BULLETIN OF THE POLISH ACADEMY OF SCIENCES
TECHNICAL SCIENCES
Vol. 52, No. 3, 2004

Evolutionary algorithm for a reconstruction of NOE paths in NMR spectra of RNA chains

J. BŁAŻEWICZ[1], M. SZACHNIUK[1]* and A. WOJTOWICZ[2]

[1] Institute of Bioorganic Chemistry, Polish Academy of Sciences, 12/14 Noskowskiego St., 61–704 Poznań, Poland
[2] Institute of Computing Science, Poznan University of Technology, 3a Piotrowo St., 60–965 Poznań, Poland

Abstract. Resonance assignment remains one of the hardest stages in RNA tertiary structure determination with the use of Nuclear Magnetic Resonance spectroscopy. We propose an evolutionary algorithm being a tool for an automatization of the procedure. NOE pathway, which determines the assignments, is constructed during an analysis of possible connections between resonances within aromatic and anomeric region of 2D-NOESY spectra resulting from appropriate NMR experiments. Computational tests demonstrate the performance of the evolutionary algorithm as compared with the exact branch-and-cut procedure applied for the experimental and simulated spectral data for RNA molecules.

Keywords: NMR, NOE path, RNA tertiary structure, evolutionary algorithm, branch-and-cut algorithm.

1. Introduction

Cognition of biomolecule structures has been now one of the most fundamental tasks in the biochemical research area. Structural analysis of proteins and nucleic acids contributes for clarifying their biological functions, drug design, identification of the new diseases, raising the new specimens of plants and animals etc. Initially, the research in that area concentrated on proteins and deoxyribonucleic acids (DNA). However, knowledge gained when studying these molecules appeared insufficient to answer all the questions that have arisen for years. Thus, the research has been extended for the molecules of the ribonucleic acid (RNA), which transmits genetic information from DNA into proteins and controls certain chemical processes in the cell. Regarding RNA functional variety as well as its quick degradation under in vitro conditions, studying the structures of these molecules proved to be more difficult than the examination of proteins and deoxyribonucleic acids. Consequently, development of analytical methods dedicated to the exploration of RNA has been less dynamic than the spread of processing protein and DNA structures.

The subjects of RNA structural analysis are: primary structure determined by the sequence of nucleoside monophosphates in the chain, secondary structure describing one- and two-strand fragments as well as the formation of loops or helices, and tertiary structure, which characterizes the three-dimensional shape of the entire chain. An elucidation of molecule tertiary structures has become possible owing to the development of crystallographic methods, Raman spectroscopy, fluorescence, nuclear magnetic resonance (NMR) spectroscopy as well as some other analytical methods. Recent years, yielded a quick spread of NMR spectroscopy, which has been now a well established method for structure determination of biomolecules in solution [1]. The elucidation procedure using NMR is composed of two general stages: experimental, where multidimensional correlation spectra are acquired and computational, where spectra are analyzed and structure is determined. Types of NMR experiments differ for proteins [2] and nucleic acids [3, 4]. Nevertheless, in all methods of NMR structure analysis raw experimental data are exposed to the action of processing, peak-picking, assignment, restraints determination, structure generation and refinement. The procedure assigning the observed NMR signals to the corresponding protons and other nuclei is a bottleneck of the RNA structure elucidation process. The assignment is usually based on the analysis of two dimensional (2D) spectra resulting from NMR experiments like NOESY (Nuclear Overhauser Enhancement SpectroscopY), COSY (COrrelated SpectroscopY) and TOCSY (TOtal Correlation SpectroscopY). For short DNA and RNA duplexes the assignment is performed manually in accordance with the experimentator's knowledge and intuition. However, for the longer nucleic chains, due to a considerable large number of signals and their overlapping, the assignment step becomes troublesome. Therefore, it has been of a great need to facilitate NMR structural analysis of biopolymers by an introduction of automatic procedures at this level.

At present, automatization of NMR spectra analysis makes the strong impact on the determination of protein structures [5]. Several programs exist which automatize the process of their assignment [6–11]. Unfortunately, these programs cannot be applied for an automatic assignment of the nucleic acids spectra. Distinctive patterns of NH peptide bond resonances, for several amino acid residues within protein structure, make their recognition via automatic assignment much easier than in case of nucleic acids, especially RNA. To our knowledge, only two papers exist concerning an automatic pathway analysis applied for RNA duplexes [12, 13]. The first algorithm is based on the Reduced Adjacency Matrix and Backtracking procedures. The second method uses Branch-and-Cut

* e-mail: Marta.Szachniuk@cs.put.poznan.pl

procedure idea. Both are exact algorithms, applicable for an analysis of short unbreaked RNA duplexes.

In this paper, our attention is focused on the development of an approximation algorithm for an automatic generation of pathways between H6/H8 and H1' resonances, known as the NOE (Nuclear Overhauser Effect) signals, observed for RNA duplexes during the NOESY experiment. The NOE peaks illustrated in the 2D-NOESY spectrum resulting from the experiment are connected to form the path, called the NOE pathway. The method takes into account the specificity of the data, thus, basing on the combinatorial model of a NOESY graph [13]. Exact analytical algorithm applied for the long nucleic chains may provide too many solutions, preventing from performance of the next steps in the elucidation process. As it is crucial to generate the pathways as close as possible to the original one, we have considered an application of metaheuristics, known as successful in tackling the difficult combinatorial problems. The experiments have shown a good quality of the evolutionary approach providing the new analytical tool for resolving the problem of structure reconstruction.

An organization of the paper is as follows. Section 2 discusses the combinatorial model of the problem in question. Section 3 outlines the basis of the general evolutionary algorithm structure and presents the details about the new evolutionary algorithm dedicated to the problem of the NOE paths reconstruction. In Section 4, the results of computational experiments are given, while Section 5 sums up the results of application of the evolutionary approach to the problem of NOE paths assignment and points out the directions for further research.

2. Combinatorial model of the problem

Our aim has been to facilitate the NMR analysis of RNA molecules, especially their fragments known as the helical motives in ribonucleic acids structure. One of the major analytical steps is an identification of the sequence-specific connectivity H8/H6$_{(i)}$–H1'$_{(i)}$–H8/H6$_{(i+1)}$ pathway, represented as NOE path in the 2D–NOESY spectrum of RNA duplex [1]. Formation of such a path is possible because each aromatic H6/H8 proton of nucleotide residue is in close proximity to two anomeric protons: its own and the preceding H1' proton. A short exemplary H8/H6$_{(i)}$–H1'$_{(i)}$–H8/H6$_{(i+1)}$ pathway going through the four-nucleotide strand r(CGUA) is illustrated in Fig. 1, where main NOE interactions between protons of our interest are marked with arrows.

The NOE interactions between protons are represented as cross-peaks in the 2D-NOESY spectrum generated for the molecule during the appropriate NMR experiment. The whole spectrum contains nine characteristic regions of the correlated signals. In the search for NOE connectivity pathway H8/H6$_{(i)}$–H1'$_{(i)}$–H8/H6$_{(i+1)}$ we focus only on the aromatic/anomeric region, which

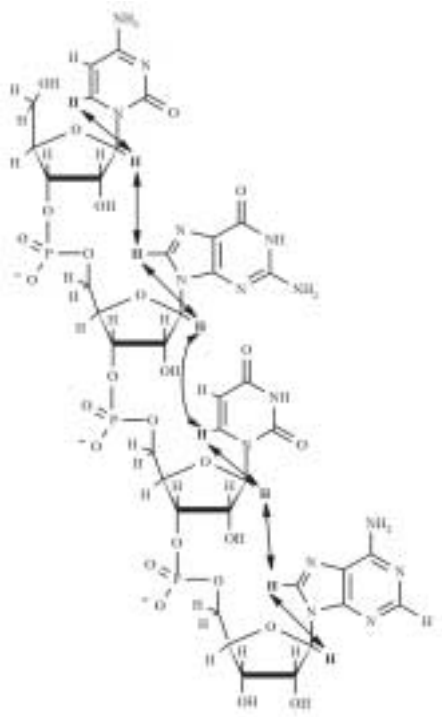


Fig. 1. Main NOE interactions in r(CGUA)

borders interactions between protons of our interest. Figure 2 presents an exemplary aromatic/anomeric region of 2D–NOESY spectrum for r(GAGGUCUC)$_2$. It includes 24 numbered cross-peaks representing the NOE interactions generated by the pairs of protons. Each proton of the analyzed molecule can be described by its resonance frequency known as the chemical shift and expressed in parts per million (ppm). Thus, for example peak 6 from Fig. 2, having coordinates (8.02, 5.66) represents the NOE signal generated by H8 (8.02 ppm) and H1' (5.66 ppm) protons belonging to G1 nucleotide.

The path is composed of intranucleotide (with higher intensity) and internucleotide (with lower intensity) interactions. They give rise to the alternately appearing cross-peaks. In case of the ideal A-RNA duplexes, the NOE pathway starts with the intranucleotide interaction at 5' end of the strand and length of the path equals $2 \cdot n - 1$, where n is a number of residues in RNA chain. Each proton, except for the starting and terminal ones, belonging to the pathway gives cross-peaks with two other protons. Every cross-peak is characterized by the two coordinates of its centre, widths in both directions and the value of signal intensity. However, if the fine structure of a cross-peak is not considered, it can be defined as the point with two coordinates only, specified by the values of the chemical shifts of the corresponding protons. Therefore, every two consecutive points in the NOE pathway have exactly one coordinate in common and consecutive connections within the pathway lay vertically or horizontally. Figure 3 demonstrates the NOE pathway found in the spectrum of r(GAGGUCUC)$_2$.

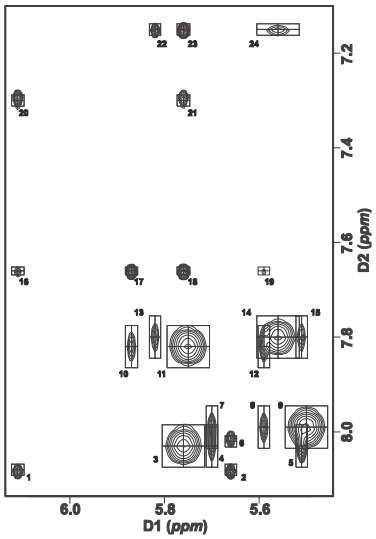


Fig. 2. Aromatic/anomeric region of the 2D-NOESY spectra for r(GAGGUCUC)$_2$

Only one NOE pathway exists for each RNA molecule forming self-complementary chain. It satisfies all the conditions mentioned in the previous paragraphs. We will call this path an original solution.

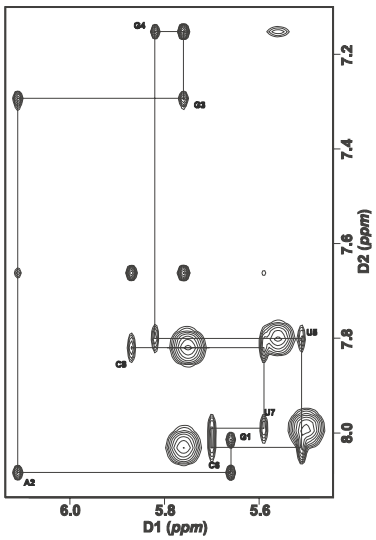


Fig. 3. NOE pathway for r(GAGGUCUC)$_2$

Respecting the above description of the problem, we proposed its graph-theoretic model [13] being a background for the complexity analysis and for the construction of the algorithms solving the problem. Sequential assignments of H6/H8–H1' correspond to searching for a path between vertices of a graph, thus, converting 2D–NOESY spectrum to a graph structure seemed an attractive idea. We defined a NOESY undirected graph G = (V,E) situated on a plane following the succeeding prerequisities:

1. every vertex $v \in V$, where V is the set of vertices, represents one cross-peak from the hypothetical NOESY spectrum,
2. vertices are weighted: weight 1 is assigned to every vertex representing intranucleotide NOE signal, weight 0 — to every vertex representing internucleotide NOE signal,
3. number of vertices in a graph equals the number of cross-peaks in the spectrum,
4. every edge $e \in E$, where E is the set of edges represents a possible connection between two cross-peaks with different intensity having one common coordinate,
5. number of edges in a graph equals the number of all possible correct connections (i.e. lines between two cross-peaks of different intensities having one common coordinate) that can be drawn in the spectrum [13].

As the NOE interactions are illustrated only in the aromatic/anomeric region of the spectrum, the definition of the graph also concerns only this part of 2D–NOESY, but — for the simplicity — we call it the spectrum.

Basing on the aromatic/anomeric region of the 2D-NOESY spectrum and regarding the characteristics of the NOE pathway one can create an appropriate NOESY graph being compatible with the above definition. Figure 4 shows an exemplary NOESY graph corresponding to the spectrum of r(GAGGUCUC)$_2$ illustrated in Figure 2.

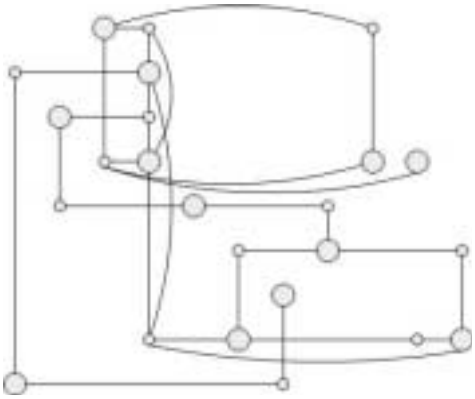
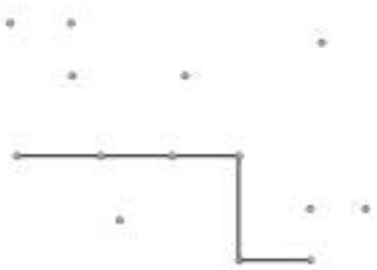


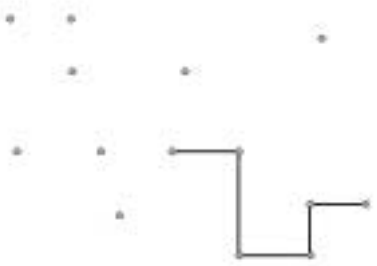
Fig. 4. NOESY graph corresponding to the spectrum of r(GAGGUCUC)$_2$

Converting an aromatic/anomeric region of the spectrum to a graph requires formulation of the NOE pathway problem in terms of graph theory. Thus, an appropriate path in the graph, being the corresponding solution of the problem in the theoretical model, must satisfy the following conditions [13]: every vertex and edge may occur in the path at most once, every two neighboring edges are perpendicular, no two edges lie on the same horizontal or vertical line, the length of a path, measured as a number of constituent edges, equals $2|V_1| - 2$, where $|V_1|$ is a number of intranucleotide signals. We can see the conformity between the problems of the NOE path and Hamiltonian path in a graph. However, the problem of NOE assignments assumes additional constraints on the search space of the algorithms:

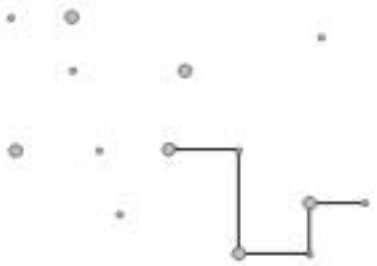
a) an edge can join only vertices having exactly one common coordinate, thus, only horizontal and vertical edges are correct,



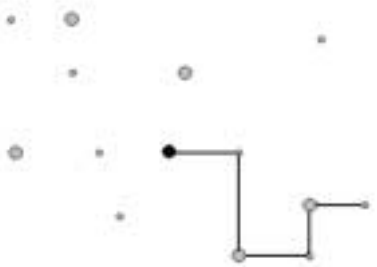
b) horizontal and vertical edges occur alternatively in the path,



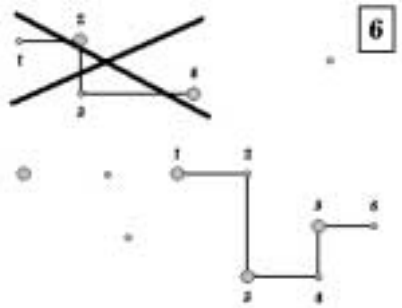
c) every edge connects two vertices having different intensities (inter-intra),



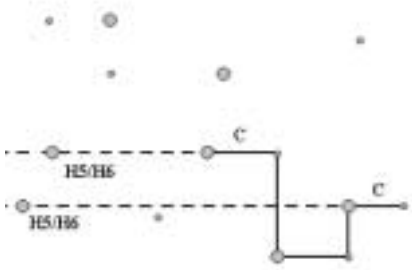
d) for some instances we know the position of some vertices in a sequence; in particular the starting points of a path are very probable to be known,



e) for some instances we know the NOE path length,



f) for the molecules including citidine, every vertex representing signal intra generated by citidine protons have the same coordinate as one of the vertices representing H5–H6 signals.



There are many cases, which demand an additional expert information for a proper interpretation of the spectral data. Such a knowledge, given to the algorithm, extends the chances of practical appliance of the proposed combinatorial model for the non-ideal instances. We have proposed supplying the following information if needed:

a) spectral resolution
   If the resolution of the spectrum is small, then location and dimension of the cross-peak should be considered within error range.

b) distance between doublet cross-peaks
   Sometimes the NMR signal is so strong, that the spectrometer registers it as two signals, called the doublet. An expert can distinguish such a doublet in the spectrum and define a value of the distance between two cross-peaks, that represent one splitting signal. Next, an algorithm finds the doublet and calculates one signal out of it: computes an average position, appropriate widths and mates the volume (intensity).

c) overlapping in a specified spectral region
   Signal overlapping appears when we analyze the spectra of longer chains. It provokes occurring many cross-peaks with the same value of one centre coordinate. For such instances the NOE pathway can include more than one edge lying on the same horizontal or vertical line within the specified overlapping region. Thus, an algorithm generates and accepts solutions, which allow for such edges occurence.
   By indication, that a NOESY graph contains NOE

pathway if and only if there exists a Hamiltonian path of the desired properties in the corresponding graph, it has been proved, that the problem of the NOE path construction in the NOESY graph in its decision version is strongly NP-complete [13, 14]. Hence, no polynomial-time exact algorithm is likely to exist for this problem. Recently the branch-and-cut algorithm has been implemented for the problem in question [13]. An examination of the results obtained after adopting this exact algorithm to a set of many various data made us try the other approach, that could improve the process of NOE assignments in case of longer RNA chains and the noised spectra. Consequently, a new algorithm for solving the problem will be proposed in the next section.

3. Evolutionary algorithm

An idea of evolutionary algorithms for combinatorial optimization problems, inspired by Darwin's theory of evolution, was introduced by John Holland in 1975. Computer procedures employing the mechanics of natural selection and genetics to evolve solutions to combinatorial problems appeared very effective in providing near optimal solutions with a reasonable computational effort. Thus, the recent years brought an increasing popularity of evolutionary methods as well as of the other metaheuristics, designed for complex combinatorial optimization problems [15].

Evolutionary algorithms use biologically derived techniques such as inheritance, mutation, natural selection and recombination. To their basic components one can subsume population (set of solutions), chromosomes (individuals), fitness of the chromosomes, process of reproduction (selection of parents and children generation), replacement (death of the individuals) and generation completion. Typically evolution starts from a population of completely random individuals (solutions), represented by chromosomes, and happens in generations. Traditionally, solutions occur as binary strings of 0s and 1s, but different encodings are also possible. Each individual is characterized by its fitness. Each generation is defined by population size, as well as the birth and death processes. In every generation, multiple individuals are stochastically selected from the current population, and next — modified through mutation or recombination to form a new population, which becomes current in the following iteration of the algorithm. Solutions which form the offspring are selected according to their fitness — the more suitable they are the more chances they have to reproduce. This is motivated by a hope, that the new population will be better than the old one. In such a manner, an approximation algorithm evolves towards better solutions. The procedure stops when the desired stopping criterion, like number of populations or improvement of the best solution, is reached. As a result of this simulated evolution one obtains highly evolved solution to the original problem, that is the best chromosome picked out of the final population.

Evolutionary algorithms have been widely applied to many different optimization problems. Examination of their performance has shown that the computational effectiveness depended on the values given to algorithm parameters like population size, initial population, genetic operators, fitness and stopping criteria [16–18]. Our implementation of the evolutionary algorithm complies with the typical structure characteristics described above as well as the nature of the problem of the NOE pathway reconstruction on the basis of 2D-NOESY spectra of RNA molecules. All components of the proposed algorithm are introduced in the following paragraphs.

Input data: We consider an aromatic/anomeric region of two-dimensional NOESY spectrum resulting from NMR experiment performed for RNA molecule. The spectrum contains cross-peaks, which represent NOE interactions between the atoms. In particular, the considered region borders the interactions between protons: H6, H8, H1', H5. Every cross-peak is characterized by the two coordinates $D1$, $D2$ of its center, widths $dD1$, $dD2$ in both directions and volume Vol of the signal corresponding to its intensity.

Size of the population: Population size P is kept constant through the generations. It is the parameter, which can be changed if necessary. We assumed its value between 250 an 1000 individuals.

Initial population: The initial population partially consists of the individuals generated randomly (but satisfying some predefined criteria) and partially — of the solutions generated by the greedy algorithm beginning the search from various starting points.

Individual encoding: An individual is represented by a vector of size n, where n is a maximum length of the NOE path for the given molecule. The value of n can be derived from the molecule primary structure and equals $2N - 1$, where N denotes the number of nucleotides in the RNA chain. The vector is composed of the sequence of vertex numbers written in the order of their occurrence in the path.

Fitness: Fitness of each individual is determined by the associated value of the goal function, being one of the most crucial component of the algorithm.

Goal function: The function comprises the knowledge about desirable features of the problem solutions, thus assuring metaheuristic efficiency. It assembles the criteria, which serve the evaluation of individuals (paths). The value of the function basically depends on acceptability criteria, path's length and the sum of the acceptable edge deviations. Particularly, when a couple of solutions having the same function value appear, an aleatory element is considered as an additional function component. This additional element tries to prevent the search process to enter a local optimum. The criteria aggregated in the goal function can be divided into two groups of acceptability (more important) and optimization (less

important) criteria. The function penalizes the solution, if:

a) the solution does not contain the given signal, which is to appear in the specified position (when such a knowledge is defined in the input),
b) the solution contains neighbouring vertices, which can not be joined by an edge (e.g they do not have one common coordinate or have the same intensity),
c) the solution contains neighbouring edges, that are not perpendicular,
d) the solution contains edges lying on the same horizontal or vertical line,
e) the solution has not got the maximum length (when length of the path is not defined in the input and should be maximized),
f) the sum of edges deviation within the path is greater than zero.

Proper definition of the penalty values prevents their substitution between acceptability and optimization criteria. Goal function value is minimized during optimization.

Selection: The aim of the selection step is to eliminate the solutions and passing the good ones from one generation to the other. Thus, basing on fitness values, the procedure picks individuals from the current population and builds from them the mating pool for the reproduction step. We have adopted the roulette system of selection, which demands a calculation of a fitness of each individual, a total fitness of the whole population, a probability of each individual selection and a cumulative distribution of each solution. Afterwards, algorithm draws $P/2$ values out of the interval $\langle 0, 1\rangle$ and removes from the current population all individuals having cumulative distribution values corresponding to the fated ones. Thus, higher goal function values result in a greater probability of being removed from the population with no chance to get into the mating pool.

Reproduction: New solutions (offspring) are generated with the use of crossover and mutation operators applied to the mating pool.

Crossover: Crossover phase in our algorithm is based on the two operators *OX* and *merge* applied to the individuals selected according to the roulette mechanism. Solutions having better (smaller) values of the goal function are chosen for reproduction with higher likelihood. An offspring is added to the population of the next generation in place of the individuals removed in the selection step. Crossing operators are responsible for carrying valuable schemes to the next generations. Thus, their proper definitions provide the algorithm convergence to the optimum. We have introduced the following operators:

a) OX operator

In the problem of NOE pathways, like in the other problems of that kind [19], the quality of solutions depends mostly on the features of edges, which have to satisfy a number of particular conditions. The OX operator disrupts relatively small number of edges, thus, letting to preserve many features of the parents [20]. It is used by the algorithm as the only operator in cases, where the length of the NOE pathway is known a priori. At the beginning, the OX operator qualifies a random sequence of vertices of one parent and places it adequately in the offspring sequence. Next, the empty places of the new solution are filled with the vertices of the second parent according to their succession. No vertices reduplicate within the generated sequence.

b) Merge operator

If the length of the NOE pathway is not known a priori, we propose the usage of two operators: OX and merge. Merge operator improves the quality of solutions in the final population. It keeps valuable schemes of the short parent sequences and generates long, more desirable offspring solutions. The operator copies the whole sequence of one parent into the new solution. Subsequently, in the second parent sequence it finds a vertex, which has occupied the closing position of the first parent. Then, operator copies the subsequent vertices of the second parent to the offspring sequence. Copying from the second parent stops, when procedure finds a vertex which has been already put into the generated solution.

Mutation: During the mutation phase each solution can be a subject of up to five independent mutation operators. In practice, five operators are used if the pathway length is a priori unknown. If the length is known, only three of them mutate. In theory, the probability of utilization equals 0.1 for each mutation operator and the probability of mutation equals 0.3 or 0.5 for each solution. Practically, likelihood values are smaller. We have defined the following mutation operators:

a) Swapping operator
   Swap operator selects two random vertices and exchanges their positions within the sequence.
b) Replacement operator
   Replacement operator draws a random vertex from the sequence and replaces it with a random vertex from behind the sequence.
c) Inversion operator
   Inversion operator selects two positions in the solution sequence. Next, it rewrites the vertices positioned between the fated places in the reverse order.
d) Addition operator
   Addition operator is used if the pathway length is unknown. The operator inserts an additional unemployed vertex into the random position of the sequence.
e) Deletion operator
   Deletion operator is used if the pathway length is unknown. The operator removes a randomly selected vertex from the sequence.

Creating the new generation: The next generation is formed out of the best parent solutions and all individuals from the offspring population.

Stopping criterion: Stopping criterion has been de-

fined as the number of succeeding generations without improvement of the best individuals. Its value has been set experimentally to 250 iterations.

The proposed algorithm performs in pursuance of the following steps:

1. Generate initial population $t = 0$: create P individuals, where each individual is a permutation of n signals given in the input.
2. Calculate fitness of each individual in population t and find the best individual in this population.
3. Repeat steps 4–6 until the stopping criteria are not satisfied:
4. Basing on the fitness values of the individuals select parents for the new population $t + 1$ according to the roulette system. For each pair of the selected parents apply crossover operators *OX* and *Merge*.
5. Mutate individuals: considering offspring and parents populations pick an individual and apply the chosen mutation operators.
6. Evaluate fitness for each individual in the current offspring and parent population. Create the new generation out of the best parents and all the offspring solutions.

4. Computational tests

Both, evolutionary and exact algorithms 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. As a testing set we used a group of experimental and simulated 2D-NOESY spectra. The input data are the same as used in [13]. 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. The 2D-NOESY spectrum of d(GACTAGTC)$_2$ was acquired on Bruker AVANCE 600 MHz. 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 [21–24] and three dimensional structures from Protein Data Bank. 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 the consistency of paths generated by each algorithm with the original solution. All the molecules formed self-complementary chains, so one pathway (an original solution), correct from the biochemical point of view, existed for each of them. Exact algorithm has been designed in such a way, that it enumerates all feasible solutions to each instance of the problem. Evolutionary algorithm may also produce feasible paths as well as one optimal, being the best feasible solution. The cardinality of a final feasible solution set depended on the expert information reducing the search space. In the case of an expert information deficiency the algorithms generated a set of feasible solutions, which satisfied all the conditions of the NOE path. Additionally, the evolutionary algorithm optimalized path length and edge deviations, thus giving the optimal solution.

Three tests were performed for every molecule. In the first test algorithms used all available expert knowledge. In the second test we checked how the algorithms worked if some important data lacked – no path length was defined. In the third case the algorithms did not consider any expert knowledge. Table 1 summarizes the experimental results of the algorithms. The first column contains the molecule sequence and the number of the analyzed cross-peaks (instance size) for each instance of the problem. In the second column, numbers of tests (1–3) are given. The number of feasible paths generated by the exact algorithm for every molecule has been placed in the third column. As, the exact algorithm looks through the whole search space, it always finds an original solution as well as the other feasible solutions if they exist. In the next four columns we have placed the results obtained by the evolutionary algorithm run for the different population sizes $P \in \{250, 500, 750, 1000\}$, i.e. values of optimal solution precision. The precision is given as two numbers o/v, where o determines the length of maximum sub-path in the optimal solution that covers the original vertex sequence and v is the number of vertices in the original path.

Computational experiments have proved, that the proposed evolutionary algorithm in most cases gives very good results. The precision of this method have appeared high enough to consider it as an alternative approach in solving the problem of NOE pathways reconstruction. This especially concerns the instances, for which an exact algorithm is hardly effective because of an enormous number of feasible paths generated. The results of test number 3 for all the analyzed molecules deserve the special attention. In this test both algorithms operated on the minimum expert knowledge, what means that only the information required for a proper interpretation of the input spectral data has been supplied. Thus, the information about spectral resolution, doublets or overlapping has been defined, while any additional, like path length, volume intervals, H5–H6 signals, known signal positions within the path, signals rejection has not been provided. Such an additional information is easy to define for the spectra of short RNA chains, where the 2D–NOESY spectra are not overcrowded. Unfortunately, the longer chain is analyzed, the more packed spectrum results from the NMR experiment. An extreme number of cross-peaks located within the same spectral region prevents the experimentator to define additional information just from the spectral data and results in many overlapping signals. Thus, supplying any additional data to the algorithms solving the problem of assignments appears hard and the experimentators rather induce to — the less risky — trying the algorithms without an expert information. Unfortunately, computational analysis with the

use of exact algorithm in such cases appears completely ineffective and disqualifies this method here. Of course, we are sure that the exact algorithm finds the original solution, but looking through the generated set of 3192 feasible paths (see test 3 for the seventh molecule) in order to situate the original one is a hopeless job and harder than manual reconstruction of the NOE path. Thus, it seems a good idea to apply an evolutionary algorithm for solving such instances of the problem. Even if the evolutionary method finds only half of the original pathway it facilitates the problem to a very large degree. Having the partial assignment an experimentator is able to complete the NOE pathway in a measurable time with not an extreme effort. It seems possible to analyze manually the set of up to 20 feasible solutions in order to find the original one among them. But greater number of paths discourages an ordinary researcher to look through them. So, if we revise experimental results placed in Table 1, we will see that even the best case for the test 3, which is the forth molecule r(CGCG$^{\mathrm{F}}$CG)$_2$ (63 feasible solutions generated by the exact algorithm) makes us rather consider the optimal path generated by the evolutionary algorithm.

Another important part of the tests is measuring the time of computation. Table 2 contains the times of computation for each analyzed instance.

Table 1
Number of feasible solutions and optimal solution precision

Molecule instance size	Test	Exact algorithm	Evolutionary algorithm			
			P = 250	P = 500	P = 750	P = 1000
1. r(CGCGCG)$_2$	1	1	9/11	11/11	11/11	11/11
17 peaks	2	2	11/11	9/11	11/11	11/11
	3	140	4/11	10/11	10/11	10/11
2. 2'-OMe(CGCGCG)$_2$	1	2	11/11	11/11	11/11	11/11
17 peaks	2	4	5/11	11/11	6/11	9/11
	3	776	5/11	11/11	11/11	7/11
3. r(CGCG$^{\mathrm{F}}$CG)$_2$	1	3	11/11	9/11	11/11	9/11
16 peaks	2	21	8/11	6/11	11/11	9/11
	3	72	9/11	9/11	7/11	7/11
4. r(CGCG$^{\mathrm{F}}$CG)$_2$	1	2	9/11	9/11	9/11	9/11
22 peaks	2	6	10/11	9/11	9/11	7/11
	3	63	10/11	9/11	6/11	8/11
5. d(GACTAGTC)$_2$	1	4	7/15	12/15	8/15	6/15
26 peaks	2	8	6/15	8/15	6/15	6/15
	3	240	5/15	5/15	7/15	5/15
6. r(GAGGUCUC)$_2$	1	1	14/15	14/15	14/15	14/15
24 peaks	2	1	9/15	9/15	14/15	14/15
	3	160	6/15	5/15	5/15	5/15
7. r(GGCAGGCC)$_2$	1	2	13/15	15/15	13/15	15/15
26 peaks	2	2	10/15	13/15	12/15	15/15
	3	3192	4/15	3/15	8/15	7/15
8. r(GGAGUUCC)$_2$	1	1	14/15	15/15	14/15	15/15
25 peaks	2	1	14/15	14/15	14/15	14/15
	3	843	5/15	12/15	13/15	6/15
9. r(GGAGUUCC)$_2$	1	1	14/15	14/15	14/15	15/15
26 peaks	2	1	10/15	8/15	13/15	8/15
	3	1134	8/15	5/15	4/15	5/15
10. r(GGCGAGCC)$_2$	1	4	10/10	10/10	10/10	10/10
20 peaks	2	8	10/10	10/10	10/10	10/10
	3	64	6/10	10/10	8/10	8/10

Table 2
Time of computations [s]

Molecule instance size	Test	Exact algorithm	Evolutionary algorithm			
			P = 250	P = 500	P = 750	P = 1000
1. r(CGCGCG)$_2$	1	1	1	3	8	11
17 peaks	2	4	3	7	18	31
	3	5	3	10	21	48
2.2'-OMe(CGCGCG)$_2$	1	1	1	3	5	13
17 peaks	2	2	2	9	15	32
	3	60	2	12	38	63
3. r(CGCG$^{\mathrm{F}}$CG)$_2$	1	2	2	4	16	17
16 peaks	2	3	2	8	20	64
	3	4	1	6	17	45
4. r(CGCG$^{\mathrm{F}}$CG)$_2$	1	1	2	5	5	19
22 peaks	2	1	2	9	22	58
	3	2	2	4	7	20
5. d(GACTAGTC)$_2$	1	1	7	22	74	129
26 peaks	2	1	2	7	22	64
	3	5	4	6	27	90
6. r(GAGGUCUC)$_2$	1	1	4	18	44	80
24 peaks	2	1	2	8	16	52
	3	30	4	17	73	132
7. r(GGCAGGCC)$_2$	1	1	4	6	10	16
26 peaks	2	1	2	10	37	57
	3	2453	2	8	27	62
8. r(GGAGUUCC)$_2$	1	1	5	18	44	80
25 peaks	2	1	1	6	17	45
	3	170	2	6	22	63
9. r(GGAGUUCC)$_2$	1	1	5	19	43	80
26 peaks	2	1	3	8	14	33
	3	573	3	5	34	31
10. r(GGCGAGCC)$_2$	1	1	1	2	4	12
20 peaks	2	1	2	5	12	26
	3	5	2	5	21	24

We can observe, that both — exact as well as evolutionary — algorithms work quite fast. In most cases we obtain the results before the end of the first minute. This is very important, especially when we recall that, the problem of NOE pathways reconstruction is hardly NP-complete. Fortunately, the NOESY graphs created upon the 2D–NOESY spectra belong to the class of the sparse graphs ($E \ll V^2$, where E is the number of edges, V — a number of vertices), thus, the cardinality of the edge set is rather small, which considerably reduces the time of computations. Computational times of evolutionary method in the worst cases are usually better or similar to these obtained by the exact procedure. However, the time deviations are very small, what drives us to the conclusion that the time of computations plays a peripheral role in solving the problem of NOE pathway assignments with automatic methods.

5. Conclusions

In the paper, we have considered the problem of the reconstruction of NOE pathways in 2D-NOESY spectra of RNA molecules. Basing on the combinatorial model of the problem, we have proposed the evolutionary algorithm and applied it to the collection of spectral data gathered from the NMR experiment for different RNA molecules. During computational experiments we have compared the results obtained by the exact branch-and-cut algorithm and the new evolutionary method. An evolutionary approach gives very good results and for most instances the obtained solution coincides with the prevalent number of vertices in the original NOE path. Evolutionary algorithm appears very useful in the situation of an expert knowledge deficiency. The large number of feasible NOE pathways returned by the exact algorithm for the instances without the additional expert information, makes

it hard to use such an approach in practice. Thus, joining the analysis of the solutions generated by the evolutionary algorithm with the manual method of assignment seems a better idea.

As a continuation of the research reported in this paper, one may consider the analysis of spectra which contain a lot of noise signals. The typical exact method can hardly cope with such cases, especially when an expert is not able to define which cross-peaks should not be considered during the NOE pathway reconstruction. The evolutionary method applied for the noised instances can also facilitate the separation of correct signals and the noised ones, if the cross-peak appearance in the reconstructed solutions is the basis of its evaluation. Experimenting with the reduction of the search space by an elimination of the signals with the least rates can be helpful in considering longer RNA chains.

The steps for improving evolutionary algorithm can be also undertaken. For example, it seems a good idea to try one of the other selection strategies, that have been successfully used for graph problems: tournament selection, elitist recombination etc. Using some other crossover operators, like edge recombination operator or asymmetric edge recombination operator can also improve optimization procedure and result in more precise optimal solutions.

Acknowledgments. The research was partially supported by the grant from the State Committee for Scientific Research, Poland.

References

[1] K. Wüthrich, *NMR of Proteins and Nucleic Acids*, John Willey & Sons, New York, 1986.

[2] J. Cavanach, W. J. Fairbrother, A. G. Palmer III and N. J. Skelton, *Protein NMR Spectroscopy: Principles and Practice*, Academic Press, San Diego, 1996.

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

[4] S. S. Wijmenga and B. N. M van Buuren, "The use of NMR methods for conformational studies of nucleic acids", *Prog. NMR Spectrosc.* 33, 287–387 (1998).

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

[6] H. S. Atreya, S. C. Sahu, K. V. Chary and G. Govil, "A tracked approach for automated NMR assignments in protein (TATAPRO)", *J. Biomol. NMR* 17, 125–36 (2000).

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

[8] J. A. Lukin, A. P. Gove, S. N. Talukdar and C. Ho, "Automated probabilistic method for assigning backbone resonances of (^{13}C,^{15}N)-labeled proteins", *J. Biomol. NMR* 9, 151–166 (1997).

[9] H. N. B. Moseley, D. Monleon and G. T. Montelione, "Automatic determination of protein backbone resonance assignments from triple-resonance NMR data", *Methods in Enzymology* 339, 91–108 (2001).

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

[11] J. P. Linge, M. Habeck, W. Rieping and M. Nilges, "ARIA: automated NOE assignment and NMR structure calculation", *Bioinformatics* 19/2, 315–316 (2003).

[12] M. W. Roggenbuck, T. J. Hyman and P. N. Borer, "Path Analysis in NMR Spectra: Application to an RNA Octamer", *Structure & Methods* 3, 309–317 (1990).

[13] 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, 163–180 (2004).

[14] M. Szachniuk, R. W. Adamiak, P. Formanowicz, Z. Gdaniec, M. Kasprzak, M. Popenda and J. Błażewicz, "A combinatorial analysis of 2D NMR spectra of RNA duplexes", *Curr. Comp. Biol.* 345–346 (2003).

[15] V. J. Rayward-Smith, I. H. Osman, C. R. Reeves and G. D. Smith, *Modern Heuristic Search Methods*, John Willey & Sons, Chichester, 1996.

[16] C. R. Reeves, *Modern Heuristic Techniques for Combinatorial Problems*, McGraw-Hill, London, 1993.

[17] I. H. Osman and J. P. Kelly, *Meta-Heuristics: Theory and Applications*, Kluwer Academic Publishers, Boston, 1995.

[18] E. H. L. Aarts and J. K. Lenstra, *Local Search in Combinatorial Optimization*, John Willey & Sons, Chichester, 1997.

[19] A. Homaifar and S. Guan, *A New Approach to the Traveling Salesman Problem by Genetic Algorithm*, Technical Report, North Carolina A & T State University, 1991.

[20] L. Davis, "Applying adaptive algorithms to epistatic domains", *Proc. of the International Joint Conference on Artificial Intelligence*, 162–164 (1985).

[21] J. A. McDowell and D. H. Turner, "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 (1996).

[22] M. Wu, Jr. J. Santa Lucia and D. H. Turner, "Solution structure of $(rGGCAGGCC)_2$ by 2-D NMR and the iterative relaxation matrix approach", *Biochemistry* 36, 4449–4460 (1997).

[23] J. A. McDowell, L. He, X. Chen and D. H. Turner, "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 (1997).

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