Int. J. Appl. Math. Comput. Sci., 2020, Vol. 30, No. 2, 315–324
DOI: 10.34768/amcs-2020-0024

NEW MODELS AND ALGORITHMS FOR RNA PSEUDOKNOT ORDER ASSIGNMENT

TOMASZ ZOK[a,b], JAN BADURA[a], SYLWESTER SWAT[a], KACPER FIGURSKI[b], MARIUSZ POPENDA[c], MACIEJ ANTCZAK[a,c,*]

[a] Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
e-mail: {tzok,jbadura,sswat,mantczak}@cs.put.poznan.pl

[b] Poznan Supercomputing and Networking Center
Jana Pawla II 10, 61-131 Poznan, Poland
e-mail: kfigurski@man.poznan.pl

[c] Institute of Bioorganic Chemistry
Polish Academy of Sciences
Noskowskiego 12/14, 61-704 Poznan, Poland
e-mail: marpop@ibch.poznan.pl

The pseudoknot is a specific motif of the RNA structure that highly influences the overall shape and stability of a molecule. It occurs when nucleotides of two disjoint single-stranded fragments of the same chain, separated by a helical fragment, interact with each other and form base pairs. Pseudoknots are characterized by great topological diversity, and their systematic description is still a challenge. In our previous work, we have introduced the *pseudoknot order*: a new coefficient representing the topological complexity of the pseudoknotted RNA structure. It is defined as the minimum number of base pair set decompositions, aimed to obtain the unknotted RNA structure. We have suggested how it can be useful in the interpretation and understanding of a hierarchy of RNA folding. However, it is not trivial to unambiguously identify pseudoknots and determine their orders in an RNA structure. Therefore, since the introduction of this coefficient, we have worked on the method to reliably assign pseudoknot orders in correspondence to the mechanisms that control the biological process leading to their formation in the molecule. Here, we introduce a novel graph coloring-based model for the problem of pseudoknot order assignment. We show a specialized heuristic operating on the proposed model and an alternative integer programming algorithm. The performance of both approaches is compared with that of state-of-the-art algorithms which so far have been most efficient in solving the problem in question. We summarize the results of computational experiments that evaluate our new methods in terms of classification quality on a representative data set originating from the non-redundant RNA 3D structure repository.

Keywords: RNA pseudoknot order, conflict graph, vertex coloring, maximum independent set, integer programming.

1. Introduction

Recent decades have seen close intermingling of computational and life sciences. Their collaboration has contributed to solving many problems related to various aspects of life at the molecular level, as well as the development of algorithms inspired by natural phenomena. It has also led to the evolution of new specialized research areas, such as structural bioinformatics, in which a key issue of this paper is anchored.

Primary research problems in the field of structural bioinformatics include (i) determining the 3D structure of molecules (Adamiak *et al.*, 2004; Blazewicz *et al.*, 2005; Slabinski *et al.*, 2007), (ii) 3D structure prediction (Kuang *et al.*, 2004; Seetin and Mathews, 2011; Leontis and Westhof, 2012; Antczak *et al.*, 2016), (iii) quality assessment of 3D structures (Zemla, 2003; Kuang *et al.*, 2004; Parisien *et al.*, 2009; Lukasiak *et al.*, 2015;

[*] Corresponding author.

Wiedemann and Milostan, 2017; Magnus *et al*., 2020), (iv) clustering molecules in order to identify common properties at various structural levels (Cheng *et al*., 2013; Zok *et al*., 2015), (v) discovering a relationship between the 3D fold and its biological function (Saenger, 1984; Blazewicz *et al*., 2011; Szostak *et al*., 2014; Rebis *et al*., 2015), (vi) analysing conformation changes and their influence on biological processes (Leontis and Westhof, 2012; Morimura *et al*., 2013; Adrjanowicz *et al*., 2016), (vii) performing molecular dynamics (Sarzynska and Kulinski, 2005; Popenda *et al*., 2009), (viii) searching for common structural characteristics (Pugalenthi *et al*., 2007; Wiedemann *et al*., 2017; Miskiewicz and Szachniuk, 2018; Popenda *et al*., 2020). All of these problems require solutions that involve sophisticated, theoretical models and efficient computational methods dedicated to a specific molecule: DNA, RNA or protein.

Biological molecule structures are analyzed at various levels of detail, taking into account selected topological characteristics (Kuppusamy and Mahendran, 2016). For example, structural motifs can be searched for in a sequence (i.e., a primary structure), in an intramolecular interaction network (i.e., a secondary structure), or in a full-atom 3D model (i.e., a tertiary structure). Their reliable identification and analysis often bring significant insight into the understanding of biological processes (Rybarczyk *et al*., 2017). Structural motifs play various important biological roles, e.g., serving as binding sites in molecular complexes, indicating cleavage sites in the biogenesis of specialized molecules, blocking or initiating some cellular processes, etc. Despite intensive studies of molecular motifs carried so far, many questions have remained unanswered. Some of them concern pseudoknots, which are specific RNA motifs that play crucial roles in various types of biological processes (Miao and Westhof, 2017).

The first recollection of a pseudoknot motif in the RNA structure comes from the work of Rietveld *et al*. (1982). Since then, motifs of this type have been studied with various intensity and from different perspectives (Staple and Butcher, 2005; Pasquali *et al*., 2005; Aalberts, 2005; Pillsbury *et al*., 2005; Rødland, 2006; Reidys *et al*., 2011; Bon *et al*., 2012; Vernizzi *et al*., 2016). Among others, significant efforts have been made to predict pseudoknotted RNA 3D structures, visualize pseudoknots in various representations of an RNA secondary structure, and to investigate and describe their topology. Within the latter issue, three leading approaches were developed. Kucharík *et al*. (2016) distinguished four specific pseudoknot topologies; H-type (interaction between a loop and a non-looped single strand), K-type (loop-loop interaction), L-type and M-type (more complex pseudoknot topologies), according to type and vicinity of fragments participating in the motif's formation. Bon *et al*. (2008) as well as Chiu and Chen (2012) introduced a genus concept that was applied to categorize entangled structure topologies. Finally, the pseudoknot order has been defined to quantify structural complexity of pseudoknotted RNA structures (Antczak *et al*., 2014).

Here, we describe the pseudoknot order assignment problem (abbreviated as POA). We present two novel algorithmic approaches solving this problem. The first one, named MIS (maximum independent set-based algorithm), applies an optimization algorithm built on a new graph model of POA. An alternative solution, abbreviated as MILP, is a mixed-integer linear programming-based model solved using CPLEX. Computational experiments show that both the proposed approaches provide better results than state-of-the-art algorithms, especially for large, pseudoknotted RNA structures.

2. Pseudoknot order assignment problem

In this work, we focus on the RNA secondary structure, which is represented as a network of interactions occurring within an RNA molecule. A single interaction results in the formation of a pair of bounded bases. Therefore, the most common representation of a secondary structure is a list of base pairs. Every base is represented by a unique, serial number that corresponds to its position in an RNA chain. Thus, a base pair is represented as a pair of numbers.

Let us consider two base pairs (i, j) and (k, l) that belong to RNA structure M, where $i < j$, $k < l$, $i < k$. They are found in one of the following relationships:

1. $i < k < l < j$: base pair (k, l) is nested in (i, j),
2. $i < j < k < l$: base pairs are disjoint,
3. $i < k < j < l$: base pairs interlace and form a pseudoknot.

The pseudoknot is thus composed of at least two base pairs that interlace. Additionally, this motif may contain any number of nested and disjoint base pairs. At least one base pair interlaces with the remaining ones.

The pseudoknot order ($psorder$) has been introduced to assess the topological complexity of motifs of this kind. According to Antczak *et al*. (2014), it is calculated as the minimum number of decompositions of the entire base pair set, which lead to obtaining the unknotted structure (i.e., a set including only nested and disjoint pairs). Further analysis by Antczak *et al*. (2018) has shown that this coefficient can be useful in interpretation and understanding of the RNA folding hierarchy. We assume that at the preliminary step of RNA folding only nested (cf. Fig. 1(a)) and disjoint (cf. Fig. 1(b)) base pairs are formed. Their $psorder = 0$. Next, topologically simple pseudoknots (with $psorder = 1$) are formed.

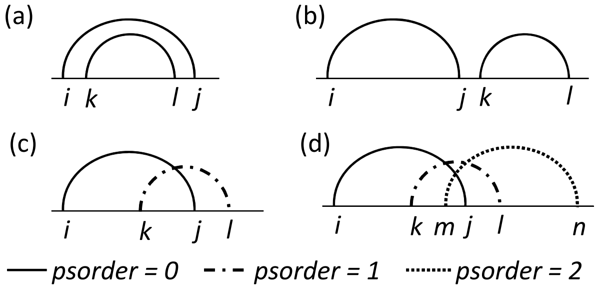

Fig. 1. Arc diagrams that represent nested base pairs (a), disjoint base pairs (b), a pseudoknot of order 1 (c), a pseudoknot of order 2 (d).

They are induced by base pairs that interlace with any of previously formed pairs (cf. Fig. 1(c)). Further, base pairs that interlace with simple pseudoknots form more topologically complex motif with $psorder = 2$ (cf. Fig. 1(d)), etc. Concluding, the pseudoknot of the n-th order contains at least one base pair that interlaces with n previously formed base pairs having pseudoknot orders $n-1, n-2, \ldots, 0$. In the problem of pseudoknot order assignment, having the input RNA secondary structure, we trace back the RNA folding process designating orders for all base pairs involved in pseudoknot formation. If two base pairs interlace, we need to assign $psorder = n$ to one of them, and $psorder = n-1$ to the other ($n \geq 1$).

Let us define the pseudoknot order assignment problem in the following way. Having a set of base pairs $B = \bigcup_{k\geq 1}(i_k, j_k)$, assign $psorder \geq 0$ to every base pair in B according to a bilevel optimization procedure (Antczak *et al.*, 2018) considering the following phases: (i) first, maximize the number of base pairs with $psorder = 0$ (they form a core of RNA structure); (ii) next, minimize the sum of values of positive ($psorder \geq 1$) pseudoknot orders assigned to all other base pairs.

3. Criterion function

Solving the POA problem requires the use of a criterion function that evaluates the quality of feasible solutions. Here, we apply the function $fscore$ from the work of Antczak *et al.* (2018). It operates on a compressed representation of an RNA secondary structure in which series of neighboring base pairs are treated as units called regions.

As mentioned in Section 2, the pseudoknot is often formed by double-stranded fragments (consisting of nested base pairs) that intertwine within the structure. Thus, we can represent the secondary structure by enumerating its double-stranded fragments (regions) instead of single base pairs. A region is represented by its outermost base pair and the number of pairs nested in it, uninterrupted by any unpaired base. The latter value denotes the region length. Thus, the RNA secondary structure can be encoded as a set R of triples $r = (i, j, l)$, where i and j represent bases of the outermost pair of region r, and l is the length of region r. Having such a representation of the structure, we solve the POA problem by assigning pseudoknot orders to regions in R.

We state that $fscore$ is represented by a pair of integers $(fcore, fpk)$. The first component denotes the size of a core structure and is computed according to

$$fcore = \sum_{r_k \in R} (1 - x_k) \cdot l(r_k), \tag{1}$$

where $l(r_k)$ denotes the length of region r_k and x_k is a decision variable:

$$x_k = \begin{cases} 0 & \text{if } psorder(r_k) = 0, \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

The second component is a sum of regions' lengths weighted by their $psorder$ and is calculated according to

$$fpk = -1 \cdot \sum_{r_k \in R} psorder(r_k) \cdot l(r_k). \tag{3}$$

We use the following domination rule (Antczak *et al.*, 2018) to decide which feasible solution is better:

$$\begin{gathered} fscore_i > fscore_j \\ \equiv \\ fcore_i > fcore_j \vee (fcore_i = fcore_j \wedge fpk_i > fpk_j). \end{gathered} \tag{4}$$

Thus, the first component $fcore$ takes precedence when looking for the maximum value of $fscore$. Only if two feasible solutions have the same $fcore$ values is fpk used. For example, $fscore$ of $(5, -2)$ is better than $(4, -3)$ (cf. Fig. 2(a)) because the former one includes a higher number of base pairs in the core structure. Moreover, $(5, -4)$ is better than $(5, -5)$ (cf. Fig. 2(b)) because we first minimize the $psorder$ for the whole structure and, next, we maximize the number of base pairs with the lower $psorder$.

4. Graph-based solution

Graph theory has proved its effectiveness in modeling a large number of problems derived from life sciences (Gan *et al.*, 2003; Giuliani *et al.*, 2008; Blazewicz *et al.*, 2013; 2018; Szachniuk *et al.*, 2014; 2015; Wojciechowski *et al.*, 2016; Radom *et al.*, 2017). Graph-based models are often applied, e.g., to show a discrete view of biomolecular architecture, including an RNA secondary structure (Simon, 2005; Lai *et al.*, 2012; Schlick, 2018). In our proposal, we have developed a new graph model for the POA problem.

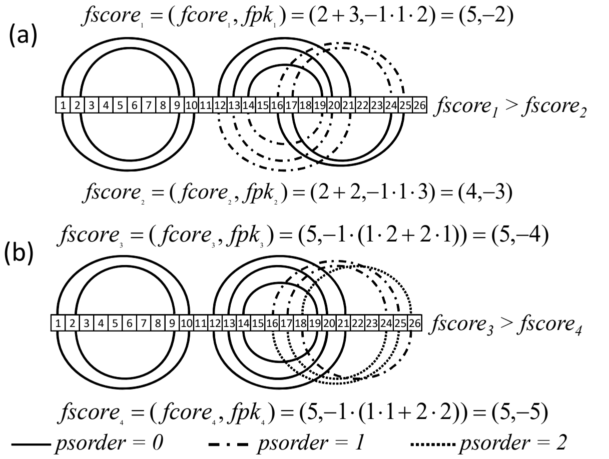
(a)
$fscore_1 = (fcore_1, fpk_1) = (2+3, -1\cdot1\cdot2) = (5, -2)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
$fscore_1 > fscore_2$
$fscore_2 = (fcore_2, fpk_2) = (2+2, -1\cdot1\cdot3) = (4, -3)$
(b)
$fscore_3 = (fcore_3, fpk_3) = (5, -1\cdot(1\cdot2+2\cdot1)) = (5, -4)$
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
$fscore_3 > fscore_4$
$fscore_4 = (fcore_4, fpk_4) = (5, -1\cdot(1\cdot1+2\cdot2)) = (5, -5)$
psorder = 0
psorder = 1
psorder = 2

Fig. 2. Criterion function interpretation examples.

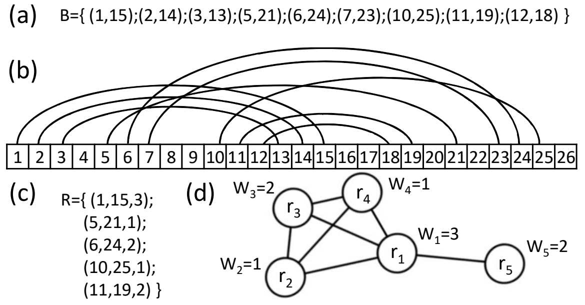
(a) B={ (1,15);(2,14);(3,13);(5,21);(6,24);(7,23);(10,25);(11,19);(12,18) }
(b)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
(c) R={ (1,15,3); (5,21,1); (6,24,2); (10,25,1); (11,19,2) }
(d)
W3=2
r3
r4
W4=1
W1=3
r1
W5=2
r5
W2=1
r2

Fig. 3. Example RNA secondary structure represented as a set of base pairs (a), an arc diagram (b), a set of triples (regions) (c), a conflict graph with weighted vertices (d).

4.1. Graph-based model of the problem.

Let $G = (V, E)$ be an undirected graph, where V is a set of weighted vertices and E a set of edges. G is an *RNA conflict graph* if the following conditions are satisfied. Every vertex $v_k \in V$ represents region $r_k \in R$ of RNA molecule M, where R is a set of triples encoding the secondary structure of M. Every vertex $v_k \in V$ is associated with weight $w_k = l(r_k)$, where $l(r_k)$ is the number of nested base pairs forming r_k (i.e., the length of r_k). If two regions, r_i and r_j, of RNA structure M interlace (i.e., they are *in conflict*), their corresponding vertices, $v_i, v_j \in V$, are connected with an edge $e(v_i, v_j) \in E$ (cf. Fig. 2(d)).

A feasible solution of the POA problem is a vertex-coloring of a given graph G with nc colors. Colors are represented by natural numbers following the formula $psorder(r_i) = c(v_i) - 1$, where $c(v_i)$ is the color of vertex v_i. Thus, the optimum solution is a coloring scheme which (i) maximizes the sum of weights of vertices assigned with color 1 and (ii) minimizes the sum of products of vertices' weights and their colors for all vertices for which $c(v_i) > 1$.

4.2. MIS algorithm.

To explain the MIS (*maximum independent set*-based) algorithm, we need to recall several definitions from graph theory. An *independent set* in a given graph $G = (V, E)$ is a subset $P \subset V$ in which no nodes are connected by an edge from E. Set P is *maximal* if there is no vertex $v \in V \setminus P$ such that $P \cup \{v\}$ is also an independent set in G. Finally, a *maximum independent set* is a maximal independent set with the biggest size (counted by the number of vertices included) (Tarjan and Trojanowski, 1977). Let us note that every maximum independent set is also maximal. However, the opposite statement is not always true (cf. Fig. 3).

Graph coloring and independent set determination problems are closely related because in a feasible coloring scheme vertices with the same color represent an independent set. In the POA problem, the objective is to find a coloring scheme which inherently optimizes color assignments. In the proposed MIS algorithm the solution is built by constructing maximal independent sets which are then translated to colors and indirectly to pseudoknot orders of underlying RNA base pairs.

The POA problem can be solved by finding partition $P = (P_1, P_2, \ldots, P_{nc})$ of vertices of $G = (V, E)$ such that, for each $i \in [1, nc]$, P_i is an independent set in G and $\bigcup_{i=1}^{nc} P_i = V$. Obviously, if $f : V \rightarrow \{c_1, c_2, \ldots, c_{nc}\}$ is the coloring scheme of a given graph G, then $P_i = f^{-1}(i)$.

The complexity of the input graph G determines a scenario followed during solving the POA problem. Thus, MIS selects the appropriate processing scheme based on the number of vertices $|V|$ of a given graph G. If $|V| > N$ (by default $N = 50$), the graph is considered complex. Otherwise, the graph is treated as simple. In the case of simple graphs, an exhaustive search algorithm identifies an optimal solution (if it exists) in which all P_i sets are maximum independent sets of their corresponding graphs. The heuristic procedure linearly depends on the sum of vertex and edge numbers in a given graph G. Otherwise, the best solution maximizing $fscore$ is found. For complex graphs, the first set P_1, representing regions with $psorder = 0$, is found by a heuristic approach. Next, MIS recursively solves the POA problem for a reduced graph $G[V \setminus P_1]$ and incrementally constructs the solution based on partial results. In this case, the Bron–Kerbosch algorithm, which is exponential, is used for identification of all maximal independent sets in a given graph G. So, in the worst case, the MIS algorithm is also exponential. However, in practice, its computational efficiency is good enough (cf. Table 2).

CGO: Complex graph-oriented scenario. Although maximal independent sets for a given graph G are found in polynomial time (Luby, 1986), it remains difficult to select the promising one that is also the maximum independent set. To identify the most promising partial

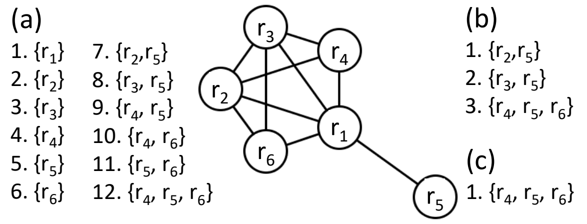

Fig. 4. Independent sets in example graph G: all independent sets (a), maximal independent sets (b), maximum independent set (c).

solution, MIS applies the following steps: (i) construct set $S = \{S_1, S_2, \ldots, S_t\}$ of random maximal independent sets of graph G (Luby, 1986), where t is a positive integer, (ii) identify intersection K of all sets in S, (iii) assign $psorder = 0$ to all regions corresponding to vertices in K that are treated as solved. Next, the hybrid algorithm (Antczak *et al.*, 2018) finds the final solution focusing only on still unsolved regions. Note that if we find only random maximal independent sets S_i in graph G, we cannot be certain if any of them are the maximum independent set. However, computational experiments show that, if t is large enough (e.g., 10^4), the maximum independent set is usually found. Furthermore, the size of K is usually large. Thus, it significantly reduces the search space of the hybrid algorithm, which next leads to significant improvement of the solution.

SGO: Simple graph-oriented scenario. In this scenario, MIS exhaustively searches all partitions $P = (P_1, \ldots, P_{nc})$, where P_1 is the maximal independent set of a given graph $G = (V, E)$, P_i is the maximal independent set of $G[V \setminus (\bigcup_{j=1}^{i-1} P_i)]$ for $2 \leq i \leq nc$, and $|P_i|$ is not smaller than the size of the maximum independent set of the corresponding graph decreased by 1 for $i \in [1, nc]$. This is done as follows: (i) find all maximal independent sets Z in a given graph G using the Bron–Kerbosch algorithm (Bron and Kerbosch, 1973), (ii) for each maximal independent set $I \in Z$ satisfying the size constraint, identify recursively the current solution for graph $G[V \setminus I]$ using the hybrid algorithm (Antczak *et al.*, 2018), (iii) assess the current solution using $fscore$ and remember it if it is the best one so far.

Note that using this routine to identify all partitions $P = (P_1, \ldots, P_{nc})$ may not lead to finding the optimum with respect to $fscore$. Instead of searching for all maximal independent sets of G, one can only consider independent sets satisfying the size constraint. Thus, larger graphs can be also effectively processed.

5. MILP-based solution

The POA problem can be modeled as a *mixed-integer linear programming* (MILP) one. It requires finding an upper-bound on the pseudoknot order for input RNA using the first-come-first-served algorithm (Antczak *et al.*, 2018). Then, the MILP-based model is defined as follows:

1. *Goal function*
   The objective of this problem is to maximize the pseudoknot classification quality in the POA problem, concerning the $fscore$ function:
   $$\max \sum_{r_k \in R} y_{1,r_k} \cdot l(r_k) \cdot C_{psorder} - \sum_{i=2}^{nc} \sum_{r_k \in R} y_{i,r_k} \cdot l(r_k) \cdot psorder(r_k), \tag{5}$$
   where $C_{psorder} = 10$ is a constant value selected to exceed the $psorder$ that can be assigned to any base pair in known, experimentally determined RNA 3D structures.

2. *Decision variables*
   $$y_{i,r_k} = \begin{cases} 1 & \text{if } psorder(r_k) = i - 1, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

3. *Constraints*
   Every region must have exactly one value of the pseudoknot order assigned:
   $$\forall r_k \in R \sum_{i \in [1,nc]} y_{i,r_k} = 1. \tag{7}$$
   Two interlacing regions $r_k, r_l \in R$ are assigned different values of $psorder$:
   $$\forall e(r_k, r_l) \in E \quad \forall i \in [1, nc] \quad y_{i,r_k} + y_{i,r_l} \leq 1. \tag{8}$$

It is well known that MILP is NP-hard. In the worst case, the algorithm verifies all potential combination of values for the integral variables and then solves the resulting LP. Nevertheless, the proposed algorithm is efficient in practice (cf. Table 2) taking into account the sizes of known experimentally determined RNA 3D structures.

6. Computational experiments

The performance of the MIS and MILP algorithms was assessed in the context of state-of-the-art methods, i.e., hybrid (HYB) and dynamic programming (DP)-based (Antczak *et al.*, 2018). The HYB algorithm uses an exhaustive search method for small instances and random walk for larger ones. DP optimally solves the POA problem restricted to the $fcore$ criterion function only (i.e., DP does not optimize fpk component of $fscore$). The MILP model was solved using the CPLEX solver

(ver. 12.7). The MIS algorithm was developed in C++ and integrated selected procedures from the HYB method (Antczak *et al.*, 2018). All algorithms were executed and assessed in a coherent computational infrastructure.

A repository of complex pseudoknotted RNAs for computational tests was constructed based on a non-redundant set of representative RNA 3D structures (Leontis and Zirbel, 2012) with 491 entries. The set was processed by 3DNA/DSSR with the helix analysis option (Lu and Olson, 2008) to get the secondary structure (the base pair list) for every RNA considered. Next, all RNAs without pseudoknots were removed. Finally, we obtained the test set consisting of 240 pseudoknotted RNA secondary structures.

Every RNA secondary structure from the test set was processed using all the algorithms considered. In many cases, the HYB method (Antczak *et al.*, 2018), as well as both new algorithms, returned the same solution. It was the case for all small or topologically simple structures. However, for large and complex pseudoknotted RNAs, the diversity of generated solutions was high. The latter examples created a set of 48 structures with pseudoknots of up to the 7th order, on which we compared the algorithms' performance.

In the first experiment, solutions generated by all the algorithms were evaluated based on the criterion function $fscore$, and their all-against-all comparison was made. The results of this experiment are presented in Table 1. It shows how many times the algorithm from the first column won a duel with algorithms listed in columns 2–5. It also presents the total number of duels and battles won. For example, we can see that MILP won 20 duels with DP (i.e., in 20 cases its solution dominated over the solution found by DP), 13 duels with HYB and 3 duels with MIS. The number of all duels won by MILP equals 36. In four cases, it was better than every other algorithm.

In the second experiment, we evaluated the generated solutions based on the maximum $psorder$ value assigned by each algorithm for every input structure. Again, all-against-all comparison was performed and its results are presented in Table 2. The maximum value of the $psorder$ of a winner (in the duel) must be lower as compared to the loser. Let us notice that the results in Tables 1 and 2 differ. Winning the duel in the first experiment (Table 1) is easier than in the second one (Table 2).

The results confirm that the MILP-based method outperforms all the others. It is the only approach to win a battle. Considering the total number of duels won, its dominance over the the MIS algorithm is not so significant as far as the $fscore$ value is concerned. However, if the maximum value of $psorder$ is concerned, MILP dominates the MIS method. Note that MILP algorithm won four battles, however, only for three of them it achieved the lowest value of the maximum pseudoknot order. The MIS approach is ranked the second best. It can be easily noticed from the analysis of duels that both proposed approaches stand out among state-of-the-art algorithms. Moreover, they did not lose a single battle.

Although the proposed algorithms are exponential in the worst case, their computational efficiency (Table 3) confirms that they can be successfully applied in practice.

In the third experiment, we performed a single case test to analyze one large, pseudoknotted RNA structure. We selected the yeast mitochondrial ribosome structure (PDBid: 5MRC, chain A) (Desai *et al.*, 2017) to compare all the algorithms performance. This RNA molecule contains 1027 base pairs. Among them, 971 pairs were identified by all the algorithms to form the core structure (their $psorder = 0$). All the algorithms also found the same set of 31 base pairs with $psorder = 1$. However, for higher-order pseudoknots we can see the differences in the generated solutions. The maximum pseudoknot order assigned by MIS and MILP equals 5, and it is lower by 1 and 2 than in solutions generated by the HYB and DP approaches, respectively. Table 4 shows detailed information about solutions found by all the algorithms for the analyzed structure. For each algorithm we can see to how many base pairs it assigned $psorder = 0, 1, 2, 3, 4, 5, 6, 7$.

Following the criterion function $fscore$, the lower the pseudoknot order value, the more reliable the RNA secondary structure. We believe that the pseudoknot order represents the number of tiers in the RNA folding hierarchy (Antczak *et al.*, 2018). Therefore, solutions found by both MILP and MIS ($psorder = 5$) are more convincing than those found by HYB ($psorder = 6$) and DP ($psorder = 7$). From a biological point of view, the higher the number of base pairs formed during initial tiers of RNA folding, the better. So, for $psorder = 2$, MILP and MIS formed 12 base pairs but HYB and DP formed only 11 and 10 of them, respectively. Moreover, this difference is next propagated into base pairs formed in tiers with a higher value of $psorder$ assigned.

7. Conclusions

Discovering and classifying specific motifs in molecular structures is an important challenge of many modern bioinformatics problems. Understanding unique properties and spatial neighborhood of structural patterns allows one to realise their impact on the functioning of biological molecules and can be successfully applied in the service of biotechnology and biomedicine. One of the most interesting motifs are pseudoknots, often found in complex and biologically important RNA molecules. Their topological characteristics can be described by the pseudoknot order coefficient, introduced and discussed in our previous works (Antczak *et al.*, 2014; 2018; Zok *et al.*, 2018). Here, we presented new algorithmic

Table 1. Number of duels and battles won by each algorithm as compared to the others (concerning the $fscore$ value).

	DP	HYB	MILP	MIS	# Duels won	# Battles won
DP	–	9	0	0	9	0
HYB	10	–	0	0	10	0
MILP	20	13	–	3	36	4
MIS	20	13	0	–	33	0
# Duels lost	50	35	0	3	–	–
# Battles lost	10	9	0	0	–	–

Table 2. Number of duels and battles won by each algorithm as compared to the others (concerning the maximal value of $psorder$).

	DP	HYB	MILP	MIS	# Duels won	# Battles won
DP	–	1	0	0	1	0
HYB	7	–	0	0	7	0
MILP	12	7	–	3	22	3
MIS	9	4	0	–	13	0
# Duels lost	28	12	0	3	–	–
# Battles lost	7	1	0	0	–	–

Table 3. Comparison of computational efficiency for algorithms considered: average processing time (standard deviation) in [ms].

DP	HYB	MILP	MIS
1097 (22)	1198 (97)	69 (65)	49833 (75347)

Table 4. Base pair distribution in solutions generated by all the algorithms for the 5MRC structure.

psorder	0	1	2	3	4	5	6	7
DP	971	31	10	8	3	2	1	1
HYB	971	31	11	7	5	1	1	0
MILP	971	31	12	7	4	2	0	0
MIS	971	31	12	7	4	2	0	0

approaches solving the pseudoknot order assignment problem in a set of base pairs that represent the secondary structure of an RNA molecule. They outperform state-of-the-art algorithms in terms of the classification quality, especially for large and topologically complex RNA structures. Therefore, they can better reflect the biological nature of the problem in question.

We believe that the proposed algorithms will significantly contribute to further development of 3D RNA structures prediction methods handling pseudoknots, especially for approaches based on the input RNA secondary structure (Purzycka *et al.*, 2015; Shi *et al.*, 2018; Lim and Brown, 2018; Szachniuk, 2019). In the future, we plan to address the issues common to all dynamic systems, including RNAs, which is uncertainty and errors, both in data acquisition and processing (Kropat *et al.*, 2016). Furthermore, recent developments indicate that there are conformational switches in the RNA folding pathway (Sun *et al.*, 2018). This suggests that pseudoknot order assignment can be analyzed as a time-dependent problem (Gebert *et al.*, 2006; Weber *et al.*, 2009; 2011; Kruthika *et al.*, 2017). Appropriate adaptation of this perspective can lead to more robust solutions.

8. Code availability

Source codes of the proposed algorithms are available on request from the corresponding author.

Acknowledgment

This research was carried out in the European Centre for Bioinformatics and Genomics, Poznan University of Technology (Poland). It was supported by the National Science Centre, Poland (2016/23/B/ST6/03931), and statutory funds of the Poznan University of Technology (0311/SBAD/0684). The authors thank Prof. Marta Szachniuk for critical proofreading and discussions.

References

Aalberts, D.P. (2005). Asymmetry in RNA pseudoknots: Observation and theory, *Nucleic Acids Research* **33**(7): 2210–2214.

Adamiak, R., Blazewicz, J., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M. and Szachniuk, M. (2004). An algorithm for an automatic NOE pathways analysis in 2D NMR spectra of RNA duplexes, *Journal of Computational Biology* **42**(11): 163–180.

Adrjanowicz, K., Kaminski, K., Tarnacka, M., Szutkowski, K., Popenda, L., Bartkowiak, G. and Paluch, M. (2016). The effect of hydrogen bonding propensity and enantiomeric composition on the dynamics of supercooled ketoprofen—Dielectric, rheological and NMR studies, *Physical Chemistry Chemical Physics* **18**(15): 10585–10593.

Antczak, M., Popenda, M., Zok, T., Sarzynska, J., Ratajczak, T., Tomczyk, K., Adamiak, R.W. and Szachniuk, M. (2016). New functionality of RNAComposer: Application to shape the axis of miR160 precursor structure, *Acta Biochimica Polonica* **63**(4): 737–744.

Antczak, M., Popenda, M., Zok, T., Zurkowski, M., Adamiak, R.W. and Szachniuk, M. (2018). New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation, *Bioinformatics* **34**(8): 1304–1312.

Antczak, M., Zok, T., Popenda, M., Lukasiak, P., Adamiak, R.W., Blazewicz, J. and Szachniuk, M. (2014). RNApdbee—A webserver to derive secondary structures from PDB files of knotted and unknotted RNAs, *Nucleic Acids Research* **42**(W1): W368–W372.

Blazewicz, J., Figlerowicz, M., Kasprzak, M., Nowacka, M. and Rybarczyk, A. (2011). RNA partial degradation problem: Motivation, complexity, algorithm, *Journal of Computational Biology* **18**(6): 821–834.

Blazewicz, J., Frohmberg, W., Gawron, P., Kasprzak, M., Kierzynka, M., Swiercz, A. and Wojciechowski, P. (2013). DNA sequence assembly involving an acyclic graph model, *Foundations of Computing and Decision Sciences* **38**(1): 25–34.

Blazewicz, J., Kasprzak, M., Kierzynka, M., Frohmberg, W., Swiercz, A., Wojciechowski, P. and Zurkowski, P. (2018). Graph algorithms for DNA sequencing—Origins, current models and the future, *European Journal of Operational Research* **264**(3): 799–812.

Blazewicz, J., Szachniuk, M. and Wojtowicz, A. (2005). RNA tertiary structure determination: NOE pathways construction by tabu search, *Bioinformatics* **21**(10): 2356–2361.

Bon, M., Micheletti, C. and Orland, H. (2012). McGenus: a Monte Carlo algorithm to predict RNA secondary structures with pseudoknots, *Nucleic Acids Research* **41**(3): 1895–1900.

Bon, M., Vernizzi, G., Orland, H. and Zee, A. (2008). Topological classification of RNA structures, *Journal of Molecular Biology* **379**(4): 900–911.

Bron, C. and Kerbosch, J. (1973). Algorithm 457: Finding all cliques of an undirected graph, *Communications of the ACM* **16**(9): 575–577.

Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. and Corander, J. (2013). Hierarchical and spatially explicit clustering of DNA sequences with BAPS software, *Molecular Biology and Evolution* **30**(5): 1224–1228.

Chiu, J.K.H. and Chen, Y.-P.P. (2012). Conformational features of topologically classified RNA secondary structures, *PLoS ONE* **7**(7): e39907.

Desai, N., Brown, A.A. and Ramakrishnan, V. (2017). The structure of the yeast mitochondrial ribosome, *Science* **355**(6324): 528–531.

Gan, H.H., Pasquali, S. and Schlick, T. (2003). Exploring the repertoire of RNA secondary motifs using graph theory: Implications for RNA design, *Nucleic Acids Research* **31**(11): 2926–2943.

Gebert, J., Lätsch, M., Pickl, S.W., Weber, G. and Wünschiers, R. (2006). An algorithm to analyze stability of gene-expression patterns, *Discrete Applied Mathematics* **154**(7): 1140–1156.

Giuliani, A., Krishnan, A., Zbilut, J. and Tomita, M. (2008). Proteins as networks: Usefulness of graph theory in protein science, *Current Protein & Peptide Science* **9**(1): 28–38.

Kropat, E., Özmen, A., Weber, G., Meyer-Nieberg, S. and Defterli, O. (2016). Fuzzy prediction strategies for gene-environment networks—Fuzzy regression analysis for two-modal regulatory systems, *RAIRO Operations Research* **50**(2): 413–435.

Kruthika, H.A., Mahindrakar, A.D. and Pasumarthy, R. (2017). Stability analysis of nonlinear time-delayed systems with application to biological models, *International Journal of Applied Mathematics and Computer Science* **27**(1): 91–103, DOI: 10.1515/amcs-2017-0007.

Kuang, R., Leslie, C.S. and Yang, A.-S. (2004). Protein backbone angle prediction with machine learning approaches, *Bioinformatics* **20**(10): 1612–1621.

Kucharík, M., Hofacker, I.L., Stadler, P.F. and Qin, J. (2016). Pseudoknots in RNA folding landscapes, *Bioinformatics* **32**(2): 187–194.

Kuppusamy, L. and Mahendran, A. (2016). Modelling DNA and RNA secondary structures using matrix insertion–deletion systems, *International Journal of Applied Mathematics and Computer Science* **26**(1): 245–258, DOI: 10.1515/amcs-2016-0017.

Lai, D., Proctor, J.R., Zhu, J.Y.A. and Meyer, I.M. (2012). R-CHIE: A web server and R package for visualizing RNA secondary structures, *Nucleic Acids Research* **40**(12): e95.

Leontis, N.B. and Zirbel, C.L. (2012). Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking, *in* N. Leontis and E. Westhof (Eds), *Nucleic Acids and Molecular Biology*, Springer Nature, Berlin/Heidelberg, pp. 281–298.

Leontis, N. and Westhof, E. (2012). *RNA 3D Structure Analysis and Prediction*, Springer, Berlin/New York, NY.

Lim, C.S. and Brown, C.M. (2018). Know your enemy: Successful bioinformatic approaches to predict functional RNA structures in viral RNAs, *Frontiers in Microbiology* **8**: 2582.

Lu, X.-J. and Olson, W.K. (2008). 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures, *Nature Protocols* **3**(7): 1213–1227.

Luby, M. (1986). A simple parallel algorithm for the maximal independent set problem, *SIAM Journal on Computing* **15**(4): 1036–1053.

Lukasiak, P., Antczak, M., Ratajczak, T., Szachniuk, M., Popenda, M., Adamiak, R.W. and Blazewicz,

J. (2015). RNAssess—A web server for quality assessment of RNA 3D structures, *Nucleic Acids Research* **43**(W1): W502–W506.

Magnus, M., Antczak, M., Zok, T., Wiedemann, J., Lukasiak, P., Cao, Y., Bujnicki, J.M., Westhof, E., Szachniuk, M. and Miao, Z. (2020). RNA-Puzzles toolkit: A computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools, *Nucleic Acids Research* **48**(2): 576–588.

Miao, Z. and Westhof, E. (2017). RNA structure: Advances and assessment of 3D structure prediction, *Annual Review of Biophysics* **46**: 483–503.

Miskiewicz, J. and Szachniuk, M. (2018). Discovering structural motifs in miRNA precursors from the *Viridiplantae* kingdom, *Molecules* **23**(6): 1367.

Morimura, H., Tanaka, S.-I., Ishitobi, H., Mikami, T., Kamachi, Y., Kondoh, H. and Inouye, Y. (2013). Nano-analysis of DNA conformation changes induced by transcription factor complex binding using plasmonic nanodimers, *ACS Nano* **7**(12): 10733–10740.

Parisien, M., Cruz, J.A., Westhof, E. and Major, F. (2009). New metrics for comparing and assessing discrepancies between RNA 3D structures and models, *RNA* **15**(10): 1875–1885.

Pasquali, S., Gan, H. and Schlick, T. (2005). Modular RNA architecture revealed by computational analysis of existing pseudoknots and ribosomal RNAs, *Nucleic Acids Research* **33**(4): 1384–1398.

Pillsbury, M., Orland, H. and Zee, A. (2005). Steepest descent calculation of RNA pseudoknots, *Physical Review E* **72**(1).

Popenda, L., Bielecki, L., Gdaniec, Z. and Adamiak, R.W. (2009). Structure and dynamics of adenosine bulged RNA duplex reveals formation of the dinucleotide platform in the C:G-A triple, *Arkivoc* **2009**(3): 130–144.

Popenda, M., Miskiewicz, J., Sarzynska, J., Zok, T. and Szachniuk, M. (2020). Topology-based classification of tetrads and quadruplex structures, *Bioinformatics* **36**(4): 1129–1134.

Pugalenthi, G., Suganthan, P.N., Sowdhamini, R. and Chakrabarti, S. (2007). SMotif: A server for structural motifs in proteins, *Bioinformatics* **23**(5): 637–638.

Purzycka, K., Popenda, M., Szachniuk, M., Antczak, M., Lukasiak, P., Blazewicz, J. and Adamiak, R. (2015). Automated 3D RNA structure prediction using the RNAComposer method for riboswitches, *in* S.J. Chen and D.H. Burke Aguero (Eds), *Methods in Enzymology*, Vol. 553, Elsevier, San Diego, CA, pp. 3–34.

Radom, M., Rybarczyk, A., Szawulak, B., Andrzejewski, H., Chabelski, P., Kozak, A. and Formanowicz, P. (2017). Holmes: A graphical tool for development, simulation and analysis of Petri net based models of complex biological systems, *Bioinformatics* **33**(23): 3822–3823.

Rebis, T., Lijewski, S., Nowicka, J., Popenda, L., Sobotta, L., Jurga, S., Mielcarek, J., Milczarek, G. and Goslinski, T. (2015). Electrochemical properties of metallated porphyrazines possessing isophthaloxybutylsulfanyl substituents: Application in the electrocatalytic oxidation of hydrazine, *Electrochimica Acta* **168**: 216–224.

Reidys, C.M., Huang, F.W.D., Andersen, J.E., Penner, R.C., Stadler, P.F. and Nebel, M.E. (2011). Topology and prediction of RNA pseudoknots, *Bioinformatics* **27**(8): 1076–1085.

Rietveld, K., Poelgeest, R.V., Pleij, C., Boom, J.V. and Bosch, L. (1982). The tRNA-like structure at the 3*ı* terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA, *Nucleic Acids Research* **10**(6): 1929–1946.

Rødland, E.A. (2006). Pseudoknots in RNA secondary structures: Representation, enumeration, and prevalence, *Journal of Computational Biology* **13**(6): 1197–1213.

Rybarczyk, A., Hertz, A., Kasprzak, M. and Blazewicz, J. (2017). Tabu search for the RNA partial degradation problem, *International Journal of Applied Mathematics and Computer Science* **27**(2): 401–415, DOI: 10.1515/amcs-2017-0028.

Saenger, W. (1984). *Principles of Nucleic Acid Structure*, Springer-Verlag, London.

Sarzynska, J. and Kulinski, T. (2005). Dynamics and stability of GCAA tetraloops with 2-aminopurine and purine substitutions, *Journal of Biomolecular Structure and Dynamics* **22**(4): 425–439.

Schlick, T. (2018). Adventures with RNA graphs, *Methods* **143**: 16–33.

Seetin, M. and Mathews, D. (2011). Automated RNA tertiary structure prediction from secondary structure and low-resolution restraints, *Journal of Computational Chemistry* **32**(10): 2232–2244.

Shi, Y.-Z., Jin, L., Feng, C.-J., Tan, Y.-L. and Tan, Z.-J. (2018). Predicting 3D structure and stability of RNA pseudoknots in monovalent and divalent ion solutions, *PLOS Computational Biology* **14**(6): e1006222.

Simon, M. (2005). *Emergent Computation. Emphasizing Bioinformatics*, Springer New York, New York, NY.

Slabinski, L., Jaroszewski, L., Rodrigues, A.P., Rychlewski, L., Wilson, I.A., Lesley, S.A. and Godzik, A. (2007). The challenge of protein structure determination-lessons from structural genomics, *Protein Science* **16**(11): 2472–2482.

Staple, D.W. and Butcher, S.E. (2005). Pseudoknots: RNA structures with diverse functions, *PLoS Biology* **3**(6): e213.

Sun, T.-t., Zhao, C. and Chen, S.-J. (2018). Predicting cotranscriptional folding kinetics for riboswitch, *The Journal of Physical Chemistry B* **122**(30): 7484–7496.

Szachniuk, M. (2019). RNApolis: Computational platform for RNA structure analysis, *Foundations of Computing and Decision Sciences* **44**(2): 241–257.

Szachniuk, M., Cola, M.C.D., Felici, G. and Blazewicz, J. (2014). The orderly colored longest path problem—A survey of applications and new algorithms, *RAIRO—Operations Research* **48**(1): 25–51.

Szachniuk, M., Cola, M.C.D., Felici, G., de Werra, D. and Blazewicz, J. (2015). Optimal pathway reconstruction on 3D NMR maps, *Discrete Applied Mathematics* **182**: 134–149.

Szostak, N., Royo, F., Rybarczyk, A., Szachniuk, M., Blazewicz, J., del Sol, A. and Falcon-Perez, J.M. (2014). Sorting signal targeting mRNA into hepatic extracellular vesicles, *RNA Biology* **11**(7): 836–844.

Tarjan, R.E. and Trojanowski, A.E. (1977). Finding a maximum independent set, *SIAM Journal on Computing* **6**(3): 537–546.

Vernizzi, G., Orland, H. and Zee, A. (2016). Classification and predictions of RNA pseudoknots based on topological invariants, *Physical Review E* **94**(4).

Weber, G., Defterli, O., Gök, S.Z.A. and Kropat, E. (2011). Modeling, inference and optimization of regulatory networks based on time series data, *European Journal of Operational Research* **211**(1): 1–14.

Weber, G., Kropat, E., Akteke-Öztürk, B. and Görgülü, Z. (2009). A survey on OR and mathematical methods applied on gene-environment networks, *CEJOR* **17**(3): 315–341.

Wiedemann, J. and Milostan, M. (2017). StructAnalyzer—A tool for sequence vs. structure similarity analysis, *Acta Biochimica Polonica* **63**(4): 753–757.

Wiedemann, J., Zok, T., Milostan, M. and Szachniuk, M. (2017). LCS-TA to identify similar fragments in RNA 3D structures, *BMC Bioinformatics* **18**(1): 456.

Wojciechowski, P., Frohmberg, W., Kierzynka, M., Zurkowski, P. and Blazewicz, J. (2016). G-MAPSEQ—A new method for mapping reads to a reference genome, *Foundations of Computing and Decision Sciences* **41**(2): 123–142.

Zemla, A. (2003). LGA: A method for finding 3D similarities in protein structures, *Nucleic Acids Research* **31**(13): 3370–3374.

Zok, T., Antczak, M., Riedel, M., Nebel, D., Villmann, T., Lukasiak, P., Blazewicz, J. and Szachniuk, M. (2015). Building the library of RNA 3D nucleotide conformations using the clustering approach, *International Journal of Applied Mathematics and Computer Science* **25**(3): 689–700, DOI: 10.1515/amcs-2015-0050.

Zok, T., Antczak, M., Zurkowski, M., Popenda, M., Blazewicz, J., Adamiak, R.W. and Szachniuk, M. (2018). RNApdbee 2.0: Multifunctional tool for RNA structure annotation, *Nucleic Acids Research* **46**(W1): W30–W35.

Tomasz Zok is an assistant professor at the Insitute of Computing Science, Poznan University of Technology. His research interests include computational methods to support analysis of the biomolecule structure and dynamics, operations research, high-throughput, and high-performance computing. His works have been published in computing and life sciences journals. He has been awarded in several national contests. He holds the MSc (2011) and PhD (2018) degrees in computing science.

Jan Badura is a PhD student at the Poznan University of Technology. His main research interests are in combinatorial optimization and hyperheuristics. He obtained a Master's degree in computing science (with distinction) in 2017. He has represented the PUT in numerous programming contests, such as ACM ICPC and Polish Championships in Group Programming.

Sylwester Swat is a PhD student at the Poznan University of Technology. His research interests include combinatorial optimization, graph theory, and data structures. He obtained a Master's degree in mathematics in 2018. He has succeeded in several programming contests related to algorithmics and group programming. He was awarded by the Minister of Science and Higher Education as an outstanding student in 2015.

Kacper Figurski is a BSc student at Adam Mickiewicz University in Poznan and works as a software engineer at the Poznan Supercomputing and Networking Center. His research interests include bioinformatics, algorithm solving molecular biology-inspired problems, web applications, and software engineering. He has succeeded in several programming contests related to life sciences.

Mariusz Popenda is a research associate at the Institute of Bioorganic Chemistry, Polish Academy of Sciences. His research interests include the RNA structure and dynamics, structural bioinformatics, algorithms solving molecular biology-inspired problems. He has authored papers published in top scientific journals on computing and life sciences, and holds a PhD (1999) in chemistry.

Maciej Antczak is an assistant professor at the Institute of Computing Science, Poznan University of Technology. His research interests include algorithms for bioinformatics and molecular biology, combinatorial optimization, operations research, high-throughput computing, software engineering, and artificial intelligence. He has authored papers published in top scientific journals on computing and life sciences, and holds a PhD (2013) and a DSc (2019) in computing science.

Received: 25 July 2019
Revised: 24 January 2020
Accepted: 12 March 2020