BIOINFORMATICS ORIGINAL PAPER
Vol. 21 no. 10 2005, pages 2356–2361
doi:10.1093/bioinformatics/bti351

Structural bioinformatics

RNA tertiary structure determination: NOE pathways construction by tabu search

Jacek Blazewicz[1,2], Marta Szachniuk[1,*] and Adam Wojtowicz[2]

[1]Istitute of Bioorganic Chemistry, Polish Academy of Sciences, 61-704 Poznan, Poland and
[2]Institute of Computing Science, Poznan University of Technology, 60-965 Poznan, Poland

Received on January 23, 2005; revised on February 20, 2005; accepted on February 22, 2005
Advance Access publication February 24, 2005

ABSTRACT

Motivation: Liquid state nuclear magnetic resonance (NMR) spectroscopy has now been well established as a method for RNA tertiary structure determination. Most of the steps involved in the determination of RNA molecules are performed using computer programs. They however, do not apply to resonance assignment being the starting point of the whole procedure. We propose a tabu search algorithm as a tool for automating this step. Nuclear overhause effect (NOE) pathway, which determines the assignment, is constructed during an analysis of possible connections between resonances within aromatic/anomeric region of two-dimensional NOESY spectrum resulting from appropriate NMR experiment.

Results: Computational tests demonstrate the superior performance of the tabu search algorithm as compared with the exact enumerative approach and genetic procedure applied to the experimental and simulated spectral data for RNA molecules.

Availability: The software package can be obtained upon request from Marta Szachniuk.

Contact: mszachniuk@cs.put.poznan.pl

1 INTRODUCTION

In the recent years, a determination of biomolecule structures has become one of the most fundamental tasks in structural biology and biochemistry. Initially, the researchers concentrated on proteins and deoxyribonucleic acids (DNA). However, studying these molecules alone appeared insufficient to answer all the questions posed for years and the research has been extended to the molecules of the ribonucleic acid (RNA), which transmits genetic information from DNA to proteins and controls certain chemical processes in the cell.

Primary, secondary and tertiary structures are the subjects of RNA structural analysis. The primary structure depends on the number and sequence of nucleotides in the chain, secondary structure describes one- and two-strand fragments and the formation of loops or helices, finally, tertiary structure characterizes the spatial shape of the entire chain. A quick spread of nuclear magnetic resonance (NMR) spectroscopy in the last decades of the 20th century has resulted in this method gaining superiority over others, as far as structure determination of biomolecules in solution is concerned (Wüthrich, 1986). The elucidation procedure using NMR involves two general stages: experimental, where multidimensional correlation spectra are acquired; and computational, where spectra are analyzed and the structure is determined. In all the methods of NMR structure analysis the raw experimental data are processed using the following procedures of peak-picking, assignment, restraints determination, structure generation and refinement (Varani and Tinoco, 1991). The assignment of the observed NMR signals to the corresponding protons and other nuclei is a drawback of RNA structure elucidation process and has recently become one of the most important spectral problem. The assignment is usually based on the analysis of two-dimensional (2D) spectra resulting from NMR experiments. For short DNA and RNA duplexes it is performed manualy in accordance with the experimenter's knowledge and intuition. However, for the longer nucleic chains the assignment step becomes complex, owing to a large number of signals and their overlapping in the spectrum. Therefore, it has become necessary to facilitate NMR structural analysis of biopolymers by automating the procedures at this level.

As far as proteins are concerned, the above process has resulted in the designing of several automatic methods for their assignment (Atreya *et al*., 2000; Balley-Kellogg *et al*., 2004; Linge *et al*., 2003; Moseley and Montelione, 1999). However, these methods cannot be applied for nucleic acids spectra, because of the differences in NMR experiments for proteins and nucleic acids, and in their assignment procedures, based on different sets of signals. To our knowledge, only three papers (Adamiak *et al*., 2004; Roggenbuck *et al*., 1990; Blazewicz *et al*., 2004) have dealt with procedures for automatic generation of pathways between H6/H8 and H1′ resonances, known as the nuclear overhause effect (NOE) signals for RNA molecules (leading to a construction of the NOE pathway). The first two are concerned with exact backtracking and enumerative algorithms, applicable for an analysis of short unbroken RNA duplexes. In the third paper, the genetic algorithm in the construction of optimal solution in the NMR spectra of RNA simplexes and duplexes, has been considered. Because of their limited applicability owing to time constraints and the excessive number of solutions generated by exact algorithms, as well as the unsatisfactory precision of the optimal solution constructed by the genetic method, we have focused in this paper on the development of a tabu search approximation algorithm for an automatic generation of the NOE pathway. The method takes into account the specificity of the data and it is based on the combinatorial model of the NOESY graph (Adamiak *et al*., 2004; Szachniuk *et al*., 2003). The enumerative analytical algorithm (Adamiak *et al*., 2004) applied for the long nucleic chains may construct too many feasible

*To whom correspondence should be addressed.

solutions, preventing the performance of the subsequent steps in the determination process. As it is crucial to generate the pathways as close as possible to the original one, we have considered an application of metaheuristics. The experiments have shown a good quality of the tabu search algorithm with a clear superiority over another heuristic approach based on the genetic algorithm (Blazewicz *et al*., 2004).

The paper is organized as follows. In Section 2 we present basic ideas of a combinatorial model of the problem. The new tabu search algorithm dedicated to the problem of NOE paths reconstruction, is proposed in Section 3. Section 4 outlines the results of computational experiments comparing exact enumerative, genetic and tabu search algorithms. Section 5 sums up the results of tabu search application in solving the problem of the NOE path construction and points out the directions for further research.

2 SYSTEMS AND METHODS

One of the first analytical step in the tertiary structure elucidation, is the identification of the sequence-specific connectivity H8/H6$_{(i)}$–H1$'_{(i)}$–H8/H6$_{(i+1)}$ pathway, represented as the NOE pathway in the 2D-NOESY spectrum of RNA molecules (Wüthrich, 1986). A formation of such a path is possible because each aromatic H6/H8 proton of a nucleotide residue is in close proximity to two anomeric protons: its own and the preceding H1′ one. For each RNA simplex and self-complementary RNA duplex one NOE path exists. For non-complementary duplexes two NOE paths exist. We will call them the original paths. Finding them is the aim of the presented algorithm.

The NOE interactions between protons are represented as cross-peaks in the 2D-NOESY spectrum generated for the molecule during the appropriate NMR experiment. In the search for NOE connectivity pathway we focus on the aromatic/anomeric region [(5–6) × (7–8) ppm] of the spectrum, which borders interactions between protons of our interest (H6, H8, H1′). The path is composed of intranucleotide and internucleotide interactions, which give rise to the alternately appearing cross-peaks. In the ideal cases, the NOE pathway starts with the intranucleotide interaction at 5′ end of the strand and its length equals $2M-1$, where M is a number of residues (nucleotides) in the RNA chain. Each proton belonging to the pathway, except for the starting and terminal ones, gives cross-peaks with two other protons. Every cross-peak is characterized by the two coordinates of its center, widths in both dimensions and the value of signal intensity. Every two consecutive points in the NOE pathway have exactly one coordinate in common and consecutive connections within the pathway lie vertically or horizontally. Figure 1 demonstrates an exemplary NOE pathway found in the analyzed region of the 2D-NOESY spectrum.

Respecting the biochemical description of the problem, we proposed a graph-theoretic model (Adamiak *et al*., 2004; Szachniuk *et al*., 2003), as a background for the complexity analysis and for the construction of the algorithms solving the problem. The process of sequential assignments of H6/H8–H1′ corresponds to the construction of a path between vertices of a graph. Thus, converting 2D-NOESY spectrum to a certain graph structure seems to be an attractive idea. Cross-peaks are obvious candidates for graph vertices. Possible connections, which can be suggested during NOE pathway reconstruction, define the edges of the graph. The following definition (Adamiak *et al*., 2004; Blazewicz *et al*., 2004; Szachniuk *et al*., 2003) characterizes a new type of a graph representing the selected region of 2D-NOESY spectra and NOE sequence properties.

DEFINITION 1. *(NOESY graph). Let $G=(V,E)$, where V is a set of vertices, E is a set of edges, be an undirected graph situated on a plane. We will call G a NOESY graph, if the following conditions are satisfied:*

(1) every vertex $v \in V$ represents one cross-peak from a hypothetical spectrum S corresponding to G, and has the following properties of the cross-peak: a number, two coordinates and widths in two dimensions;

(2) a number $|V|$ of vertices in graph G equals a number N of cross-peaks in spectrum S;

(3) every vertex $v_i \in V$, $i=1..N$, is weighted and has a weight $w_i \in \{0,1\}$: $w_i=0$ if the i-th cross-peak represents internucleotide signal, $w_i=1$ if the i-th cross-peak represents intranucleotide signal; thus $V=V_0 \cup V_1$, where $V_0=\{v_i : w_i=0, i=0..N\}$ and $V_1=\{v_i : w_i=1, i=0..N\}$;

(4) every edge $e \in E$ represents a potential connection between two vertices of V having different weights and exactly one common coordinate; and

(5) a number $|E|$ of edges in graph G equals a number of all possible connections (i.e. lines between two cross-peaks of different intensity intervals having exactly one common coordinate) that can be drafted in spectrum S.

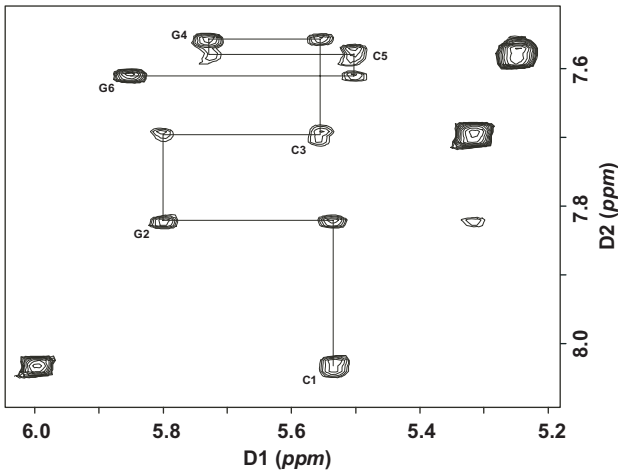

Fig. 1. NOE connectivity pathway for r(CGCGCG)$_2$.

Having any 2D-NOESY spectrum S, one can construct NOESY graph G corresponding to aromatic/anomeric region of S. Figure 2 shows the relationship between the [(5–6) × (7–8)] region of the 2D-NOESY spectrum of r(CGCGCG)$_2$ (Fig. 2a) and the corresponding NOESY graph (Fig. 2b) obtained according to Definition 1.

After converting spectrum S to graph G, an appropriate connectivity pathway can be looked for in G. This, however, requires a formulation of the NOE pathway problem in terms of graph theory.

DEFINITION 2. *(NOE path). Let $P_G=v_1, v_2, \ldots, v_l$ be a sequence of vertices of the NOESY graph $G=(V,E)$. We will call P_G the NOE path in G, if the following conditions are satisfied (Adamiak et al., 2004; Blazewicz et al., 2004; Szachniuk et al., 2003, 2004):*

(1) $v_1 \in V_1$,

(2) every vertex $v_i \in V$ and every edge $e_j \in E$ of G occurs in path P_G at most once,

(3) vertices with different weights appear alternately in P_G,

(4) every two neighboring edges of P_G are perpendicular,

(5) no two edges of P_G occur on the same horizontal or vertical line and

(6) a length of P_G equals $2|V_1|-1$.

NOE pathway constructed in the 2D-NOESY spectrum S of RNA molecule is the solution (i.e. the original pathway) for the assignment problem. NOE path P_G found in NOESY graph G corresponding to spectrum S, is the

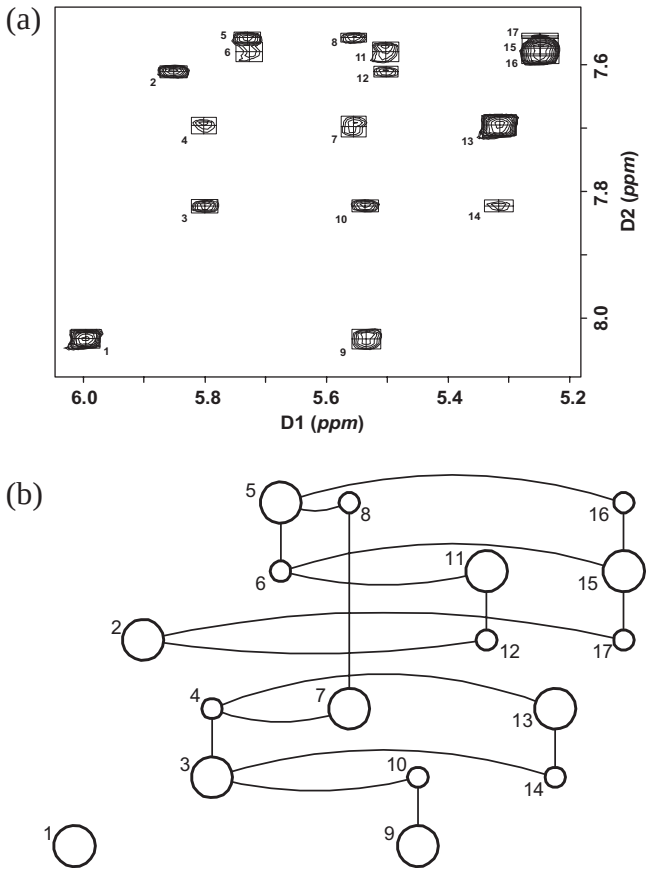

Fig. 2. Relationship between (**a**) NOESY spectrum (5–6) × (7–8) region and (**b**) the corresponding NOESY graph.

appropriate solution for the same problem in the theoretical model. Thus, a reconstruction of NOE path P_G (similar to Hamiltonian Path as far as complexity of the problem is concerned) results in solving the problem of the H6/H8–H1′ assignment.

For many instances, additional information, which reduces the search space, is available and is considered by algorithms while verifying path correctness (Adamiak *et al*., 2004; Blazewicz *et al*., 2004). Such information extend the application of the proposed combinatorial model to the non-ideal instances. The following information can be added to the model, if available: a spectral resolution, distance between splitting signals that form doublets, overlapping in a specified spectral region (the algorithm can generate and accept solutions, which include edges positioned on the same horizontal or vertical line).

It has been proved that the problem of the NOE path construction in the NOESY graph in its optimization version is strongly NP-hard (Adamiak *et al*., 2004) even for the ideal case. Hence, no polynomial-time exact algorithm is likely to exist for this problem. Recently, the exact enumerative (Adamiak *et al*., 2004) and genetic (Blazewicz *et al*., 2004) algorithms have been implemented for the problem in question. An examination of the results obtained after adopting these algorithms to sets of real data revealed certain drawbacks of these approaches (e.g. an enormous set of feasible solutions were generated by enumerative algorithm for many instances, poor precision of optimal solution when supplemental data were not provided). Thus, a need has arisen for another approach that could improve the process of NOE signal assignments in case of longer RNA chains and the noised spectra. Consequently, a new algorithm for solving the problem, based on tabu search approach, is proposed in the next section.

3 ALGORITHM

The heuristic algorithm for NOE pathways construction presented in this paper is based on tabu search method, which is an extended version of a local search procedure (Glover and Laguna, 1997). In the tabu method, for every solution x in the search space X, a neighborhood $N(x) \subset X$ is defined in such a way that every neighboring solution $x' \in N(x)$ can be reached from x in one move. A move is an elementary operation of the method. Usually, the search space X contains only feasible solutions for a considered problem. Since the tabu search method is used to solve optimization problems, every solution $x \in X$ must be evaluated according to some criterion function f. Thus, the goal is to find an element x^* in X having the optimal (i.e. minimal or maximal) value of the criterion function. At each iteration i, an algorithm chooses current solution x_i and searches its neighborhood $N(x_i)$ in order to find a local optimum, i.e. solution $x_{i+1} = \max_{x'_i \in N(x_i)}\{f(x'_i)\}$ for maximization problems or solution $x_{i+1} = \min_{x'_i \in N(x_i)}\{f(x'_i)\}$ for minimization problems. Special mechanisms, like tabu list, prevent the algorithm from getting stuck in loops and in their search procedure. A tabu list stores recent moves made by the algorithm and none of them, nor any of their reverses can be performed unless they lead to the solution better than the best one already found. Moreover, some specific situations allow for performing random or almost random moves, which cause the algorithm to jump to the other parts of the search space X. This description presents a general framework of the tabu search method and can be enriched with some additional components specific to the requirements of the considered problem.

The proposed tabu search algorithm is based on the above general tabu approach but it is extended by adding an elite structure that stores the most promising solutions. They are used as base solutions in the succeeding iterations of the search if the neighborhood of a new solution appears to be worse. The elite structure stores whole solutions together with their versions of tabu list.

The search space X is constructed on the basis of fundamental and supplemental input data. Fundamental data come from 2D-NOESY experiment and describe cross-peaks in the spectrum, i.e. cross-peak center coordinates, widths in two dimensions and NOE signal intensity. The supplemental data contain the analyzed structure and the obtained spectral information and include analyzed molecule sequence, NOE pathway length, intensity intervals, spectrum resolution, overlapping signals, doublets, additional signal rejection, pathway potential starting points, known signal positions within the pathway and H5–H6 interactions. Every solution $x \in X$ is a vector of at most n cross-peaks, where $n = 2M - 1$ and M is a number of residues in the analyzed RNA molecule. Solution x in X, being a NOE pathway, has the following properties: each cross-peak is unique within x, every two neighboring edges of x are perpendicular, no two edges of x occur on the same horizontal or vertical line. A move of tabu method is feasible if it constructs a solution obeying these constraints.

The tabu list has been designed as a queue storing q last moves leading to the base solution considered in the current iteration. Its length, equal to $18 + I_S/2$, has been selected experimentally and defined as a linear function of the problem instance size I_S being a number of cross-peaks in an aromatic/anomeric region of the analyzed 2D-NOESY spectrum.

Initial solution is generated by a random procedure limited by the general feasibility rules or by a greedy algorithm constructing the solution by adding cross-peaks one by one and starting

from various cross-peaks of a NOESY spectrum aromatic/anomeric region. Finally, if the path we look for has a maximum length (owing to some predefined conditions) and the greedy procedure returns a solution shorter than $2M-1$, then the vector is complemented with random cross-peaks.

Four different moves can be performed in order to generate the neighborhood $N(x)$ of the base solution x: swapping two selected cross-peaks of the base solution, exchanging one cross-peak from x with an unused cross-peak from the spectrum, inserting an unused cross-peak into any position of the vector storing solution x or deleting a selected cross-peak from x.

The algorithm tends to maximize a number of cross-peaks in the solution and minimizes edge deviations, inconsistency in neighboring cross-peaks alternative appearances as well as cross-peaks incompatibility with such predefined conditions like known positions within the path or H5–H6 signals. A random factor also slightly influences the evaluation of solutions. It has been introduced in order to increase the probability of leaving the local optimum and to differentiate solutions with the same scores. The global criterion function f has been defined as a weighted sum combining a set of different criteria:

$$f = \frac{1}{n}\left(\sum_{i=1}^{7} w_i y_i + r\right).$$

The components of the criterion function f are defined as follows:

- n denotes path length; if the length has not been predefined by the user, then $n = 1$;
- r denotes a random factor, $r \in \langle 0, 0.001\rangle$;
- $y_1 \in \{0,1\}$ ($y_1 = 1$ if the predefined starting cross-peak is not present on the first/last position of solution x, $y_1 = 0$ if starting cross-peak is not predefined or the predefined starting cross-peak is present on the first/last position of x);
- $y_2 = \sum_{i=1}^{n-1} a_{i,i+1}$,
  where $a_{i,i+1} \in \{0,1\}$ ($a_{i,i+1} = 1$ if the i-th and the $(i+1)$-st cross-peaks have intensities in the same interval, otherwise $a_{i,i+1} = 0$);
- $y_3 = \sum_{j=2}^{m-1} b_{j-1,j,j+1}$,
  where $b_{j-1,j,j+1} \in \{0, 0.8, 1\}$ (the value of $b_{j-1,j,j+1}$ depends on deviation of edges between j-th and $(j-1)$-st as well as j-th and $(j+1)$-st cross-peaks from horizontal/vertical position);
- $y_4 = \sum_{j=1}^{m}\sum_{i=1}^{m} c_{ji}$,
  where $c_{ji} \in \{0,1\}$ ($c_{ji} = 1$ if the j-th and the i-th edges are located on the same horizontal/vertical line, otherwise $c_{ji} = 0$);
- $y_5 = \sum_{i=1}^{n} d_i$,
  where $d_i \in \{0,1\}$ ($d_i = 1$ if the i-th cross-peak does not correspond to any predefined H5–H6 cross-peak, otherwise $d_i = 0$);
- $y_6 = \sum_{i=1}^{n-1} e_{i,i+1}$,
  where $e_{i,i+1}$ has a value corresponding to an acceptable horizontal/vertical deviation of an edge between the i-th and the $(i+1)$-st cross-peak in the solution; and
- $y_7 = N - n$.

In the above formulae N denotes the total number of cross-peaks in the considered region of the spectrum, n stands for the length (i.e. the number of cross-peaks) of the current solution x and m denotes the number of edges in x. Weighting factors in function f have been set to the following values: $w_1 = 100\,000$, $w_2 = 10\,000$, $w_3 = 10\,000$, $w_4 = 10\,000$, $w_5 = 1000$, $w_6 = 1$ and $w_7 = 1$. Optimization in the algorithm means minimization of the global criterion function value.

The new base solution, as the starting point of the succeeding iteration of the tabu algorithm, is selected according to an aspiration criterion. The latter is constructed on the basis of the following assumptions. Let x'_T denote the best neighboring solution of x obtained by a move deposited on tabu list T. Next, let us denote the best neighboring solution of x obtained by a move, which is not deposited on tabu list T by x'_{nT}. The aspiration criterion says:

IF $f(x'_T) \le f(x'_{nT}) < f(\text{best})$

OR $f(\text{best}) < f(x'_T) \le f(x'_{nT})$

THEN new base solution $= x'_{nT}$

ELSE IF $f(x'_T) < f(\text{best}) \le f(x'_{nT})$

THEN new base solution $= x'_T$.

In the above statement f denotes the global criterion function and $f(\text{best})$ stands for the value of the best solution found so far, i.e. a minimum. The remaining cases for aspiration criterion are typical and are the same as for a standard version of the tabu search algorithm.

The method stops when a global optimum has been found or 500 iterations without an improvement of the criterion function value have been performed.

4 RESULTS AND DISCUSSION

In the experiments, three methods—tabu search, genetic and exact enumerative algorithms have been analyzed. All of them were tested on Indigo 2 Silicon Graphics workstation (1133 MHz, 64 MB) in IRIX 6.5 environment. The algorithms were implemented in ANSI C programming language and tested on the real RNA data. The manual assignment of NOE resonances is very tedious and time-consuming as a result of the large number of cross-peaks and possible large number of existing alternative pathways. As a result, RNA chains analyzed in laboratories are generally rather small and match those considered in this paper. As a testing set we used a group of experimental and simulated 2D-NOESY spectra for the following molecules: I, r(CGCGCG)$_2$; II, 2′-O-Me(CGCGCG)$_2$; III and IV, r(CGCG$^{\mathrm{F}}$CG)$_2$; V, d(GACTAGTC)$_2$; VI, r(GAGGUCUC)$_2$; VII, r(GGCAGGCC)$_2$; VIII and IX, r(GGAGUUCC)$_2$; and X, r(GGCGAGCC)$_2$. The input data are the same as used by Adamiak *et al*. (2004) and Blazewicz *et al*. (2004). Experimental spectra of r(CGCGCG)$_2$, 2′-O-Me(CGCGCG)$_2$ and r(CGCG$^{\mathrm{F}}$CG)$_2$ in D_2O at 30°C were recorded on Varian Unity + 500 MHz spectrometer. Standard pulse sequence (Jeener *et al*., 1979) $\pi/2 - t_1 - \pi/2 - \tau_{\mathrm{m}} - \pi/2 - t_2$ was applied with mixing time $\tau_{\mathrm{m}} = 150$ ms. Spectra were acquired with 1K complex data points in t_2 and 1K real points in the t_1 dimension, with spectral width set to 3.7 kHz. After digital filtration by Gaussian functions, filling zero in t_1 dimension and base correction in t_2, data were collected in 1K × 1K matrices with the final digital resolution of 3.5 Hz/point in both dimensions. The 2D-NOESY spectrum of d(GACTAGTC)$_2$ was acquired on Bruker AVANCE 600 MHz. This spectrum was recorded with mixing time $\tau_{\mathrm{m}} = 400$ ms, 1K

Table 1. Time of computations (s) by exact, genetic and tabu search algorithms and a number of feasible pathways

Molecule	Test	Number of feasible solutions	Exact alg. computing time (s)	Genetic algorithm computing time (s), $p = 250$	Tabu search computing time (s)
I	T1	1	1	1	0.5
	T2	140	5	3	1
II	T1	2	1	1	0.7
	T2	776	60	2	1
III	T1	3	2	2	0.6
	T2	72	4	4	1.1
IV	T1	2	1	1	0.5
	T2	63	4	2	1
V	T1	4	1	7	0.5
	T2	240	5	4	0.9
VI	T1	1	1	4	0.7
	T2	160	30	4	1.3
VII	T1	2	1	4	0.4
	T2	3192	2453	2	1
VIII	T1	1	1	5	0.6
	T2	843	170	2	1
IX	T1	1	1	5	0.6
	T2	1134	573	3	1
X	T1	4	1	1	0.4
	T2	64	5	2	0.8

real points in t_1, 1K complex points in t_2 and spectral width of 6.0 kHz in both dimensions. After processing, the final digital resolution was equal to 6 Hz/points in both dimensions. The spectra of r(GAGGUCUC)$_2$, r(GGCAGGCC)$_2$, r(GGAGUUCC)$_2$ and r(GGCGAGCC)$_2$ were simulated using Matrix Doubling method of Felix software based on published ^{1}H chemical shifts (McDowell *et al*., 1997; McDowell and Turner, 1996; Santa Lucia and Turner, 1993; Wu *et al*., 1997) and 3D structures from Protein Data Bank. Volumes (intensities) of NOE cross-peaks for $\tau_m = 0.3$ ms were calculated from the full relaxation matrix, where a correlation time was set to 2 ns. The Lorentzian line shape functions were used for simulated NOE cross-peaks. The widths of these functions depended on the sums of coupling constants calculated from the duplex structures based on Karplus equation using Lankhorst and Haasnoot parameters (Jeener *et al*., 1979; Haasnoot *et al*., 1980; Lankhorst *et al*., 1984). Numeric data for computational experiments were obtained after peak-picking procedure of Felix Accelrys.

All the instances had been already solved manually, so we could verify whether or not each algorithm found original solution. All the molecules formed self-complementary chains, so one original pathway existed for each of them. The tests have shown, that for most instances a number of feasible solutions existed.

Two tests T1 and T2 were performed for every molecule and each algorithm. In the first test (T1) algorithms used all available expert (supplemental) knowledge, while in the second test (T2) the minimum amount of information were considered. We have examined the time taken for computations by all the algorithms. Table 1 shows computation time for test T2 (which is a worse case), when a small amount of supplemental data has been provided. In the case of exact enumerative algorithm all the feasible solutions have been generated (their number *No* for each instance is shown in Table 1), while genetic (GA) and tabu search (TS)

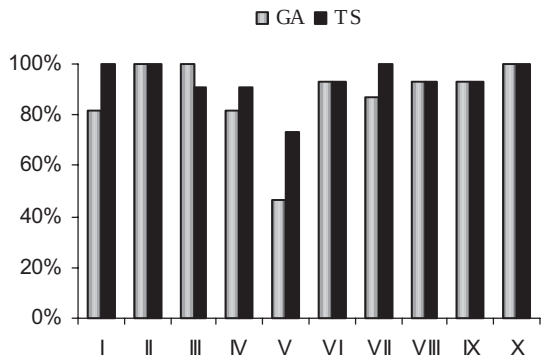

Fig. 3. Precision in test T1.

methods have constructed one optimal solution for each instance. An analysis of computation time for genetic algorithm was made for the population size $p = 250$ (Blazewicz *et al*., 2004). We have also analyzed the precision of optimal solution found by genetic and tabu algorithms (Figs 3 and 4) in tests T1 and T2. The value of precision was calculated as a percentage of the cross-peaks in the original path covered by the generated optimal solution.

We see that the proposed tabu search algorithm is the fastest of all the designed methods. This will be crucial in case of an analysis of longer nucleic chains, for which manual assignment is a difficult and tiresome work, usually impossible to be done in days or even weeks. What is more, in most cases, the precision of optimal solutions constructed by TS is better than those of the genetic algorithm. This is especially true for the instances with no supplemental data provided, for which also an enumerative algorithm is hardly effective

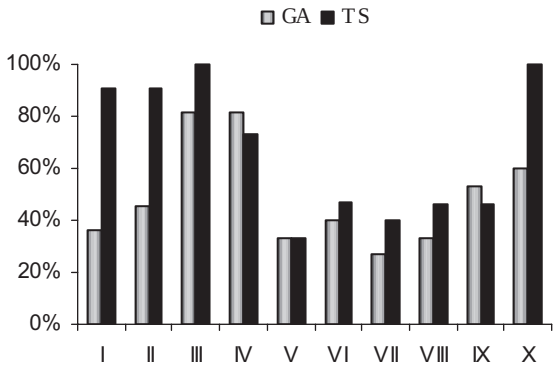

Fig. 4. Precision in test T2.

because of the huge number of feasible solutions that are generated. The results depend strongly on the input data, especially on the availability of supplemental data, which were not provided for one of the tests (T2). In the latter, the algorithms operated on minimum expert knowledge, which means that the information required for a proper interpretation of the input spectral data only, has been supplied. Thus, the information about spectral resolution, doublets or overlappings has been given, while no additional information, such as path length, intensity intervals, H5–H6 signals, known signal positions within the path, signal rejections, etc. have been provided. Such additional information is easy to define for the spectra of short RNA chains, where the 2D-NOESY spectra are not overcrowded. Unfortunately, longer the chain that is analyzed, the more packed is the spectrum obtained in the NMR experiment. Too many cross-peaks located within the same spectral region prevents the experimenter from defining the additional information just from the spectral data. This in turn results in many overlapping signals. Thus, supplying any additional data to the algorithms solving the problem appears hard and the experimenters try rather, the less risky, algorithms without the expert information. Unfortunately, computational analysis with the use of enumerative algorithm in such cases appears rather ineffective and disqualifies this method here. Of course, one may be sure that the enumerative algorithm finds the original solution, but looking through the generated set of at least 60 feasible paths in order to locate this original one is a hopeless job and harder than a manual reconstruction of the NOE path. Thus, it seems beneficial to apply genetic or tabu search algorithm for solving such instances of the problem. Even if the heuristics find only half of the original pathway it facilitates the problem to a very large degree. Having the partial assignment, an experimenter is able to complete the NOE pathway in a reasonable time without too much of an effort. On the other hand, in case of the lack of knowledge (test T2), the obtained path maybe of lower than expected precision. In these cases, additional NMR experiments may help to verify the goodness of the solution obtained by the tabu search approach.

5 CONCLUSIONS

In this paper, the problem of the reconstruction of NOE pathways in 2D-NOESY spectra of RNA molecules has been considered. The tabu search algorithm, based on the combinatorial model of the problem, has been proposed and applied to the collection of spectral data gathered from the NMR experiments for different RNA molecules. During computational experiments we have compared the results obtained by the tabu method against those obtained by exact enumerative and genetic algorithms, respectively. Tabu search method gives superior results and for most instances the obtained solutions coincide with the majority of vertices in the original NOE path. Tabu search is also the fastest of all the tested methods designed for an automatic assignment of NOE pathways. In case of an expert knowledge deficiency, the tabu approach narrows down the final solution set indisputably well, thus, appearing very useful in practical situations. The large number of possible NOE pathways returned by the enumerative algorithm for the instances without the additional expert information, makes this approach hard to use.

As a continuation of the research reported in this paper, one may consider the analysis of spectra which contain a lot of noise signals as well as 3D spectra of RNA molecules. Since tabu search is a quick and precise method, applying it to more complicated cases and instances as well as to an analysis of long nucleic chains seems promising.

ACKNOWLEDGEMENT

The research has been partially supported by the grant 3T11F00227 from the Ministry of Science, Poland.

REFERENCES

Adamiak,R.W. *et al*. (2004) an algorithm for an automatic NOE pathways analysis of 2D NMR spectra of RNA duplexes. *J. Comput. Biol.*, **11**, 163–180.

Atreya,H.S. *et al*. (2000) A tracked approach for automated NMR assignments in protein (TATAPRO). *J. Biomol. NMR*, **17**, 125–36.

Balley-Kellogg,C. *et al*. (2004) A random graph approach to NMR sequential assignment. *Curr. Comput. Mol. Biol.*, 58–67.

Blazewicz,J. *et al*. (2004) Evolutionary approach to NOE paths assignment in RNA structure elucidation. In *Proceedings of the IEEE Symposium on Computer International in Bioinformatics and Computational Biology*, 206–213.

Glover,F. and Laguna,M. (1997) *Tabu Search*. Kluwer Academic Publishers, Boston, MA.

Haasnoot,C.A.G. *et al*. (1980) The relationship between proton–proton NMR coupling constants and substituent electronesgativities—I. *Tetrahedron Lett.*, **36**, 2783–2792.

Jeener,J. *et al*. (1979) Investigation of exchange processes by 2D NMR spectroscopy. *J. Chem. Phys.*, **71**, 4546–4593.

Lankhorst,P.P. *et al*. (1984) Carbon-13 NMR in conformational analysis of nucleic acid fragment. *J. Biomol. Struct. Dyn.*, **1**, 1387–1405.

Linge,J.P. *et al*. (2003) ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics*, **19**, 315–316.

McDowell,J.A. and Turner,D.H. (1996) Investigation of the structural basis for thermodynamic stabilities of tandem GU mismatches: solution structure of $(rGAGGUCUC)_2$ by 2-D NMR and simulated annealing. *Biochemistry*, **35**, 14077–14089.

McDowell,J.A. *et al*. (1997) Investigation of the structural basis for thermodynamic stabilities of tandem GU wobble pairs: NMR Structures of $(rGGAGUUCC)_2$ and $(rGGAUGUCC)_2$. *Biochemistry*, **36**, 8030–8038.

Moseley,H.N.B. and Montelione,G.T. (1999) Automated analysis of NMR assignments and structures for proteins. *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Roggenbuck,M.W. *et al*. (1990) Path analysis in NMR spectra: application to an RNA octamer. *Structure Methods*, **3**, 309–317.

Santa Lucia,J. and Turner,D.H. (1993) Structure of $(rGGCGAGCC)_2$ in solution from NMR and restrained molecular dynamics. *Biochemistry*, **32**, 12612–12623.

Szachniuk,M. *et al*. (2003) A combinatorial analysis of 2D NMR spectra of RNA duplexes. *Curr. Comput. Biol.*, 345–346.

Szachniuk,M. Popenda,M., Adamiak,R.W. and Blazewicz,J. (2004) The method of assignment of the magnetization transfer pathway between H6/H8–H1′ protons in the 2-dimensional NOESY spectra in Nuclear Magnetic Resonance spectroscopy of nucleic acids. Polish Patent Pending Application P364736.

Varani,G. and Tinoco,I.,Jr (1991) RNA structure and NMR spectroscopy. *Q. Rev. Biophys.*, **24**, 479–532.

Wu,M. *et al*. (1997) Solution structure of $(rGGCAGGCC)_2$ by 2-D NMR and the iterative relaxation matrix approach. *Biochemistry*, **36**, 4449–4460.

Wüthrich,K. (1986) *NMR of Proteins and Nucleic Acids*. John Willey & Sons, NY.