BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES, Vol. 63, No. 1, 2015
DOI: 10.1515/bpasts-2015-0037

Assigning NMR spectra of irregular RNAs by heuristic algorithms

M. SZACHNIUK[1,2*]

[1] Institute of Bioorganic Chemistry, Polish Academy of Sciences 12/14 Noskowskiego St., 61-704 Poznan, Poland
[2] Institute of Computing Science, Poznan University of Technology, 2 Piotrowo St., 60-965 Poznan, Poland

Abstract. Computer-aided analysis and preprocessing of spectral data is a prerequisite for any study of molecular structures by Nuclear Magnetic Resonance (NMR) spectroscopy. The data processing stage usually involves a considerable dedication of time and expert knowledge to cope with peak picking, resonance signal assignment and calculation of structure parameters. A significant part of the latter step is performed in an automated way. However, in peak picking and resonance assignment a multistage manual assistance is still essential. The work presented here is focused on the theoretical modeling and analyzing the assignment problem by applying heuristic approaches to the NMR spectra recorded for RNA structures containing irregular regions.

Key words: bioinformatics, NMR, tabu search, beam search, RNA structure.

1. Introduction

The problems of NMR spectra interpretation and assignment arise during the processing of data obtained from Nuclear Magnetic Resonance spectroscopy experiments. Such experiments are widely used in structural molecular biology to determine three-dimensional shapes of molecules, to analyze their conformations (in vitro NMR) or to track the movements of molecules (in vivo NMR) [1, 2]. NMR spectroscopy takes advantage of the fact, that many nuclei of biological molecules can produce their own magnetic fields. These fields are manipulated by NMR spectrometer, which generates electromagnetic waves and detects their absorption by the nuclei jumping into higher energy states. This phenomenon, called resonance, has a corresponding frequency, which is unique for a particular type of nucleus [3]. The frequencies of resonance signals and intensities of electromagnetic wave absorbance are recorded during NMR experiment and presented as its output data. The assignment of the recorded signals to molecule nuclei explains to us looking at the molecule structure on the atomic level, through the resonances registered in the NMR spectrum [4]. However, the assignment procedure is challenging, especially dealing with big molecules or their small irregular fragments. Let us explain that a typical RNA contains duplexes (i.e. regular regions) formed by canonical Watson-Crick pairs which stabilize the whole structure, and irregular regions, such as loops and single stranded regions which influence structure flexibility.

The technology for determination of three-dimensional structures using NMR has developed significantly over the last decades, especially for proteins. This has also given an impetus to the design of novel computational methods for the resonance assignment. Several tools supporting the interpretation of NMR spectral data for proteins and algorithms dedicated to their assignment have been published. They employ different algorithmic approaches, such as best-first search [5], exhaustive and heuristic search [6–8] branch-and-bound [9], simulated annealing [10], genetic algorithms [11], Monte Carlo optimization [12, 13], and other [14, 15]. Despite the fact that a lot has been already done in the area, and – what is more – further computation provides for protein structures showing the quality comparable with this of models solved using X-ray crystallography [16], the assignment procedure is still a weak point of structure elucidation process. The same concerns NMR-based determination of nucleic acid structures. An experimental assessment of the three-dimensional RNA structures remains difficult, which results in a relatively small number of known RNAs. Currently (July 2014), the Protein Data Bank stores 1080 RNA structures, as compared to ca 94000 proteins. 44% of PDB-deposited RNAs have been solved using NMR. For proteins this number reaches 10%. Although the general concept of structure determination using NMR is similar for proteins and nucleic acids, there are many differences in the particular steps of the process. First, some NMR experiments are more useful for proteins, some – for nucleic acids. Next, since different resonance signals are observed for protein nuclei and for RNA nuclei, their assignments also follow the other rules. Summing up, it has appeared that algorithms designed for protein assignment cannot be directly applied to assign RNA resonances [17]. Thus, a few methods have been developed especially for RNA assignment. They take advantage of exhaustive search [18, 19], evolutionary algorithm [2, 20], tabu [21, 22] and beam search [23] strategies, and they have been applied to process the spectra recorded for regular duplex structures and small bulged duplexes.

This paper is focused on the problem of resonance assignment considered for structures of irregular regions, generally understood as loops – the most common motifs occurring in RNA. A typical RNA loop is composed of one-, two-, or n-strands closed by one, two or n canonical base pairs respectively. In the literature these motifs are known as hairpin apical loops (one-stranded), internal loops and bulges (two-stranded)

*e-mail: mszachniuk@cs.put.poznan.pl

or n-way junctions (n-stranded). Their pictorial representation is shown in Fig. 1.

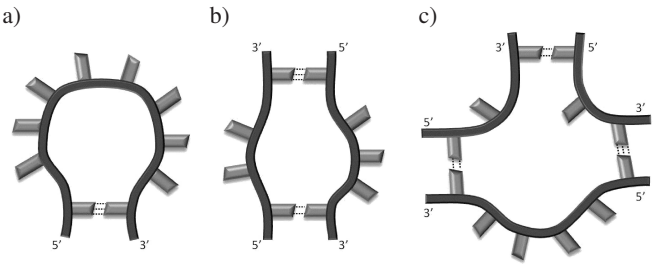

Fig. 1. Schematic representation of: a) hairpin loop, b) internal loop and c) 3-way junction

The problem of resonance assignment upon the spectra of irregular RNA regions is modeled by means of the graph theory with respect to its regular counterpart. Next, it is analyzed using two heuristic algorithms that proved to perform well for regular RNA molecules. Their versions supplemented with minor modifications are run to find the disjoint pathways of magnetization transfer within the spectra of RNAs with internal loops.

2. Problem description and graph-model

In this section the problem of RNA resonance assignment in two-dimensional NMR spectra is described for both, regular and irregular case. Next, the problem formulation is presented on the basis of the graph theory.

2.1. RNA assignment problem. An elucidation of molecular structures on the primary, secondary and tertiary level of their organization is usually composed of two general phases: the experimental and the computational one (cf. [5, 8, 24]). Such is also NMR-based determination of three-dimensional shapes of RNA molecules. Here, the experimental part depends on the type of NMR spectrometer (solution or solid-state NMR) and on the interaction types to be observed during the experiment. It also influences the implementation of the second, computational phase. The results presented here are related to the study of RNA in solution, aimed at the observation of the Nuclear Overhauser Effect (NOE). Thus, further part of the paper is focused on this particular case.

The NOE phenomenon can be exploited using 2D NMR spectroscopy in solution, during NOESY (Nuclear Overhauser Effect SpectroscopY), HOESY (Heteronuclear Overhauser Effect SpectroscopY), ROESY (Rotational frame nuclear Overhauser Effect SpectroscopY), TRNOE (TRansferred Nuclear Overhauser Effect), and DPFGSE-NOE (Double Pulsed Field Gradient Spin Echo Nuclear Overhauser Effect) experiments, to characterize and refine the three-dimensional molecular structures. The NOE signal observed during NMR experiment shows nucleus-nucleus through space magnetic interaction occurring between atoms which are in close proximity to each other. Such signal is recorded by NMR spectrometer and visualized as a cross-peak in the spectrum (Fig. 2).

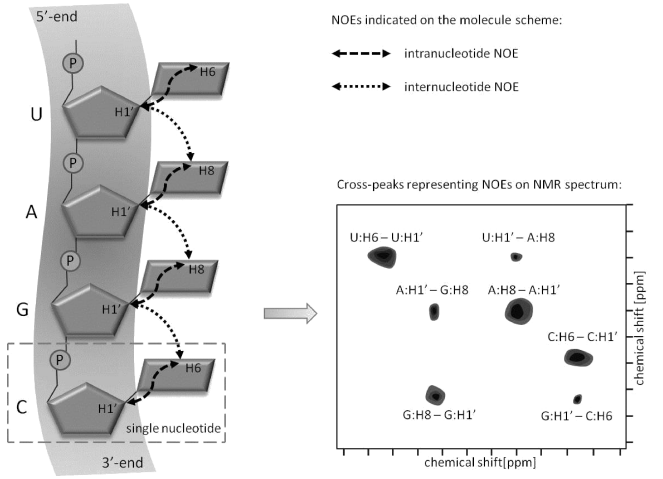

Fig. 2. NOE interactions between H1' and H6/H8 nuclei in r(UAGC) and their representation in 2D NMR spectrum

NMR machine records the frequency and the intensity of absorbance for each detected NOE signal. The frequencies (given in Hz) are next converted to chemical shift values (in ppm – parts per million). This makes experimental results independent of the measuring device parameters. On the basis of the NOE signal intensity the distance between interacting atoms can be computed [25, 26]. However, an estimation of the distance is possible only if it is known which atoms generated the signal under consideration. NMR experiment outputs the information about NOEs, but it is the task of an experimenter to find the relation between these signals and molecule atoms (i.e. assign signals to atoms). Resonance assignment is, thus, the preliminary step in the computational phase of structure determination process. Its key point to find a specific arrangement of cross-peaks (which are signals' representatives on the spectrum) and its association with molecule sequence. The general idea of the assignment is shown in Fig. 3.

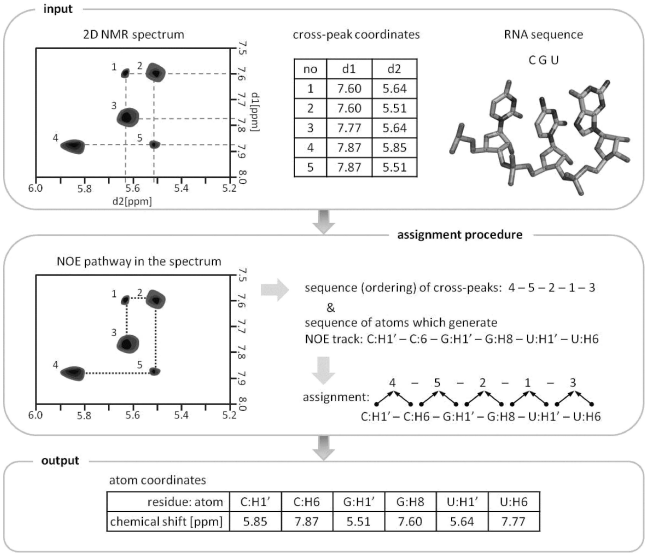

Fig. 3. General idea of the resonance assignment procedure

Cross-peak arrangement is reflected in a path, called NOE pathway, which should ideally go through all peaks of the

spectrum [17, 26]. It corresponds to the trail designated along RNA strand by NOEs occurring between consecutive nuclei of the selected atoms (e.g. H1', H6 and H8 as shown in Fig. 2). In the case of one-stranded RNA or regular RNA duplex structure (i.e. RNA molecule consisting of two self-complementary strands), a reconstruction of NOE pathway that corresponds to one strand is a good starting point for the calculation of the whole structure. Based on the NOE assignments for a single strand, one is able to find the positions of the large set of molecule atoms – also these included in the complementary strand when the duplex structure is analyzed [17]. This latter operation is possible, because two strands of the duplex structure are arranged in parallel and such are the NOE pathways for these strands (in practice, the two NOE paths of the duplex overlap along the entire length). However, when dealing with K-stranded irregular structures ($K > 1$), e.g. structures containing loops, separate pathways must be constructed by the assignment procedure for each of K fragments (strands) involved in the RNA loop.

A single NOE pathway starts from a cross-peak representing intranucleotide NOE signal. In the ideal case we are able to differentiate between intra- and internucleotide cross-peaks based on their volumes (volume of the cross-peak represents the corresponding signal intensity), which can be easily separated into two groups: big (for intranucleotide NOEs) and small volumes (for internucleotide NOEs). However, in reality this is often impossible. The pathway visits consecutive cross-peaks following the Manhattan manner: each new passage is perpendicular to the preceding one. If intra cross-peaks can be distinguished from the inter ones, the pathway should pass intra/inter cross-peaks alternately. Moreover, no cycles are allowed (each cross-peak is visited at most once) and the maximum length of the pathway is known, since it results from the strand length. In the ideal spectrum recorded for a single stranded RNA molecule or for a regular duplex, the NOE pathway is Hamiltonian and covers all cross-peaks of the spectrum (i.e. its length equals $2 \cdot r-1$, where r is a number of nucleotides in one RNA strand). The ideal case for irregular structures assumes that NOE pathways reconstructed in the spectrum are disjoint and together they make use of all the existing cross-peaks. The disjoint NOE pathways drawn in one spectrum might be of the same or different lengths. This depends on the size of the involved strands. However, the real NMR spectra are often subject to positive or negative errors, which means that some cross-peaks can be redundant (positive errors) and some can be missing (negative errors) from the spectrum. Thus, one may specify the maximum length of each reconstructed pathway, but with no certainty that this length will be achieved.

2.2. Graph-based model of the problem. The problem of RNA assignment pathway reconstruction can be modeled by means of graph theory with respect to its biochemical origin. The search space of the problem, i.e. the two-dimensional NMR spectrum S containing N cross-peaks is represented by the NOESY graph G [26]. Each cross-peak is associated with two coordinates (x, y), widths (dx, dy), and volume (m), which are used to characterize vertices and edges of G.

Definition 1. Let $G = (V, E)$ be an undirected graph placed on the Euclidean plane $X \times Y$, where V is set of vertices (position of vertex $v_i \in V$ on the plane is defined by two coordinates x_i, y_i) and E denotes a set of edges. G is a *NOESY graph* representing spectrum S of N cross-peaks, if it satisfies the following conditions:

1. $|V| = N$
2. for each $i = 1...N$: $v_i(x_i, y_i) \in V$ represents cross-peak $c_i(x_i, y_i) \in S$
3. for each $i = 1...N$: $v_i \in V$ has weight $w_i = \{0, 1\}$, where $w_i = 0$ if $c_i \in S$ is an inter cross-peak, $w_i = 1$ if $c_i \in S$ is an intra cross-peak
4. for each i, $j = 1...N$: $e(v_i, v_j) \in E$ if $(x_i = x_j, y_i \neq y_j)$ or $(x_i \neq x_j, y_i = y_j)$.

Let us notice that the above definition is appropriate if the intra and inter cross-peaks can be distinguished. However, it is not always guaranteed for the real instances, thus, the modification of the NOESY graph can be required - either vertex weights are omitted or they are equal to the volumes of the corresponding cross-peaks.

Once the NMR spectrum is represented as the NOESY graph, one can look for NOE pathway(s) between graph vertices. Let us recall the basic formulation of single NOE pathway [21].

Definition 2. Let $G = (V, E)$ be a NOESY graph with an edge set E and a vertex set $V = V_0 \cup V_1$, where $v_i \in V_0$ if $w_i = 0$ and $v_i \in V_1$ if $w_i = 1$, $i = 1...N$. A series of vertices $P = (v_1, \ldots, v_l)$, $v_i \in V$, $i = 1...l$, $l = 2...N$, is called *NOE pathway* in G if P contains no cycles, two vertices v_i, $v_j \in V$ are neighbors in P if $w_i \neq w_j$, two edges e_i, $e_j \in E$ are neighbors in P if they are perpendicular, and any two edges of P do not occur on the same horizontal/vertical line.

In the absence of positive and negative errors in the NOESY graph $G = (V, E)$, i.e. if G represents an ideal NMR spectrum S, the following conditions are satisfied by NOE pathways in G:

$$\bigcup_{i=1}^{K} P_i = V, \tag{1}$$

$$\bigcap_{i=1}^{K} P_i = \emptyset, \tag{2}$$

where $K \geq i$ refers to the number of strands in the analyzed RNA molecule, which equals the number of NOE pathways, and P_i denotes the i-th of the K pathways in graph G. Let us notice, that in the case of regular RNA structure $K = 1$, thus, formula (1) is reduced to equation $P_1 = V$ (meaning that P_1 is a Hamiltonian path in G), while formula (2) does not apply. An example ideal case including a NOESY graph G with two NOE pathways, the corresponding spectrum S and a scheme of molecule related to S is shown in Fig. 4. Vertex color represents its weight (vertices with $w = 0$ are white, while grey is for $w = 1$).

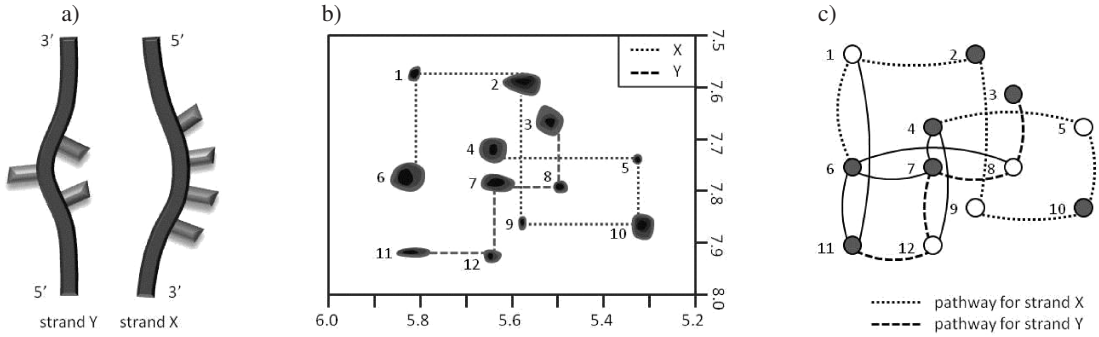

Fig. 4. a) Schematic view of RNA internal loop, b) a fragment of an ideal NOESY spectrum for this motif and c) a corresponding NOESY graph. Two NOE pathways are marked in both spectrum and graph

In the case of the real spectral data which usually include positive and negative errors, formula (2) remains valid, while formula (1) is modified:

$$\bigcup_{i=1}^{K} P_i \leq V. \tag{3}$$

Thus, NOE pathways in graph representing the real, non-ideal case, must be disjoint but they may not cover all the vertices of the graph.

3. Heuristics for reading the spectra of RNAs

The problem of NOE pathway reconstruction in the NOESY graph has been proven to be strongly NP-hard for regular RNAs [17, 27]. It remains computationally intractable also for irregular cases. Therefore, no exact algorithm is likely to solve every instance of this problem in a polynomial time and thus, the use of approximation algorithms is strongly recommended.

Several heuristic approaches have been proposed to solve the assignment problem for NMR spectra of protein and RNA structures [6–12, 17–23]. According to the author's knowledge, they were primarily dedicated to regular cases and only one of them [22] has been tested on the structures containing bulges. The latter approach, based upon tabu search, has been used in this work. The second heuristics presented here applies beam search strategy and has been recently published in [23]. Like tabu search, it has shown a very good performance in reading the spectra of RNA duplexes. Here, their performance will be analyzed mainly with reference to the irregular cases. Both methods search for the assignment pathway(s) and try to optimize the distance between an acceptable and an ideal pathway, using similar goal functions to evaluate solutions. The definitions of acceptability and ideality of a solution, and the goal function are introduced in the following subsection. The tabu and beam search strategies are outlined in Subsecs. 3.2 and 3.3.

3.1. Optimization and goal function. The graph-based model of NOE assignment, described in Sec. 2, determines the search problem without any element of optimization. However, this model works only for ideal cases, which are rare if one deals with the real experimental data recorded in NMR spectra. In the following paragraphs, the NOESY graph satisfying Definition 1 will be called an *ideal NOESY graph*, while the NOE pathway meeting all requirements of Definition 2 will be called an *ideal NOE path*. Both of them constitute an ideal version of the NOE pathway problem, which reflects the assignment based upon the ideal NMR spectrum. In such (ideal) case it is assumed that: (i) NMR spectrum contains no positive/negative errors, (ii) all desirable NOE pathways consume in total all cross-peaks of the spectrum, (iii) cross-peaks do not overlap in the spectrum and can be easily separated, (iv) each cross-peak can be unequivocally classified as either intra- or intermolecular one, (v) NMR spectrum contains pairs of collinear cross-peaks lying along either horizontal or vertical axis, (vi) there are two cross-peaks that have exactly one collinear partner in the spectrum (either horizontal or vertical), each of the remaining cross-peaks from the pool has two partners (one aligned horizontally and the second aligned vertically). Most experimental NMR spectra do not meet all of the above objectives, thus, they are not ideal and cannot be represented by ideal NOESY graphs. However, also non-ideal data can lead to proper cross-peak assignment, as it has been repeatedly shown by NMR experts. An observation of human experts dealing with the real data has resulted in the definition of graph and pathway acceptability, which reflect the RNA assignment problem embedded in the real NMR spectrum S that can be either ideal on non-ideal.

Definition 3. Let $G = (V, E)$ be an undirected graph placed on the Euclidean plane $X \times Y$, where V is set of vertices (position of vertex $v_i \in V$ on the plane is defined by two coordinates x_i, y_i) and E denotes a set of edges. G is an *acceptable NOESY graph* representing spectrum S of N cross-peaks, if:

1. $|V| = N$
2. for each $i = 1...N$: $v_i(x_i, y_i) \in V$ represents cross-peak $c_i(x_i, y_i) \in S$
3. for each $i = 1...N$: $v_i \in V$ has weight w_i corresponding to the volume of $c_i \in S$.

The new graph formulation requires an adaptation of the NOE pathway definition, so that it is consistent with the solutions to the real assignment problem.

Definition 4. Let $G = (V, E)$ be an acceptable NOESY graph with an edge set E and a vertex set V. A series of vertices $P = (v_1, \ldots, v_l)$, $v_i \in V$, $i = 1...l$, $l = 2...N$, is called an *acceptable NOE pathway* in G, if it does not contain cycles and two edges e_i, $e_j \in E$ are neighbors in P if they are not parallel.

An analysis of the real instances, solved by human experts in the NMR field, has shown that even in non-ideal cases there has been a strong pressure to look for the pathways that are close to ideal solutions. That observation contributed to the definition of a goal function that could be used to evaluate NOE pathways based on their estimated distance from the ideal. Next, the heuristic methods have been proposed, which optimize the goal function value during their search routine. Let us recall that the *NOE evaluation function* $f(P)$ to evaluate the acceptable NOE pathway P is a weighted sum consisting of six components [17, 21]:

$$f(P) = \frac{1}{n} \sum_{i=1}^{6} \phi_i d_i + r, \tag{4}$$

where n stands for pathway length (measured as the number of vertices in the path), r is a random factor, d_i are disorder parameters, and ϕ_i are weight factors that reflect the impact of the corresponding d_i on the overall function value. In particular, the disorder parameters correspond to the penalty, which is imposed if:

- pathway P does not start from the user-defined vertex (y_1),
- weights of the neighboring vertices (in P) belong to the same interval (y_2),
- the neighboring edges (in P) are not perpendicular (y_3),
- pathway P contains collinear edges (y_4),
- vertices in P do not correspond to the user-defined H5-H6 cross-peaks (y_5),
- edges in P are neither oriented horizontally nor vertically (y_6).

The random element (r) is introduced to prevent the search process from entering a local optimum. The goal function is minimized during the process of searching for the best solution, that is the ordering of vertices which is closest to the ideal NOE pathway.

3.2. Tabu strategy. Tabu search (TS) has been first introduced by Glover [28] as a novel approach to solve mathematical optimization problems using a local search procedure. Its advantage over the ordinary local search is based on the use of a memory structure (known as tabu list), which stores the recent moves or solutions, in order to prevent the algorithm from oscillating around the previously visited states and sticking in suboptimal regions. Since this heuristics introduction, it has been successfully applied to solve various optimization problems. In the area of bioinformatics, tabu method has been used for example in prediction of protein structures, multiple sequence alignment, DNA sequencing, signal assignment on NMR spectra of proteins and ribonucleic acids. The algorithm presented in the latter work [21] has been designed to solve NMR assignment problem based on the two-dimensional spectra of regular RNA duplexes, and run also for few bulged duplexes. Here, its performance is summarized for irregular RNA structures in general, when the search procedure is supplemented to handle such cases. Let us now outline the idea and the structure of this algorithm.

The tabu search algorithm for the resonance signal assignment has been based on the classical tabu approach. Apart from the tabu list, it has been supplemented with an elite structure to store the most promising solutions (i.e. acceptable NOE pathways with the relatively low value of objective function) and their tabu lists. The computation starts from an initial pathway (being the first base solution), which is constructed by a greedy component of the method. Next, the algorithm generates the neighborhood of the base solution (i.e. a collection of the subsequent solutions) and explores it iteratively in order to find the improved solution, and move there in the search space exploration. The neighboring solutions are generated using several transition operators: swap, exchange, insert and delete. Each solution is verified according to its acceptability and evaluated by a global criterion function (formula (3)). The most promising one becomes the base solution for the next turn, provided it is not on the tabu list, and the transition from current base solution to the new one is not prohibited. If no improvement of the objective function value is observed for 500 iterations or the global optimum is found (i.e. $f(P) = 0$), the algorithm stops. It is worth noting that the tabu list has a limited capacity. For the considered implementation, its length has been experimentally determined and set to $N/2 + 18$, where N is the number of NOESY graph vertices. Further details concerning the algorithm can be found in [21, 22].

Let us remind that the above described version of the method finds a single pathway, which – in practice – corresponds to each of the two self-complementary strands of RNA duplex. However, in case of processing irregular RNA regions, each RNA strand is associated with one, separate NOE pathway, thus, the number of pathways must equal the number of strands. In order to adapt the method to the irregular case, minor modifications must have been introduced. First, the algorithm may be provided with the number K of strands, which equals the number of required solutions. Second, pathway lengths need to be defined on the input (at least $K-1$ lengths are needed). Third, the resulting set of solutions must satisfy formulas (1)–(2), thus, vertices forming the K-th pathway are annotated unavailable before the algorithm starts computation for the $(K+1)$-th strand.

3.3. Beam search. Beam search (BS) strategy has been developed as a memory-bounded heuristics based upon the best-first search idea. It constructs a tree-structured search space progressing in a layer by layer manner and expanding nodes of the current layer in the breadth-first order. At each level of the search tree, the algorithm estimates a promise of the succeeding layer nodes, by using the built-in evaluation function. The most promising nodes are selected for a further branch, while the other ones are permanently discarded. The number of expanded states is controlled by beam width B, which value is fixed in the initiating phase of the search process.

Usually, the algorithm terminates when it reaches the goal node, performs a predefined number of iterations, runs out of memory, or if there are no nodes to be branched.

So far, beam search has been successfully applied to solve various optimization problems, like speech recognition, image processing, segment matching, alignment of chromatographic signals, task planning, NMR assignment, etc. In [23], the beam search strategy has been adapted to solve the resonance assignment problem for two dimensional NMR spectra of regular RNAs. The referred algorithm starts from constructing a root node (zero level of the search tree) and a set of single edges (i.e. one-edge pathways) which build the tree's first level. The nodes of the succeeding level are constructed by extending every assignment pathway with an edge. All nodes of current level are evaluated and B best nodes are selected for branching in the next iteration. Beam width B has been experimentally set to n^2/s, where n is the number of cross-peaks (as well as the cardinality of NOESY graph's vertex set) and s stands for currently processed level of the search tree. The algorithm has been combined with a multilayer perceptron (MLP) that estimates the quality of nodes and, thus, defines the beam priority. In typical applications, MLP is used as a classifier and outputs one-bit binary value. However, in [23] it has been implemented to evaluate nodes (i.e. acceptable pathways being partial solutions to the problem), that is, to compute a priority value which allows to rank the nodes. Thus, the MLP's output is a value between 0.0 and 1.0, computed on the connections of the neural network between the hidden and the output layer. The presented beam search method has been first tailored to process the spectra of the regular structures and the MLP trained on dataset containing only regular cases. The minor modification of beam search allows to handle also a selection of irregular structures. In the irregular case, the algorithm constructs a set of solutions of the required length (path length should be provided). Within this set, the pairs of disjoint pathways are annotated ($K = 2$). The solution of irregular case (i.e. pair of pathways) is acceptable if it satisfies formulas (2) and (3).

4. Processing of irregular experimental data

This section shows how the presented TS (tabu search) and BS (beam search) heuristics read the real NMR spectra of irregular RNA regions. Their performance is pictured using example 2D NMR spectrum that has been recorded on Varian Unity 400 MHz spectrometer and preprocessed using peak-picking procedure of the Accelrys Felix package.

Figure 5 illustrates the secondary structure of an example irregular instance (hereinafter referred to as AA-loop) and its NMR spectrum. The structure contains one two-nucleotide internal loop and two helical fragments located on both ends. The NMR spectrum reveals 77 under-diagonal cross-peaks corresponding to the NOE signals observed during 2D NOESY experiment for this instance.

Taking into account the structure of AA-loop and assuming that the analyzed molecule has no missing atoms, the assignment problem solution should be composed of two paths which, respectively, have lengths 23 and 19. Thus, an ideal spectrum might contain only 42 cross-peaks, and the two disjoint paths could consume all of them. However, the real spectrum includes the information about additional signals that have appeared during the experiment (some of them, e.g. doublets and fuzzy peaks, constituting positive errors) which are insignificant in the backbone assignment. Moreover, further analysis shows that cross-peak locations do not allow to construct the NOESY graph sufficient to contain highly evaluated disjoint pathways of the expected lengths, that could be clearly indicated as global optimum.

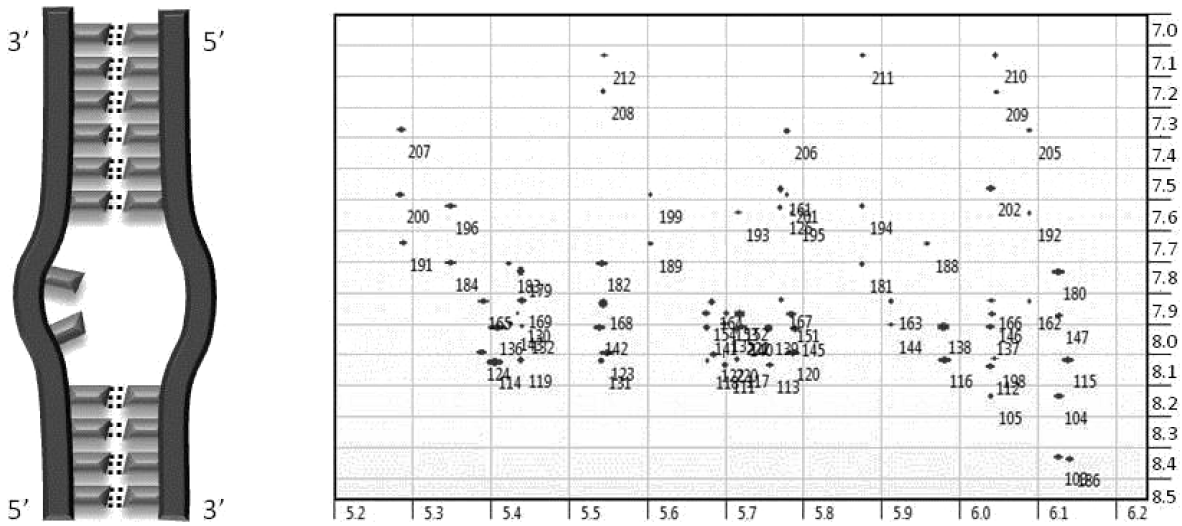

Fig. 5. AA-loop instance: schematic representation of the secondary structure and visualization of NMR spectral data provided by beam search software (Ref. 23)

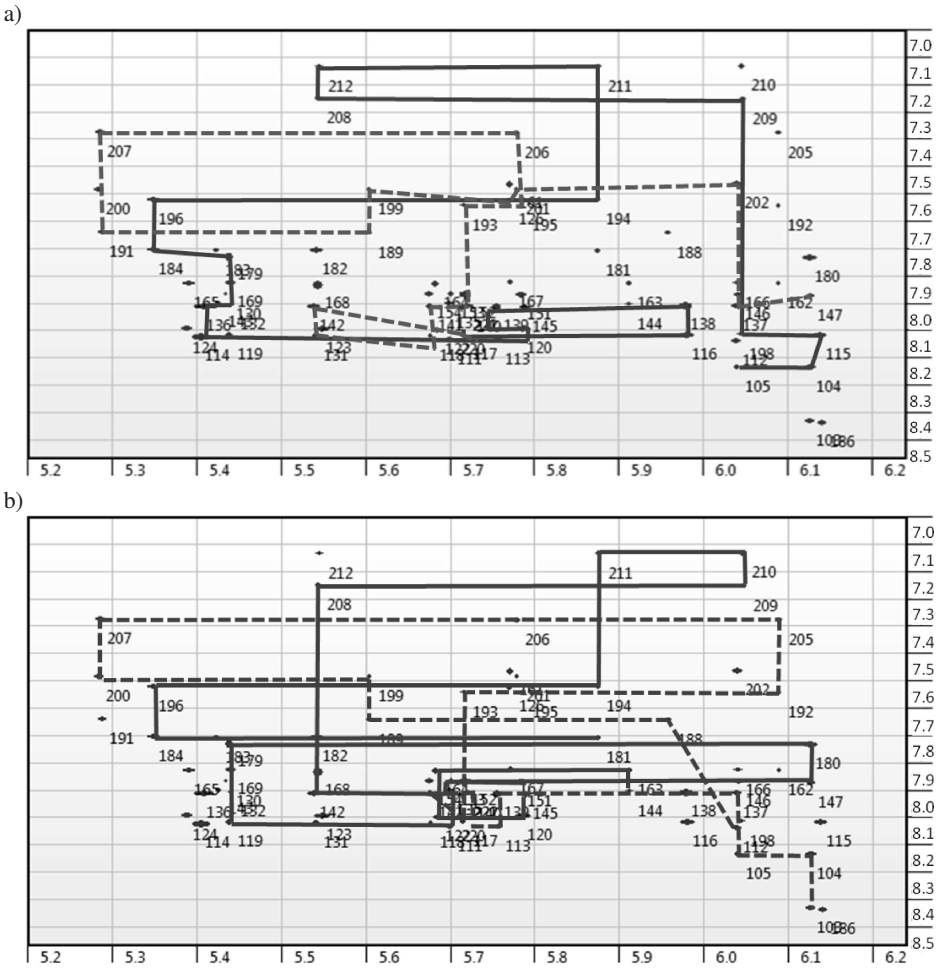

Fig. 6. Optimal solutions found by tabu search for AA-loop instance in experiment A and B: a) solution A: $P_1(23)$-solid line, $P_2(19)$-dashed line, b) solution B: $P_1(19)$-dashed line, $P_2(23)$-solid line

The AA-loop instance has been processed by both heuristic algorithms. In the first experiment (named experiment A), tabu search has been run with its default parameters. The algorithm, forced to construct 23- and 19-long pathways, has output them within less than 2 seconds. The first optimal solution $P_1(23)$, with length equal to 23, has been found after 121 iterations. $P_1(23)$ passes through the following vertices of the NOESY graph (or cross-peaks of the corresponding spectrum): 220, 111, 116, 138, 140, 152, 151, 120, 124, 165, 169, 179, 184, 196, 194, 211, 212, 208, 209, 198, 115, 104, 105. In the next step, all of these vertices have been annotated as unavailable and the algorithm was run to construct the second path. The latter one has been returned after 785 iterations, out of which the final 500 have not brought any improvement, thus, making the stopping condition executed. The second pathway $P_2(19)$ follows the route through 19 cross-peaks: 117, 222, 142, 131, 118, 141, 221, 193, 195, 206, 207, 191, 189, 199, 126, 161, 202, 146, 147. Solution of experiment A found by tabu search has been visualized in Fi.6a. The first pathway, $P_1(23)$, is drawn with a solid line, while the second one, $P_2(19)$ – with a dashed line. It can be easily observed that $P_1(23)$ is closer to the ideal path model than $P_2(19)$, which contains the edges with visible deviation from vertical and horizontal alignment. The main reason of this difference results from the fact that $P_1(23)$ has been constructed first. Thus, the algorithm was able to look for the optimal solution within the whole search space, using the complete set of cross-peaks and having a lot of flexibility in the neighborhood generation at every step. The search for $P_2(19)$ has been done within the reduced set of cross-peaks where the possibility to improve the pathway course has been significantly limited.

In the second experiment (denoted as experiment B), tabu search has been executed with the reverse order of path construction, in order to check the alternative solution. In this run, the shorter path, denoted as $P_1(19)$, was constructed first, revealing the following route, returned after 707 iterations: 198, 188, 189, 199, 200, 207, 205, 192, 193, 118, 154, 153, 111, 113, 139, 137, 105, 104, 103. The longer path $P_2(23)$ has been found in 392 iterations, after reducing the set of available cross-peaks. It has passed through 23 cross-peaks in the following order: 144, 163, 164, 122, 120, 145, 222, 152, 147, 180, 179, 119, 117, 221, 142, 208, 209, 210, 211, 194, 196, 184, 181. Both pathways forming the solution of experiment

B have been presented in Fig. 6b. $P_1(19)$ is drawn in a dashed line, while $P_2(23)$ – in a solid line. Comparing solutions from both experiments, it can be observed that $P_1(23)$ and $P_2(23)$ have only 3 common edges (which constitutes 13% of the entire path) and 9 common vertices (i.e. 40% of the path). As for the shorter paths, $P_1(19)$ and $P_2(19)$ overlap in 5% of edges (1 common edge) and 21% of vertices (4 common vertices). In turn, if the entire solutions A and B are to be compared, it appears that A: $\{P_1(23), P_2(19)\}$ and B:$\{P_1(19), P_2(23)\}$ overlap in 15% of edges (6 common edges) and almost 50% of vertices (20 common vertices). Considering NOE path definition and its evaluation by the goal function, solution B is found to be better than A, due to the smaller number of edge deviations.

The AA-loop instance has been processed for the third time in experiment C. This time the beam search algorithm has been run a number of times, with different values of input parameters which define the pathway length, the threshold and the acceptable angle divergence (determining how much an edge can be inclined from the vertical/horizontal). First test was executed with the divergence set to zero and the required maximum length of the pathway equal to 23. In this assay, beam search has been able to reconstruct half-solution, i.e. 23-peak long pathway without its disjoint partner. The same results have been obtained in the following steps, for shorter pathways (lengths > 13) – the algorithm could find only half-solution to the problem. The maximum length of disjoint pathways forming one solution, constructed by beam search, was equal to 13. The algorithm returned around 25 different solutions of this length, i.e. couples consisting of 13-peak long pathways. All beam search solutions have overlapped with those reconstructed by tabu search to a various extent (see Fig. 7 that enumerates selected pairs and presents one of them drawn in the spectrum). In contrast to the tabu method, our implementation of beam search does not release unsatisfactory solutions, i.e. pathways with goal function value above the threshold, which are significantly distorted, contain many collinear edges, etc. Thus, another set of experiments has been run to find whether longer solutions could be found. Upon increasing the value of threshold and an acceptable angle divergence parameter, beam search has been able to construct solutions consisting of paths passing through 15 vertices each. However, longer paths could not be generated by this algorithm in the reasonable time – note that with the increase in the above mentioned parameter values, the time of computation extends significantly. Analyzing the results of beam search, one must be aware that in experiment C the algorithm has been performed with perceptron trained on spectral data obtained for regular structures. Due to a very limited access to irregular structures, it was still not possible to run the training process on irregular datasets. This is indeed the direction for further discussion on beam search improvement aiming to adjust it better to process the real spectral data for irregular RNA instances.

a)

$P_1(13)$	$P_2(13)$
105, 104, 103, 186, 115, 116, 138, 132, 130, 146, 112, 111, 135	192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136
131, 111, 135, 132, 119, 115, 186, 103, 104, 105, 112, 113, 139	192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136
192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136	119, 111, 135, 132, 130, 146, 105, 104, 103, 186, 115, 116, 138
142, 123, 115, 186, 103, 104, 105, 112, 111, 135, 132, 119, 113	192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136
153, 111, 112, 105, 104, 103, 186, 115, 116, 138, 132, 119, 113	192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136
162, 163, 144, 132, 119, 115, 186, 103, 104, 105, 112, 111, 135	192, 126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136
126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136, 132	130, 119, 111, 135, 137, 105, 104, 103, 186, 115, 117, 152, 146
208, 123, 115, 147, 130, 119, 111, 135, 137, 105, 104, 103, 186	126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136, 132
126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136, 132	211, 181, 179, 119, 111, 135, 137, 105, 104, 103, 186, 115, 113
169, 119, 111, 135, 137, 105, 104, 103, 186, 115, 117, 152, 130	126, 161, 199, 189, 191, 200, 201, 120, 122, 118, 114, 136, 132

b)

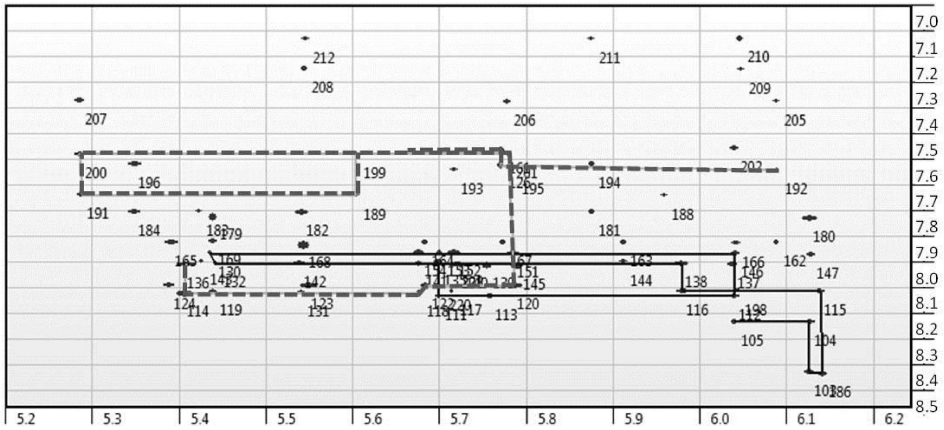

Fig. 7. Short optimal solutions found by beam search for AA-loop instance in experiment C: a) a list of selected solution, b) a visualization of two paths from the shaded row. $P_1(13)$-solid line, $P_2(13)$-dashed line

5. Concluding remarks

The paper has focused on the problem of automated reading of structural information included in the NMR spectra of irregular RNA regions and signal assignment. The theoretical graph model of the resonance signal assignment problem has been discussed and its regular version extended with regard to the irregular one. Two heuristic algorithms, tabu search and beam search, have been tried on example NMR data obtained from NMR experiment run for RNA structure with internal loop. Both heuristics have been previously tested for regular structures and proved their utility.

First experiments with the processing of irregular data show that the algorithms tailored to deal with regular structures can be helpful to the experimenters, by proposing various solutions that might be used to create final assignment. However, if more accurate results are to be obtained, these algorithms should not be directly used on data recorded for irregular molecules. The algorithms provide quite good results, but they should be more adjusted to the processing of irregular spectra. Thus, a modification of these methods should be continued in the following aspects. First, it would be interesting to see how the tabu search performs if the goal function evaluates the entire solution instead of single pathways. To enable this, the procedure should search for K paths at the same time, and consider them as a whole. Next, the set of irregular spectra should be collected and used to train the MLP component of beam search. Moreover, it would be necessary to make a separate analysis of noise in NMR spectra. Since the problem of fuzzy data occurs often, especially for bigger RNA molecules, it would be challenging to preprocess the data and reduce the noise automatically or to recognize it during pathway reconstruction procedure. Finally, beam search should be adjusted to search for pathways of various lengths in order to cover all the instances that can occur while processing irregular structures.

Acknowledgements. The research was supported by the grant 2012/05/B/ST6/03026 from the National Science Center, Poland. The author thanks Adam Wojtowicz and Mikolaj Malaczynski for their contribution to the implementation of the algorithms, and Lukasz Popenda for providing spectral data and expert knowledge.

REFERENCES

[1] A.H. Kwan, M. Mobli, P.R. Gooley, G.F. King, and J.P. Mackay, "Macromolecular NMR spectroscopy for the non-spectroscopist", *FEBS J.* 278 (5), 687–703 (2011).

[2] J. Blazewicz, M. Szachniuk, and A. Wojtowicz, "Evolutionary approach to NOE paths assignment in RNA structure elucidation", *Proc. 2004 IEEE Symp. on Computational Intelligence in Bioinformatics and Computational Biology* 1, 206–213 (2004).

[3] G. Brix, H. Kolem, W.R. Nitz, M. Bock, A. Huppertz, C.J. Zech, and O. Dietrich, "Basics of magnetic resonance imaging and magnetic resonance spectroscopy" in eds. M.F. Reiser, W. Semmler, and H. Hricak, *Magnetic Resonance Tomography*, Springer, Berlin, 2008.

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

[5] D. Zimmerman, C. Kulikowski, Y. Huang, W. Feng, M.S. Tashiro, S. Shimotakahara, C. Chien, R. Powers, and G.T. Montelione, "Automated analysis of protein NMR assignments using methods from artificial intelligence", *J. Mol. Biol.* 269 (4), 592–610 (1997).

[6] H.S. Atreya, S.C. Sahu, K.V.R. Chary, and G. Govil, "A tracked approach for automated NMR assignments in proteins (TATAPRO)", *J. Biomol. NMR* 17 (2), 125–136 (2000).

[7] G. Cavuslar, B. Catay, and M.S. Apaydin, "A tabu search approach for the NMR protein structure-based assignment problem", *IEEE/ACM Trans. Comput. Biol. Bioinform.* 9 (6), 1621–1628 (2012).

[8] P. Guntert, M. Saltzmann, D. Braun, and K. Wuthrich, "Sequence-specific NMR assignment of proteins by global fragment mapping with program Mapper", *J. Biomol. NMR* 18 (2), 129–137 (2000).

[9] D. Stratmann, C. van Heijenoort, and E. Guittet, "NOEnet – use of NOE networks for NMR resonance assignment of proteins with known 3D structure", *Bioinformatics* 25 (4), 474–481 (2009).

[10] X. Wan and G. Lin, "CISA: combined NMR resonance connectivity information determination and sequential assignment", *IEEE ACM T. Comput. Bi.* 4 (3), 336–348 (2007).

[11] N.E.G Buchler, E.P.R. Zuiderweg, H. Wang, and R.A. Goldstein, "Protein heteronuclear NMR assignments using mean-field simulated annealing", *J. Mol. Resonance* 125 (1), 34–42 (1997).

[12] C. Bartels, P. Guntert, M. Billeter, and K. Wuthrich, "GARANT – A general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra", *J. Comp. Chem.* 18 (1), 139–149 (1997).

[13] M. Leutner, R.M. Gschwind, J. Liermann, C. Schwarz, C. Gemmecker, and H. Kessler, "Automated backbone assignment of labeled proteins using the threshold accepting algorithm", *J. Biomol. NMR* 11 (1), 31–43 (1998).

[14] T.K. Hitchens, J.A. Lurkin, Y. Zhan, S.A. McCallum, and G.S. Rule, "MONTE: An automated Monte Carlo based approach to nuclear magnetic resonance assignment of proteins", *J. Biomol. NMR* 25 (1), 1–9 (2003).

[15] M.S. Apaydin, B. Catay, N. Patrick, and B.R. Donald, "NVR-BIP: nuclear vector replacement using binary integer programming for NMR structure-based assignments", *Computer J.* 54 (5), 708–716 (2011).

[16] T. Zok, M. Popenda, and M. Szachniuk, "MCQ4Structures to compute similarity of molecule structures", *Central Eur. J. Operations Research* 22 (3), 457–474 (2014).

[17] R.W. Adamiak, J. Blazewicz, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda, and M. Szachniuk, "An algorithm for an automatic NOE pathways analysis of 2D NMR spectra of RNA duplexes", *J. Comp. Biol.* 11 (1), 163–180 (2004).

[18] M.W. Roggenbuck, T.J. Hyman, and P.N. Borer, "Path analysis in NMR spectra: application to an RNA octamer", *Structure and Methods* 3, 309–317 (1990).

[19] M. Szachniuk, M. Popenda, R.W. Adamiak, and J. Blazewicz, "An assignment walk through 3D NMR spectrum", *Proc. 2009 IEEE Symp. on Computational Intelligence in Bioinformatics and Computational Biology* 1, 215–219 (2009).

[20] J. Blazewicz, M. Szachniuk, and A. Wojtowicz, "Evolutionary algorithm for a reconstruction of NOE paths in NMR spectra of RNA chains", *Bull. Pol. Ac.: Tech.* 53 (3), 221–230 (2004).

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

[22] M. Szachniuk, L. Popenda, Z. Gdaniec, R.W. Adamiak, and J. Blazewicz, "NMR analysis of RNA bulged structures: tabu search application in NOE signal assignment", *Proc. 2005 IEEE Symp. on Computational Intelligence in Bioinformatics and Computational Biology* 1, 172–178 (2005).

[23] M. Szachniuk, M. Malaczynski, E. Pesch, E.K. Burke, and J. Blazewicz, "MLP accompanied beam search for the resonance assignment problem", *J. Heuristics* 19 (3), 443–464 (2013).

[24] W. Frohmberg, M. Kierzynka, J. Blazewicz, P. Gawron, and P. Wojciechowski, "G-DNA – a highly efficient multi-GPU/MPI tool for aligning nucleotide reads", *Bull. Pol. Ac.: Tech.* 61 (4), 989–992 (2013).

[25] F.A.L. Anet and A.J.R. Bourn, "Nuclear Magnetic Resonance spectral assignments from Nuclear Overhauser effects", *J. Am. Chem. Soc.* 87 (22), 5250–5251 (1965).

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

[27] M. Szachniuk, M.C. De Cola, G. Felici, and J. Blazewicz, "The orderly colored longest path problem – a survey of applications and new algorithms", *RAIRO Operations Research* 48 (1), 25–51 (2014).

[28] F. Glover and M. Laguna, *Tabu Search*, Kluwer Academic Publishers, Boston, 1997.