Journal of Discrete Algorithms 6 (2008) 618–626

Contents lists available at ScienceDirect

Journal of Discrete Algorithms

www.elsevier.com/locate/jda

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

Guillaume Blin [a], Guillaume Fertin [b], Danny Hermelin [c,*], Stéphane Vialette [a]

[a] IGM-LabInfo, UMR CNRS 8049, Université de Marne-la-Vallée, France
[b] Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR CNRS 6241 Université de Nantes, 2 rue de la Houssinière, 44322 Nantes Cedex 3, France
[c] Department of Computer Science, University of Haifa, Mount Carmel, Haifa, Israel

ARTICLE INFO

Article history:
Received 4 September 2006
Accepted 27 March 2008
Available online 12 April 2008

Keywords:
mRNA optimization
Protein similarity
Selenocysteine insertion
Parameterized complexity
Fixed-parameter tractability

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 mRNA Structure Optimization (MRSO) problem. Since MRSO is known to be **APX**-hard, Bongartz [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), 2004, pp. 174–183] suggested to attack the problem using the approach of parameterized complexity. In this paper we propose three fixed-parameter algorithms that apply for several interesting parameters of MRSO. We believe these algorithms to be relevant for practical applications today, as well as for possible future applications. Furthermore, our results extend the known tractability borderline of MRSO, and provide new research horizons for further improvements of this sort.

1. Introduction

Perhaps the most significant process in molecular biology known today is the transformation of genetic information encoded in DNA into proteins. In this process, segments of DNA are transcribed into messenger RNA (mRNA) molecules, which in turn serve as blueprints for manufacturing proteins. This protein blueprint is provided by triplets of nucleotides known as codons, which compose the mRNA nucleotide sequence, where each codon encodes a specific amino acid. An mRNA is thus translated into a protein by reading each of its codons in sequential fashion, and creating a chain of amino acids which forms the target protein. Recently, biologists found out that according to the folding structure of an mRNA molecule, a certain codon might encode for different amino acids. This folding structure is captured in many ways, in what is called the mRNA secondary structure, the set of all hydrogen bonds, or base pairings, formed by the molecule's nucleotides.

In [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 mRNA Structure Optimization (MRSO) problem. The initial motivation of MRSO is concerned with selenocysteine

☆ A preliminary version of this work can be found in [G. Blin, G. Fertin, D. Hermelin, S. Vialette, Fixed-parameter algorithms for protein similarity search under mRNA structure constraints, in: Proc. of the 31st International Workshop on Graph-Theoretic Concepts in Computer Science (WG), 2005, pp. 271–282].

* Corresponding author.
E-mail addresses: gblin@univ-mlv.fr (G. Blin), guillaume.fertin@univ-nantes.fr (G. Fertin), danny@cri.haifa.ac.il (D. Hermelin), vialette@univ-mlv.fr (S. Vialette).

1570-8667/$ – see front matter
doi:10.1016/j.jda.2008.03.004

Fig. 1. The translation of UGA into selenocysteine. Termination of translation is inhibited in the presence of the SECIS element.

insertion, i.e. generating new amino acid sequences containing selenocysteine. This rare amino acid was discovered as the 21st amino acid [6], 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 [6] that in case of selenocysteine, termination of translation is inhibited in the presence of a sequence of nucleotides, the SECIS element, which forms a hairpin-like structure in the 3′-region after the UGA codon (see Fig 1). It is argued in [3] 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 for this type of structure the linear-time algorithm presented in [3] provides an optimal solution. However, it is reasonable to assume that the discovery of selenocysteine will lead to the discovery of several other amino acids of similar kind, some of which are likely to require more complex secondary structures. Even today, similar problems occur in programmed frameshifts which allow to encode two different amino acid sequences in one mRNA sequence [17,18]. This motivates the investigation of MRSO for more elaborate secondary structures, as suggested also by [3,9], and is the starting point of our study.

Previous results. 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 outerplanar graph (as it is the case of 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 [3], and Bongartz recently showed that in fact the problem is **APX**-hard [9]. An algorithm for approximating MRSO within ratio 2 is given in [3]. A slightly slower but somewhat simpler 4-approximation algorithm is given in [9]. We mention also that an extension of MRSO, where insertions and deletions are allowed in the amino acid sequence is presented in [2].

Parameterized complexity. Since MRSO for general secondary structures is known to be **APX**-hard [9], Bongartz proposed in [9] to attack the problem using the approach of parameterized complexity [11]. Parameterized complexity is an approach to complexity theory which offers an alternative method of analyzing computational problems 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 $f(k)n^{\mathcal{O}(1)}$ time complexity are called *fixed-parameter tractable*, where k is the parameter, f can be any arbitrary function depending only on k, and n denotes the overall input size. The best general reference here is [11].

Our contribution. In the last decade, parameterized complexity has proven to be useful in several applications within computational biology [8]. In this paper we follow this trend by presenting fixed-parameter algorithms for several interesting parameters of MRSO. We believe these algorithms to be relevant for practical applications today, as well as for several future applications. Furthermore, our results extend the known tractability borderline of MRSO, and provide new research horizons for further improvements of this sort.

The paper is organized as follows. In the next section we briefly discuss basic notations and definitions that we will use throughout. In Section 3, we present a fixed-parameter algorithm for two natural parameters of MRSO, namely the number of degree three vertices, and the number of edge crossings in the given implied structure graph (see Definition 2 in the following section). Also, we show that MRSO remains **NP**-complete even when the implied structure graph is quite restricted. In Section 4, we consider the cutwidth of the implied structure graph as a parameter, and show that the problem is fixed-parameter tractable when parameterized by this parameter. Following this, in Section 5, we show that a boolean variant of MRSO is fixed-parameter solvable when parameterized by the score of the optimal solution. We summarize and discuss possible future directions of research in Section 6.

2. Preliminaries

An mRNA molecule is viewed as 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*. Hydrogen bonds can only be formed between complementary nucleotides in an mRNA folding. A *codon* of an mRNA sequence is a segment of three nucleotides, i.e. a string in Σ^3. Thus, an mRNA sequence $S = s_1 \dots s_{3n}$ 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 $S = s_1 \dots s_{3n}$, we wish to evaluate the codon-wise similarity between S and another *target* mRNA sequence $T = t_1 \dots t_{3n}$. For this, we are provided with a set of n functions, $\mathcal{F} = f_1, \dots, f_n$, called *similarity functions* of S, such that for all i, $1 \leqslant i \leqslant n$, 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

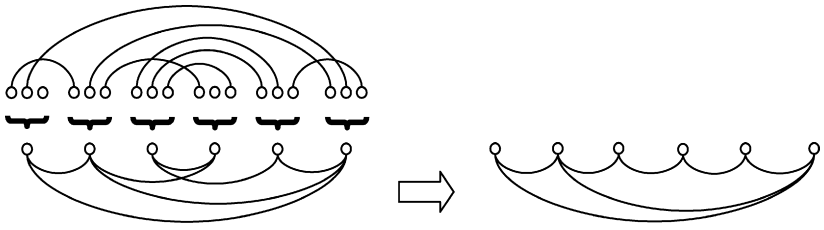

Fig. 2. (a) An example of an implied structure graph obtained from the set of structure constraints $\{\{1,4\},\{2,17\},\{5,16\},\{6,10\},\{7,14\},\{8,13\},\{9,12\},\{15,18\}\}$, where the ordering of both nucleotides and codons is from left to right. The set of edges in the implied structure graph is $\{\{1,2\},\{1,6\},\{2,4\},\{2,6\},\{3,4\},\{3,5\},\{5,6\}\}$. (b) The implied structure graph is outerplanar since swapping the two middle vertices yields an ordering of the vertices with no edge crossings.

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 constraints* $\Gamma \subseteq \{\{i,j\} \mid 1 \leqslant i < j \leqslant 3n\}$ 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. It is assumed that each nucleotide can pair with at most one other nucleotide in any folding, hence 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 i, $1 \leqslant i \leqslant 3n-2$.

Given a set of structure constraints $\Gamma \subseteq \{\{i,j\} \mid 1 \leqslant i < j \leqslant 3n\}$, and an arbitrary target mRNA sequence $T = t_1 \ldots t_{3n}$, we say that nucleotides t_i and t_j in T 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***) [3]).* Let $\mathcal{F}$ be a set of n similarity functions for a source mRNA sequence of length $3n$, and let $\Gamma \subseteq \{\{i,j\} \mid 1 \leqslant i < j \leqslant 3n\}$ be a set of structure constraints. 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 linear ordering. Therefore, we can view Γ as a linear graph with $3n$ vertices which has a maximum degree of one. However, since we are really interested in codon-wise similarity, we use a more suitable representation of Γ:

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

$$\mathbf{V}(G_\Gamma) = \{1,2,\ldots,n\}, \quad \text{and}$$
$$\mathbf{E}(G_\Gamma) = \big\{\{i,j\} \mid \exists\{x,y\} \in \Gamma : x \in \{3i-2, 3i-1, 3i\} \wedge y \in \{3j-2, 3j-1, 3j\}\big\}.$$

In this way, vertex i in $\mathbf{V}(G_\Gamma)$ corresponds to the ith codon of the target mRNA sequence, and $i,j \in \mathbf{V}(G_\Gamma)$ are connected in $\mathbf{E}(G_\Gamma)$ if there are any structure constraints in Γ between the ith and jth codons of the sequence. Note that there can be at most three structure constraints between any pair of codons, therefore G_Γ has maximum degree of three, i.e. it is a *subcubic* graph (see Fig. 2). Also note that, while this representation may seem lossy, in fact Γ can be retained from G_Γ by adding up to three labels for each edge in $\mathbf{E}(G_\Gamma)$.

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 V as its vertex set, and $\mathbf{E}(G_\Gamma) \cap (V \times V)$ as its edge set. Similarly, given a subset of edges $E \subseteq \mathbf{E}(G_\Gamma)$, $G_\Gamma[E]$ denotes the subgraph of G_Γ with vertex set $\{i \mid \{i,j\} \in E\}$ and edge set E. Also, we use $G_\Gamma[i,\ldots,j]$ to denote the subgraph of G_Γ induced by $\{i,\ldots,j\} \subseteq \mathbf{V}(G_\Gamma)$. Two edges $\{i,j\}$ and $\{i',j'\}$ *cross* in G_Γ if either $i < i' < j < j'$ or $i' < i < j' < j$. Note that two crossing edges might not cross under a different ordering of $\mathbf{V}(G_\Gamma)$. If there exists an ordering of $\mathbf{V}(G_\Gamma)$ which introduces no edge crossings then G_Γ is *outerplanar*. Recall that if G_Γ is outerplanar, algorithm $\mathcal{A}_{\mathrm{OP}}$ [3] can be used to solve MRSO in $\mathcal{O}(n)$ time.

A *codon assignment* for G_Γ is a mapping from some $V \subseteq \mathbf{V}(G_\Gamma)$ to Σ^3. An assignment for a pair of vertices $i,j \in \mathbf{V}(G_\Gamma)$, $i \to t_{3i-2}t_{3i-1}t_{3i}$ and $j \to t_{3j-2}t_{3j-1}t_{3j}$, is compatible with respect to G_Γ, if either $\{i,j\} \notin \mathbf{E}(G_\Gamma)$ or $t_{i'}$ and $t_{j'}$ are complementary nucleotides for any $\{i',j'\} \in \Gamma \cap \{3i-2, 3i-1, 3i\} \times \{3j-2, 3j-1, 3j\}$. More generally, an assignment $\phi : V \to \Sigma^3$ for some $V \subseteq \mathbf{V}(G_\Gamma)$ is compatible with respect to G_Γ, if for any $i,j \in V$, the assignment $i \to \phi(i)$ and $j \to \phi(j)$ is compatible with respect to G_Γ. Henceforth, we consider instances for MRSO of the form $(G_\Gamma, \mathcal{F})$. Our goal in this setting

is then to find an assignment $\phi : \mathbf{V}(G_\Gamma) \to \Sigma^3$ (i.e. a target mRNA sequence $T = \phi(1)\ldots\phi(n)$), 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. These are the number of edge crossings in G_Γ, and the number of degree three vertices in G_Γ. We use χ and δ to denote these two parameters respectively throughout the section.

Our initial interest in parameters χ and δ stems from the fact that we believe them to be small in many practical applications. Consider parameter χ, the number of edge crossings in G_Γ. This parameter was previously suggested in [9]. Indeed, almost all currently known mRNAs have secondary structures which induce outerplanar formations, i.e. formations with no edge crossings. Furthermore, many secondary structure prediction algorithms restrict their search space to structures with bounded edge crossings, since prediction with unbounded edge crossings usually becomes **NP**-hard, and is anyhow assumed unnatural (see for instance [1]). As for parameter δ, the number of degree three vertices, 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 geometric constraints imposed on any such folding. Also, pairs of hydrogen bonds of the form $\{i, j\}$ and $\{i+1, j-1\}$, called *stacking pairs*, tend to contribute quite substantially to the overall stability of the folding structure of the mRNA [16,21]. A secondary structure is hence expected to have a relatively high number of stacking pairs, and therefore to induce an implied structure graph with a relatively small number of degree three vertices.

It turns out that MRSO is polynomial-time solvable when either χ or δ are fixed. To show this, we will first present an initial algorithm, and later demonstrate how it can be applied for both cases. We will need the following definition (see also Fig. 4(a) for an example):

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 $\mathcal{P}$ respectively, such that $E_t \cup E_b = \mathbf{E}(G_\Gamma)$, $E_t \cap E_b = \emptyset$ and $E_t \neq \emptyset$. Furthermore, $\mathcal{P}$ is said to be *nice*, if the subgraph $G_\Gamma[E_t]$ is outerplanar.

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.

At the heart of algorithm $\mathcal{A}_{\text{NEB}}$ lies 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 for the instance $(G_\Gamma, \mathcal{F}')$, where $\mathcal{F}' = f', f_2, \ldots, f_n$, will then give us our desired mRNA. The following definition generalizes this example.

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 [3] for outerplanar implied structure graphs, as a subprocedure in its computation. At its core, $\mathcal{A}_{\text{NEB}}$ is basically an exhaustive search procedure 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. Any solution returned by $\mathcal{A}_{\text{OP}}$ is guaranteed to be compatible with G_Γ since it is simultaneously compatible with both $G_\Gamma[E_b]$ and $G_\Gamma[E_t]$. Finally, $\mathcal{A}_{\text{NEB}}$ terminates by outputting the maximum solution over all target mRNAs returned by $\mathcal{A}_{\text{OP}}$. A schematic description of $\mathcal{A}_{\text{NEB}}$ is given in Fig. 3.

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, $\mathcal{A}_{\text{NEB}}$ computes an optimal target mRNA sequence for this instance in $\mathcal{O}(2^{12\beta}n)$ time, where $n = |\mathbf{V}(G_\Gamma)|$ and $\beta = |E_b|$.*

Proof. Consider the schematic description of $\mathcal{A}_{\text{NEB}}$ in Fig. 3 and let $V_b = \{i \mid \{i, j\} \in E_b\}$ be the subset of vertices incident to E_b. Any assignment $\phi : V_b \to \Sigma^3$ enumerated in the algorithm is verified for compatibility with respect to $G_\Gamma[E_b]$. Hence, by the correctness of $\mathcal{A}_{\text{OP}}$, any target mRNA outputted by $\mathcal{A}_{\text{NEB}}$ with a similarity score higher than $-\infty$ is compatible with respect to G_Γ. Furthermore, by the optimality of $\mathcal{A}_{\text{OP}}$, and since all possible codon assignments to V_b are considered by $\mathcal{A}_{\text{NEB}}$, this target mRNA is optimal with respect to $\mathcal{F}$.

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

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 $\mathcal{P} = (E_t, E_b)$.
Result : An optimal target mRNA sequence T which is compatible with respect to G_Γ.
begin
  foreach *possible codon assignment* ϕ *to vertices incident to edges in* E_b **do**
    if ϕ *is compatible with respect to* $G_\Gamma[\mathcal{E}_b]$ **then**
      (a) Construct $\mathcal{F}_\phi$, the similarity functions corresponding to ϕ.
      (b) Invoke $A_{\mathrm{OP}}(G_\Gamma[E_t], \mathcal{F}_\phi)$.
    end
  end
  return *the target mRNA sequence found in Step (b) with the highest similarity score.*
end

Fig. 3. Algorithm $\mathcal{A}_{\mathrm{NEB}}$.

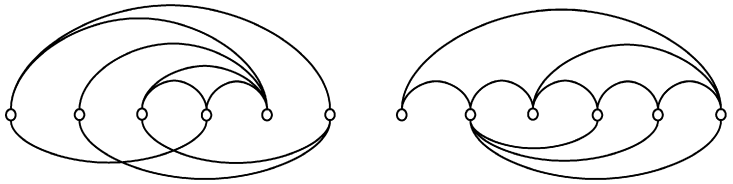

Fig. 4. The difference between nice edge bipartitions and 2-page embeddings. In (a) a nice edge bipartition of a subcubic graph. In (b), a 2-page embedding of a graph.

For the time complexity bound, note that the number of codon assignments enumerated by the algorithm is $|\Sigma^3|^{|V_b|} \leqslant 64^{2\beta} = 2^{12\beta}$. Furthermore, constructing any such assignment and checking it for compatibility with respect to $G_\Gamma[E_b]$ can be done in at most $\mathcal{O}(n)$ time. Therefore, since each call to $\mathcal{A}_{\mathrm{OP}}$ also requires $\mathcal{O}(n)$ time [3], the overall time complexity of $\mathcal{A}_{\mathrm{NEB}}$ is $\mathcal{O}(2^{12\beta}n)$. □

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 corollary.

Corollary 1. MRSO *is solvable in* $\mathcal{O}(2^{12\chi}n)$ *time.*

Next consider parameter δ. Constructing a nice edge bipartition with δ bottom edges is immediate after establishing 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 connected 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.

Corollary 2. MRSO *is solvable in* $\mathcal{O}(2^{12\delta}n)$ *time.*

3.1. Implied structure graphs with page-number two

In light of algorithm $\mathcal{A}_{\mathrm{NEB}}$ and Lemma 1, a natural question to ask is whether MRSO is polynomial-time solvable in case we are provided an edge bipartition in which both parts have no edge crossings under the same vertex ordering. Such

would be the case if G_Γ had page-number two. In general, the *page-number* of a given graph G is the partitioning of $\mathbf{E}(G)$ into the smallest number of subsets possible, such that each subset of edges in the partition has no edge crossings under the same vertex ordering (see for instance Fig. 4(b)). If we could solve MRSO for page-number two graphs, we might also hope that MRSO becomes fixed-parameter tractable when parameterized by the page-number of G_Γ. Unfortunately, this is not the case, as MRSO is **NP**-complete already for page-number two implied structure graphs.

Lemma 3. MRSO *is **NP**-complete 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 [5]. The proof is a direct extension of the **APX**-completeness proof for MRSO given in [9].

Let an instance of the Maximum Independent Set problem be given by a cubic planar bridgeless connected graph G of order n. According to [4], 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 MRSO. The implied structure graph G_Γ is merely the input graph G and the set of similarity functions $f_i : \Sigma^3 \to \mathbb{Q}$, $1 \leqslant i \leqslant n$, is defined as follows:

$$\forall i, 1 \leqslant i \leqslant n, \quad f_i(C) = \begin{cases} 1 & C = AAA \\ 0 & C \neq AAA. \end{cases}$$

Quoting [9], 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 [9], Theorem 3. □

Corollary 3. *Unless **P** = **NP**,* MRSO *is not fixed parameter tractable when parameterized by the page-number of the given implied structure graph.*

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 section we explore the fixed-parameter tractability of MRSO when parameterized by the cutwidth of G_Γ. Our motivation for this is twofold. First, although cutwidth is perhaps not as natural as the two previously discussed parameters, it has been studied quite considerably for other problems dealing with RNAs [12,13,19]. Second, as we shall soon see, the fact that MRSO is polynomial-time solvable in case G_Γ has bounded cutwidth implies polynomial-time solvability in several other interesting cases. In particular, it implies that MRSO can be solved in polynomial time in case G_Γ is either chordal, circular-arc, or k-outerplanar for any constant k.

Let ψ denote the cutwidth of G_Γ. To show that MRSO is fixed parameter tractable when parameterized by ψ, we present an algorithm which we call $\mathcal{A}_{\text{CUT}}$. This algorithm works by recursively partitioning G_Γ into two subgraphs $G_\Gamma[1, \ldots, p]$ and $G_\Gamma[p+1, \ldots, n]$, 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 optimal and compatible with respect to G_Γ, $\mathcal{A}_{\text{CUT}}$ exhaustively searches through all possible codon assignments which are compatible with the subset of edges connecting vertices in $G_\Gamma[1, \ldots, p]$ to vertices in $G_\Gamma[p+1, \ldots, n]$.

In order to maintain compatibility throughout the recursion, we distinguish in $\mathcal{A}_{\text{CUT}}$ between vertices which were assigned a codon in a previous recursive step, and those which have not yet been assigned one. We enforce two invariants. First, all assigned vertices are compatible throughout the entire execution of the algorithm. And second, once a vertex is assigned at some recursive step of the algorithm, no assignments are enumerated for this vertex in any subsequent step. Enforcing both invariants is done using corresponding similarity functions (recall Definition 4). In what follows, we call a similarity function f *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 distinguish between assigned and unassigned vertices along the recursion, and also to propagate codon assignments of assigned vertices. In this way, in a particular recursive step, vertex $i \in \mathbf{V}(G_\Gamma)$ is considered assigned if f_i is degenerate and it is assigned the unique codon C such that $f_i(C) > -\infty$. A schematic description of $\mathcal{A}_{\text{CUT}}$ is given in Fig. 5.

Lemma 4. *Given an instance $(G_\Gamma, \mathcal{F})$ for* MRSO*, algorithm $\mathcal{A}_{\text{CUT}}$ computes an optimal target mRNA sequence for this instance in $\mathcal{O}(2^{12\psi} n)$ time, where $n = |\mathbf{V}(G_\Gamma)|$ and ψ is the cutwidth of G_Γ.*

Proof. Consider the schematic description of $\mathcal{A}_{\text{CUT}}$ in Fig. 5. We prove the correctness and optimality of the algorithm by induction on its recursion depth. At the recursive basis, the solution returned is optimal and compatible by construction. For the inductive step, assume T' and T'' are the two target mRNAs computed at steps (a) and (b) respectively. Then T' and T'' are compatible with respect to $G_\Gamma[1, p]$ and $G_\Gamma[p+1, n]$ respectively. Hence, since by construction $T'T''$ is compatible with

Algorithm $\mathcal{A}_{\mathrm{CUT}}(G_\Gamma, \mathcal{F})$

Data : An implied structure graph G_Γ 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_Γ.
begin
1. **if** $\mathbf{E}(G_\Gamma) = \emptyset$ **then return** T that maximizes $\mathcal{F}$.
2. Select the smallest $p \in \{1, \ldots, n-1\}$ with p-cutwidth greater than zero.
3. Set $E_p = \{\{i, j\} \in \mathbf{E}(G_\Gamma) \mid 1 \leq i \leq p,\ p+1 \leq j \leq n\}$.
4. Set $V_p = \{i \mid \{i, j\} \in E_p\}$ to be the vertices incident to E_p.
5. Let $A_p = \{i \in V_p \mid f_i \text{ is degenerate}\}$ be the assigned vertices in V_p.
6. Define $\phi^{A_p} : A_p \to \Sigma^3$ such that $\phi^{A_p}(i) = C \Leftrightarrow f_i(C) > -\infty$.
7. **foreach** *possible codon assignment* $\phi^{V_p - A_p} : V_p - A_p \to \Sigma^3$ **do**
  if $\phi = \phi^{A_p} \cup \phi^{V_p - A_p}$ *is compatible with respect to* $G_\Gamma[E_p]$ **then**
    (a) $T' \leftarrow \mathcal{A}_{\mathrm{CUT}}(G_\Gamma[1, \ldots, p], f_1^\phi, \ldots, f_p^\phi)$.
    (b) $T'' \leftarrow \mathcal{A}_{\mathrm{CUT}}(G_\Gamma[p+1, \ldots, n], f_{p+1}^\phi, \ldots, f_n^\phi)$.
  end
end
return *the highest similarity scoring target mRNA sequence* $T = T'T''$ *found in step 7.*
end

Fig. 5. Algorithm $\mathcal{A}_{\mathrm{CUT}}$.

respect to $G_\Gamma[E_p]$, it is also compatible with respect to G_Γ. Furthermore, since the algorithm considers all assignments to vertices in V_p with score higher than $-\infty$, and since T' and T'' are both optimal for $G_\Gamma[1, \ldots, p]$ and $G_\Gamma[p+1, \ldots, n]$ respectively, the target mRNA returned at this step must be optimal as well.

For the time complexity bound of $\mathcal{A}_{\mathrm{CUT}}$, note that the number of different subgraphs of G_Γ that $\mathcal{A}_{\mathrm{CUT}}$ encounters is $\mathcal{O}(n)$. Since for each implied structure graph, the algorithm enumerates at most $|\Sigma^3|^{|V_p|} \leqslant 2^{12\psi}$ different codon assignments, the total number of subproblems considered by $\mathcal{A}_{\mathrm{CUT}}$ is bounded by $\mathcal{O}(2^{12\psi}n)$. Hence, since the computation in each subproblem is linear, the overall time complexity of $\mathcal{A}_{\mathrm{CUT}}$ is $\mathcal{O}(2^{12\psi}n)$. □

As a first consequence of Lemma 4, we get the next corollary:

Corollary 4. MRSO *is polynomial-time solvable in case* $\psi = \mathcal{O}(\lg n)$.

We next consider the implications of Corollary 4. The treewidth [24] of a graph is a graph parameter that has been studied extensively in the literature. Informally, it measures in some sense the degree of tree-likeness of a given graph. In [20] (via [10]), the authors showed that for a graph 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 $\mathcal{O}(\lg n)$. This implies the following statement:

Corollary 5. MRSO *is polynomial-time solvable in case* G_Γ *has constant treewidth.*

Note that the treewidth of any outerplanar graph is bounded by two [7], and so the algorithm above generalizes $\mathcal{A}_{\mathrm{OP}}$, although the time complexity bound of $\mathcal{A}_{\mathrm{OP}}$ is better. In [7], Bodlaender gives a list of several other interesting graph classes which are subclasses of constant treewidth graphs. Among many others, we state the three which we feel are the most relevant to our application.

Corollary 6. MRSO *is polynomial-time solvable in case* G_Γ *is either a chordal graph, a circular arc graph, or a k-outerplanar graph where k is any constant.*

5. Boolean similarity functions

In the following section we suggest a relaxation on the similarity functions provided with an MRSO instance. Namely, we suggest considering instances restricted to *boolean similarity functions*, i.e. functions of the form $f_i : \Sigma^3 \to \{0, 1\}$. We let $\mathrm{MRSO}_{\mathbb{B}}$ denote the MRSO problem restricted to instances with this type of similarity functions.

Boolean similarity functions can model the case where we are only interested in the number of exact amino acid matches between our source and target proteins. This relaxation might still make sense in certain real-life applications, however at first glance it doesn't seem useful since $\mathrm{MRSO}_{\mathbb{B}}$ remains **NP**-hard (recall the proof of Lemma 3). Nevertheless, using a simple

combinatorial argument, we can obtain an exact algorithm for $\text{MRSO}_{\mathbb{B}}$ when the problem is parameterized by the optimal score of the given instance.

Let $(G_\Gamma, \mathcal{F})$ be an arbitrary instance of $\text{MRSO}_{\mathbb{B}}$, and let κ denote the similarity score of an optimal target mRNA for this instance. In what follows we say that the ith codon C of a target mRNA is *correct* if $f_i(C) = 1$, and otherwise we say it is *incorrect*. In these terms, κ measures the number of correct codons of an optimal target mRNA. Note that we can assume without loss of generality that every vertex can be assigned a correct codon, since otherwise, we can solve the sub-instance $(G'_\Gamma, \mathcal{F}')$ obtained by deleting all vertices which do not have a correct codon assignment from G_Γ and all their corresponding similarity functions 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.

Our result is based on the simple observation that since G_Γ is subcubic, any maximal (according to inclusion) independent set is of size at least $n/4$. To see this, consider the greedy algorithm which computes an independent set of G_Γ by repeatedly selecting an arbitrary vertex to add to its solution, and then omitting this vertex from G_Γ along with all of its neighbors. Any maximal independence set in G_Γ can be computed by this algorithm. Furthermore, since at any iteration of the algorithm, one vertex is added to the solution and at most four are omitted, the independent set found is of size at least $n/4$.

By the above observation, we can devise a simple algorithm for MRSO by considering $\kappa \leqslant n/4$ and $\kappa > n/4$ as two separate cases. If $\kappa \leqslant n/4$, we can find an independent set $I \subset \mathbf{V}(G_\Gamma)$ of size κ using the greedy algorithm above, and assign all the vertices in I correct codons. Again, such an assignment can always be extended to an assignment for $\mathbf{V}(G_\Gamma)$ with score at least κ, since Γ has maximum degree one. Hence, in this case we can find a target mRNA with κ codons in $\mathcal{O}(\kappa)$ time. If $\kappa > n/4$, then $\binom{n}{\kappa} < \binom{4\kappa}{\kappa} < 2^{3.25\kappa}$ (using Stirling's formula), and so we can exhaustively search for a κ-subset of $\mathbf{V}(G_\Gamma)$ which allows a pairwise compatible assignment of κ correct codons. The amount of time required to search and verify all assignments for each subset is bounded by $\mathcal{O}(2^{6\kappa}\kappa)$, and so the amount of time required by the entire procedure is bounded by $\mathcal{O}(2^{9.25\kappa}\kappa)$. Accounting also for the linear amount of work we inevitably must spend on processing our input, we obtain the following corollary.

Corollary 7. $\text{MRSO}_{\mathbb{B}}$ *is solvable in* $\mathcal{O}(2^{9.25\kappa}\kappa + n)$ *time.*

It is interesting to note that due to the boolean nature of our model, the above result also applies if we parameterize $\text{MRSO}_{\mathbb{B}}$ by the number of incorrect codons of an optimal solution. Let $\bar{\kappa}$ denote this parameter. Then $\bar{\kappa} = n - \kappa$. Also, for all $1 \leqslant i \leqslant n$, define the complementary of f_i, denoted as $\bar{f}_i$, by:

$$\bar{f}_i(C) = \begin{cases} 1 & f_i(C) = 0 \\ 0 & f_i(C) = 1. \end{cases}$$

It is now easy to verify that the number of correct codons in an optimal solution for $(G_\Gamma, \overline{\mathcal{F}})$, where $\overline{\mathcal{F}} = \bar{f}_1, \ldots, \bar{f}_n$, equals the number of incorrect codons in an optimal solution for $(G_\Gamma, \mathcal{F})$.

Corollary 8. $\text{MRSO}_{\mathbb{B}}$ *is solvable in* $\mathcal{O}(2^{9.25\bar{\kappa}}\bar{\kappa} + n)$ *time.*

6. Discussion and open problems

In this paper we considered the problem of computing a target mRNA of maximal codon-wise similarity to a given source mRNA that additionally satisfies some secondary structure constraints, the MRSO problem. We proved that MRSO is fixed-parameter tractable when parameterized by the number of degree-three vertices, the number of edge crossings, and the cutwidth of the given implied structure graph. The latter result implies that MRSO can be solved in polynomial time in case G_Γ is either chordal, circular-arc, k-outerplanar for any constant k, or in case G_Γ has constant treewidth. Also, we showed that for instances restricted to boolean similarity functions, one can easily devise a fixed-parameter algorithm when the number of correct or incorrect codons in an optimal solution is taken as a parameter. We believe our results to be relevant for practical applications today, as well as for possible future applications.

There are many interesting issues and related problems arising from our study. Below we state a few of them:

- Both the number of edge crossings and the cutwidth of G_Γ are parameters which rely on the particular ordering of $\mathbf{V}(G_\Gamma)$. It is therefore natural to ask whether one can find an ordering of G_Γ which minimizes these two parameters. For the second parameter, this question has been studied in the literature under the name Cutwidth [15]. The first problem has surprisingly not been considered to the best of our knowledge, however close variants such as minimizing the number of edge crossings in a planar embedding of a graph [14,15], or the number of edge crossings in a 2-page embedding of a graph [22], have been previously considered. The Cutwidth problem is **NP**-complete even in subcubic graphs [23], but is fixed-parameter tractable (when parameterized by the number of edge crossings in the optimal solution) in the general case [25,26]. It is not known whether such an algorithm exists for the first problem, and so it is an interesting open problem to determine whether such an algorithm exists, especially for the case where the input graph is restricted to be subcubic. Note that both problems become trivial if the input graph has maximum degree two.

- In light of algorithm $\mathcal{A}_{\text{NEB}}$ and Lemma 1, the following problem comes to mind: Given a subcubic graph G, find a linear embedding of G together with a nice edge bipartition $\mathcal{P} = (E_t, E_b)$ with a minimal number of bottom edges. It would be interesting to design an efficient fixed-parameter algorithm for this problem, perhaps by focusing on the subcubic case. We also note that one can use the framework of algorithm $\mathcal{A}_{\text{NEB}}$, together with the results of Corollaries 5 and 6, to devise fixed-parameter algorithms for different parameters of MRSO. As an example, one can define a bipartition of $\mathbf{E}(G_\Gamma)$ where the upper edges induce, for instance, a chordal graph.
- In Section 5, we introduced $\text{MRSO}_{\mathbb{B}}$, the restricted variant of MRSO, in which all similarity functions are boolean. Although boolean similarity functions allow a simple and relatively fast fixed-parameter algorithm, they can indeed be too restrictive for some applications. We do believe that it might be worth considering similarity functions of the form $f_i : \Sigma^3 \to \{0, 1, -\infty\}$ since these capture most of the information necessary in most practical settings. 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 the target mRNA.

References

[1] T. Akutsu, Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots, Discrete Applied Mathematics 104 (2000) 45–62.
[2] R. Backofen, A. Busch, Computational design of new and recombinant selenoproteins, in: Proc. of the 15th Annual Symposium on Combinatorial Pattern Matching (CPM), 2004, pp. 270–284.
[3] R. Backofen, N.S. Narayanaswamy, F. Swidan, Protein similarity search under mRNA structural constraints: Application to targeted selenocystein insertion, Silico Biology 2 (3) (2002) 275–290.
[4] F. Bernhart, P.C. Kainen, The book thickness of a graph, Journal of Combinatorial Theory Series B 27 (3) (1979) 320–331.
[5] T.C. Biedl, G. Kant, M. Kaufmann, On triangulating planar graphs under the four-connectivity constraints, Algorithmica 19 (1997) 427–446.
[6] A. Böch, K. Forchhammer, J. Heider, C. Baron, Selenoprotein synthesis: A review, Trends in Biochemical Sciences 16 (2) (1991) 463–467.
[7] H.L. Bodlaender, Classes of graphs with bounded tree-width, Research Report, Utrecht University, 1986.
[8] H.L. Bodlaender, R.G. Downey, M.R. Fellows, M.T. Hallett, H.T. Wareham, Parameterized complexity analysis in computational biology, Computer Applications in the Biosciences 11 (1995) 49–57.
[9] 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), 2004, pp. 174–183.
[10] F.R. Chung, P.D. Seymour, Graphs with small bandwidth and cutwidth, Discrete Mathematics 75 (1–3) (1989) 113–119.
[11] R. Downey, M. Fellows, Parameterized Complexity, Springer-Verlag, 1999.
[12] P.A. Evans, Finding common subsequences with arcs and pseudoknots, in: Proc. of the 10th Annual Symposium on Combinatorial Pattern Matching (CPM), 1999, pp. 270–280.
[13] P.A. Evans, H.T. Wareham, Exact algorithms for computing d pairwise alignments and 3-medians from structure-annotated sequences (extended abstract), in: Proc. of the 6th Pacific Symposium on Biocomputing (PSB), 2001, pp. 559–570.
[14] M.R. Garey, D.S. Johnson, Crossing number is NP-complete, SIAM Journal on Algebraic and Discrete Methods 4 (1983).
[15] M.R. Garey, D.S. Johnson, Computers and Intractability, Freeman, San Francisco, 1979.
[16] S. Ieong, M.Y. Kao, T.W. Lam, W.K. Sung, S.M. Yiu, Predicting RNA secondary structures with arbitrary pseudoknots by maximizing the number of stacking pairs, in: Proc. of the 2nd Symposium on BioInformatics and BioEngineering (BIBE), 2002, pp. 183–190.
[17] T. Jacks, M. Power, F. Masiarz, P. Luciw, P. Barr, H. Varmus, Characterization of ribosomal frameshifting in HIV-1 gag-pol expression, Nature 331 (1988) 280–283.
[18] T. Jacks, H. Varmus, Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting, Science 230 (1985) 1237–1242.
[19] T. Jiang, G. Lin, B. Ma, K. Zhang, The longest common subsequence problem for arc-annotated sequences, in: Proc. of the 11th Annual Symposium on Combinatorial Pattern Matching (CPM), 2000, pp. 154–165.
[20] E. Korach, N. Solel, Tree-width, path-width, and cutwidth, Discrete Applied Mathematics 43 (1) (1993) 97–101.
[21] R.B. Lyngsø, C.N.S. Pedersen, RNA pseudoknot prediction in energy based models, Journal of Computational Biology 7 (2000) 409–428.
[22] S. Masuda, K. Nakajima, T. Kashiwabara, T. Fujisawa, Crossing minimization in linear embeddings of graphs, IEEE Transactions on Circuits and Systems 39 (1990).
[23] B. Monien, I.H. Sudborough, Min cut is NP-complete for edge weighted trees, Theoretical Computer Science 58 (1988).
[24] N. Robertson, P.D. Seymour, Graph minors II: Algorithmic aspects of tree-width, Journal of Algorithms 7 (1986) 309–322.
[25] D.M. Thilikos, M. Serna, H.L. Bodlaender, Cutwidth I: A linear time fixed parameter algorithm, ALGORITHMS: Journal of Algorithms 56 (2005).
[26] D.M. Thilikos, M. Serna, H.L. Bodlaender, Cutwidth II: Algorithms for partial w-trees of bounded degree, ALGORITHMS: Journal of Algorithms 56 (2005).