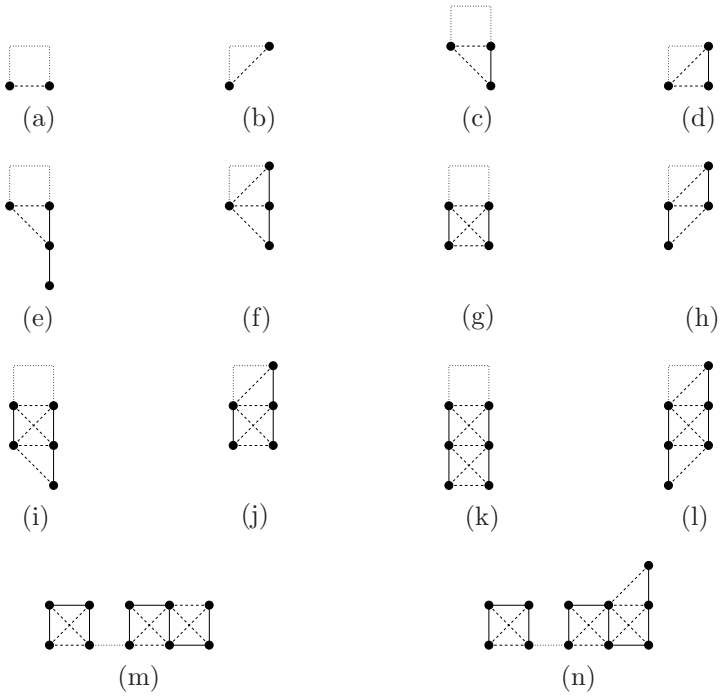

FIGURE 11. Transitions between blocks; solid lines denote binding edges, dashed lines denote contact edges, and parts of the embedding which are not considered (including all zeros) are indicated by dotted lines.

- y is framed by two transitions of type Figure 11c or d.

  We determine the contribution to the overall contact weight of this particular transition and then multiply it by two to account for the second transition. The ones in y are incident to contacts of total weight $1+\alpha$. Moreover, the involved singleton contributes also $1+\alpha$, but according to the worst-case scenario discussed in the previous case, we only counted α for this transition of this singleton. Thus, we may count an additional 1 for the contact weight of y in this case. Summing up, y contributes $2+\alpha$ for each transition it participates in, and hence $2\cdot(2+\alpha)$ altogether. Finally, on average each one in y contributes $\frac{1}{2}\cdot 2\cdot(2+\alpha) = 2+\alpha$.
- y is framed by two transitions of type Figure 11g.

  Here, the ones in y contribute $2\cdot(1+\alpha)$ for each transition. Multiplying by two (to account for two transitions) and dividing by two (to compute the average) will give again a contact weight of $2\cdot(1+\alpha)$ guaranteed by each one in y in this case. (As no singleton participates in this transition, we do not count any additional contact weight.)

- y is framed by two transitions of type Figure 11h.

  Here, the ones in y contribute $1 + 2\alpha$ for each transition. Following the same argument as above, this leads to a contact weight of $1 + 2\alpha$ for each one in y on average.
- y is framed by two transitions of type Figure 11i or (j).

  Here, the ones in y contribute $2 + 3\alpha$ for each transition. Following the same argument as above, this leads to a contact weight of $2 + 3\alpha$ for each one in y on average.

This implies that we can guarantee a contact weight of at least

$$\min\{2 + \alpha, 2 \cdot (1 + \alpha), 1 + 2\alpha, 2 + 3\alpha\} = 1 + 2\alpha$$

for each one occuring in a pair.

(3) Let z denote a triple of ones. In this case, we have to consider different kinds of transitions, namely all types of transitions where a triple may participate in, *i.e.*, those shown in Figure 11e, f, i–l. Again, we only have to consider pairs of the same transitions to detect the worst-case.

- z is framed by two transitions of type Figure 11e.

  Here, the ones in z contribute $1 + \alpha$ for each transition. Moreover, for each transition we can guarantee an additional contact weight of 1 for the involved singleton. Thus, both transitions contribute $4 + 2\alpha$ in total. Hence, each one in z contributes $\frac{4}{3} + \frac{2}{3}\alpha$ on average.
- z is framed by two transitions of type Figure 11f.

  Here, the ones in z contribute $1 + 2\alpha$ for each transition. Moreover, for each transition we can guarantee an additional contact weight of $1 + \alpha$ for the involved singleton. Thus, both transitions contribute $4 + 6\alpha$ in total. Hence, each one in z contributes $\frac{4}{3} + 2\alpha$ on average.
- z is framed by two transitions of type Figure 11i or j.

  Here, the ones in z contribute $2 + 3\alpha$ for each transition. Thus, both transitions contribute $4+6\alpha$ in total. Hence, each one in z contributes $\frac{4}{3} + 2\alpha$ on average.
- z is framed by two transitions of type Figure 11k.

  Here, the ones in z contribute $3 + 4\alpha$ for each transition. Thus, both transitions contribute $6+8\alpha$ in total. Hence, each one in z contributes $2 + \frac{8}{3}\alpha$ on average.
- z is framed by two transitions of type Figure 11l.

  Here, the ones in z contribute $2 + 4\alpha$ for each transition. Thus, both transitions contribute $4+8\alpha$ in total. Hence, each one in z contributes $\frac{4}{3} + \frac{8}{3}\alpha$ on average.

For triples of ones, we can thus guarantee a contact weight of at least

$$\min\{\tfrac{4}{3} + \tfrac{2}{3}\alpha, \tfrac{4}{3} + 2\alpha, \tfrac{4}{3} + 2\alpha, \tfrac{4}{3} + \tfrac{8}{3}\alpha\} = \tfrac{4}{3} + \tfrac{2}{3}\alpha$$

for each one on average.

(4) Let w_e denote blocks of ones of length m, where $m \geq 4$ and m is even (see Fig. 11m). To compute the average contribution of each one in this case, we can simply determine the total contribution to the contact weight by inner ones and then additionally add half of the worst-case contribution of a pair for each of the two borders. Let therefore $c_p = 2 + 4\alpha$ denote the minimum contribution for a pair of ones(for the whole pair and not for each one in a pair on average). To easily compute the contributed contact weight, observe that the folding shown in Figure 11m contains $\frac{m-2}{2}$ square shaped regions, and each of these contributes two diagonal contact edges and one horizontal contact edge. Additionally, we have to consider the contact edges incident to the borders of the folding. Then, we obtain an average contact weight of

$$\frac{\frac{m-2}{2} \cdot (4\alpha + 2) + 2 \cdot \frac{1}{2} \cdot c_p}{m} = 1 + 2\alpha.$$

(5) Let w_o denote blocks of ones of length m, where $m \geq 5$ and m is odd (see Fig. 11n). To compute the average contribution of each one in this case, we can simply determine the total contribution to the contact weight by inner ones and then additionally add half of the worst-case contribution of a pair for one border and half of the worst-case contribution for a triple for the other border. Let, therefore, $c_p = 2 + 4\alpha$ denote the minimum contribution for a pair (for the whole pair and not for each one in a pair on average) of ones, and similarly let $c_t = 4 + 2\alpha$ denote the minimum contribution for a triple of ones. Again, we count the number of square shaped regions in the folding shown in Figure 11n, which is $\frac{m-3}{2}$ here. Then, we obtain an average contact weight of

$$\frac{\frac{m-3}{2} \cdot (4\alpha + 2) + 2\alpha + \frac{1}{2} \cdot c_p + \frac{1}{2} \cdot c_t}{m} \geq 1 + 2\alpha - \frac{1}{5}\alpha = 1 + \frac{9}{5}\alpha.$$

Here, the worst-case occurs for blocks of length 5.

Now, in what follows, we analyze the singletons and non-singletons separately. For singletons we can in the worst-case only guarantee a contact weight of 2α for each one. To estimate the least contribution of contact weight made by other blocks, we have to compute the minimum of all the above computed average contributions per one, *i.e.*, we have to determine

$$\min\left\{1 + 2\alpha, \frac{4}{3} + \frac{2}{3}\alpha, 1 + 2\alpha, 1 + \frac{9}{5}\alpha\right\} = \begin{cases} 1 + \frac{9}{5}\alpha & \text{, if } 0 \leq \alpha \leq \frac{5}{17} \\ \frac{4}{3} + \frac{2}{3}\alpha & \text{, if } \frac{5}{17} < \alpha \leq 1. \end{cases}$$

Namely, except for singletons, the worst case will either occur for the situation shown in Figure 11n or for the one shown in Figure 11e, depending on the choice of α.

Altogether we can conclude that algorithm Meander guarantees at least an overall contact weight of $l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}$, where l_1 denotes the number of singletons, and n denotes the overall number of ones in the input. □

Clearly, this is a quite rough estimation and the algorithm will do better in many cases.

However, a more detailed analysis leads to rather involved complications. So for instance, if we would consider each type of block and its guaranteed contact weight separately, thus establishing an overall contact weight of $l_1 \cdot 2\alpha + l_{\text{even}} \cdot (1 + 2\alpha) + l_3 \cdot (\frac{4}{3} + \frac{2}{3}\alpha) + l_{\text{odd}} \cdot (1 + 2\alpha, 1 + \frac{9}{5}\alpha)$, where $l_{\text{even}}, l_3, l_{\text{odd}}$ denote the number of ones in block of even length, the number of ones in triples, and the number of ones in odd length blocks longer than 5, respectively, this would imply certain ratios between the number of these blocks that can maximally occur in the folding. This is due to the observation, that a worst case contribution of *e.g.* a triple requires the presence of two bordering singletons, etc. Therefore, as these ratios have a reasonable influence on the overall contact weight contributed by the folding, it is necessary to take them into account for identifying the worst case scenario, but performing such an analysis seems to be rather cumbersome.

Moreover, we have only taken into account the additional contact weights guaranteed by neighboring blocks for the case of singletons. Clearly, this analysis might be extended to other transitions and blocks as well, *e.g.* if a transition of type 2-3 occurs.

However, both of the suggested improvements will (most probably) not help in our estimation against the worst-case szenario.

Theorem 4.5. *Algorithm Meander is a linear-time δ-approximation algorithm for the α-DC-HP$_{2d}$ problem, where $\delta = 1 + \frac{1}{\alpha}$ for $0 \leq \alpha \leq \sqrt{\frac{2}{5}}$ and $\delta = \frac{3+6\alpha}{2+\alpha}$ for $\sqrt{\frac{2}{5}} < \alpha \leq 1$.*

Proof. The linear running time of algorithm Meander is a direct consequence of the sequential application of the embedding patterns given in Figure 11a–n.

To compute the approximation ratio of algorithm Meander, we combine the result from Lemma 4.4 with the upper bound established in Theorem 3.6.

In this way we obtain

$$\frac{4\alpha \cdot n + \min\{2\mu, 2n - 2k \cdot \min\{\frac{2}{3}, \alpha\}\}}{l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}}, \tag{9}$$

where, as usual, k denotes the number of blocks, l_1 denotes the number of singletons, and n denotes the total number of ones in the given input.

Since $\mu = 2 \cdot \min\{\mathit{odds}(p), \mathit{evens}(p)\} \leq 2 \cdot \frac{n}{2} = n$, we can estimate this ratio from above by

$$\frac{4\alpha \cdot n + 2n - 2k \cdot \min\{\frac{2}{3}, \alpha\}}{l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}}. \tag{10}$$

Moreover, the number k of blocks in our input, is at least the number of singletons l_1. This leads to the following upper bound on the ratio

$$R_{\text{Meander}}(n, l_1, \alpha) = \frac{4\alpha \cdot n + 2n - 2l_1 \cdot \min\{\frac{2}{3}, \alpha\}}{l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}}. \tag{11}$$

To further simplify this term, we use the abbreviations $\beta = \min\{\alpha, \frac{2}{3}\}$ and $\gamma = \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}$.

$$R_{\text{Meander}}(n, l_1, \alpha) = \frac{4\alpha \cdot n + 2n - 2l_1 \cdot \beta}{l_1 \cdot 2\alpha + (n - l_1) \cdot \gamma}. \tag{12}$$

We are now looking for the worst-case ratio according to the parameter l_1. To achieve this, we compute the derivative of $R_{\text{Meander}}(n, l_1, \alpha)$ with respect to l_1. According to the quotient rule this gives

$$\begin{aligned}
&R_{\text{Meander}}(n, l_1, \alpha)\frac{\mathrm{d}}{\mathrm{d}l_1} \\
&\quad = \frac{-2\beta(2l_1\alpha + (n - l_1) \cdot \gamma) - (2\alpha - \gamma)(4n\alpha + 2n - 2l_1\beta)}{(l_1 \cdot 2\alpha + (n - l_1) \cdot \gamma)^2} \\
&\quad = \frac{-4l_1\alpha\beta - 2n\beta\gamma + 2l_1\beta\gamma - 8n\alpha^2 - 4n\alpha + 4l_1\alpha\beta + 4n\alpha\gamma + 2n\gamma - 2l_1\beta\gamma}{(l_1 \cdot 2\alpha + (n - l_1) \cdot \gamma)^2} \\
&\quad = \frac{-2n\beta\gamma - 8n\alpha^2 - 4n\alpha + 4n\alpha\gamma + 2n\gamma}{(l_1 \cdot 2\alpha + (n - l_1) \cdot \gamma)^2} \\
&\quad =: f.
\end{aligned}$$

As the denominator of this fraction is always positive, monotonicity of our original approximation ratio depends on the sign of the numerator. To determine, in which intervals the numerator becomes greater than or equal to zero, we distinguish three cases according to the value of α, which will enable us to determine the values of β and γ respectively.

Case $0 \leq \alpha \leq \frac{5}{17}$, thus $\beta = \alpha$ and $\gamma = 1 + \frac{9}{5}\alpha$.

Plugging in these values of β and γ in the numerator of the fraction denoted by f and asking whether it is greater equal zero (*i.e.*, whether the approximation

ratio is increasing in l_1), we obtain

$$\begin{aligned}
& -2n\alpha\left(1+\frac{9}{5}\alpha\right) - 8n\alpha^2 - 4n\alpha + 4n\alpha\left(1+\frac{9}{5}\alpha\right) + 2n\left(1+\frac{9}{5}\alpha\right) \geq 0 \\
\iff \quad & -2n\alpha - \frac{18}{5}n\alpha^2 - 8n\alpha^2 - 4n\alpha + 4n\alpha + \frac{36}{5}n\alpha^2 + 2n + \frac{18}{5}n\alpha \geq 0 \\
\iff \quad & -\frac{22}{5}n\alpha^2 + \frac{8}{5}n\alpha + 2n \geq 0 \\
\iff \quad & \alpha^2 - \frac{4}{11}\alpha - \frac{5}{11} \leq 0 \\
\iff \quad & \alpha^2 - \frac{4}{11}\alpha + \frac{4}{121} - \frac{4}{121} - \frac{5}{11} \leq 0 \\
\iff \quad & \left(\alpha - \frac{2}{11}\right)^2 \leq \frac{59}{121}.
\end{aligned}$$

Thus,

$$\begin{aligned}
-\frac{\sqrt{59}}{11} \quad &\leq \quad \alpha - \frac{2}{11} \quad \leq \quad \frac{\sqrt{59}}{11} \\
\iff \quad -\frac{\sqrt{59}-2}{11} \quad &\leq \quad \alpha \quad \leq \quad \frac{\sqrt{59}+2}{11} \approx 0.88.
\end{aligned}$$

This implies that $R_{\text{Meander}}(n, l_1, \alpha)$ from equation (12) is increasing in l_1 for $(-\frac{\sqrt{59}-2}{11}) < 0 \leq \alpha \leq \frac{5}{17} < (\frac{\sqrt{59}+2}{11})$ and therefore, to determine the worst-case ratio, we set $l_1 = n$. Then, we obtain

$$\begin{aligned}
R_{\text{Meander}}\left(n, l_1 = n, 0 \leq \alpha \leq \frac{5}{17}\right) \quad &\leq \quad \frac{4\alpha \cdot n + 2n - 2n \cdot \alpha}{n \cdot 2\alpha} \\
&= \quad \frac{2n\alpha + 2n}{2n\alpha} \\
&= \quad 1 + \frac{1}{\alpha}.
\end{aligned}$$

Thus, the approximation ratio for $\alpha \to 0$ tends to infinity, which is what we already expected since in this case algorithm Meander might only guarantee diagonal contacts.

Case $\frac{5}{17} < \alpha \leq \frac{2}{3}$, thus $\beta = \alpha$ and $\gamma = \frac{4}{3} + \frac{2}{3}\alpha$.

Plugging in these values of β and γ in the numerator of the fraction denoted by f and asking whether it is greater equal zero (*i.e.* whether the approximation

ratio is increasing in l_1), we obtain

$$\begin{aligned}
& -2n\alpha\left(\frac{4}{3}+\frac{2}{3}\alpha\right)-8n\alpha^2-4n\alpha+4n\alpha\left(\frac{4}{3}+\frac{2}{3}\alpha\right)+2n\left(\frac{4}{3}+\frac{2}{3}\alpha\right)\geq 0\\
\iff\quad & -\frac{8}{3}n\alpha-\frac{4}{3}n\alpha^2-8n\alpha^2-4n\alpha+\frac{16}{3}n\alpha+\frac{8}{3}n\alpha^2+\frac{8}{3}n+\frac{4}{3}n\alpha\geq 0\\
\iff\quad & -\frac{20}{3}n\alpha^2+\frac{8}{3}n\geq 0\\
\iff\quad & \alpha^2-\frac{2}{5}n\leq 0.
\end{aligned}$$

Thus,

$$-\sqrt{\frac{2}{5}}\leq\alpha\leq\sqrt{\frac{2}{5}}.$$

This implies that $R_{\text{Meander}}(n,l_1,\alpha)$ from equation (12) is increasing in l_1 for $\left(-\sqrt{\frac{2}{5}}\right)<\frac{5}{17}<\alpha\leq\sqrt{\frac{2}{5}}<\left(\frac{2}{3}\right)$ and decreasing in l_1 for $\sqrt{\frac{2}{5}}<\alpha\leq\frac{2}{3}$. Therefore, we distinguish these two intervals and determine the worst-case ratio for the first one setting $l_1=n$. Then, we obtain

$$\begin{aligned}
R_{\text{Meander}}\left(n,l_1=n,\frac{5}{17}<\alpha\leq\sqrt{\frac{2}{5}}\right) &\leq \frac{4\alpha\cdot n+2n-2n\cdot\alpha}{n\cdot 2\alpha}\\
&= \frac{2n\alpha+2n}{2n\alpha}\\
&= 1+\frac{1}{\alpha}
\end{aligned}$$

as above.

For the second one, we set $l_1=0$ and obtain

$$\begin{aligned}
R_{\text{Meander}}\left(n,l_1=0,\sqrt{\frac{2}{5}}<\alpha\leq\frac{2}{3}\right) &\leq \frac{4\alpha\cdot n+2n}{n\frac{4}{3}+n\frac{2}{3}\alpha} && (13)\\
&= \frac{3+6\alpha}{2+\alpha}. && (14)
\end{aligned}$$

Finally, we have to consider the remaining case.

Case $\frac{2}{3}<\alpha\leq 1$, thus $\beta=\frac{2}{3}$ and $\gamma=\frac{4}{3}+\frac{2}{3}\alpha$.

Plugging in these values of β and γ in the numerator of the fraction denoted by f and asking whether it is greater equal zero (*i.e.* whether the approximation

ratio is increasing in l_1), we obtain

$$\begin{aligned}
& -2n\frac{2}{3}\left(\frac{4}{3}+\frac{2}{3}\alpha\right) - 8n\alpha^2 - 4n\alpha + 4n\alpha\left(\frac{4}{3}+\frac{2}{3}\alpha\right) + 2n\left(\frac{4}{3}+\frac{2}{3}\alpha\right) \geq 0 \\
\iff \quad & -\frac{16}{9}n - \frac{8}{9}n\alpha - 8n\alpha^2 - 4n\alpha + \frac{16}{3}n\alpha + \frac{8}{3}n\alpha^2 + \frac{8}{3}n + \frac{4}{3}n\alpha \geq 0 \\
\iff \quad & \frac{8}{9}n + \frac{16}{9}n\alpha - \frac{16}{3}n\alpha^2 \geq 0 \\
\iff \quad & \alpha^2 - \frac{1}{3}\alpha - \frac{1}{6} \leq 0 \\
\iff \quad & \alpha^2 - \frac{1}{3}\alpha + \frac{1}{36} - \frac{1}{36} - \frac{1}{6} \leq 0 \\
\iff \quad & \left(\alpha - \frac{1}{6}\right)^2 - \frac{7}{36} \leq 0.
\end{aligned}$$

Thus,

$$\begin{aligned}
-\sqrt{\frac{7}{36}} \quad &\leq \quad \alpha - \frac{1}{6} \quad \leq \quad \sqrt{\frac{7}{36}} \\
\iff \quad -\frac{1-\sqrt{7}}{6} \quad &\leq \quad \alpha \quad \leq \quad \frac{1+\sqrt{7}}{6}.
\end{aligned}$$

As $\frac{1+\sqrt{7}}{6} < \frac{2}{3}$ this implies that $R_{\text{Meander}}(n, l_1, \alpha)$ from equation (12) is decreasing in l_1 for $\frac{2}{3} < \alpha \leq 1$. Therefore, we obtain the worst-case by setting $l_1 = 0$, yielding

$$\begin{aligned}
R_{\text{Meander}}\left(n, l_1 = 0, \frac{2}{3} < \alpha \leq 1\right) \quad &\leq \quad \frac{4\alpha \cdot n + 2n}{\frac{4}{3}n + \frac{2}{3}n\alpha} \\
&= \quad \frac{3+6\alpha}{2+\alpha}.
\end{aligned}$$

This completes the proof. □

According to the approximation ratios determined in the previous theorem, we can directly infer the following approximation ratios for particular values of α.

Corollary 4.6. *For particular values of α, algorithm Meander guarantees the following approximation ratio $\varrho(\alpha)$:*

α	$\alpha \to 0$	$\alpha = \frac{1}{4}$	$\alpha = \frac{5}{17}$	$\alpha = \frac{1}{2}$	$\alpha = \sqrt{\frac{2}{5}}$	$\alpha = \frac{2}{3}$	$\alpha = 1$
$\varrho(\alpha)$	$\to \infty$	5	4.4	3	$1 + \sqrt{\frac{5}{2}} \approx 2.58$	2.625	3

In particular, according our estimation, the minimal approximation ratio is achieved by algorithm Meander for $\alpha = \sqrt{\frac{2}{5}}$. □

Please note that we estimated the overall contact weight achieved by algorithm Meander quite roughly in the previous proof, since we considered the worst-case complexity. On the other hand, the algorithm will perform significantly better for many inputs.

Moreover, one might come up with more clever folding strategies for long sequences of ones guaranteeing more contacts. However, since the worst-case approximation ratio is essentially determined by very short blocks of ones, we will not be able to prove a better approximation guarantee in general, while there is no doubt that this would be a useful heuristic modification of the algorithm.

4.4. Comparing Newman's algorithm to algorithm Meander

According to Theorem 4.3 and Lemma 4.4, Newman's algorithm improves over algorithm Meander if

$$\mu \geq \frac{l_1 \cdot 2\alpha + (n - l_1) \cdot \min\{1 + \frac{9}{5}\alpha, \frac{4}{3} + \frac{2}{3}\alpha\}}{\frac{2}{3} + \frac{1}{3}\alpha}.$$

For the case $\alpha = 0$ this implies that $\mu \geq \frac{3}{2}(n - l_1)$. Thus, Newman's algorithm outperforms algorithm Meander only if the number of singletons is quite high even in this case.

An interesting difference between the two studied algorithms besides their approximation ratio is that Newman's algorithm is folding-point based, which is a kind of *global* property of the input utilized by the algorithm. The aim is to find the best point to "cut" the input string and to align the resulting two strands to each other.

On the other hand, algorithm Meander focusses on *local* properties and tries to compute a folding optimizing local regions of the input. Moreover, algorithm Meander considers every one in the input and does not restrict itself to horizontal/vertical contacts, while Newman's algorithm only considers ones that might establish horizontal/vertical contacts (which is not surprising since Newman's algorithm was originally designed for the HP problem, where we ignore diagonal contacts at all).

Therefore, a further investigation and combination of the results of both algorithm might yield valuable insights in the structure of the corresponding protein.

5. Conclusion

There are numerous issues for further research in this area. On the theoretical side, for many problems their complexity, in particular with respect to approximation, remains unclear. Moreover, there is a practical need not only for improved algorithms but also for further refined models to successively come closer to the real problem setting. For example, it is straightforward to extend the α-DC-HP$_{2d}$ problem to the 3-dimensional case and to additionally consider also spatial diagonals as potential contact edges. In this context it would be meaningful to introduce

a further parameter β to measure the binding forces along these spatial contacts. A possible approach would also be the introduction of vertex weights to account for the different hydrophobicity of the amino acids.

References

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

[2] C.B. Anfinsen, Principles that govern the folding of protein chains. *Science* **181** (1973) 223–230.

[3] C.B. Anfinsen, E. Haber, M. Sela and F.H. White, The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA* **47** (1961) 1309–1314.

[4] H.-J. Böckenhauer and D. Bongartz, Protein folding in the HP model on grid lattices with diagonals. *Discrete Appl. Math.* **155** (2007) 230–256. Extended Abstract in *Proc. of the 29th International Symposium on Mathematical Foundations of Computer Science (MFCS'04). Lect. Notes Comput. Sci.* **3153** (2004) 227–238.

[5] V. Chandra, A. DattaSharma and V.S.A. Kumar, The algorithmics of folding proteins on lattices. *Discrete Appl. Math.* **127** (2003) 145–161.

[6] P. Crescenzi, D. Goldman, C. Papadimitriou, A. Piccolboni and M. Yannakakis, On the complexity of protein folding. *J. Comput. Biol.* **5** (1998) 423-466. Extended Abstract in *Proc. of the 30th Annual ACM Symposium on the Theory of Computing (STOC 1998)* (1998) 597–603.

[7] K.A. Dill, Theory for the folding and stability of globular proteins. *Biochemistry* **24** (1985) 1501.

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

[9] W.E. Hart and S. Istrail, Fast protein folding in the hydrophobic-hydrophilic model within three-eights of optimal. *J. Comput. Biol.* **3** (1996) 53–96.

[10] A. Newman, A New Algorithm for Protein Folding in the HP Model, in *Proc. of the 13th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA'02)* (2002) 876–884.

Communicated by J. Hromkovic.
Received September 1st, 2006. Accepted November 4, 2006.