Fixed-parameter algorithms for protein similarity search under mRNA structure constraints

Guillaume Blin[1], Guillaume Fertin[1],
Danny Hermelin[2], and Stéphane Vialette[3]

[1] Laboratoire d'Informatique de Nantes-Atlantique (LINA), FRE CNRS 2729
Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3 - France
{blin,fertin}@lina.univ-nantes.fr

[2] Department of Computer Science, University of Haifa, Mount Carmel, Haifa - Israel
danny@cri.haifa.ac.il

[3] Laboratoire de Recherche en Informatique (LRI), UMR CNRS 8623
Faculté des Sciences d'Orsay - Université Paris-Sud, 91405 Orsay - France
vialette@lri.fr

Abstract. In the context of protein engineering, we consider the problem of computing an mRNA sequence of maximal codon-wise similarity to a given mRNA (and consequently, to a given protein) that additionally satisfies some secondary structure constraints, the so-called MRSO problem introduced in [2]. Since the MRSO problem is known to be **APX**-hard [7], Bongartz proposed in [7] to attack the problem using the concept of parameterized complexity. In this paper we devise fixed-parameter algorithms for MRSO for several interesting parameters.

1 Introduction

In [2, 3], Backofen *et al.* introduced the problem of computing an mRNA sequence of maximum codon-wise similarity to a given mRNA (and consequently, to a given protein) that additionally satisfies some secondary structure constraints, the so-called MRSO problem.

The initial motivation of MRSO is concerned with selenocysteine insertion, *i.e.* generating new amino acid sequences containing selenocysteine. This rare amino acid was discovered as the 21st amino acid [5], giving another clue to the complexity and flexibility of the mRNA translation mechanism. Selenocysteine is encoded by the UGA codon, which is usually a stop codon encoding the end of translation. It has been shown [5] that in case of selenocysteine, termination of translation is inhibited in the presence of a sequence of nucleotides which forms a hairpin like structure in the $3'$-region after the UGA codon. It is argued in [2] that modifying existing proteins by incorporating selenocysteine instead of a catalytic cysteine is an important problem for catalytic activity enhancement and X-ray crystallography.

Selenocysteine insertion is concerned with a restricted type of secondary structure, *i.e.* a secondary structure without pseudo-knots, and hence the linear-time algorithm presented in [2] provides an optimal solution. However, similar problems occur with complex secondary structures, *e.g.* for programmed frameshifts which allow to encode two different amino acid sequences in one mRNA sequence [10, 9]. This motivates the investigation of MRSO for more elaborate secondary structures [2, 7], and is the starting point of our study.

For the MRSO problem, it has been shown in [2] that there exists a linear-time algorithm if the considered secondary structure corresponds to an outer-planar graph (as it is the case for Selenocysteine insertion). In this paper, we refer to this algorithm as $\mathcal{A}_{\mathrm{OP}}$. For the general case, the problem was proved to be **NP**-complete in [2], and Bongartz showed recently that the problem is in fact **APX**-hard [7]. An algorithm for approximating the MRSO problem within ratio 2 is given in [2]. A slightly slower but somewhat simpler algorithm for approximating the MRSO problem within ratio 4 is given in [7]. We mention also that an extension of the MRSO problem, where insertions and deletions are allowed in the amino acid sequence, is presented in [1].

Since the MRSO problem for general secondary structures is known to be **APX**-hard [7], Bongartz proposes in [7] to attack the problem using the concept of parameterized complexity [8]. Parameterized complexity is an approach to complexity theory which offers a means of analyzing algorithms in terms of their tractability. For many hard problems, the seemingly unavoidable combinatorial explosion can be restricted to a small part of the input, the *parameter*, so that the problems can be solved in polynomial-time when the parameter is fixed.

In the last decade, parameterized complexity has proved to be useful in computational biology [6]. Consequently, since MRSO is **APX**-hard [7], Bongartz proposed in [7] to attack the problem using the concept of parameterized complexity [8]. In this paper we adopt this suggested approach. Our main contribution is new polynomial-time algorithms for MRSO when certain parameters of the problem's input are fixed.

2 Preliminaries

An mRNA is a string over the alphabet $\Sigma = \{A, C, G, U\}$, where Σ represents the four different types of nucleotides in the molecule. The pairs $\{A, U\}$, $\{G, C\}$, and $\{G, U\}$ are known as *complementary nucleotide pairs.* Note that hydrogen bonds can only be formed between complementary nucleotides in an mRNA folding. A *codon* of an mRNA sequence is a sequence of three consecutive nucleotides, *i.e.* a string in Σ^3. Thus, an mRNA sequence $S = s_1 \cdots s_{3n}$ is a concatenation of n consecutive codons, where the ith codon of S is $s_{3i-2}s_{3i-1}s_{3i}$.

Given a *source* mRNA sequence $S = s_1 \ldots s_{3n}$, we wish to evaluate the codon-wise similarity of S and another *target* mRNA sequence $T = t_1 \ldots t_{3n}$. For this, we are provided with a set of n functions, $\mathcal{F} = f_1, \ldots, f_n$, called *similarity functions* of S, such that for all $1 \leq i \leq n$, each function f_i is of the form $f_i : \Sigma^3 \to \mathbb{Q}$. Thus, f_i assigns a value to the ith codon of T according to its level of similarity in comparison with the ith codon of S. The total level of similarity between S and T is then given by $\sum_{i=1}^{n} f_i(t_{3i-2}t_{3i-1}t_{3i})$. Note that given a set of similarity functions $\mathcal{F} = f_1, \ldots, f_n$ for S, one does not need to know anything else about S in order to compute the similarity score of S and T.

The *structure constrains* $\Gamma \subseteq \{\{i,j\} \,|\, 1 \leq i < j \leq 3n\}$ for a target mRNA sequence T of length $3n$, are pairings between distinct integers in $\{1, 2, \ldots, 3n\}$. These represent necessary hydrogen bonds in the folding of T. Since we assume that each nucleotide can pair with at most one other nucleotide in any folding, each integer appears in at most one pair in Γ. Furthermore, there are no pairs of the form $\{i, i+1\}$ or $\{i, i+2\}$ in Γ, for all $1 \leq i \leq 3n-2$.

Given a set of structure constrains $\Gamma \subseteq \{\{i,j\} \,|\, 1 \leq i < j \leq 3n\}$, and an arbitrary target mRNA sequence $T = t_1 \cdots t_{3n}$, we say that nucleotides t_i and t_j are *compatible* with respect to Γ, if either $\{t_i, t_j\}$ is a complementary nucleotide pair or $\{i,j\} \notin \Gamma$. The entire sequence T is compatible with respect to Γ, if all pairs of nucleotides in T are compatible with respect to Γ.

Definition 1 (mRNA Structure Optimization (MRSO**) [2]).** *Let $\mathcal{F}$ be a set of n similarity functions for a source mRNA sequence of length $3n$, and let $\Gamma \subseteq \{\{i,j\} \,|\, 1 \leq i < j \leq 3n\}$ be a set of structure constrains. The* MRSO *problem asks to find a target mRNA sequence which is compatible with respect to Γ, and which achieves the highest possible similarity score with respect to $\mathcal{F}$.*

It is convenient to formalize MRSO in a slightly different manner using graph theoretic concepts. For a graph G, we let $\mathbf{V}(G)$ denote the set of vertices of G, and $\mathbf{E}(G)$ the set of edges of G. A linear graph G is a graph with $\mathbf{V}(G) = \{1, \ldots, |\mathbf{V}(G)|\}$. That is, it is a graph with vertices which have a fixed ordering. Therefore, we now view Γ as a linear graph with $3n$ vertices and a maximum degree of one. As we are really interested in codon-wise similarity, we use a more suitable representation of Γ.

Definition 2 (Implied structure graph [2]). *Let $\Gamma \subseteq \{\{i,j\} \,|\, 1 \leq i < j \leq 3n\}$ be a set of structure constrains for a target mRNA sequence of length $3n$. The implied structure graph of Γ, is the linear graph G_Γ with:*

$\mathbf{V}(G_\Gamma) = \{1, 2, \ldots, n\}$, *and*

$$\mathbf{E}(G_\Gamma) = \Big\{ \{i,j\} \;\Big|\; \exists \{x,y\} \in \Gamma : x \in \{3i-2, 3i-1, 3i\} \;\wedge\; y \in \{3j-2, 3j-1, 3j\} \Big\}.$$

Hence, G_Γ is a subcubic graph (*i.e.* a graph with a maximum degree of three) where vertex i in $\mathbf{V}(G_\Gamma)$ corresponds to the ith codon of a target mRNA sequence, and $i, j \in \mathbf{V}(G_\Gamma)$ are connected in $\mathbf{E}(G_\Gamma)$ if there are any structure constrains in Γ between the ith and jth codons of the sequence. Note that there can be at most three structure constrains between any pair of codons.

Given a subset of vertices $V \subseteq \mathbf{V}(G_\Gamma)$, we let $G_\Gamma[V]$ denote the subgraph of G_Γ *induced* by V, *i.e.* the subgraph with vertex set V and edge set $\mathbf{E}(G_\Gamma) \cap (V \times V)$. Similarly, given a subset of edges $E \subseteq \mathbf{E}(G_\Gamma)$, $G_\Gamma[E]$ denotes the subgraph of G_Γ with vertex set $\{i : \{i,j\} \in \mathbf{E}(G_\Gamma)\}$ and edge set E. Furthermore, we let $G_\Gamma[i,j]$ denote the subgraph of G_Γ induced by $\{i, \ldots, j\} \subseteq \mathbf{V}(G_\Gamma)$.

Henceforth, we speak of *codon assignments* for the vertices of G_Γ, *i.e.* mappings from some $V \subseteq \mathbf{V}(G_\Gamma)$ to Σ^3. An assignment for a pair of vertices $i, j \in \mathbf{V}(G_\Gamma)$, $i \rightarrow t_{3i-2}t_{3i-1}t_{3i}$ and $j \rightarrow t_{3j-2}t_{3j-1}t_{3j}$, is compatible with respect to G_Γ, if either $\{i, j\} \notin \mathbf{E}(G_\Gamma)$ or for any $\{x, y\} \in \Gamma \cap \{3i-2, 3i-1, 3i\} \times \{3j-2, 3j-1, 3j\}$, t_x and t_y are complementary nucleotides. More generally, an assignment $\phi : V \rightarrow \Sigma^3$ for some $V \subseteq \mathbf{V}(G_\Gamma)$ is compatible with respect to G_Γ, if for any $i, j \in V$, the assignment $i \rightarrow \phi(i)$ and $j \rightarrow \phi(j)$ is compatible with respect to G_Γ. Our goal in this setting, is to find an assignment $\phi : \mathbf{V}(G_\Gamma) \rightarrow \Sigma^3$ (*i.e.* a target mRNA sequence $T = \phi(1) \cdots \phi(n)$), which is compatible with G_Γ, and which maximizes $\sum_{i=1}^{n} f_i(\phi(i))$.

3 Two natural parameters for MRSO

Our discussion begins by considering two natural parameters for MRSO. Let $(G_\Gamma, \mathcal{F})$ be an instance of MRSO. The two parameters we consider are the number of edge crossings and the number of degree three vertices in G_Γ, as parameters for MRSO. We let χ and δ denote these two parameters respectfully.

Our initial interest in parameters χ and δ arises from the fact that we believe them to be small in many practical applications. Consider parameter χ. It is widely believed that many natural mRNA secondary structures form an outerplanar formation, *i.e.* a formation containing no edge crossings. Consequently, exploring this parameter was suggested explicitly in [7]. As for parameter δ, recall that a vertex of degree three in G_Γ represents a codon with three nucleotides, each pairing with complementary nucleotides in three different codons. Although this situation can occur in a folding of an mRNA molecule, it can be expected to be quite rare due to the natural geometric and thermodynamic constrains imposed on any such folding.

It turns out that MRSO is in polynomial-time solvable when either χ or δ are fixed. To show this, we will first describe a general algorithm, and later demonstrate how it can be applied for both cases. We will need the following definition:

Definition 3 (Nice edge bipartition). *Let G_Γ be an implied structure graph with n vertices. An* edge bipartition *$\mathcal{P} = (E_t, E_b)$ of G_Γ is a partitioning of the edges in G_Γ into E_t and E_b, the* top *and* bottom *edges of $\mathcal{P}$, such that $E_t \cup E_b = \mathbf{E}(G_\Gamma)$, $E_t \cap E_b = \emptyset$ and $E_t \neq \emptyset$. If the subgraph $G_\Gamma[E_t]$ is outerplanar then $\mathcal{P}$ is* nice.

Our initial algorithm is called $\mathcal{A}_{\text{NEB}}$. This algorithm will apply only for cases where a nice edge bipartition of G_Γ with a fixed number of bottom edges is given alongside the input. Following the description of $\mathcal{A}_{\text{NEB}}$, we show that when considering either χ or δ to be fixed, one can easily obtain such a bipartition.

The heart of algorithm $\mathcal{A}_{\text{NEB}}$ is the following simple observation. Suppose we want to find the highest scoring compatible mRNA sequence which starts with codon AAA. For this, we can replace the similarity function $f_1 \in \mathcal{F}$ by a different function f', where $f'(AAA) = f_1(AAA)$ and $f'(C) = -\infty$ for all codons $C \neq AAA$. Solving MRSO with the instance $(G_\Gamma, \mathcal{F}')$, where $\mathcal{F}' = f', f_2, \ldots, f_n$, gives us the desired mRNA. We extend this example in the following definition:

Definition 4 (Corresponding similarity functions). *Let $(G_\Gamma, \mathcal{F})$ be an instance of* MRSO *with $\mathcal{F} = f_1, \ldots, f_n$. Also, let $\phi : V \to \Sigma^3$ be a codon assignment for some $V \subseteq \mathbf{V}(G_\Gamma)$. The corresponding set of similarity functions of assignment ϕ, denoted $\mathcal{F}_\phi = f_1^\phi, \ldots, f_n^\phi$, is defined as follows:*

- *For all $i \in V : f_i^\phi(\phi(i)) = f_i(\phi(i))$, and $f_i^\phi(C) = -\infty$ for any $C \neq \phi(i)$.*
- *For all $j \in \mathbf{V}(G_\Gamma) - V : f_j^\phi = f_j$.*

Algorithm $\mathcal{A}_{\text{NEB}}$ uses $\mathcal{A}_{\text{OP}}$, the algorithm given in [2] for outerplanar implied structure graphs, as a subprocedure. At its core, $\mathcal{A}_{\text{NEB}}$ is basically an exhaustive search algorithm that searches through all possible codon assignments for vertices which are incident to edges in E_b. For each such assignment, $\mathcal{A}_{\text{NEB}}$ first checks if the assignment is compatible with respect to $G_\Gamma[E_b]$, and if so, it invokes $\mathcal{A}_{\text{OP}}$ with the set of similarity functions corresponding to this assignment. Finally, $\mathcal{A}_{\text{NEB}}$ outputs the maximum solution over all target mRNAs returned by $\mathcal{A}_{\text{OP}}$. A schematic description of $\mathcal{A}_{\text{NEB}}$ is given in Figure 1.

```
Algorithm $\mathcal{A}_{\text{NEB}}(G_\Gamma, \mathcal{F}, \mathcal{P})$
  Data   : An implied structure graph $G_\Gamma$ of order $n$, a set of similarity functions
           $\mathcal{F} = f_1, \ldots, f_n$ and a nice edge bipartition $\mathcal{P} = (E_t, E_b)$.
  Result : An optimal target mRNA sequence $t = t_1 t_2 \ldots t_n$ which is compatible
           with $G_\Gamma$.
  begin
      foreach possible codon assignment $\phi$ to vertices incident to edges in $E_b$ do
          if $\phi$ is compatible with respect to $G_\Gamma[\mathcal{E}_b]$ then
              $(a)$ Construct $\mathcal{F}_\phi$, the similarity functions corresponding to $\phi$.
              $(b)$ Invoke $A_{\text{OP}}(G_\Gamma[E_t], \mathcal{F}_\phi)$.
          end
      end
      return the target mRNA sequence found in Step (b) with the highest
      similarity score.
  end
```

Fig. 1. Algorithm $\mathcal{A}_{\text{NEB}}$.

Lemma 1. *Given an instance $(G_\Gamma, \mathcal{F})$ for* MRSO *accompanied by a nice edge bipartition $\mathcal{P} = (E_t, E_b)$ of G_Γ, $\mathcal{A}_{NEB}$ computes an optimal target mRNA sequence for this instance in $\mathcal{O}(64^{2\epsilon} n)$ time, where $n = |\mathbf{V}(G_\Gamma)|$ and $\epsilon = |E_b|$.*

Proof. Consider the schematic description of $\mathcal{A}_{\text{NEB}}$ in Figure 1. Any assignment enumerated in the algorithm is verified for compatibility with respect to $G_\Gamma[E_b]$.

Hence, by the correctness of $\mathcal{A}_{\mathrm{OP}}$, any target mRNA outputted by $\mathcal{A}_{\mathrm{NEB}}$ with a similarity score higher than $-\infty$ is compatible with respect to G_Γ. Furthermore, all possible codon assignments to vertices which are incident to edges in E_b are considered by $\mathcal{A}_{\mathrm{NEB}}$. Therefore, by the optimality of $\mathcal{A}_{\mathrm{OP}}$, this target mRNA must be optimal with respect to $\mathcal{F}$.

For the time complexity bound, consider any vertex in G_Γ. The number of possible codons assignments to this vertex is $|\Sigma^3| = 64$. Therefore, the number of assignments enumerated in the algorithm is bounded by $\mathcal{O}(64^{2\epsilon})$. Furthermore, constructing any such assignment and checking it for compatibility with respect to $G_\Gamma[E_b]$ can be done in $\mathcal{O}(n)$ time. Hence, since each call to $\mathcal{A}_{\mathrm{OP}}$ requires $\mathcal{O}(n)$ time, the overall time complexity of $\mathcal{A}_{\mathrm{NEB}}$ is bounded by $\mathcal{O}(64^{2\epsilon}n)$. □

We now return to our two parameters χ and δ, starting with χ. Recall that if $\chi = 0$ then G_Γ is outerplanar. Hence, a nice edge bipartition with χ bottom edges is available by definition. To see this, consider an edge bipartition with one bottom edge for each pair of edge crossings in G_Γ. Such an edge bipartition is nice, has at most χ bottom edges, and can be constructed in linear time. We therefore obtain the following proposition.

Proposition 1. MRSO *is polynomial-time solvable in case* $\chi = \mathcal{O}(\lg|\mathbf{V}(G_\Gamma)|)$.

Proof. According to the above discussion, G_Γ has a nice edge bipartition with at most χ bottom edges and this partitioning can be constructed in $\mathcal{O}(n)$ time. Thus, by Lemma 1, algorithm $\mathcal{A}_{\mathrm{NEB}}$ can be applied to solve MRSO in $\mathcal{O}(64^{2\delta}n)$ time, and so proposition above follows. □

Next consider parameter δ. Constructing a nice edge bipartition with δ bottom edges is immediate when considering the following easy lemma.

Lemma 2. *If G is a graph with maximum degree 2, then G is outerplanar.*

Proof. If G is a graph with maximum degree 2, then every component in G is either a path or a cycle. Since paths and cycles are outerplanar, the lemma immediately follows. □

Consider an edge bipartition of G_Γ such that for each degree three vertex $i \in \mathbf{V}(G_\Gamma)$, exactly one edge incident to i is a bottom edge. Clearly, such a bipartition has at most δ bottom edges and can be constructed in linear time. Let $\mathcal{P} = (E_t, E_b)$ be an edge bipartition obtained in this fashion. Since G_Γ is subcubic, every vertex is incident to at most two top edges in $\mathcal{P}$. Thus, by Lemma 2, $G[E_t]$ is outerplanar and $\mathcal{P}$ is nice.

Proposition 2. MRSO *is polynomial-time solvable in case* $\delta = \mathcal{O}(\lg|\mathbf{V}(G_\Gamma)|)$.

Proof. Replace δ with χ in the proof of Proposition 1. □

4 The cutwidth of G_Γ

Let $(G_\Gamma, \mathcal{F})$ be an instance of MRSO with $\mathbf{V}(G_\Gamma) = \{1, \ldots, n\}$. For $p \in \{1, \ldots, n-1\}$, the *p-cutwidth* of G_Γ is defined as the number of edges connecting vertices in $\{1, \ldots, p\}$ to vertices in $\{p+1, \ldots, n\}$. The *cutwidth* of G_Γ is defined as the maximum p-cutwidth over all $p \in \{1, \ldots, n-1\}$. In the following we consider the cutwidth of G_Γ as a parameter for MRSO. We begin by showing that the problem is polynomial-time solvable in case G_Γ has a cutwidth which is bounded by $\mathcal{O}(\lg n)$. Following this, we show that this result implies that MRSO is polynomial-time solvable for several other interesting cases. We let ψ denote the cutwidth of G_Γ throughout the section.

For obtaining our initial result, we present an algorithm which we call $\mathcal{A}_{\text{CUT}}$. This algorithm works by recursively partitioning G_Γ into two subgraphs $G_\Gamma[1, p]$ and $G_\Gamma[p+1, n]$, and then concatenating two optimal target mRNA sequences $T' = C_1, \ldots, C_p$ and $T'' = C_{p+1}, \ldots, C_n$ which are compatible with respect to these two subgraphs. To ensure that the concatenated solution $T = T'T''$ is also compatible with respect to G_Γ, the algorithm enumerates all codon assignments between connected vertices of the two subgraphs. In order to prevent unnecessary assignments from being enumerated, the algorithm distinguishes between vertices which were assigned a codon in a previous recursive step, and those which have yet been assigned one.

As in $\mathcal{A}_{\text{NEB}}$, algorithm $\mathcal{A}_{\text{CUT}}$ uses corresponding similarity functions (Definition 4) to enforce codon assignments. A similarity function f is *degenerate*, if there is some codon C such that $f(C) > -\infty$, and $f(C') = -\infty$ for any other codon $C' \in \Sigma^3$, $C' \neq C$. In $\mathcal{A}_{\text{CUT}}$, we use degenerate similarity functions both to recognize the assigned vertices along the recursion, and also to propagate their corresponding codon assignment. A schematic description of $\mathcal{A}_{\text{CUT}}$ is given in Figure 2.

Lemma 3. *Given an instance $(G_\Gamma, \mathcal{F})$ for* MRSO*, algorithm $\mathcal{A}_{CUT}$ computes an optimal target mRNA sequence for this instance in $\mathcal{O}(64^{2\psi} n)$ time, where $n = |\mathbf{V}(G_\Gamma)|$ and ψ is the natural cutwidth of G_Γ.*

Proof.

Corollary 1. MRSO *is polynomial-time solvable in case $\psi = \mathcal{O}(\lg |\mathbf{V}(G_\Gamma)|)$.*

We now consider the implications of corollary 1. The treewidth [] of a graph is a graph property that has been studied extensively in the literature. In particular [] (via []) showed that for graphs with n vertices, constant maximum degree, and constant treewidth, one can obtain an ordering of the vertices such that the linear graph under this ordering has cutwidth bounded by $\mathcal{O}(\lg n)$.

Corollary 2. MRSO *is polynomial-time solvable in case G_Γ has constant treewidth.*

In [], Bodlaender gives a list of several interesting graph classes which are subclasses of the class of constant treewidth graphs. We state a few of these classes in the following corollary.

```
Algorithm $\mathcal{A}_{\text{CUT}}(G_\Gamma, \mathcal{F})$
  Data   : An implied structure graph $G_\Gamma$ with $\mathbf{V}(G_\Gamma) = \{1, \ldots, n\}$, and a set of
           similarity functions $\mathcal{F} = f_1, \ldots, f_n$.
  Result : An optimal target mRNA sequence $T$ which is compatible with
           respect to $G_\Gamma$.
  begin
      1. if $\mathbf{E}(G_\Gamma) = \emptyset$ then return $T$ that maximizes $\mathcal{F}$.
      2. Select $p \in \{1, \ldots, n-1\}$ with maximum $p$-cutwidth.
      3. Let $E = \{\{i, j\} \in \mathbf{E}(G_\Gamma) \mid 1 \leq i \leq p,\ p+1 \leq j \leq n\}$ and
      $V = \{i \in \mathbf{V}(G_\Gamma) \mid \{i, j\} \in E\}$ be the vertices incident to $E$.
      4. Set $A = \{i \in V \mid f_i$ is degenerate$\}$.
      5. Define $\phi^A : A \to \Sigma^3$ such that $\phi^A(i) = C \Leftrightarrow f_i(C) > -\infty$.
      6. foreach possible codon assignment $\phi^{V-A} : V - A \to \Sigma^3$ do
          if $\phi = \phi^A \cup \phi^{V-A}$ is compatible with respect to $G_\Gamma[E]$ then
              (a) $T' \leftarrow \mathcal{A}_{\text{CUT}}(G_\Gamma[1, p], f_1^\phi, \ldots, f_p^\phi)$.
              (b) $T'' \leftarrow \mathcal{A}_{\text{CUT}}(G_\Gamma[p+1, n], f_{p+1}^\phi, \ldots, f_n^\phi)$.
          end
      end
      return the highest similarity scoring target mRNA sequence $T = T'T''$
      found in step 6.
  end
```

Fig. 2. Algorithm $\mathcal{A}_{\text{CUT}}$.

Corollary 3. MRSO *is polynomial-time solvable in case* G_Γ *is either a chordal graph, an interval graph, circular arc graph, or a* k*-outerplanar graph for any constant* k*.*

5 Planar implied structure graphs

Since for any fixed k, MRSO is polynomial-time solvable in case G_Γ is k-outerplanar, a natural question to ask is whether the problem is still tractable when the implied structure graph is planar. In this section we provide a negative answer to this question by proving that MRSO remains **NP**-hard even for a restrictive class of implied structure graphs.

Given a graph G, the *page-number* of G is the smallest partitioning of $\mathbf{E}(G)$ possible, such that each subset of edges in the partition forms an edge-induced outerplanar subgraph under the same vertex ordering. Clearly the page-number of an outerplanar graph is one. For planar graphs however, there are graphs with page-number four [12]. We show that the MRSO problem is **NP**-complete even for cases where the given implied structure graph has page number two.

Proposition 3. MRSO *is* **NP***-complete even when restricted to implied structure graphs with page-number two.*

Proof. We describe a reduction from the Maximum Independent Set problem, which is known to be **NP**-complete even when restricted to cubic planar bridgeless connected graphs [4]. The proof is a direct extension of the **APX**-completeness proof for MRSO given in [7].

Let an instance of the Maximum Independent Set problem be given by a cubic planar bridgeless connected graphs G of order n. According to [11], there exists a linear-time algorithm for finding a 2-page embedding of a cubic planar bridgeless graph, and hence there is no loss of generality in assuming that G is given in the form of a linear graph with page-number two. We now turn to defining the corresponding instance of the MRSO problem. The implied structure graph G_Γ is merely the input graph G and the set of similarity functions $f_i : \Sigma^3 \to \mathbb{Q}$, $1 \leq i \leq n$, is defined as follows:

$$\forall i,\ 1 \leq i \leq n, \quad f_i(t_{3i-2}t_{3i-1}t_{3i}) = \begin{cases} 1 & \text{if } t_{3i-2}t_{3i-1}t_{3i} = AAA \\ 0 & \text{otherwise} \end{cases}$$

Quoting [7], the idea of the reduction is simply to identify the set of vertices which are assigned to AAA in a solution for the corresponding instance of the MRSO problem, with an independent set in G. Correctness of the proof now follows directly from [7], Theorem 3. □

6 Parameterizing by the similarity score

We next turn to consider the score of the optimum solution as a parameter for MRSO. For this, we suggest a relaxation on the similarity functions of an MRSO instance. More specifically, we consider instances with similarity functions of the form $f_i : \Sigma^3 \to \mathbb{N}$. We call similarity functions of this sort *natural similarity functions*, and denote MRSO$_\mathbb{N}$ the MRSO problem restricted to instances with this type of similarity functions. Most of the interest in restrictive similarity functions stems from the following proposition.

Proposition 4. MRSO$_\mathbb{N}$ *is polynomial-time solvable in case the similarity score of the optimal solution is fixed.*

Proof. Let $(G_\Gamma, \mathcal{F})$ be an instance of MRSO$_\mathbb{N}$ and let κ denote the similarity score of the optimal target mRNA of this instance. Set $n = \mathbf{V}(G_\Gamma)|$. We may assume with out loss of generality that for all $1 \leq i \leq n$, $f_i(C) > 0$ for some codon $C \in \Sigma^3$. Otherwise, if there exists any function $f_i \in \mathcal{F}$ which fails to meet this requirement, we solve the sub-instance $(G'_\Gamma, \mathcal{F}')$ obtained by deleting i from G_Γ and f_i from $\mathcal{F}$. Any feasible solution for $(G'_\Gamma, \mathcal{F}')$ can then be extended to a feasible solution of the same score for the original instance since Γ has maximum degree one. We present an algorithm which searches for a target mRNA string T, by focusing on finding κ pairwise compatible codons with respect to G_Γ. The proof is divided into two separate parts depending on $\alpha(G_\Gamma)$, the cardinality of a maximum independent set in G_Γ.

Suppose $\kappa \leq \alpha(G_\Gamma)$. Let $V \subseteq \mathbf{V}(G_\Gamma)$ be an independent set of size κ in G_Γ. Since G_Γ is at most cubic, such a subset V can be found in $\mathcal{O}(4^\kappa n)$ time using the bounded search tree technique []. We define a string T of length $3n$

as follows. For each $i \in V$, assign codon $C_i \in \Sigma^3$ such that $f_i(C_i) \geq 1$. This is always possible since V is an independent set in G_Γ, and since for all $1 \leq i \leq n$, $f_i(C) > 0$ for some $C \in \Sigma^3$. For each $j \in \mathbf{V}(G_\Gamma) - V$, assign codon C_j which is compatible with all codons assigned to vertices in V with respect to G_Γ. Again this is always possible since Γ has maximum degree one. We check at once that $T = C_1 C_2 \dots C_n$ is compatible with respect to G_Γ and $\sum_{i=1}^{n} f_i(C_i) \geq |V| = \kappa$.

Now suppose $\kappa > \alpha(G_\Gamma)$. Since G_Γ is at most cubic, we have $\alpha(G_\Gamma) \geq \frac{n}{4}$, and hence $\kappa > \frac{n}{4}$. Here, the algorithm is by direct enumeration. More precisely, the algorithm tries in turn to obtain a solution mRNA string T by finding ℓ pairwise compatible codons, where ℓ ranges from 1 to κ. So, let $\ell \in \{1, 2, \dots, \kappa\}$. We search through all ℓ-subsets of $\mathbf{V}(G_\Gamma)$ for an ℓ-subset with an assignment which is compatible with respect to G_Γ. Such an exhaustive search can be executed in $\mathcal{O}(\binom{n}{\ell} 64^\ell)$ time. Summing-up over ℓ and neglecting the time to check $\kappa > \alpha(G_\Gamma)$, *i.e.*, $\mathcal{O}(4^\kappa)$, we obtain $\mathcal{O}(\sum_{\ell=1}^{\kappa} \binom{n}{\ell} 64^\ell)$, which is $\mathcal{O}(2^{\mathcal{O}(\kappa)} \kappa^{\kappa+1})$ since G_Γ is at most cubic and $\kappa > \alpha(G_\Gamma) \geq \frac{n}{4}$.

Hence, $\mathrm{MRSO}_{\mathbb{N}}$ can be solved in $\mathcal{O}(2^{\mathcal{O}(\kappa)} \kappa^{\kappa+1} + 4^\kappa n)$ time, and the proposition above follows. □

Note that all hardness results obtained for MRSO still hold for MRSO under natural similarity functions. Nevertheless, using a simple combinatorial argument, we can easily obtain an optimal algorithm if we consider the score of the optimal solution for $\mathrm{MRSO}_{\mathbb{N}}$ to be fixed. Even so, it is a challenging problem to investigate the parameterized complexity of the MRSO problem for more general similarity functions. We do believe that it might be worth considering similarity functions of the form $f_i : \Sigma^3 \to \mathbb{N} \cup \{-\infty\}$ since these capture most of the information necessary in most practical applications. Here, the $-\infty$ value can be used in case a certain codon (*e.g.* a stop codon) is not acceptable in a certain position of T.

Acknowledgments

The authors would like to thank Gad Landau for his support and valuable advice.

References

1. R. Backofen and A. Busch. Computational design of new and recombinant selenoproteins. In *Proc. of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM)*, volume 3109 of *LNCS*, pages 270–284, 2004.
2. R. Backofen, N.S. Narayanaswamy, and F. Swidan. On the complexity of protein similarity search under mRNA structure constraints. In *Proc. of the 19th Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 2285 of *LNCS*, pages 274–286, 2002.
3. R. Backofen, N.S. Narayanaswamy, and F. Swidan. Protein similarity search under mRNA structural constraints: application to targeted selenocystein insertion. *In Silico Biology*, 2(3):275–290, 2002.

4. T.C. Biedl, G. Kant, and M. Kaufmann. On triangulating planar graphs under the four-connectivity constraints. *Algorithmica*, 19:427–446, 1997.
5. A. Böch, K. Forchhammer, J. Heider, and C. Baron. Selenoprotein synthesis: a review. *Trends in Biochemical Sciences*, 16(2):463–467, 1991.
6. H.L. Bodlaender, R.G. Downey, M.R. Fellows, M.T. Hallett, and H.T. Wareham. Parameterized complexity analysis in computational biology. *Computer Applications in the Biosciences*, 11:49–57, 1995.
7. D. Bongartz. Some notes on the complexity of protein similarity search under mRNA structure constraints. In *Proc. of the 30th Conference on Current Trends in Theory and Practice of Computer Science (SOFSEM)*, volume 2932 of *LNCS*, pages 174–183, 2004.
8. R. Downey and M. Fellows. *Parameterized Complexity.* Springer-Verlag, 1999.
9. T. Jacks, M. Power F. Masiarz, P. Luciw, P. Barr, and H. Varmus. Characterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature*, 331:280–283, 1988.
10. T. Jacks and H. Varmus. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. *Science*, 230:1237–1242, 1985.
11. G. Lin, Z-Z. Chen, T. Jiang, and J. Wen. The longest common subsequence problem for sequences with nested arc annotations. *Journal of Computer and System Sciences*, 65(3):465–480, 2002. Special issue on computational biology.
12. M. Yannakakis. Embedding planar graphs in four pages. *Journal of Computer and System Sciences*, 38:36–67, 1986.