Bicolored Path Embedding Problems in Protein Folding Models

Tianfeng Feng[1], Ryuhei Uehara[2], and Giovanni Viglietta[3]

1 **School of Information Science, Japan Advanced Institute of Science and Technology (JAIST)**
ftflluy@jaist.ac.jp
2 **School of Information Science, Japan Advanced Institute of Science and Technology (JAIST)**
uehara@jaist.ac.jp
3 **School of Information Science, Japan Advanced Institute of Science and Technology (JAIST)**
johnny@jaist.ac.jp

Abstract

In this paper, we introduce a path embedding problem inspired by the well-known HP model of protein folding. A graph is said *bicolored* if each vertex is assigned a label in the set {red, blue}. For a given bicolored path P and a given bicolored graph G, our problem asks whether we can embed P into G in such a way as to match the colors of the vertices, or not.

We first show that our problem is NP-complete even if G is a dense graph of the same size as P. We then study the special case where G is a grid graph (a typical scenario in protein folding models), showing that the path embedding problem remains NP-complete even if P is monochromatic, or if G and P have the same size. By contrast, we prove that the path embedding problem becomes tractable if the grid graph G has fixed height. Finally, we show the NP-hardness of a maximization problem directly inspired by the HP model of protein folding.

1 Introduction

The protein folding problem asks how a protein's amino acid sequence dictates its three-dimensional atomic structure. This problem has wide applications and a long history dating back to the 1960s [7]. From the viewpoint of theoretical computer science, there is ongoing research aiming at revealing insights into reality by working on simplified abstract models.

One of the most popular such models is the *hydrophobic-polar* (HP) model [6, 8]. A protein in the HP model is represented as an abstract open chain, where each link has unit length and each joint is marked either H (hydrophobic, i.e., non-polar) or P (hydrophilic, i.e., polar). A protein is usually envisioned as a path embedded in a grid within the 2D or 3D lattice, where each joint in the chain maps to a point on the lattice, and each link maps to a single edge. The HP model of energy specifies that a chain desires to maximize the number of *H-H contacts*, which are pairs of H nodes that are adjacent on the lattice but not adjacent along the chain. The optimal folding problem in the HP model asks to find an embedding of a sequence of Hs and Ps on the 2D square lattice that maximizes the number of H-H contacts. This problem is known to be NP-hard in general [3].

Previous results on the HP model mostly concern the 2D square lattice, and some techniques rely on the properties of parity in a lattice (see [4, Sec. 9.3] for a comprehensive survey). However, such parity-related observations have no meaning in the original protein folding problem that we aim to model. Also, the number of H-H contacts is not the only possible measure that may be used to capture the intricate physical and chemical laws that

37th European Workshop on Computational Geometry, St. Petersburg, Russia, April 7–9, 2021.
This is an extended abstract of a presentation given at EuroCG'21. It has been made public for the benefit of the community and should be considered a preprint rather than a formally reviewed paper. Thus, this work is expected to appear eventually in more final form at a conference with formal proceedings and/or in a journal.

describe how a real protein folds. These facts have taken us to a new variant of the protein folding problem within the HP model, which we named *bicolored path embedding problem*.

In our model, we combine the basic ideas of protein folding with the complementary problem of *protein design*, where the goal is to synthesize a protein of a given shape (and function) from an amino acid sequence. Thus, we provide the "blueprint" of the folded shape of a protein, in the form of an input (grid) graph G with colors assigned to its vertices, and we ask if a given colored path P can be (injectively) embedded in G in such a way that vertex colors match. In other terms, we are effectively asking whether a given amino acid sequence can fold into (part of) a protein with prescribed structure. Since the HP model has nodes of only two types, we assume both G and P to be *bicolored*, say, with colors "red" and "blue".[1]

In Section 2, we prove that the bicolored path embedding problem is NP-complete even if G is a dense graph of the same *size* as P (i.e., with the same number of vertices).[2] In Section 3, we consider the case where G is a (square) grid graph, which is the standard assumption in the HP model. We first prove that the path embedding problem remains NP-complete even if P is monochromatic (e.g., all its vertices are blue), and then we prove that the problem is NP-complete even if G and P have the same size. Next, we contrast these hardness results with a polynomial-time algorithm for the case where G is a grid of fixed height: thus, the bicolored path embedding problem, parameterized according to the height of G, is in XP. In Section 4, we show that maximizing red-red contacts in the bicolored path embedding problem (defined in the same way as H-H contacts in the HP model) is also NP-hard.[3]

2 Bijective embedding in a dense graph

Let us first consider the case where the bicolored blueprint G is a "precise" description of a protein, i.e., it has to be matched exactly by the amino acid sequence represented by the bicolored path P. In other words, G and P have the same number of vertices, and the embedding should therefore be bijective. We will show that the embedding problem is NP-hard even if G is a dense graph (intuitively, a blueprint with many edges should allow greater leeway in the construction of an embedding of P) by a reduction from the strongly NP-complete *3-Partition* problem [9]. We recall that the input to the 3-Partition problem is a multiset of $3m$ positive integers $\{a_1, a_2, \ldots, a_{3m}\}$, and the goal is to decide whether it can be partitioned into m multisets of equal sum S. Our reduction is sketched in Figure 1.

The path P has length $m \cdot (S+1)$, and is made up of m consecutive copies of a sub-path denoted by P_{S+1}, which in turn consist of a red vertex followed by S blue vertices. The blueprint G contains a complete bipartite graph $K_{6m,m}$ with m red vertices on one side and $6m$ blue vertices on the other side. These blue vertices are further connected in all possible ways, forming a clique Q of size $6m$ (the gray box in the figure). Additionally, for each a_i, we construct a clique of $a_i - 2$ blue vertices (in the 3-Partition problem, we can safely assume that $a_i > 2$), and we connect two of its vertices with two vertices of Q.

---

[1] With regard to bicolored and monochromatic graphs, we do *not* adhere to established terminology from classical graph coloring theory; for the purposes of this paper, a coloring of a graph is simply a labeling of its vertices, with no extra constraints. In particular, adjacent vertices may have the same color.

[2] Formally, an infinite collection of graphs $\mathcal{S}$ is said to be a set of *dense graphs* if there is a positive constant α such that, for any large-enough n, every graph in $\mathcal{S}$ with n vertices has at least $\alpha \cdot n^2$ edges.

[3] We remark that, in previous work, it has been established that the problem of maximizing H-H contacts is NP-hard when G is not given, and P can be embedded in any way on a grid [3].

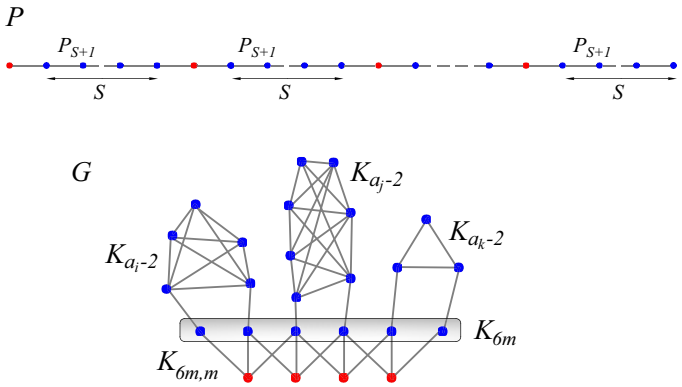

Figure 1 Sketch of the NP-hardness reduction from the 3-Partition problem. For clarity, the edges of the clique K_{6m}, as well as some edges of the complete bipartite graph $K_{6m,m}$, have been omitted.

It is easy to see that the graph G is dense, it has the same size as P, and there is an embedding of P into G if and only if the a_i's can be partitioned into multisets of sum S.

▶ **Theorem 1.** *The bicolored path embedding problem is NP-complete even if the blueprint G is a dense graph of the same size as the path P.* ◀

3 Embedding in a grid graph

In this section we focus on blueprint graphs G which are *grid graphs*, i.e., they are obtained from regular tilings of the plane (sometimes these are also called *lattice graphs*). This is the typical setting of the standard HP model.

3.1 Monochromatic path

If the path P only consists of blue vertices, there is a simple NP-hardness reduction from the *Hamiltonian path* problem (i.e., given a graph, decide if there is a walk that visits each vertex exactly once), which is known to be NP-complete even if the graph G is an induced subgraph of a square grid graph, a triangular grid graph, or a hexagonal grid graph [1, 2, 13].

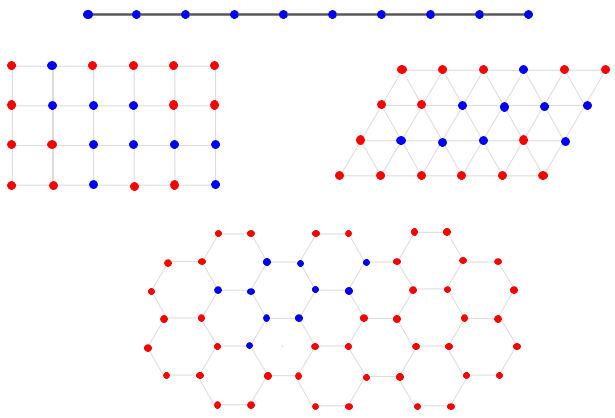

Figure 2 NP-hardness reduction from the Hamiltonian path problem for several grid graphs

Our reduction is sketched in Figure 2: given a graph G' on n vertices, which is an induced subgraph of a grid graph, we color all its vertices blue, and we "complete" it to a grid graph G by adding red vertices. Obviously, we can embed a path P of n blue vertices into G if and only if G' has a Hamiltonian path.

▶ **Theorem 2.** *The bicolored path embedding problem is NP-complete even if the blueprint G is a (square, triangular, or hexagonal) grid graph, and P is a monochromatic path.* ◀

3.2 Bijective embedding in a square grid graph

Let us turn again to bijective embeddings, this time in the case where the blueprint G is a square grid graph. We will give another NP-hardness reduction from the Hamiltonian path problem. We start from a square grid graph $R(m', n')$ with an induced subgraph G' (which is an instance of the Hamiltonian path problem), and we construct the blueprint G by "expanding" each vertex v of $R(m', n')$ into a $(k+2) \times (k+2)$ block B_v (where k is a large-enough even constant, defined later). If v is not a vertex of G', then all vertices of B_v are blue; if v is a vertex of G', then B_v is illustrated in Figure 3: its four central vertices are red, and all other vertices are blue. The size of G is therefore $(k+2) \cdot m' \times (k+2) \cdot n'$; the full construction is shown in Figure 4.

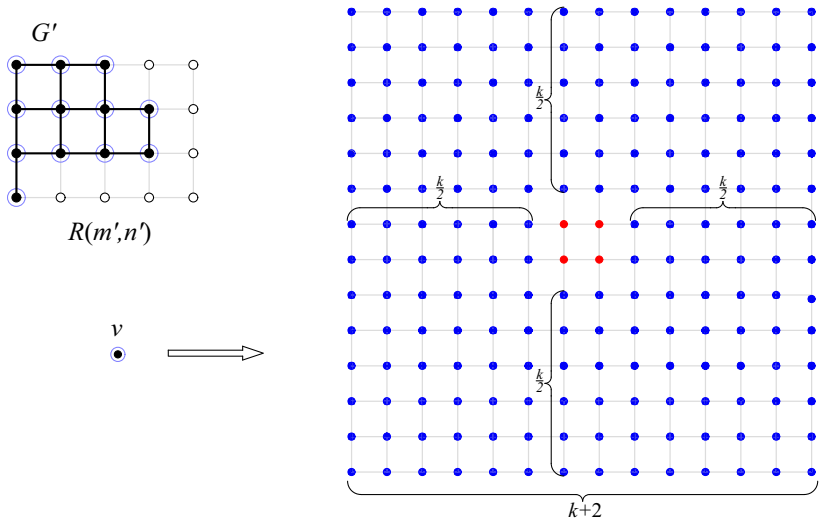

Figure 3 Transformation of a vertex v of G' into the $(k+2) \times (k+2)$ block B_v

The path P is sketched in Figure 5: there is a copy of the subpath P' for each vertex of G', and then a final trail of blue vertices such that the total length of P matches the size of G. Now, to embed P into G, we have to start from a set of four red vertices in some block B_v, and then move to another set of four red vertices in some other block B_w. Since we must traverse exactly $2k$ blue vertices between these two red sets, this is possible only if v and w are adjacent in G' (note that a "diagonal" move would take $2k+1$ steps on blue vertices). Thus, embedding P into G is impossible if G' is not Hamiltonian.

Assume now that G' is Hamiltonian. We can embed all copies of P' into G by "mimicking" a Hamiltonian path in G' and moving from one set of red vertices to the next by covering the $2 \times k$ rectangle between them in a zig-zag fashion. Eventually, the region of G covered by all the copies of P' looks like a winding "tube" of width 2, as sketched in Figure 6.

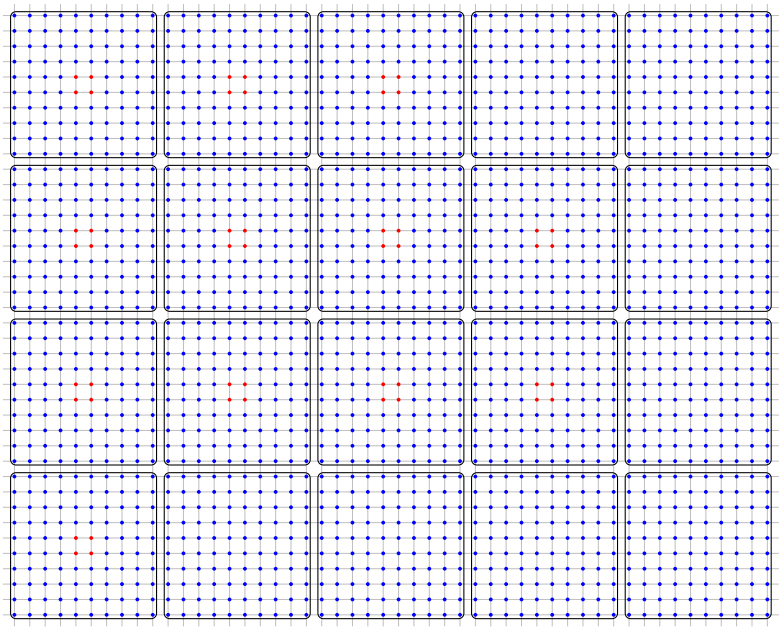

Figure 4 Complete construction of G: each block represents a vertex in the original graph

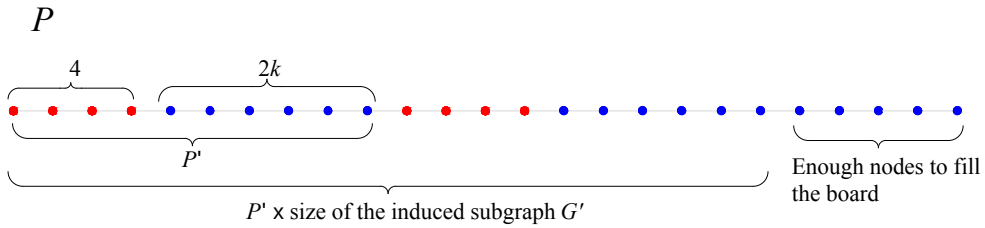

Figure 5 Construction of the path P

Now we have to cover the remaining part of G with the trailing sequence of blue vertices of P. In order to do that, we partition this region into maximal "horizontal rectangles", i.e., in such a way that no two rectangles touch each other along vertical edges, as shown in Figure 6. Then we do a depth-first traversal of these rectangles. When we visit a rectangle, we cover it as shown in Figure 7: we further divide it into smaller rectangular "tiles", one for each unvisited neighboring rectangle. After covering a tile, we visit its adjacent rectangle in the partition, and then we move to the next tile when we backtrack from that rectangle.

Constructing the tiles such that each of them can be covered completely before moving on to the next rectangle is indeed possible. In [10], the grid graphs containing a Hamiltonian path with assigned endpoints have been characterized: as it turns out, if the size of a tile is even and one of its sides is longer than four vertices, then there is a Hamiltonian path in the tile with any assigned endpoints having odd distance. Because k is a large even constant, we can indeed subdivide each rectangle in the appropriate number of tiles, each of which has even size and at least one side longer than four vertices (choosing $k = 100$ suffices by a big margin). It follows that we can embed P into G.

▶ **Theorem 3.** *The bicolored path embedding problem is NP-complete even if the blueprint G is a square grid graph of the same size as the path P.* ◀

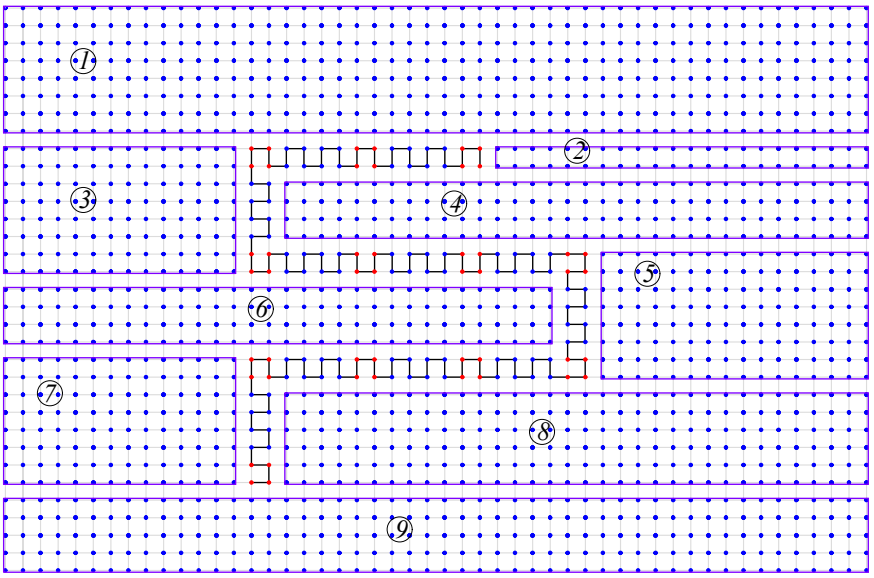

Figure 6 Partition into rectangles of the region not covered by the zig-zagging copies of P'

3.3 Fixed-height rectangular blueprint

We can contrast our previous hardness results with an embedding algorithm that runs in polynomial time, provided that the blueprint G is a grid graph of fixed height k. Thus, let G be a bicolored $m \times k$ grid, and let P be a bicolored path of n vertices. Our approach is based on dynamic programming, where a sub-problem consists of embedding part of P into a sub-grid of G going from the first column to the ath column, with $1 \leq a \leq m$. A sub-problem's specification also contains a description of the intersection between a hypothetical embedding of P and the ath column of G, illustrated in Figure 8: for each vertex w in the ath column, the sub-problem specifies which vertex v_i of P is mapped to w (if any), as well as an extra bit of information that encodes whether the left or right neighbor of v_i along P should be mapped to the left neighbor of w (if such information is incompatible with the rest of the specification, this bit is ignored). Thus, the total number of sub-problems is $2^k \cdot n^k \cdot m$ (the last factor represents the m choices of a).

The output to a sub-problem is "Yes" if an embedding satisfying the given constraints exists, "No" if it does not exist, and "N/A" if the sub-problem specifies no intersection on the ath column, and it is not possible to embed P entirely to the left of the ath colum (this implies that P should be embedded entirely to the right of the ath column, but we are still unable to determine if this is possible).

Solving a sub-problem S for column a amounts to finding a sub-problem S' for column $a-1$ with a "Yes" answer such that the specifications of S and S' are compatible. In other words, the mappings described by S and S' on columns a and $a-1$ should (i) match the colors in G and P, and (ii) match with each other: for example, if S indicates that the vertex v_{i+1} of P should be mapped to the left neighbor w' of w (where w is in column a), then S' should indicate that v_{i+1} is indeed mapped to w' (which is in column $a-1$). Thus, S can be solved by looking up at most n^k sub-problems, and each compatibility test takes $O(k)$ time.

▶ **Theorem 4.** *Given a bicolored square grid graph G of size $m \times k$ and a bicolored path P of size n, the embedding problem for G and P can be solved in $O(k \cdot 2^k \cdot n^{2k} \cdot m)$ time.* ◀

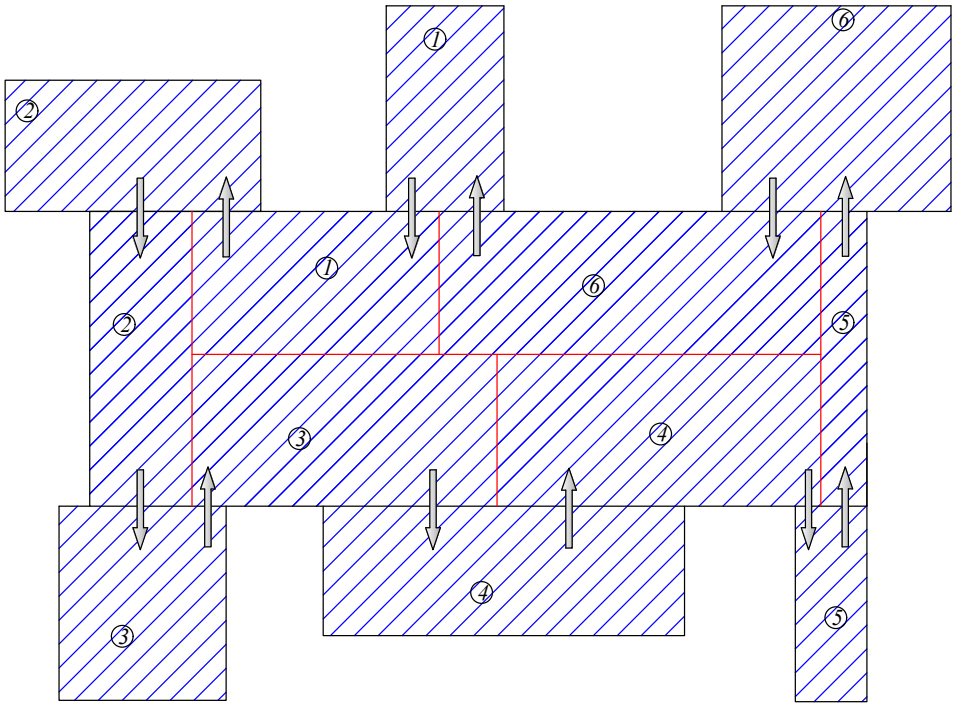

Figure 7 Traversal order of the tiles of a rectangle and its six neighboring rectangles

Note that, if k is a constant, the running time of our algorithm is $O(n^{2k}m)$, hence polynomial.

▶ **Corollary 5.** *The bicolored path embedding problem where the blueprint G is a square grid graph, parameterized according to the height of G, is in XP.* ◀

4 Maximizing red-red contacts in a grid graph

Finally, let us turn to the problem of maximizing red-red contacts in the context of the bicolored path embedding problem. Recall that, according to the HP model of energy, an amino acid chain tends to fold in a way that maximizes the number of H nodes that are close together in the folded state, even if they are not adjacent along the chain. In other words, when G and P are given, we seek an embedding of P into G that covers a large number of adjacent red vertices of G without traversing the edges between them. A red-red contact in an embedding of P is a pair of adjacent red vertices u, v in G such that the embedding of P covers both u and v, but does *not* contain the edge $\{u, v\}$.

The problem of maximizing red-red contacts in the bicolored path embedding problem is also NP-hard, even when restricted to instances where the path P is guaranteed to be embeddable into G. Figure 9 shows a reduction from the Hamiltonian path problem, where Block 2 of G is constructed as the graph in Section 3.1: that is, we are given a graph G', we color its n vertices blue, and then we "complete" it to a grid graph by adding r red vertices around it. Then we take an integer k greater than r, and we construct Block 1, which is a grid of at least k red vertices. Block 3 of G and the path P are constructed as in the figure.

Now it is easy to see that, if G' does not have a Hamiltonian path, we can only embed P in Block 3, which yields no red-red contacts. Otherwise, we can embed the blue part of P in Block 2 and the red part in Block 1, which produces a large number of red-red contacts.

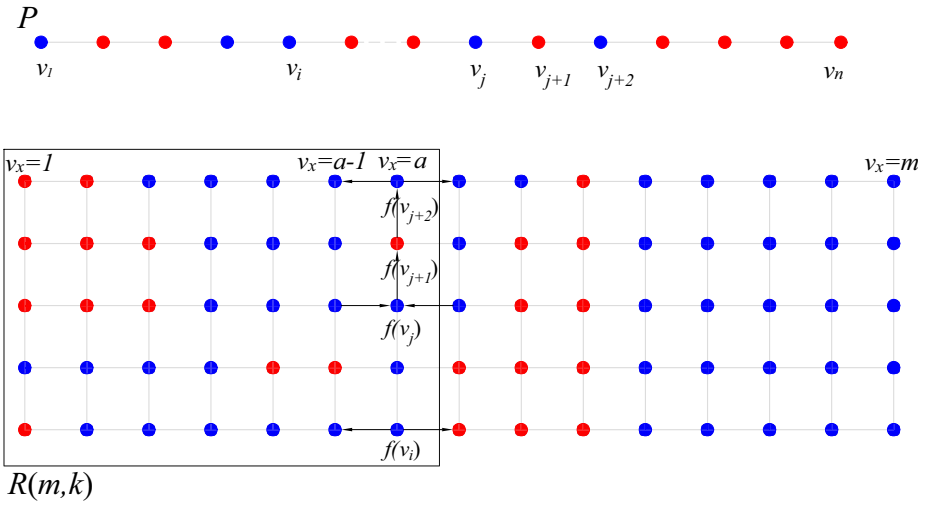

Figure 8 Illustration of the dynamic-programming algorithm for rectangular blueprints

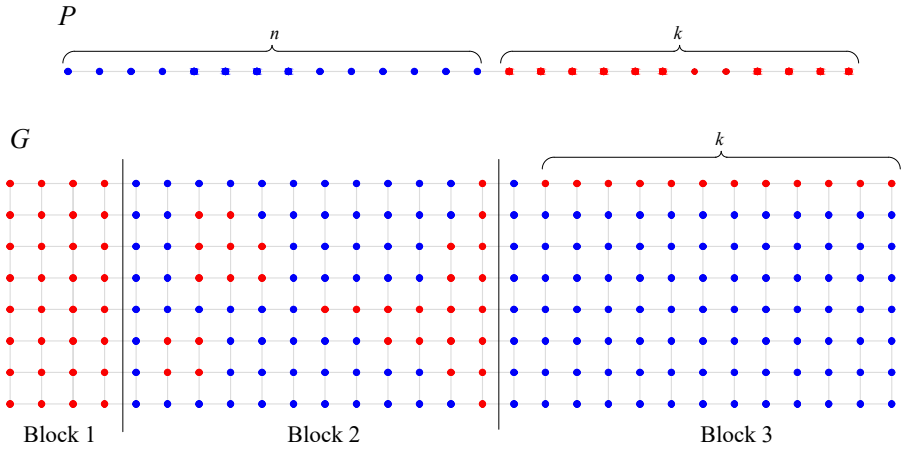

Figure 9 NP-hardness reduction for the problem of maximizing red-red contacts

▶ **Theorem 6.** *Given a bicolored grid graph G and a bicolored path P that can be embedded in G, it is NP-hard to find an embedding of P in G that maximizes red-red contacts.* ◀

This result can easily be extended to grid graphs induced by different tilings of the plane (cf. Section 3.1). Also, it shows that the related *approximation problem* is NP-hard, as well.

Acknowledgments. The authors wish to thank the anonymous reviewers for useful observations and suggestions. This work is partially supported by JSPS KAKENHI Grant Numbers 17H06287 and 18H04091.

References

1 Esther M. Arkin, Sándor P. Fekete, Kamrul Islam, Henk Meijer, Joseph S. B. Mitchell, Yurai Núñez-Rodríguez, Valentin Polishchuk, David Rappaport, and Henry Xiao. Not being (super) thin or solid is hard: A study of grid Hamiltonicity. *Computational Geometry*, 42(6–7):582–605, 2009.

2 Michael Buro. Simple Amazons endgames and their connection to Hamilton circuits in cubic subgrid graphs. In *International Conference on Computers and Games*, pages 250–261. Springer, 2000.

3 Pierluigi Crescenzi, Deborah Goldman, Christos Papadimitriou, Antonio Piccolboni, and Mihalis Yannakakis. On the complexity of protein folding. *Journal of Computational Biology*, 5(3):423–465, 1998.

4 Erik D. Demaine and Joseph O'Rourke. *Geometric folding algorithms: linkages, origami, polyhedra*. Cambridge University Press, 2007.

5 Erik D. Demaine and Mikhail Rudoy. Hamiltonicity is hard in thin or polygonal grid graphs, but easy in thin polygonal grid graphs. *arXiv:1706.10046*, 2017.

6 Ken A. Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, 24(6):1501–1509, 1985.

7 Ken A. Dill and Justin L. MacCallum. The protein-folding problem, 50 yeas on. *Science*, 338(6110):1042–1046, 2012.

8 Kit Fun Lau and Ken A. Dill. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules*, 22(10):3986–3997, 1989.

9 Michael R. Garey and David S. Johnson. *Computers and intractability*, volume 174. Freeman San Francisco, 1979.

10 Alon Itai, Christos H. Papadimitriou, and Jayme Luiz Szwarcfiter. Hamilton paths in grid graphs. *SIAM Journal on Computing*, 11(4):676–686, 1982.

11 Richard M. Karp. Reducibility among combinatorial problems. In *Complexity of Computer Computations*, pages 85–103. Springer, 1972.

12 F. Luccio and C. Mugnia. Hamiltonian paths on a rectangular chessboard. In *Proceedings of the 16th Annual Allerton Conference*, pages 161–173, 1978.

13 Christos H. Papadimitriou and Umesh V. Vazirani. On two geometric problems related to the travelling salesman problem. *Journal of Algorithms*, 5(2):231–246, 1984.