Fixed-Parameter Algorithms for Protein Similarity Search Under mRNA Structure Constraints

Guillaume Blin, Guillaume Fertin, Danny Hermelin, Stéphane Vialette

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 31905 - 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 [3]. Since the MRSO problem is known to be **APX**-hard [8], Bongartz proposed in [8] to attack the problem using the concept of parameterized complexity. We prove in this paper that the MRSO problem is fixed-parameter tractable parameterized by the number of degree 3 vertices or by the number of crossing edges in the implied structure graph. This latter result answers an open problem posed in [8]. Aiming at precisely defining the complexity landscape of the problem, we refine the **NP**-hardness result of [3] and complement this result by showing that the MRSO problem is fixed-parameter tractable parameterized by an additional parameter. Finally, we present a fixed parameter algorithm parameterized by the similarity score in a restrictive model.

Key Words: Computational biology, Parameterized Complexity, Protein Engineering.

1 Introduction

Backofen *et al.* introduced in [3, 4] 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 (see formal definition in the following section). The initial motivation of the MRSO problem is concerned with selenocystein insertion, *i.e.*, generating new amino acid sequences containing selenocystein. Selenocystein is a rare amino acid which was discovered as the 21th amino acid [6]. Proteins containing selenocystein are called selenoproteins. It has been shown [6] that in case of selenocystein, termination of translation is inhibited in the presence of a specific mRNA sequence in the 3′-region after the *UGA* codon that forms a hairpin like structure. It is argued in [3] that modifying existing proteins such that selenocystein is incorporated instead of a catalytic cystein is an important problem for catalytic activity enhancement and X-ray crystallography.

Observe that selenocystein insertion is concerned with secondary structures without pseudo-knots, and hence the linear-time algorithm presented in [3] 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 [14, 13]. This motivates the investigation of the MRSO problem for more elaborate secondary structures [3, 8], *i.e.*, secondary structures that contain pseudo-knots. We mention also that an extension of the MRSO problem, where insertions and deletions are allowed in the amino acid sequence, is presented in [2].

For the MRSO problem, it has been shown in [3] that there exists a linear-time algorithm if the considered secondary structure corresponds to an outer-planar graph (as it is the case for selenoproteins). 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 [3] and Bongartz showed recently that the problem is in fact **APX**-hard [8]. An algorithm

for approximating the MRSO problem within ratio 2 is proposed in [3]. A slightly slower but somewhat simpler algorithm for approximating the MRSO problem within ratio 4 is proposed in [8]. However, both approximation algorithms presented in [3] and in [8] suffer from the same shortcoming, as they both assume that any feasible solution has a non-negative score. Thus the optimal solution outputted by these algorithms might be biologically irrelevant, as one can not prohibit, for instance, stop codons from appearing in unwanted positions in the solution mRNA.

The computational challenge posed by **NP**-hard problems has inspired the development of a wide range of algorithmic techniques. Since the MRSO problem for general implied structure graphs is known to be **APX**-hard [8], Bongartz proposes in [8] to attack the problem using the concept of parameterized complexity. Parameterized complexity [10] 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. The parameterized problems that have algorithms of $f(k)\ n^{O(1)}$ time complexity are called *fixed-parameter tractable*, where k is the parameter, f can be an arbitrary function depending only on k, and n denotes the overall input size. We designate the class of fixed-parameter tractable problems **FPT**. In the last decade, parameterized complexity has proved to be useful in computational biology, see for example [7, 11, 1]. The best general reference here is [10].

This paper is organized as follows. After presenting some preliminaries in Section 2, we present in Section 3 fixed-parameter algorithms for two natural parameters, namely the number of degree 3 vertices and the number of crossing edges in the implied structure graph. In Section 4, we refine the **NP**-completeness result of [3] and propose a fixed-parameter algorithm parametrized by the cutwidth of a given nice edge bipartition (see Definition 3) of the implied structure graph. Finally, we present in Section 5 a fixed parameter algorithm parameterized by the similarity score in a restrictive model. Due to space constraints, several details and proofs are not presented in this paper.

2 Preliminaries

In the following section we briefly introduce notations and terminology used throughout the paper. We begin by introducing standard graph theory terminology. All graphs considered in this paper are *simple*. For a graph G we denote $\mathbf{V}(G)$ as the set of vertices in G and $\mathbf{E}(G)$ as the set of edges in G. Given a graph G and a subset $V' \subseteq \mathbf{V}(G)$ of the vertices in G, the *induced subgraph* $G[V']$ is a graph with $\mathbf{V}(G[V']) = V'$ and $\mathbf{E}(G[V']) = \{uv \in \mathbf{E}(G) : u, v \in V'\}$. Given a subset $E' \subseteq \mathbf{E}(G)$ of the edges in G, the *edge-induced subgraph* $G[E']$ is a graph with $\mathbf{E}(G[E']) = E'$ and $\mathbf{V}(G[E']) = \{v \in e : e \in E'\}$. A *linear graph* with n vertices is a vertex-labeled graph, where each vertex is labeled by a distinct label from $1, 2, \ldots, n$. Given any linear graph G, we will always assume that the labeling of G is implied in the indexing of $\mathbf{V}(G)$, *i.e.*, that v_i is the vertex with label i in $\mathbf{V}(G)$. A *linear embedding* of a graph G with n vertices is a one-to-one labeling from $\mathbf{V}(G)$ onto $1, \ldots, n$. Consequently, this labeling defines a linear ordering on the set of vertices, and can be geometrically interpreted as an embedding of the vertices on a straight line. Accordingly, an *interval* $V_{i,j}$ of a linear graph G is defined as the subset of consecutive vertices $\{v_i, v_{i+1}, \ldots, v_j\} \subseteq \mathbf{V}(G)$. Let $e_1 = v_i v_j$ and $e_2 = v_x v_y$ be two edges of an arbitrary linear graph G, such that $i < x$. We say that e_1 and e_2 *cross*, if $j < y$, and this is denoted by $e_1 \between e_2$. Now, given any graph G, an *outer-planar linear embedding* (or simply outer-planar embedding) of G is a labeling from $1, \ldots, n$ onto $\mathbf{V}(G)$, such that G with this labeling has no crossing edges. A graph with such an embedding is called *outer-planar*. It is well-known that, given a graph G, one can test in linear time if G is outer-planar, and if so, find a linear embedding without crossing edges [19, 18]. Hence, we may always assume that the outer-planar labeling of G is implied in the indexing of $\mathbf{V}(G)$ when G is outer-planar. For additional graph-theoretical notions, we use the standard notations given in [9].

We now introduce notations for describing and comparing mRNA sequences. Let $\Sigma = \{A, C, G, U\}$ be an alphabet representing the four possible nucleotides in an mRNA molecule. The pairings $\{A, U\}$

and $\{C, G\}$ are referred to as *complementary nucleotide pairs*, although all results in this paper can be extended to consider additional pairings, *e.g.* the non-standard $\{G, U\}$ pairing. Furthermore, as in most natural models, we assume hydrogen bonds can occur between any complementary nucleotides which are at least three positions apart in the mRNA sequence. A *codon* is a string of length 3 over Σ. A mRNA sequence is a concatenation of n codons, or in other words, a string in Σ^{3n}. Let $S = s_1 \dots s_{3n}$ be an mRNA sequence. Codon i, $1 \leq i \leq n$, of S is simply the substring of length 3 ending at position $3i$ in S, denoted $C_i^S = s_{3i-2}s_{3i-1}s_{3i}$. Now, suppose we wish to compute the similarity between another mRNA sequence $T = t_1 \dots t_{3n}$ and S, and we wish to do so codon-wise. For this we can provide a set of n functions, $\mathcal{F} = f_1, \dots, f_n$, 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 codon C_i^T according to its level of similarity in comparison with codon C_i^S. The total level of similarity between S and T is then defined by:

$$\text{sim}(S, T) = \sum_{i=1}^{n} f_i(C_i^T) = \sum_{i=1}^{n} f_i(t_{3i-2}, t_{3i-1}, t_{3i})$$

Notice, that given a set of similarity functions $\mathcal{F} = f_1, \dots, f_n$ for S, one does not need to know anything else about S in order to compute the similarity score of S and T. Now consider an arbitrary mRNA sequence S of length $3n$. The *secondary structure* of S, denoted $\mathcal{S}[S] \subseteq \{\{i, j\} : 1 \leq i < j \leq 3n\}$, is a set of pairings between distinct integers in $\{1, 2, \dots, 3n\}$, which represent hydrogen bonds in the folding of S. Since in our model, we assume that each nucleotide can pair with at most one other nucleotide in any folding, we assume that each integer appears in at most one pair in $\mathcal{S}[S]$. Furthermore, there are no pairs of the form $\{i, i+1\}$ or $\{i, i+2\}$ in $\mathcal{S}[S]$, for all $1 \leq i \leq 3n-2$. Consequently, the *structure graph* of S is the linear graph Γ with maximum degree 1, such that $\mathbf{V}(\Gamma) = \{u_1, u_2, \dots, u_{3n}\}$ and $\mathbf{E}(\Gamma) = \{u_i u_j : \{i, j\} \in \mathcal{S}[S]\}$ (see Figure 1). Let Γ be an arbitrary structure graph with $\mathbf{V}(\Gamma) = \{u_1, \dots, u_{3n}\}$, and let $T = t_1, \dots, t_{3n}$ be an arbitrary mRNA sequence. 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 $u_i u_j \notin \mathbf{E}(\Gamma)$. The entire sequence T is compatible with respect to Γ, if any pair of nucleotides t_i and t_j in T, $1 \leq i < j \leq 3n$, is compatible with respect to Γ. We are now in position to give a more formal definition of the mRNA Structure Optimization (MRSO) problem.

Definition 1. *Let $\mathcal{F}$ be a set of n similarity functions for a source mRNA sequence of length $3n$, and let Γ be a structure graph with $\mathbf{V}(\Gamma) = \{u_1, u_2, \dots, u_{3n}\}$. 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 $\mathcal{F}$.*

Thus, MRSO can be thought of as the problem of assigning nucleotides (or letters from $\{A, C, G, U\}$) to vertices in Γ, with the requirement that the nucleotides assigned are complementary according to Γ, and also with the requirement that such an assignment achieves the maximal similarity score with respect to $\mathcal{F}$. Since we are really interested in codon-wise similarity, we use a more convenient representation of a structure graph, namely the *implied structure graph* as defined in [3].

Definition 2. *Let Γ be a structure graph with $\mathbf{V}(\Gamma) = \{u_1, u_2, \dots, u_{3n}\}$. The implied structure graph of Γ, denoted G_Γ, is the linear graph with:*

$$\begin{aligned}\mathbf{V}(G_\Gamma) &= \{v_1, v_2, \dots, v_n\} \\ \mathbf{E}(G_\Gamma) &= \{v_x v_y : \exists u_i u_j \in \mathbf{E}(\Gamma),\ i \in \{3x-2, 3x-1, 3x\} \text{ and } \\ &\qquad\qquad j \in \{3y-2, 3y-1, 3y\}\}\end{aligned}$$

Thus, the implied structure graph may be thought of as a compressed version of a structure graph, where three consecutive vertices in the structure graph are block-wise joined into one compressed vertex, which correspondingly represents a codon or a triplet of nucleotides (see Figure 1). Notice that this

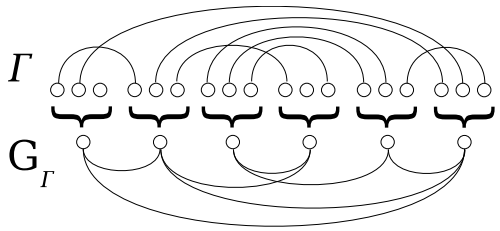

Fig. 1. A structure graph Γ compressed into its implied structure graph G_Γ.

compression is not lossless, as there might be multiple (at most three) edges in G which correspond to one edge in G_Γ (but no edge can connect two nucleotides in a single codon). Nevertheless, we may assume that the edges in the implied structure graph G_Γ are labeled with the information needed to deduce the complementary constraints defined by the edges of Γ. As this labeling will require only a constant increase in space, we assume that it is implicitly existent, without ever referring to it throughout the paper. This allows us to consider from hereafter, only implied structure graphs, and to speak of compatible codons with respect to G_Γ as opposed to compatible nucleotides with respect to Γ. Accordingly, we say that a pair of codons $C_i, C_j \in \Sigma^3$, assigned to vertices v_i and v_j respectively, is compatible with respect to G_Γ, if any pair of nucleotides $t \in C_i$ and $t' \in C_j$ are compatible with respect to Γ. Furthermore, we hereafter think of MRSO as a codon assignment problem, and consider instances for the problem of the form $(G_\Gamma, \mathcal{F})$, where G_Γ is an implied structure graph, and Γ is a structure graph which can always be inferred from G_Γ. This will prove very useful in avoiding heavy notations and improving the overall brevity of the expose.

3 Two natural parameters for MRSO

We begin the discussion in this paper by considering two natural parameters for MRSO. Namely, we consider the number of degree 3 vertices in the implied structure graph G_Γ, and the number of crossing edges in G_Γ. We denote both these parameters throughout this section by δ and χ respectively, *i.e.*, $\delta = |\{v \in \mathbf{V}(G_\Gamma) : d(v) = 3\}|$ and $\chi = |\{\{e_1, e_2\} : e_1, e_2 \in \mathbf{E}(G_\Gamma) \wedge e_1 \between e_2\}$. As it turns out, MRSO is in **FPT** when considering either one of these two parameters to be fixed (one of these two results answers an open problem posed in [8]). To show this, we will first describe a general algorithm, and later demonstrate how it can be applied for both cases. Before describing our algorithm, we first introduce the following central definition, which will prove essential throughout the entire paper.

Definition 3 (Edge bipartition and nice edge bipartition of G_Γ). *Let G_Γ be an implied structure graph with n vertices. An edge partition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of G_Γ is a partitioning of the edges in G_Γ into $\mathcal{E}_t$ and $\mathcal{E}_b$, such that $\mathcal{E}_t \cup \mathcal{E}_b = \mathbf{E}(G_\Gamma)$, $\mathcal{E}_t \cap \mathcal{E}_b = \emptyset$ and $\mathcal{E}_t \neq \emptyset$. $\mathcal{P}$ is said to be nice, if the subgraph $G_\Gamma[\mathcal{E}_t]$ is outer-planar.*

A convenient graphical representation of an edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of some implied structure graph G_Γ, is obtained by drawing the vertices on a line according to their linear embedding, and then drawing the edges in $\mathcal{E}_t$ and $\mathcal{E}_b$ above and below this line respectively (see Figure 3). Thus we will often refer to $\mathcal{E}_t$ as the *top edges* of $\mathcal{P}$, and to $\mathcal{E}_b$ as the *bottom edges* of $\mathcal{P}$. We now turn to describe our general algorithm which we call $\mathcal{A}_{\text{NEB}}$. $\mathcal{A}_{\text{NEB}}$ will apply only for cases when a nice edge bipartition of G_Γ with a fixed number of vertices incident to bottom edges is known in advance. Indeed, such an assumption may be a bit unrealistic since not all graphs have such an edge partitioning, and also since finding a nice edge bipartition with a minimal number of vertices incident to bottom edges can be a daunting task. Nonetheless, following the description of $\mathcal{A}_{\text{NEB}}$, we show that when considering either δ or χ to be fixed, one can obtain such a bipartition with very little effort.

$\mathcal{A}_{\text{NEB}}$ uses the algorithm described in [3] for the case when G_Γ is outer-planar, *i.e.*, algorithm $\mathcal{A}_{\text{OP}}$, as a sub-procedure. Recall that $\mathcal{A}_{\text{OP}}$ solves MRSO for this case in linear time. Furthermore, it is worth pointing out that $\mathcal{A}_{\text{OP}}$ can be applied even if G_Γ is outer-planar and its natural linear embedding (*i.e.* natural codon order) contains crossing edges. Now, notice that $\mathcal{A}_{\text{OP}}$ could easily be modified to find an optimal mRNA target sequence with a prespecified subset of codons. For example, given an instance $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ for MRSO, we can modify $\mathcal{A}_{\text{OP}}$ to compute the optimal mRNA sequence starting with codon AAA. For this, we simply replace our original instance with a new instance $(G_\Gamma, \mathcal{F}' = f'_1, \ldots, f'_n)$ such that $f'_1(AAA) = f_1(AAA)$, $f'_1(C_1) = -\infty$ for any codon $C_1 \neq AAA$, and $f'_i = f_i$ for all $2 \leq i \leq n$. To extend this example, in the following definition we consider any *codon assignment* for vertices in $\mathbf{V}(G_\Gamma)$, *i.e.*, any function ϕ of the form $\phi : V' \rightarrow \Sigma^3$, where V' is any subset of vertices in $\mathbf{V}(G_\Gamma)$.

Definition 4. *Let $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ be an instance of* MRSO*, and let $V' \subseteq \mathbf{V}(G_\Gamma)$. Also, let $\phi : V' \rightarrow \Sigma^3$ be a codon assignment for V'. The corresponding set of similarity functions of assignment ϕ, denoted $\mathcal{F}_\phi = f_1^\phi, \ldots, f_n^\phi$, is defined as follows: (i) $f_i^\phi = f_i$ for all i such that $v_i \in \mathbf{V}(G_\Gamma) - V'$ and (ii) $f_i^\phi(\phi(v_i)) = f_i(\phi(v_i))$, $f_i^\phi(C_i) = -\infty$ for any $C_i \neq \phi(v_i)$, for all i such that $v_i \in V'$.*

Notice that given an assignment ϕ and a set of n similarity functions $\mathcal{F}$, one can easily generate $\mathcal{F}_\phi$ in $\mathcal{O}(n)$ time. Now, given an instance $(G_\Gamma, \mathcal{F})$ of MRSO, and a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of G_Γ, let υ denote the number of vertices in G_Γ incident to edges in $\mathcal{E}_b$, *i.e.*, $\upsilon = |\mathbf{V}(G_\Gamma[\mathcal{E}_b])|$. At its core, $\mathcal{A}_{\text{NEB}}$ is basically an exhaustive search procedure that searches through all possible codon assignments to vertices incident to edges in $\mathcal{E}_b$. For each such assignment, $\mathcal{A}_{\text{NEB}}$ first checks if the assignment is compatible with respect to the edge-induced subgraph $G_\Gamma[\mathcal{E}_b]$, 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 solutions returned by $\mathcal{A}_{\text{OP}}$. A more schematic description of $\mathcal{A}_{\text{NEB}}$ is given in Figure 2.

Algorithm $\mathcal{A}_{\text{NEB}}(G_\Gamma, \mathcal{F}, \mathcal{P})$

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

Fig. 2. Algorthim $\mathcal{A}_{\text{NEB}}$.

Lemma 1. *Given an instance $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ for* MRSO *accompanied by a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of G_Γ, algorithm $\mathcal{A}_{NEB}$ computes an optimal target mRNA sequence for this instance in $\mathcal{O}(64^\upsilon n)$ time, where $\upsilon = |\{v \in e : e \in \mathcal{E}_b\}|$.*

We now consider parameter δ, the number of degree 3 vertices in the implied structure graph G_Γ. Recall that a vertex with degree 3 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. Thus, we expect that in most real-life applications, the number of degree 3 vertices in G_Γ will be relatively small in comparison with the total length of the source mRNA sequence.

Let $(G_\Gamma, \mathcal{F})$ be an instance of MRSO, and let $V' = \{v \in \mathbf{V}(G_\Gamma) : d(v) = 3\}$ be the degree 3 vertices in G_Γ. In the following we show that a nice edge bipartition of G_Γ can easily be obtained with at most $\delta = |V'|$ bottom edges. This is easily established when considering the following folklore lemma.

Lemma 2. *If G is a graph with maximum degree 2, then G is outer-planar.*

Now, consider an edge bipartition of G_Γ such that for each vertex $v \in V'$, at most one edge incident to v is a bottom edge. Clearly, such a bipartition with at most δ bottom edges exists and can be found by a simple scan of the vertices in G_Γ. Let $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ be an edge bipartition obtained in this fashion. Now, since G_Γ is at most cubic, every vertex is incident to at most two top edges in $\mathcal{P}$. Thus, by Lemma 2, $G[\mathcal{E}_t]$ is outer-planar and so one can find in linear-time, a linear embedding of G_Γ such that $\mathcal{P}$ is nice. Considering all this, we state the following proposition.

Proposition 1. *The* MRSO *problem is in* **FPT** *for parameter* $\delta = |\{v \in \mathbf{V}(G_\Gamma) : d(v) = 3\}|$.

Proof. According to the above discussion, G_Γ has a nice edge bipartition with at most δ bottom edges and this partitioning can be found in linear time. Thus, by Lemma 1, algorithm $\mathcal{A}_{\text{NEB}}$ can be applied to find an optimal solution in $\mathcal{O}(64^{2\delta}n)$ time, and so the above proposition holds. □

We now consider parameter χ. Recall that χ denotes the number of crossing edges in G_Γ. Also recall that finding an optimal solution of MRSO when parameterized by this number was posed as an open problem in [8]. Indeed, in most real life applications this parameter can be considered fixed, as most known secondary structures of real mRNA molecules consist of a very small number of pseudo-knots.

As in the case of parameter δ, a nice edge bipartition with χ bottom edges can easily be obtained in this case as well. Simply consider an edge bipartition with one bottom edge for each pair of crossing edges in G_Γ. Clearly such an edge bipartition is nice, has χ bottom edges, and can be obtained in linear time with respect to the number of vertices in G_Γ. Thus we state the following proposition.

Proposition 2. *The* MRSO *problem is in* **FPT** *for parameter* $\chi = |\{\{e_1, e_2\} : e_1, e_2 \in \mathbf{E}(G_\Gamma) \ \wedge \ e_1 \between e_2\}|$.

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

4 On the borderline between NP-completeness and P

We now explore the borderline between instances of MRSO which are tractable and instances which are intractable. We already know that in general, the MRSO problem is **NP**-complete [3], so we know there exist instances for which the problem is intractable. Furthermore, also due to [3], we know that if the implied structure graph is outer-planar, then the problem is tractable. Thus, in the following, we aim to refine this border by suggesting a tighter **NP**-completeness result, accompanied by a parameterized algorithm for a new parameter later on introduced.

Since the MRSO problem can be solved in optimal time for instances with outer-planar implied structure graphs, one can ask if the problem is still tractable when the implied structure graph is, for instance, planar. As we shall soon see, the answer to this question is negative, even for restrictive subclasses of planar graphs. More specifically, we prove in the following that the MRSO problem is **NP**-complete, even when the *page-number* of the implied structure graph is 2. The page-number of a given graph G, is the smallest partitioning of $\mathbf{E}(G)$ possible, such that each subset of edges in the partition forms an edge-induced outer-planar subgraph under the same linear embedding. Clearly the page-number of an outer-planar graph is 1. For planar graphs however, this number is bounded by 4, and furthermore, there exist examples of planar graphs which achieve this number [21]. A very interesting related result by Heath states that the edges of a planar graph can be partitioned into two parts, each inducing an outer planar graph [12]. Our proof is a direct extension of the **APX**-completeness proof provided in [8] for MRSO.

Proposition 3. *The* MRSO *problem is* **NP***-complete when restricted to implied structure graphs with page-number* 2*.*

Proposition 3 gives a tight description of **NP**-hard instances for MRSO. We now complement this result by giving a tight description of polynomial time solvable instances of MRSO. For this, we consider a new parameter for the MRSO problem. Consider any linear graph G with n vertices. The *cutwidth* of G is defined as the number $\max_{1\leq i\leq n}|\{v_xv_y \in \mathbf{E}(G) : x < i < y\}|$, *i.e.*, the maximum number of edges in $\mathbf{E}(G_\Gamma)$ cut by any position i in $1,\ldots,n$. Let $(G_\Gamma,\mathcal{F})$ be an instance of MRSO, and let $\mathcal{P} = \{\mathcal{E}_t,\mathcal{E}_b\}$ be a nice edge bipartition of G_Γ given with this instance. The cutwidth of $\mathcal{P}$ is defined as the cutwidth of the edge-induced linear subgraph $G_\Gamma[\mathcal{E}_b]$, and this value is denoted by ψ (for example see Figure 3). In the following we present a method for computing the optimal target mRNA sequence in polynomial time, in cases where a nice edge bipartition of G_Γ, with fixed cutwidth, is known in advance. Although the main motivation for this result is theoretical, it is conceivable to assume that it can be applied to practical applications as well, perhaps in conjunction with the results described in the previous section.

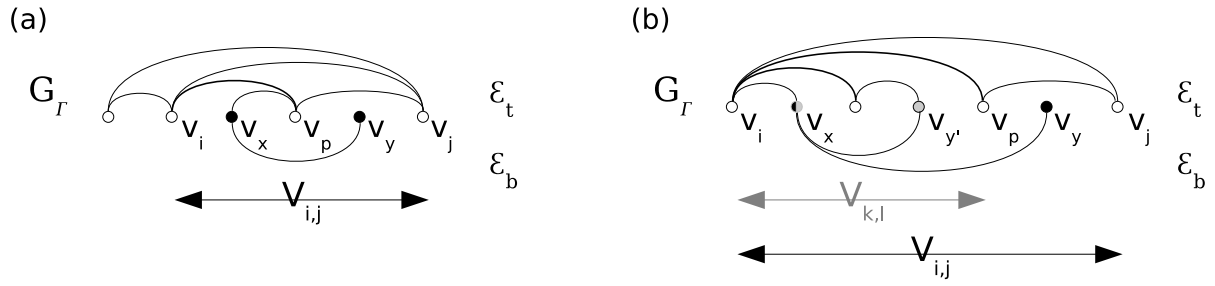

Fig. 3. Two nice edge bipartitions of G_Γ with cutwidth 1 (a) and 2 (b). In (a), edge v_xv_y is cut by interval $V_{i,j}$ with respect to $\mathcal{P}$. In (b), edge v_xv_y is cut by interval $V_{i,j}$ with respect to $\mathcal{P}$, and edge $v_xv_{y'}$ is cut by interval $V_{k,l}$ with respect to $\mathcal{P}$.

We begin by giving a general description of algorithm $\mathcal{A}_{\mathrm{OP}}$. Let $(G_\Gamma,\mathcal{F} = f_1,\ldots,f_n)$ be an instance of MRSO such that G_Γ is outer-planar. Throughout the following, we associate with any interval $V_{i,j} = \{v_i,v_{i+1},\ldots,v_j\}$ of G_Γ, the subproblem defined by the instance $(G_\Gamma[V_{i,j}],\mathcal{F}_{i,j} = f_i,\ldots,f_j)$. Thus $V_{1,n} = \mathbf{V}(G_\Gamma)$ for example, is associated with the original instance $(G_\Gamma,\mathcal{F})$. For our purposes, it suffices to describe $\mathcal{A}_{\mathrm{OP}}$ as a recursive algorithm which solves the subproblem associated with $V_{i,j}$, by solving the two subproblems associated with $V_{i,p}$ and $V_{p,j}$. Intervals $V_{i,i+1}$ of G_Γ, $1 \leq i < n$, are the base cases of this recursion, and their values are computed as the maximum value $f_i(C_i) + f_{i+1}(C_{i+1})$, over all codons $C_i,C_{i+1} \in \Sigma^3$, such that C_i and C_{i+1}, assigned to vertices v_i and v_{i+1} respectively, are compatible with respect to G_Γ. The *cutting point* of interval $V_{i,j}$, *i.e.*, index p, is determined by the edges outgoing from v_i into $V_{i,j}$. If there is no edge between v_i and a vertex in $V_{i,j}$, then $p = i+1$. Otherwise p is the largest index in $i+1,\ldots,j-1$ such that v_iv_p is an edge in $\mathbf{E}(G_\Gamma)$. Since G_Γ is outer-planar, there are no edges which connect vertices in $V_{i,p}$ to vertices in $V_{p,j}$. Thus, using a simple inductive argument, the authors in [3] proved that the recursive computation of $\mathcal{A}_{\mathrm{OP}}$ yields a feasible optimal solution. Furthermore, since G_Γ is outer-planar, the number of subproblems is bounded by $\mathcal{O}(n)$, and $\mathcal{A}_{\mathrm{OP}}$ can be implemented to run in $\mathcal{O}(n)$ time [3].

We now turn to describe our suggested modified version of $\mathcal{A}_{\mathrm{OP}}$ which we call $\mathcal{A}^*_{\mathrm{OP}}$. Let $(G_\Gamma,\mathcal{F})$ be an instance of MRSO accompanied by a nice edge bipartition $\mathcal{P}$ of G_Γ with cutwidth ψ. Algorithm $\mathcal{A}^*_{\mathrm{OP}}$ computes the optimum solution for the outer-planar edge-induced subgraph $G_\Gamma[\mathcal{E}_t]$, in a similar recursive fashion as $\mathcal{A}_{\mathrm{OP}}$, while simultaneously considering all edges in $\mathcal{E}_b$. Thus, when computing a subproblem defined by interval $V_{i,j}$ for example, we modify $\mathcal{A}_{\mathrm{OP}}$ to consider possible codon assignments for vertices

in $V_{i,j}$ which are connected in $\mathcal{E}_b$ and are about to be separated into the two subproblems defined by intervals $V_{i,p}$ and $V_{p,j}$. We develop this idea in the following.

Consider any interval $V_{i,j} \subseteq \mathbf{V}(G_\Gamma)$, $i+1 < j$, and suppose p is its cutting point in $G_\Gamma[\mathcal{E}_t]$, *i.e.*, p is the largest index in $i+1, \ldots, j-1$ such that $v_i v_p$ is an edge in $\mathcal{E}_t$, or $p = i+1$ if no such edge exists. Also, assume there exists an edge $v_x v_y \in \mathcal{E}_b$ such that $i \leq x < p < y \leq j$. Splitting $V_{i,j}$ into two intervals $V_{i,p}$ and $V_{p,j}$ will result in cutting the edge $v_x v_y$ and splitting v_x and v_y into two separate subproblems. Thus, we say that interval $V_{i,j}$ *cuts* edge $v_x v_y$ with respect to $\mathcal{P}$.

Let us begin by considering a simple case where $\mathcal{E}_b$ has only one edge $v_x v_y$ which is cut by interval $V_{i,j}$ with cutting point p in $G_\Gamma[\mathcal{E}_t]$ (see Figure 3 (a)). Suppose we wish to compute the optimal score for the subproblem associated with $V_{i,j}$. We can do this by enumerating all possible compatible codon assignments for vertices v_x and v_y with respect to G_Γ, and then computing the maximum score for subproblem $V_{i,j}$ with a set of similarity functions corresponding to each such assignment. In other words, using the notations in Definition 4, we compute the maximum score for subproblem $V_{i,j}$ over all sets of similarity functions $\mathcal{F}_\phi$, for any codon assignment ϕ of the form $\phi : \{v_x, v_y\} \to \Sigma^3$ such that $\phi(v_x)$ and $\phi(v_y)$ are compatible with respect to G_Γ. Clearly, a solution computed in this fashion is guaranteed to be feasible and optimal.

Now let us extend our example to the case where $\mathcal{P}$ has two bottom edges, $v_x v_y, v_x v_{y'} \in \mathcal{E}_b$, such that $y' < y$. Suppose interval $V_{i,j}$ of G_Γ cuts edge $v_x v_y$ with respect to $\mathcal{P}$, but does not cut edge $v_x v_{y'}$. Thus, there exists another interval $V_{k,l} \subset V_{i,j}$, $i \leq k < l < j$, such that $V_{k,l}$ cuts $v_x v_{y'}$, and furthermore, $V_{k,l}$ is a subproblem of $V_{i,j}$ (see Figure 3 (b)). Now suppose we compute the optimal score for subproblem $V_{i,j}$ as in the previous example. Thus, in this case, we will recursively compute the optimal solution for subproblem $V_{k,l}$ over all sets of similarity functions $\mathcal{F}_\phi$. Here, instead of enumerating codon assignments for both v_x and $v_{y'}$, we only enumerate assignments for $v_{y'}$, since all codon assignments for v_x are considered when computing the optimal solution for $V_{i,j}$. Furthermore, given any set of similarity functions $\mathcal{F}_\phi$, the optimal score for subproblem $V_{k,l}$ with similarity functions $\mathcal{F}_\phi$ is computed by enumerating only those assignments which assign codons to $v_{y'}$ that are compatible with codon C_x with respect to G_Γ, where C_x is the codon assigned to v_x in the current recursive call. In other words, codon C_x is the codon such that $f_x^\phi(C_x) > -\infty$, where $f_x^\phi \in \mathcal{F}_\phi$ is the similarity function corresponding to vertex v_x. If there is no such assignment, then the score of subproblem $V_{k,l}$ with the set of similarity functions $\mathcal{F}_\phi$ is $-\infty$. This ensures us that we consider only feasible solutions for subproblem $V_{k,l}$.

To generalize the discussion above, we use the following notation. Let $V_{i,j}$ be an interval of G_Γ, and let $\mathcal{F}'$ be any set of similarity functions. To compute the optimal solution for $V_{i,j}$ with $\mathcal{F}'$, we distinguish between two cases of vertices which are incident to bottom edges that are cut by $V_{i,j}$ with respect to $\mathcal{P}$. Let $v_x v_y \in \mathcal{E}_b$ be a bottom edge cut by $V_{i,j}$ with respect to $\mathcal{P}$. We say that v_x is *unassigned* in $V_{i,j}$ with respect to $\mathcal{P}$ and $\mathcal{F}'$, if there are at least two distinct codons $C_x, C'_x \in \Sigma^3$ such that $f'_x(C_x), f'_x(C'_x) > -\infty$, where $f'_x \in \mathcal{F}'$ is the similarity function corresponding to vertex v_x. Otherwise, we say that v_x is *assigned* in $V_{i,j}$ with respect to $\mathcal{P}$ and $\mathcal{F}'$. Thus in our previous example, v_x and $v_{y'}$ are respectively assigned and unassigned vertices in $V_{k,l}$ with respect to $\mathcal{P}$ and $\mathcal{F}'$. Note that vertices in $V_{i,j}$ which are not incident to bottom edges which are cut in $V_{i,j}$, are not considered assigned nor considered unassigned. Let $U_{i,j}$ and $A_{i,j}$ denote respectively, the unassigned and assigned vertices in $V_{i,j}$ with respect to $\mathcal{P}$ and $\mathcal{F}'$, for any interval $V_{i,j}$. When computing the optimal solution for $V_{i,j}$ with $\mathcal{F}'$, we enumerate only assignments ϕ of the form $\phi : U_{i,j} \to \Sigma^3$, such that for any assigned vertex $v_x \in A_{i,j}$ connected by a bottom edge to an unassigned vertex $v_y \in U_{i,j}$, $\phi(v_y)$ and C_x are compatible with respect to G_Γ, where C_x is the codon assigned to v_x in the current recursive call, *i.e,*. $f'_x(C_x) > -\infty$, $f'_x \in \mathcal{F}'$. For brevity, we denote by $\Phi[U_{i,j}]$, the set of all such assignments, *i.e.*, $\Phi[U_{i,j}]$ is the set of all codon assignments to vertices in $U_{i,j}$, which assign codons to vertices in $U_{i,j}$ that are compatible to the current codons assigned to vertices in $A_{i,j}$ with respect to G_Γ.

Thus, our suggested modified version of $\mathcal{A}_{\text{OP}}$, algorithm $\mathcal{A}^*_{\text{OP}}$, computes the optimal solution for any instance $(G_\Gamma, \mathcal{F})$ of MRSO by computing the optimal solution for $(G_\Gamma[\mathcal{E}_t], \mathcal{F})$ in a similar recursive fashion as $\mathcal{A}_{\text{OP}}$. However, for each subproblem $V_{i,j}$ encountered in this recursion, $\mathcal{A}^*_{\text{OP}}$ computes the maximum

score for $V_{i,j}$ over all sets of similarity functions $\mathcal{F}_\phi$ corresponding to codon assignments ϕ such that $\phi \in \Phi[U_{i,j}]$. Computing in this fashion ensures that only feasible solutions are considered. Furthermore, by the optimality of $\mathcal{A}_{\text{OP}}$, the solution computed by $\mathcal{A}^*_{\text{OP}}$ must be optimal, since $\mathcal{A}^*_{\text{OP}}$ eventually considers all compatible codon assignments for vertices incident to bottom edges.

Lemma 3. *Given an instance $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ for MRSO accompanied by a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of G_Γ, the optimal target mRNA sequence for this instance can be computed in $\mathcal{O}(64^{2\psi}\psi n)$ time, where ψ is the cutwidth of $G_\Gamma[\mathcal{E}_b]$.*

Lemma 3 helps establish a finer borderline between tractable and intractable instances of MRSO. Indeed, any instance $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ of MRSO, such that G_Γ has a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$, where the cutwidth of $G_\Gamma[\mathcal{E}_b]$ is bounded by $\mathcal{O}(\lg n)$, can be solved in polynomial time. However, determining whether G_Γ has such a bipartition can be quite difficult. In any case, we expect that in most practical applications, a convenient bipartition of this sort can easily be found using simple heuristic techniques. Furthermore, in some applications, one may be able to combine the above result with the results presented in the previous section. For example, given an instance with an implied structure graph containing only a few number of degree 3 vertices, one can combine the result presented in the previous section with a simple heuristic to obtain a nice edge bipartition with an even smaller cutwidth.

5 Parameterizing by the similarity score

Although the parameters introduced in the previous sections have clear biological relevance, we assume that in some practical applications these will not suffice. Thus, in the following we consider a more common parameter, namely the value of the optimum solution. For this we consider a relaxation on the similarity functions of an MRSO instance. We consider only similarity functions of the form $f_i : \Sigma^3 \to \mathbb{N}^+$, where $\Sigma = \{A, C, G, U\}$ as usual. We call similarity functions of this sort *restrictive similarity functions*, and denote MRSO_r the MRSO problem restricted to instances with restrictive similarity functions. Most of the interest in restrictive similarity functions stems from the following proposition.

Proposition 4. *The* MRSO_r *problem is in* **FPT** *for parameter k, where k is the score of the solution.*

Proof. Let $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ be an instance of MRSO_r. Denote by $\alpha(G_\Gamma)$ the size of the maximum independent set in G_Γ. We present an algorithm which searches for a target mRNA string T, by focusing on finding k pairwise compatible codons with respect to G_Γ. The proof is divided into two separate parts depending on the size of a maximum independent set in G_Γ. We may assume without loss of generality that for all $1 \leq i \leq n$, $f_i(C) > 0$ for some codon $C \in \Sigma^3$.

Suppose $k \leq \alpha(G_\Gamma)$. Let $V' \subseteq \mathbf{V}(G_\Gamma)$ be an independent set of size k in G_Γ. Since G_Γ is at most cubic, such a subset V' can be found in $\mathcal{O}(4^k n)$ time using a bounded search tree technique. We define a string T of length $3n$ as follows. For each $v_i \in V'$, 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 $\mathcal{F}$ is composed of restrictive similarity functions). For each $v_j \in \mathbf{V}(G_\Gamma) - V'$, assign codon C_j which is compatible with all previously assigned codons with respect to G_Γ (again this is always possible since Γ has maximum degree 1). We check at once that $T = C_1 C_2 \ldots C_n$ is compatible with respect to G_Γ and $\sum_{i=1}^{n} f_i(C_i) \geq |V'| = k$.

Now suppose $k > \alpha(G_\Gamma)$. Since G_Γ is at most cubic, we have $\alpha(G_\Gamma) \geq \frac{n}{4}$, and hence $k > \frac{n}{4}$. Here, the algorithm suggested is by direct enumeration. More precisely, the algorithm tries in turn to obtain a solution mRNA string T by assigning ℓ compatible codons in it, where ℓ ranges from 1 to k. So, let $\ell \in \{1, 2, \ldots, k\}$. 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 $\mathcal{O}(\binom{n}{\ell}\, n\, 64^\ell)$ time. Summing-up over ℓ and neglecting the time to check $k > \alpha(G_\Gamma)$, *i.e.*, $\mathcal{O}(4^k)$, we obtain $\mathcal{O}(n \sum_{\ell=1}^{k} \binom{n}{\ell} 64^\ell)$, which is $\mathcal{O}(2^{\mathcal{O}(k)}\, k^{k+1})$ since G_Γ is at most cubic and $k > \alpha(G_\Gamma) \geq \frac{n}{4}$.

Thus, the MRSO_r problem can be solved in $\mathcal{O}(2^{\mathcal{O}(k)}\, k^{k+1} + 4^k n)$ time, where k is the score of the solution. Hence, it follows that the MRSO_r problem is in **FPT**, and the proposition follows. □

One may argue that the restrictive model we suggest here, is too harsh and too constraining. Notice, however, that all hardness results obtained for MRSO still hold under this model. Nevertheless, using a simple combinatorial argument, we can easily obtain an optimal algorithm if we consider the score of the optimal solution for MRSO under restrictive similarity functions 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 they 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.

6 Conclusions and future work

In the context of protein engineering, we considered the problem of computing an mRNA sequence of maximal similarity to a given mRNA and a given protein that additionally satisfies some secondary structure constraints (the MRSO problem). We proved that the MRSO problem is fixed-parameter tractable parameterized by the number of degree 3 vertices, by the number of crossing edges (thus answering an open problem posed in [8]) or by the cutwidth of the implied structure graph. We believe these parameters to be relevant for practical applications. Also, we showed that the problem is in fixed-parameter tractable parameterized by the score for a restrictive class of similarity functions.

There are many interesting related problems arising in the above context. Most of them are relevant for practical applications of the MRSO problem, and hence our interest in these problems ranges from approximation to efficient fixed-parameter algorithms. Below are some of them (the first three ones ask for a preprocessing procedure prior to a fixed-parameter algorithm):

1. Proposition 2 suggests the following problem: Given an at most cubic graph G, find a linear embedding of G with a minimal number of crossing edges. This problem is known as the Outer-Planar Crossing Number problem and has been proved to be **NP**-complete for general graphs in [17]. A related problem, the so-called Maximum Outer-Planar Subgraph problem, is concerned with finding a linear embedding of G together with a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ such that $|\mathcal{E}_t|$ is maximized. This problem is **NP**-complete [20] for general graphs. However, we are not aware of any attempts to approximate the above problems. Furthermore, observe that for a given linear embedding, obtaining such a nice edge bipartition reduces to finding a maximum independent set in an overlap graph.
2. In the light of Algorithm $\mathcal{A}_{\text{PEB}}(G_\Gamma, \mathcal{F}, \mathcal{P})$ (see Section 3), the following problem is relevant for practical applications: Given an at most cubic graph G, find a linear embedding of G together with a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ with a minimal number of vertices incident to bottom edges.
3. Given an at most cubic graph G, find a a linear embedding of G together with a nice edge bipartition of G with a minimal cutwidth (in the sense of Lemma 3). The general problem of minimizing the cutwidth of an at most cubic graph is **NP**-complete [16].
4. Is the MRSO problem fixed-parameter tractable parameterized by the similarity score when restricted to similarity functions of the form $f_i : \Sigma^3 \to \mathbb{N}^+ \cup \{-\infty\}$. If so, is it still fixed-parameter stractable for any general function?

Acknowledgments

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

References

1. J. Alber, J. Gramm, J. Guo, and R. Niedermeier. Towards optimally solving the longest common subsequence problem for sequences with nested arc annotations in linear time. In *Proc. of the 13th Annual Symposium on Combinatorial Pattern Matching (CPM 2002)*, volume 2373 of *LNCS*, pages 99–114. Springer-Verlag, 2002.

2. 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.
3. 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.
4. 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.
5. T.C. Biedl, G. Kant, and M. Kaufmann. On triangulating planar graphs under the four-connectivity constraints. *Algorithmica*, 19:427–446, 1997.
6. A. Böch, K. Forchhammer, J. Heider, and C. Baron. Selenoprotein synthesis: a review. *Trends in Biochemical Sciences*, 16(2):463–467, 1991.
7. 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.
8. 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.
9. R. Diestel. *Graph Theory*. Number 173 in Graduate texts in Mathematics. Springer-Verlag, second edition, 2000.
10. R. Downey and M. Fellows. *Parameterized Complexity*. Springer-Verlag, 1999.
11. J. Gramm, J. Guo, and R. Niedermeier. Pattern matching for arc-annotated sequences. In *Proc. of the the 22nd Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS)*, volume 2556 of *LNCS*, pages 182–193, 2002.
12. L.S. Heath. Edge coloring planar graphs with two outerplanar subgraphs. In *Proc. of the 2nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 195–202.
13. 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.
14. T. Jacks and H. Varmus. Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting. *Science*, 230:1237–1242, 1985.
15. 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.
16. F. Makedon, C. Papadimitriou, and I. Sudborough. Topological bandwidth. *SIAM Journal on Algebraic and Discrete Methods*, 7(4):418–444, 1985.
17. S. Masua, T. Kashiwabara, K. Nakajima, and K. Fujisawa. On the NP-completeness of a computer network layout problem. In *Proc. IEEE International Symposium on Circuits and Systems*, pages 292–295. IEEE Press, 1987.
18. S.L. Mitchell. Linear algorithms to recognize outerplanar and maximal outerplanar graphs. *Information Processing Letters*, 9(5):229–232, 1979.
19. M.M. Syslo. Characterizations of outerplanar graphs. *Discrete Math.*, 26:47–53, 1979.
20. T. Watanbe, T. Ae, and A. Nakamura. On the NP-hardness of edge-deletion and contraction problems. *Discrete Applied Mathematics*, 6:63–78, 1983.
21. M. Yannakakis. Embedding planar graphs in four pages. *Journal of Computer and System Sciences*, 38:36–67, 1986.

Appendix (Program committee version only)

Lemma 1. *Given an instance* $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ *for* MRSO *accompanied by a nice edge bipartition* $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ *of* G_Γ*, algorithm* $\mathcal{A}_{NEB}$ *computes an optimal target mRNA sequence for this instance in* $\mathcal{O}(64^\upsilon n)$ *time, where* $\upsilon = |\{v \in e : e \in \mathcal{E}_b\}|$.

Proof. Consider the schematic description of algorithm $\mathcal{A}_{\text{NEB}}$ given in Figure 2. Since any assignment enumerated is verified for compatibility with respect to $G_\Gamma[\mathcal{E}_b]$, and by the correctness of $\mathcal{A}_{\text{OP}}$, any solution outputted by algorithm $\mathcal{A}_{\text{NEB}}$ with a score higher than $-\infty$ is feasible. Furthermore, by the optimality of $\mathcal{A}_{\text{OP}}$ this solution must be optimal. Now, consider any vertex in G_Γ. The number of possible codons assignments to this vertex is $|\Sigma^3| = 64$, so the number of assignments enumerated is bounded by $\mathcal{O}(64^\upsilon)$. Furthermore, checking any such assignment for compatibility with respect to $G_\Gamma[\mathcal{E}_b]$ can be done in $\mathcal{O}(\upsilon)$ time. As Steps (a) and (b) both require $\mathcal{O}(n)$ time, the overall time complexity of $\mathcal{A}_{\text{NEB}}$ is bounded by $\mathcal{O}(64^\upsilon(n))$, and so the above lemma holds. □

Proposition 3. *The* MRSO *problem is* **NP***-complete when restricted to implied structure graphs with page-number* 2*.*

Proof. We describe a reduction from the Maximum Independent Set problem, which is known to be **NP**-complete even when restricted to cubic planar bridegeless connected graphs [5]. The proof is a direct extension of the **APX**-completeness proof provided in [8] for MRSO.

Let an instance of the Maximum Independent Set problem be given by a cubic planar bridgeless connected graphs G of order n. According to [15], 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 2. 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 [8], 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 [8], Theorem 3. □

Lemma 3. *Given an instance* $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ *for MRSO accompanied by a nice edge bipartition* $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ *of* G_Γ*, the optimal target mRNA sequence for this instance can be computed in* $\mathcal{O}(64^{2\psi}\psi n)$ *time, where* ψ *is the cutwidth of* $G_\Gamma[\mathcal{E}_b]$.

Proof. Let $(G_\Gamma, \mathcal{F} = f_1, \ldots, f_n)$ be an arbitrary instance for MRSO accompanied by a nice edge bipartition $\mathcal{P} = \{\mathcal{E}_t, \mathcal{E}_b\}$ of G_Γ with cutwidth ψ. Compute the optimal solution using the modified version of $\mathcal{A}_{\text{OP}}$, algorithm $\mathcal{A}^*_{\text{OP}}$. By this modified version, each codon assignment to any vertex is compatible with respect to $G_\Gamma[\mathcal{E}_b]$, and by correctness of $\mathcal{A}_{\text{OP}}$, also with respect to $G_\Gamma[\mathcal{E}_t]$. Thus, $\mathcal{A}^*_{\text{OP}}$ will yield only feasible solutions for this instance. Furthermore, by the optimality of $\mathcal{A}_{\text{OP}}$, the solution computed by $\mathcal{A}^*_{\text{OP}}$ must be optimal, since $\mathcal{A}^*_{\text{OP}}$ eventually considers all compatible codon assignments for vertices incident to bottom edges. Now, by definition, any interval $V_{i,j} \subseteq \mathbf{V}(G_\Gamma)$ can cut at most ψ edges of $\mathcal{E}_b$. As a result, for each such interval $V_{i,j}$, then number of assigned and unassigned vertices, *i.e.*, $|A_{i,j} \cup U_{i,j}|$, is bounded by 2ψ. It follows from this, that for any interval $V_{i,j} \subseteq \mathbf{V}(G)$, the number of assignments in $\Phi[U_{i,j}]$ is

bounded by $64^{2\psi}$. Furthermore, each assignment $\phi : U_{i,j} \to \Sigma^3$ can be checked for compatibility with $A_{i,j}$ in $\mathcal{O}(\psi)$ time. Thus, excluding the time for computing the recursion calls, the optimal score for any interval $V_{i,j}$ can be computed in $\mathcal{O}(64^{2\psi}\psi)$ time. Since $G_\Gamma[\mathcal{E}_t]$ is outer-planar, there are at most $\mathcal{O}(n)$ subproblems, and so the total time of this computation is bounded by $\mathcal{O}(64^{2\psi}\psi n)$. □