JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 11, Number 1, 2004

Pp. 163–179

An Algorithm for an Automatic NOE Pathways Analysis of 2D NMR Spectra of RNA Duplexes

R.W. ADAMIAK,[1] J. BLAZEWICZ,[1,2] P. FORMANOWICZ,[1,2] Z. GDANIEC,[1]
M. KASPRZAK,[1,2] M. POPENDA,[1] and M. SZACHNIUK[1,2]

ABSTRACT

An algorithm is proposed to provide the tool for an automatic resonance assignment of 2D–NOESY spectra of RNA duplexes. The algorithm, based on a certain subproblem of the Hamiltonian path, reduces a number of possible connections between resonances within aromatic and anomeric region of 2D–NOESY spectra. Appropriate pathways between H6/H8 and H1′ resonances were obtained by subsequent implementation of experimental data as limiting factors. Predictive power of the algorithm was tested on both experimental and simulated data for RNA and DNA duplexes.

Key words: automatic assignments of NMR spectra, 2D–NOESY spectra, RNA duplexes, Hamiltonian paths, branch-and-cut algorithms.

1. INTRODUCTION

NUCLEAR MAGNETIC RESONANCE (NMR) SPECTROSCOPY has been now well established as a method for structure determination of biomolecules in solution (Wüthrich, 1986). The procedure is composed of two general stages: (i) experimental, where multidimensional correlation spectra are acquired, and (ii) computational, where spectra are analysed and structure is determined. Types of NMR experiments differ for proteins (Cavanach *et al.*, 1996) and nucleic acids (Varani and Tinoco Jr., 1991; Wijmenga and van Buuren, 1998). Methods utilizing uniformly ^{13}C- and ^{15}N-labeled proteins and nucleic acids are necessary for studying larger biomolecules (Mollova and Pardi, 2000; Sattler *et al.*, 1999; Varani *et al.*, 1996). The quality and quantity of the experimental data very strongly influence a computational stage. Nevertheless, in all types of NMR structure analysis, the following steps must be accomplished on raw experimental data: processing, peak picking, assignment, restraints determination, structure generation, and refinement.

The procedure assigning the observed signals to the corresponding protons and other nuclei is a bottleneck of the structure elucidation process. For nonlabeled small proteins, as well as short DNA and RNA duplexes, the assignment of NMR signals is usually based on the analysis of two-dimensional (2D) spectra like NOESY, TOCSY, and COSY. For more complex structures, both the usage of uniformly ^{13}C- and ^{15}N-labeled molecules and the application of heteronuclear 3D and 4D spectra are necessary. Due to a considerably large number of signals and their overlapping, the assignment step is troublesome. Therefore, it has been of a great need to facilitate NMR structural analysis of biopolymers by an introduction of

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

automation on this level. At present, automation of NMR spectra analysis makes strong impact on elucidation of protein structures (Moseley and Montelione, 1999). Several programs exist which automatize the process of their signal assignment (Atreya *et al.*, 2000; Leutner *et al.*, 1998; Lukin *et al.*, 1997; Moseley *et al.*, 2001; Zimmerman *et al.*, 1997). 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 the case of nucleic acids, especially RNA. To help experimentators, interactive graphic methods were proposed (Kraulis, 1989).

To our knowledge, only one report exists that concerns an automatic pathway analysis applied for the self-complementary RNA octamer duplex (Roggenbuck *et al.*, 1990). The presented algorithm was based on the reduced adjacency matrix (RAM) and backtracking (BT) procedures. No experimental results, except for one RNA octamer duplex, were reported. The number of alternative paths generated was high.

In this paper, we propose a new algorithm for an automatic generation of pathways between H6/H8 and H1′ resonances observed for short RNA duplexes in a 2D–NOESY spectra. It reduces the NOE pathways analysis to a variant of the Hamiltonian path problem. A proposed combinatorial model takes into account the specificity of the required connectivity between consecutive proton signals in the NMR spectrum. As one can expect, the general problem of finding such a path is NP-hard in the strong sense, thus, unlikely to admit a polynomial time algorithm. Hence, a branch-and-cut algorithm has been proposed, taking into account the combinatorial model and structure-specific aspects of the path generated. A representative set of NMR spectra used for an experimental validation of the algorithm proposed proves its high efficiency and surprisingly good predictive power.

The organization of the paper is as follows. Section 2 discusses the combinatorial model and gives the NP-completeness proof of the problem in question. Section 3 presents the basic algorithm for a reconstruction of the NOE path and some of its refinements. In Section 4, the results of computational experiments are given, while Section 5 points out the directions for further research.

2. COMBINATORIAL MODEL

Our aim is to facilitate the NMR analysis of short RNA duplexes, known also as the helical motifs in ribonucleic acids structure. At the beginning of structural analysis, one knows the sequence of the oligoribonucleotide strand and its potential tendency to form self-complementary duplexes. Identification of the sequence-specific connectivity $H8/H6_{(i)}–H1'_{(i)}–H8/H6_{(i+1)}$ pathway is one of the major steps in the analysis of the 2D–NOESY spectra of right-handed RNA duplexes (Wüthrich, 1986). 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 (from 5′ side) H1′ proton.

Let us consider the $r(CGCGCG)_2$ RNA duplex as an example (Popenda *et al.*, 1997). For clarity, a part of RNA single strand is shown in Fig. 1 (main NOE interactions between protons of our interest are marked with arrows). The 2D–NOESY spectrum of this duplex contains nine characteristic regions of the correlated signals (Fig. 2).

At this stage of study, we focus only on the aromatic/anomeric region of the 2D–NOESY spectra. In the case of $r(CGCGCG)_2$, the NOE connectivity pathway is composed of intranucleotide (higher intensity) and internucleotide (lower intensity) interactions (Fig. 3). They give rise to the alternately appearing cross-peaks. The signals of our interest are located in the H6/H8–H5/H1′ and H5/H1′–H6/H8 regions (rectangles 4 and 8 in Fig. 2). In the spectrum of $r(CGCGCG)_2$, strong H5–H6 cross-peaks of citidine residues are clearly visible (Fig. 3). They can be easily identified from COSY-type spectra, and they do not belong to the path.

In case of the ideal A-RNA duplexes, the NOE pathway starts with the intranucleotide protons interaction (5′ end), and its length equals $2 \cdot n - 1$, where n is the number of residues in RNA chain. Each proton, except for the terminal one, belonging to the pathway gives cross-peaks with two other protons. If the fine structure of a cross-peak is not considered, the cross-peak can be defined as the point with two coordinates 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.

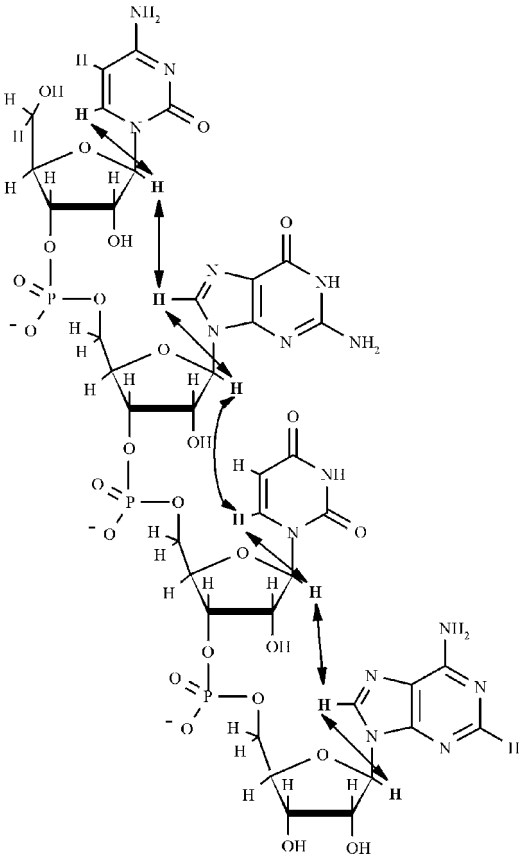

FIG. 1. Main NOE interactions in r(CGUA).

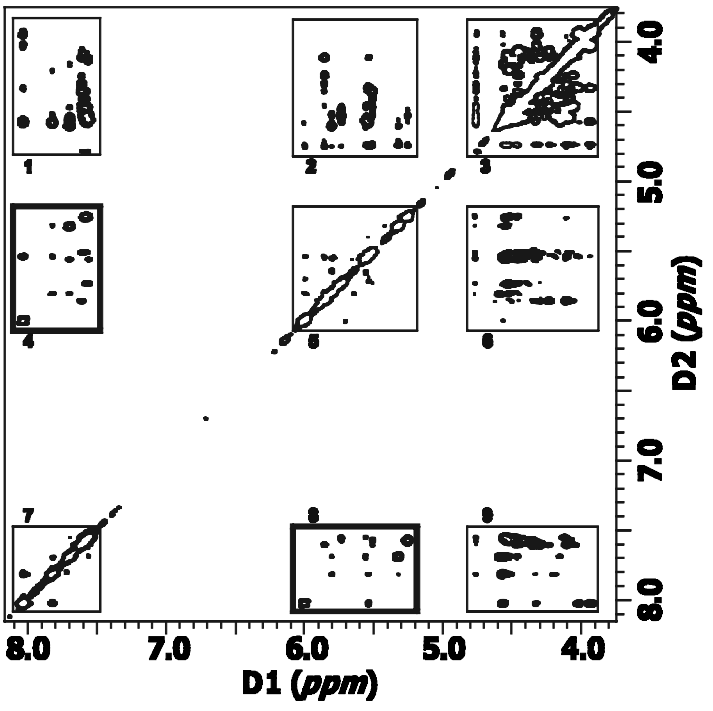

FIG. 2. 2D–NOESY spectrum of r(CGCGCG)$_2$ in D_2O.

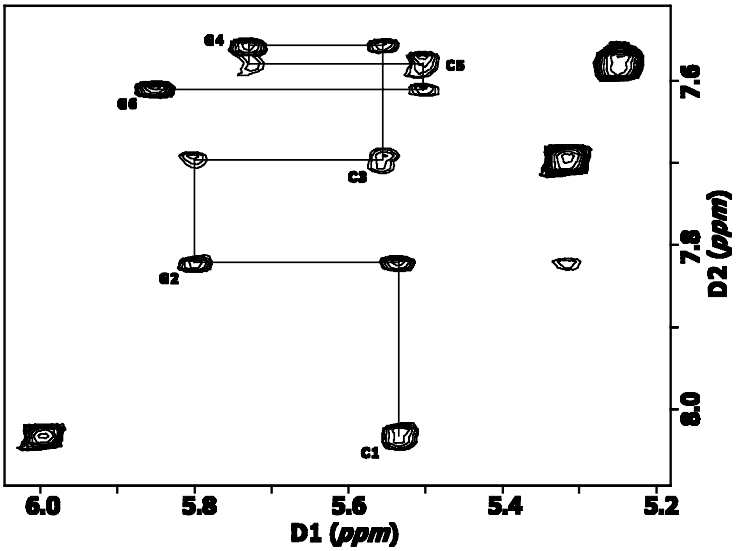

FIG. 3. NOE connectivity pathway in aromatic/anomeric region of the r(CGCGCG)$_2$ given in Fig. 2 as region 8 (Popenda *et al.*, 1997).

With respect to the above description of the problem, we propose its graph-theoretic model that can serve as a background for the complexity analysis and for the construction of the algorithm solving the problem. The process of sequential assignments of H6/H8–H1′ is similar to finding a path between vertices of a graph. Thus, converting the 2D–NOESY spectrum to a certain graph structure seems to be an attractive idea.

We will use undirected graphs $G = (V, E)$ situated on a plane, where V is a set of vertices and E is a set of edges. Because of a strict relationship between graph G and 2D–NOESY spectrum, we call G a *NOESY graph* and we define it in the following way:

1. Every vertex $v \in V$ represents one cross-peak from the spectrum.
2. Vertices are weighted: weight 1 is assigned to every vertex representing intranucleotide NOE, and weight 0 to every vertex representing internucleotide NOE.
3. The number of vertices in a graph equals the number of cross-peaks in the spectrum.
4. Every edge $e \in E$ represents a possible connection between two cross-peaks with different intensities having one common coordinate (thus, graph G includes only edges lying horizontally and vertically).
5. The 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.

Figure 4 shows the relationship between the H6/H8–H1′ region of the 2D–NOESY spectrum and the corresponding NOESY graph obtained according to the above description.

In the above example, there are seventeen cross-peaks in the spectrum. Some of them lay so close to one another (signals 5–6, 11–12 and 15–16–17) that for an inexperienced observer they seem to be the single peaks. However, they are registered as different proton signals by a peak-picking procedure. Thus, we have nine intranucleotide resonances corresponding to nine vertices with weight 1 (big circles) and eight internucleotide resonances represented by eight vertices with weight 0 (small circles) in a graph. All the edges of the graph correspond to all possible proper connections that can be drawn in the spectrum.

The aim of the spectral analysis is finding an H8/H6$_{(i)}$–H1′$_{(i)}$–H8/H6$_{(i+1)}$ pathway in 2D–NOESY spectra of RNA duplexes. Consequently, after spectrum-to-graph conversion, we should define an appropriate path in a graph that could be the corresponding solution of the problem in the theoretical model. The NOE path that is looked for in a NOESY graph may be characterized similarly to the magnetization transfer pathway in a spectrum; that is, 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, and the length of a path

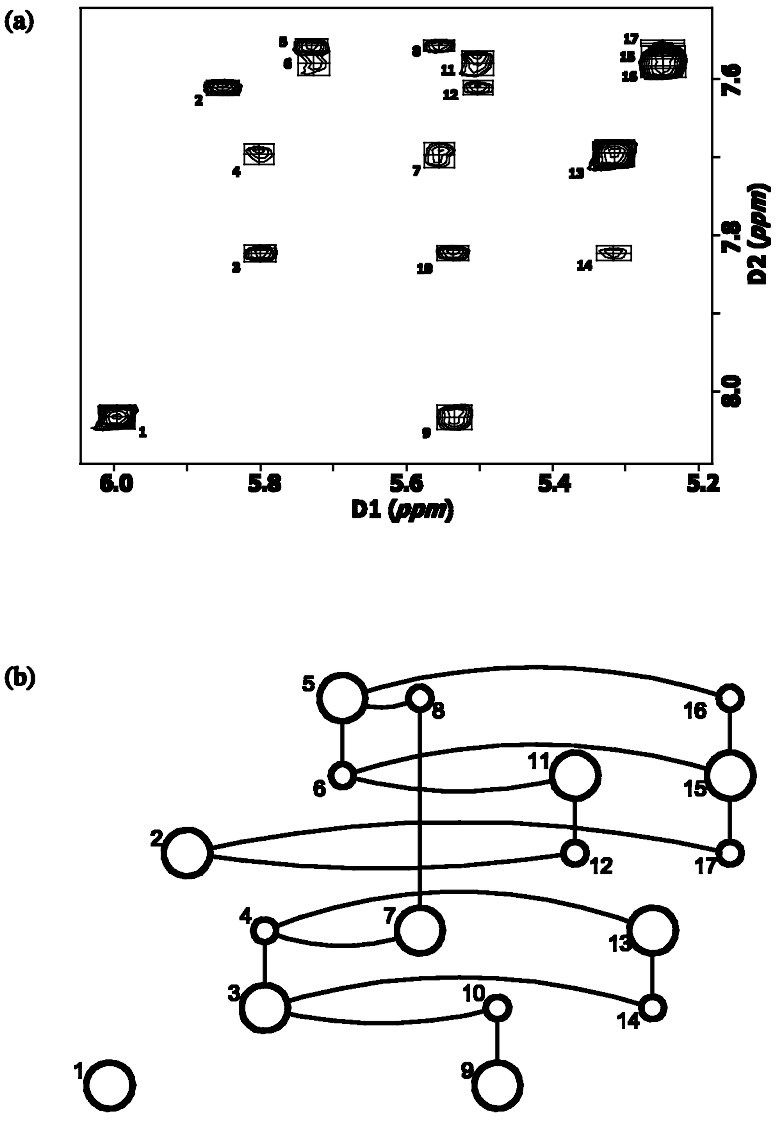

FIG. 4. (**a**) H5/H1′–H8/H6 region of the NOESY spectrum for r(CGCGCG)$_2$, (**b**) NOESY graph corresponding to the spectrum.

equals $2|V_1| - 2$ (here, length is measured as the number of edges in a path), where $|V_1|$ is the number of intranucleotide signals (we assume that all the vertices may occur in the path).

At this point, let us discuss the computational complexity of the NOE path construction in the NOESY graph. This problem will be proved to be strongly NP-hard, and thus computationally intractable.

Theorem 1. *The problem of finding the NOE path in a NOESY graph is strongly NP-hard.*

Proof. First, let us define a decision version of the simplified problem of finding the NOE path which will be proved to be strongly NP-complete. In what follows, we add two conditions to the NOE path definition. Eventually, a definition of the NOE path problem (denoted in its decision version by Π') is the following:

NOE path problem.

Instance: A NOESY graph $G' = (V', E')$: $V' = V_1 \cup V_0$ (V_1–a set of vertices with weight 1, V_0—a set of vertices with weight 0), for every $e_j = \{w_s, w_e\} \in E'$: $w_s \in V_1$, $w_e \in V_0$.

Question: Does G' contain a NOE path that is an ordering $\langle w_1, w_2, \ldots, w_m \rangle$ of the vertices of G', such that $\{w_i, w_{i+1}\} \in E'$ for all i, $1 \leq i < m$ and:

C1. $w_1 \in V_1$,
C2. every two neighboring edges of the path are perpendicular,
C3. the path is simple (every vertex and every edge occurs in the path at most once),
C4. no two edges of the path lie on the same horizontal or vertical line,
C5. $m = 2|V_1| - 2$?

To prove that $\Pi' \in$ NP, one should demonstrate a nondeterministic algorithm solving the problem in polynomial time. The algorithm needs only to guess an ordering of the vertices and check in polynomial time whether all the conditions C1–C5 from the NOE path problem definition are satisfied.

Next, let us take the Hamiltonian path problem as the known strongly NP-complete problem Π (Garey and Johnson, 1979) that will be transformed to our problem Π':

Hamiltonian path problem.

Instance: Graph $G = (V, E)$.

Question: Does G contain a Hamiltonian path that is an ordering $\langle v_1, v_2, \ldots, v_n \rangle$ of the vertices of G, where $n = |V|$, such that $\{v_i, v_{i+1}\} \in E$ for all i, $1 \leq i < n$?

We may assume that graph $G = (V, E)$ has no self-loops and no vertex with degree exceeding three and that the problem remains strongly NP-complete (Garey and Johnson, 1979). Consequently, taking an arbitrary graph $G = (V, E)$, being an instance of the Hamiltonian path problem, we construct NOESY graph $G' = (V', E')$ in the following way:

1. For every vertex $v_i \in V$, place the corresponding vertex $w_i \in V'$ on a plane at the point of coordinates (i, i) and assign to it a weight of 1 (thus, coordinates of every vertex $w_i \in V'$ satisfy the equation $f(x) = x$).
2. For every edge $e_j = (v_p, v_k) \in E$, construct a subgraph as shown in Fig. 5 and add it to graph G'.
3. Assume the following coordinates of the vertices: $w_{jt} = (p, k)$, $w_{jd} = (k, p)$ (let us observe that edges e^1_{jt} and e^2_{jt}, as well as e^1_{jd} and e^2_{jd}, respectively, are perpendicular to each other).
4. Assign weights of 0 to vertices w_{jt} and w_{jd}.

As a result, we obtain the NOESY graph $G' = (V', E')$, where $V' = V \cup \cup_{j=1..|E|}\{w_{jt}, w_{jd}\}$ and $E' = \cup_{j=1..|E|}\{e^1_{jt}, e^2_{jt}, e^1_{jd}, e^2_{jd}\}$.

Figure 6 illustrates construction of a NOESY graph for a given graph being an input of problem Π.

To complete the proof, we need to prove the following proposition:

Proposition 1. *Graph $G = (V, E)$ contains a Hamiltonian path if and only if the corresponding NOESY graph $G' = (V', E')$ contains a NOE path.*

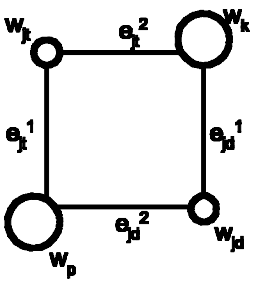

FIG. 5. NOESY subgraph.

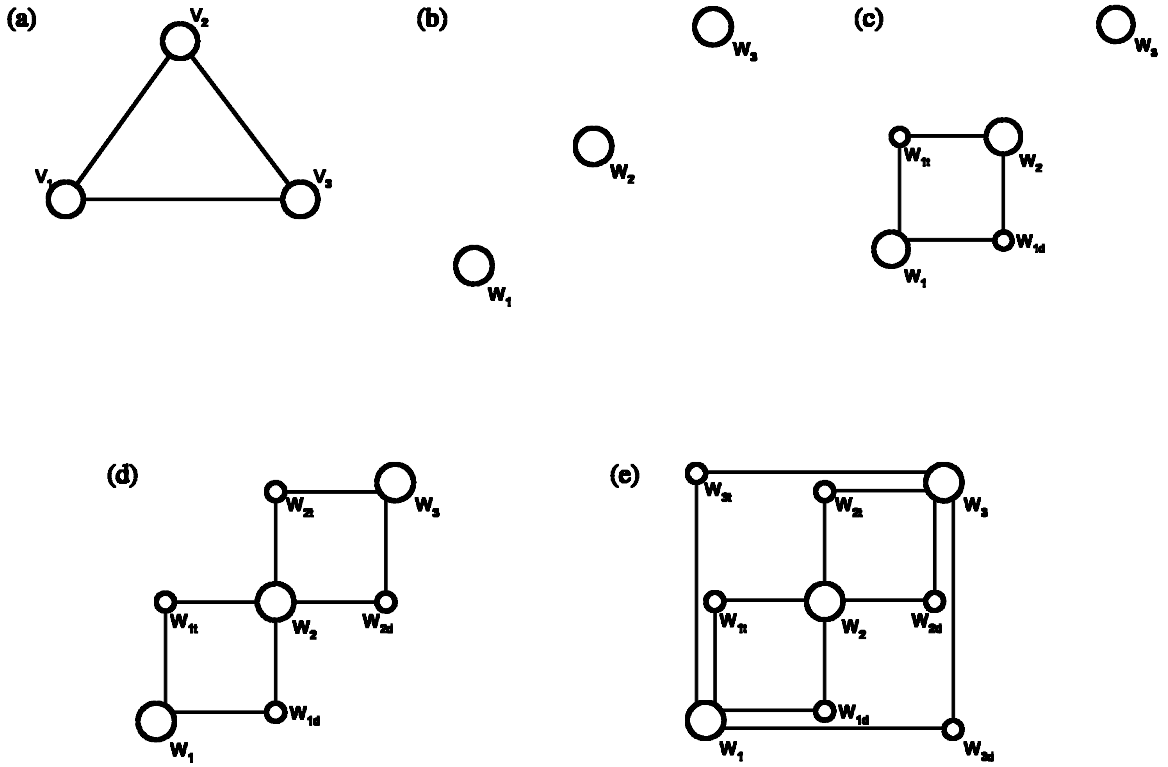

FIG. 6. NOESY graph construction. (**a**) Input base graph. (**b**)–(**e**) Succeeding steps of the construction.

Assume that graph $G = (V, E)$ contains a Hamiltonian path $v_{[1]}, v_{[2]}, \ldots, v_{[n]}$. For this path, we construct the corresponding path $w_{[1]}, w_{[2]}, \ldots, w_{[m]}$ in graph G', which satisfies conditions C1–C5 from the definition of the NOE path problem:

- For every vertex $v_i \in V$ in graph G, there exists exactly one vertex $w_i \in V_1$ in graph G', thus, $w_{[1]} \in V_1$ (condition C1).
- The Hamiltonian path satisfies condition C3; thus, the corresponding path in graph G' also satisfies this condition.
- From the construction of G', it is evident that the NOE path with property C3 must satisfy condition C4.
- The length of the Hamiltonian path (the number of edges in a path) equals $|V| - 1$. For every edge $e_j \in E$, belonging to the Hamiltonian path in graph G, there exists a subgraph in G' consisting of one vertex and two edges which we take to the NOE path. Thus, the NOE path has length $2(|V_1| - 1)$, where $|V_1| = |V|$ (condition C5).
- Constructing the NOE path, we always take the edges perpendicular to each other, which is possible because of G' construction (condition C2).

We notice that if graph G contains a Hamiltonian path, then graph G' contains a NOE path obeying properties C1–C5 from the definition of the NOE path problem.

At this point, assume that graph $G' = (V', E')$ contains a NOE path satisfying conditions C1–C5. For every vertex $w_i \in V_1$ in graph G', there exists exactly one vertex $v_i \in V$ in graph G. Additionally $|V| = |V_1|$. Thus, if in graph G' there exists a NOE path which contains all the vertices $w_i \in V_1$, then graph G contains a path covering all the vertices $v_i \in V$. Moreover, if NOE path in G' satisfies condition C3, then—following a construction of G'—one may say that the corresponding path in graph G also satisfies this condition. Summing up, we may claim that the corresponding path in G is a Hamiltonian path.

We observe that graph G' contains a NOE path if and only if graph G contains a Hamiltonian path. Thus, one can say that Proposition 1 is true and, consequently, Theorem 1 is also true because the time used for a construction of G' is bounded from the above by the input length of problem Π. ■

It should be clear that the above result implies also strong NP-hardness of the primary version of the NOE path construction problem. Hence, no polynomial-time exact algorithm is likely to exist for this problem. As a result, a new algorithm for solving it will be proposed in the next section.

3. BASIC ALGORITHM AND ITS VARIANTS

In this Section, we introduce a branch-and-cut algorithm that automatically groups H6/H8–H1′ cross-peaks of the nucleotide residues according to their position in the sequence. The algorithm is based on a Hamiltonian path construction procedure and uses domain expert knowledge to introduce additional constraints that limit the search space to the reasonable proportions. It has been implemented in C programming language and runs in a Unix environment.

The number of NOE paths and their lengths depend on RNA tertiary structure and signal overlapping. Computational analysis has shown that the number of all NOE paths in the NOESY graph reaches $2^{-(n-3)} \cdot n!$ for $n > 2$, where n is the number of graph vertices. Thus, there may be several solutions that satisfy NOE path conditions (C1–C5) and we should find them all. However, only a few of these paths are correct from the biochemical point of view, and the process of finding them relies on additional information that should be specified. Consequently, the algorithm should look through the whole search space and indicate correct paths only.

Before we start a presentation of the algorithm, let us describe the input data. The input is a text file $*.list$ generated by Accelrys FELIX software from the 2D–NOESY spectrum after a peak-picking procedure. The file contains the following information about each cross-peak: its number (*No*), two coordinates of the cross-peak ($D1$, $D2$) in ppm or Hz, its volume (*Vol*), and the widths in both dimensions ($dD1$, $dD2$) given in Hz. Additionally, the first line of $*.list$ file includes the spectrometer frequency, which is helpful in converting units (ppm to Hz). Figure 7 illustrates an example of input $*.list$ file for the considered H5/H1′–H8/H6 region of r(CGCGCG)$_2$ presented in Fig. 4a.

Additional information about the spectrum and NOE path that contains the domain expert knowledge and is consequently used to extract correct paths is placed in the second input file $*.inf$. This file is divided into several sections which may be empty or may contain the following information:

- In section ⟨VOLUMES⟩, a user can define intervals to differentiate inter- and intranucleotide cross-peaks volumes.
- Section ⟨RESOLUTION⟩ may contain the value of divergence [ppm] which depends on the digital resolution of a spectrum in both dimensions. If this parameter is given, then the cross-peaks coordinates are deviated within the given range.
- Section ⟨OVERLAPPING⟩ is filled if the lower and upper limits of the interval with overlapping signals are given.
- In section ⟨DOUBLETS⟩, one can define the distance between cross-peaks which should be interpreted as doublets.
- Section ⟨REJECT_SIGNALS⟩ contains coordinates D1, D2 of the cross-peaks which should not be considered during path construction.
- Section ⟨RNA_SEQUENCE⟩ includes the sequence of RNA (both strands in the case of non-self-complementary duplexes).
- In section ⟨PATH_LENGTH⟩, a number of cross-peaks in the expected NOE path can be defined.
- Information about cross-peaks which might be treated as starting points in the path is placed in section ⟨START_POINTS⟩.
- Section ⟨KNOWN_SIGNALS⟩ includes additional information about the cross-peaks which might help in arranging the path.
- In section ⟨H5–H6_SIGNALS⟩, a user can specify cross-peaks which can be easily identified as H5–H6 cross-peaks and, therefore, they are not taken to the final path.

Information given in the sections ⟨VOLUMES⟩ through ⟨RNA_SEQUENCE⟩ helps in making more accurate interpretation of the cross-peaks described in $*.list$ file, while this from sections ⟨PATH_LENGTH⟩ through ⟨H5–H6_SIGNALS⟩ allows reduction of the number of potential paths in the solution set.

No	D1 [ppm]	D2 [ppm]	Vol [W]	dD1 [Hz]	dD2 [Hz]	500
1	6.00	8.03	1.054	16.0	16.0	
2	5.85	7.61	0.169	9.0	7.0	
3	5.80	7.82	0.094	9.0	7.0	
4	5.80	7.72	0.042	9.0	16.0	
5	5.73	7.56	0.100	9.0	7.0	
6	5.73	7.58	0.044	9.0	16.0	
7	5.55	7.72	0.092	9.0	16.0	
8	5.55	7.56	0.049	9.0	7.0	
9	5.53	8.03	0.145	9.0	16.0	
10	5.53	7.82	0.045	9.0	7.0	
11	5.50	7.58	0.117	9.0	16.0	
12	5.50	7.61	0.051	9.0	7.0	
13	5.31	7.72	0.905	16.0	16.0	
14	5.31	7.82	0.030	16.0	7.0	
15	5.25	7.58	1.041	16.0	16.0	
16	5.25	7.56	0.037	16.0	7.0	
17	5.25	7.61	0.025	16.0	7.0	

FIG. 7. Input *rcgcgcg.list* file for region H5/H1′–H8/H6 of 2D–NOESY of r(CGCGCG)$_2$.

The proposed algorithm builds NOE paths from a chosen vertex adding one edge at a time. It looks through the search space adding edges recursively until there is no other edge that can be added. Then, the current path is verified according to the expert knowledge given in $*.inf$ file. Afterwards, the algorithm goes back, removing the edges from the path, and tries to add the other edges in place of the removed ones. The main procedure given in pseudo-code is the following:

Algorithm 1

1. **read** input files: ⟨name.list⟩ and ⟨name.inf⟩;
2. construct a set of vertices;
3. remove signals enumerated in section ⟨REJECT_SIGNALS⟩ from the set of vertices;
4. find all correct edges that can be created upon the given set of vertices;
5. **for** $i := 0$ **to** ⟨number of edges⟩ **do**
6. **begin**
7. empty the stack with current solution;
8. take the i-th edge from the set of edges;

```
 9.      if (((section ⟨START_POINTS⟩ is defined) and (the first vertex of the i-th edge is starting)) or
           (section ⟨START_POINTS⟩ is not defined))
10.      then begin
11.             put both vertices of the i-th edge on the stack with current solution;
12.             find a path starting from the second vertex of the i-th edge;
13.           end;
14.   end;
15. return ⟨set of solutions⟩;
```

The procedure that finds the path starting from the second vertex of the k-th edge (step 12 in Algorithm 1), given in pseudo-code, is the following:

Algorithm 2

```
 1. for i := 0 to ⟨number of edges⟩ do
 2.   begin
 3.     take the i-th edge from the set of edges;
 4.     if (the i-th edge does not yet belong to the current solution)
 5.     then
 6.          if ((the second vertex of the k-th edge = the first vertex of the i-th edge)
                and (the k-th edge is perpendicular to the i-th edge))
 7.        then begin
 8.             put the second vertex of the i-th edge on the stack with current solution;
 9.             find a path starting from the second vertex of the i-th edge; //recursion
10.             remove the last vertex from the stack with current solution;
11.           end;
12.   end;
13. if (the current solution is correct)
14. then add current solution to the set of solutions;
```

In the first step, Algorithm 1 reads the input files, rejects the signals that should not be considered, and creates all correct edges upon the modified set of vertices (cross-peaks). An edge is correct if it is horizontal or vertical and connects two vertices of different volumes (inter–intra). Connections are created according to the appropriate data describing cross-peaks. Additionally, if the resolution is defined, Algorithm 1 deviates the values of cross-peaks coordinates within the error range and more edges may be found. These deviated edges are not always "strictly" horizontal or vertical, but they are interpreted as if they were. If a distance for doublets is defined, Algorithm 1 finds all doublets in the set of vertices and converts them to single vertices. Then, of course, the edges are created upon the new, processed set of vertices. Next, Algorithm 1 takes every single edge from the created set and tries to build a path starting with the first vertex of this edge. If the starting vertices are defined, Algorithm 1 verifies the edge and accepts it only if its first vertex is specified as the starting one. During the search, the algorithm verifies path consistency with known signals placements and path length if these data are available. Finally, a set of correct solutions is returned.

Let us now examine the example given in Figs. 2 through 4 and see how Algorithm 1 works for r(CGCGCG)$_2$. The 2D–NOESY spectrum for this RNA duplex is shown in Fig. 2, while Fig. 4a illustrates region H5/H1′–H8/H6 of the spectrum. An appropriate text file *rcgcgcg.list* with spectral information is listed in Fig. 7. Additionally, having some expert knowledge, we decided to define it as additional information in file *rcgcgcg.inf*. There, volume intervals are defined as 0.035–0.2, which makes Algorithm 1 reject four cross-peaks numbered 1, 13, 15, 17, as we know they should not be taken into the path (instead of this, one may also specify these four cross-peaks in section ⟨REJECT_SIGNALS⟩). The difference between inter- and intranucleotide signals was hard to specify; thus, the intervals for both sets are equal. We also defined the length of the NOE path which should consist of 11 peaks. Finally, H5–H6 cross-peaks were specified, and the RNA sequence was given. This instructs Algorithm 1 to accept the paths consistent with the primary sequence, so that the peaks corresponding to citidine (these are the 1st, 5th and

9^{th} signals in the path) should have the same value of D2 coordinate as cross-peaks specified in section ⟨H5–H6_SIGNALS⟩. The steps taken by the algorithm are as follows.

No distance for doublets was given, so the set of vertices remained. Next, all possible edges were created: {(2,12), (12,2), (3,4), (4,3), (3,10), (10,3), (4,7), (7,4), (5,6), (6,5), (5,8), (8,5), (5,16), (16,5), (6,11), (11,6), (7,8), (8,7), (8,16), (16,8), (9,10), (10,9), (11,12), (12,11)}. It is important to remember that an RNA chain has two different endings (3′, 5′); thus, every possible connection is treated as two edges with opposite senses. Consequently, for every edge in the created set there exists the opposite one, and the number of edges is always even. Afterwards, the procedure started searching for correct paths and found six of them:

1: 2 12 11 6 5 8 7 4 3 10 9
2: 2 12 11 6 5 16
3: 9 10 3 4 7 8 5 6 11 12 2
4: 9 10 3 4 7 8 16
5: 16 5 6 11 12 2
6: 16 8 7 4 3 10 9

After that, a verifying procedure rejected all paths that consisted of fewer than 11 cross-peaks (the longer paths could not be found because the algorithm stops searching in the current direction if the path achieved the defined length), and two NOE paths were left:

1: 2 12 11 6 5 8 7 4 3 10 9
3: 9 10 3 4 7 8 5 6 11 12 2

One can notice that the above paths are symmetrical, so only one of them is correct from a biochemical point of view. The information about H5–H6 cross-peaks given in the *rcgcgcg.inf* file can help to choose the right NOE path. Thus, Algorithm 1 verifies path consistency with the RNA sequence and finds out whether citidine signals have the same D2 coordinate as cross-peaks specified in section ⟨H5–H6_SIGNALS⟩. It appears that only the second path is consistent, so it is returned as the only solution of our instance:

3: 9 10 3 4 7 8 5 6 11 12 2

Figure 3 illustrates the above path drawn in region H5/H1′–H8/H6 of the 2D–NOESY spectrum for r(CGCGCG)$_2$. The three biggest cross-peaks in this spectrum (with numbers 1, 13, 15) are the ones enumerated in section ⟨H5–H6_SIGNALS⟩ of the *rcgcgcg.inf* file, and one can see that they have the same D2 coordinates as citidine signals, respectively: 9,7,11 within the NOE path.

The solutions in the simple form, i.e., arrangement of the vertices (like in the above example) are written to file *paths.out.* Additionally, the program creates detailed assignment files with solutions. Figure 8 shows such a file for the analyzed example.

4. EXPERIMENTAL RESULTS

The algorithm was tested on an Indigo 2 Silion Graphics workstation (1,133MHz, 64MB) in an IRIX 6.5 environment. As a testing set, a group of experimental and simulated 2D–NOESY spectra was prepared. The 2D–NOESY spectra of r(CGCGCG)$_2$, 2′–O–Me(CGCGCG)$_2$ and r(CGCG^{F}CG)^2 in D_2O at 30°C were recorded on a Varian Unity+ 500 MHz spectrometer. A standard pulse sequence (Jeener *et al.*, 1979) $\pi/2$-t_1-$\pi/2$-τ_m-$\pi/2$-t_2 was applied with mixing time $\tau_m = 150$ 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 the t_1 dimension and a base correction in t_2, data were collected in $1K^{x}1K$ matrixes with final digital resolution of 3.5Hz/point in both dimensions.

The 2D–NOESY, DQF–COSY, and HSQC spectra of d(GACTAGTC)$_2$ were acquired on a Bruker AVANCE 600 MHz spectrometer. The analysed 2D–NOESY spectrum were recorded with mixing time $\tau_m = 400$ms, 1K real points in t_1, 1K complex points in t_2, and spectral width of 6.0kHz in both dimensions. After processing, the final digital resolution was equal to 6Hz/points in both dimensions.

No	D1 [ppm]	D2 [ppm]	Vol [W]	dD1 [Hz]	dD2 [Hz]	500		
1	6.00	8.03	1.054	16.0	16.0		none	none
2	5.85	7.61	0.169	9.0	7.0		G_6:H1'	G_6:H8
3	5.80	7.82	0.094	9.0	7.0		G_2:H1'	G_2:H8
4	5.80	7.72	0.042	9.0	16.0		G_2:H1'	C_3:H6
5	5.73	7.56	0.100	9.0	7.0		G_4:H1	G_4:H8
6	5.73	7.58	0.044	9.0	16.0		G_4:H1'	C_5:H6
7	5.55	7.72	0.092	9.0	16.0		C_3:H1'	C_3:H6
8	5.55	7.56	0.049	9.0	7.0		C_3:H1'	G_4:H8
9	5.53	8.03	0.145	9.0	16.0		C_1:H1'	C_1:H6
10	5.53	7.82	0.045	9.0	7.0		C_1:H1'	G_2:H8
11	5.50	7.58	0.117	9.0	16.0		C_5:H1'	C_5:H6
12	5.50	7.61	0.051	9.0	7.0		C_5:H1'	G_6:H8
13	5.31	7.72	0.905	16.0	16.0		none	none
14	5.31	7.82	0.030	16.0	7.0		none	none
15	5.25	7.58	1.041	16.0	16.0		none	none
16	5.25	7.56	0.037	16.0	7.0		none	none
17	5.25	7.61	0.025	16.0	7.0		none	none

FIG. 8. Output *path.details* file for r(CGCGCG)$_2$.

The spectra of r(GGCAGGCC)$_2$, r(GAGGUCUC)$_2$, r(GGCGAGCC)$_2$, and r(GCAGUGGC).r(GCCA) d(CTGC) were simulated using the Matrix Doubling method of FELIX software based on published ^{1}H chemical shifts (McDowell and Turner, 1996; SantaLucia Jr. and Turner, 1993; Szyperski *et al.*, 1999; Wu *et al.*, 1997) and three dimensional structures from the Protein Data Bank. Volumes of NOE cross-peaks for $\tau_m = 0.3$ms were calculated from the Full Relaxation Matrix, where a correlation time was set to 2ns. 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 (Lankhorst *et al.*, 1984; Haasnoot *et al.*, 1980).

To perform tests, numeric data were obtained from experimental and simulated spectra after the pick-peaking procedure (FELIX Accelrys).

All the instances had been already solved manually, so we could verify whether or not the algorithm found correct solutions. It was also possible to examine the way expert knowledge influences qualifying correct solutions and building the final solution set. A minimal expert knowledge was used in every example. Let us notice that in some cases such knowledge is necessary for an appropriate interpretation of the input data. Without additional information, the algorithm cannot find the correct solution in the spectrum enclosing

doublets or overlapping signals or in the case when spectrum resolution should be considered. Apart from the last test, all RNAs and DNAs formed self-complementary chains; thus, one pathway, being correct from biochemical point of view, existed for each of them. The last case—r(GCAGUGGC).r(GCCA)d(CTGC) structure—is the only non-self-complementary duplex tested; thus, two correct NOE pathways were found. Table 1 summarizes experimental results of Algorithm 1 tested on the above instances.

For some more complex cases, the aromatic/anomeric regions of the 2D–NOESY spectra and the correct NOE pathways calculated by Algorithm 1 are shown in Figure 9.

Analyzing the obtained results, we notice that Algorithm 1 constructed a surprisingly small number of alternative pathways in each case, thus, proving its high accuracy. On the other hand, we find the

TABLE 1. RESULTS OF TESTS

Test	*RNA/DNA duplexes and region*	*Size of an instance*	*Additional information (*.inf) based on the expert knowledge*	*Number of paths found by Algorithm 1*	*Computation time [s]*
1	r(CGCGCG)$_2$ region: H5/H1′-H8/H6/H2	17 crosspeaks	—#5 rejected signals; —RNA sequence; —path length; —#3 H5-H6 signals;	1 (1 correct)	0.01
2	2′-OMe(CGCGCG)$_2$ region: H5/H1′-H8/H6/H2 (Fig. 9a)	17 crosspeaks	—volume intervals; —interval with overlapping signals; —RNA sequence; —path length; —#3 H5-H6 signals;	2 (1 correct, Fig. 9b)	0.005
3	r(CGCG^{F}CG)$_2$ region: H8/H6/H2-H5/H1′	15 cross peaks	—#2 rejected signals; —resolution; —RNA sequence; —path length; —#2 H5-H6 signals;	1 (1 correct)	0.01
4	r(CGCG^{F}CG)$_2$ region: H5/H1′-H8/H6/H2	22 crosspeaks	—volume intervals; —distance between doublets; —resolution; —RNA sequence; —path length;	2 (1 correct)	0.01
5	d(GACTAGTC)$_2$ region: H8/H6/H2-H5/H1′ (Fig. 9c)	24 crosspeaks	—#8 rejected signals; —interval with overlapping signals; —DNA sequence; —path length; —#4 H5-H6 signals;	2 (1 correct, Fig. 9d)	0.03
6	d(GACTAGTC)$_2$ region: H5/H1′-H8/H6/H2	26 crosspeaks	—#7 rejected signals; —DNA sequence; —path length; —#4 H5-H6 signals;	6 (1 correct)	0.03
7	r(GGCAGGCC)$_2$ region: H5/H1′-H8/H6/H2 (Fig. 9e)	26 crosspeaks	—#8 rejected signals; —RNA sequence; —path length; —#5 H5-H6 signals;	2 (1 correct, Fig. 9f)	0.02
8	r(GAGGUCUC)$_2$ region: H5/H1′-H8/H6/H2	24 crosspeaks	—#6 rejected signals; —RNA sequence; —path length; —#4 H5-H6 signals;	1 (1 correct)	0.03
9	r(GGCGAGCC)$_2$ region: H8/H6/H2-H5/H1′ (Fig. 9g)	20 crosspeaks	—#5 rejected signals; —RNA sequence; —path length (broken chain);	4 (1 correct, Fig. 9h)	0.03
10	r(GCAGUGGC). r(GCCA)d(CTGC) region: H5/H1′-H8/H6/H2 (Fig. 9i)	55 crosspeaks	—volume intervals; —#7 rejected signals; —RNA sequence; —path length; —#7 H5-H6 signals	6 (2 correct, Fig. 9j)	0.06

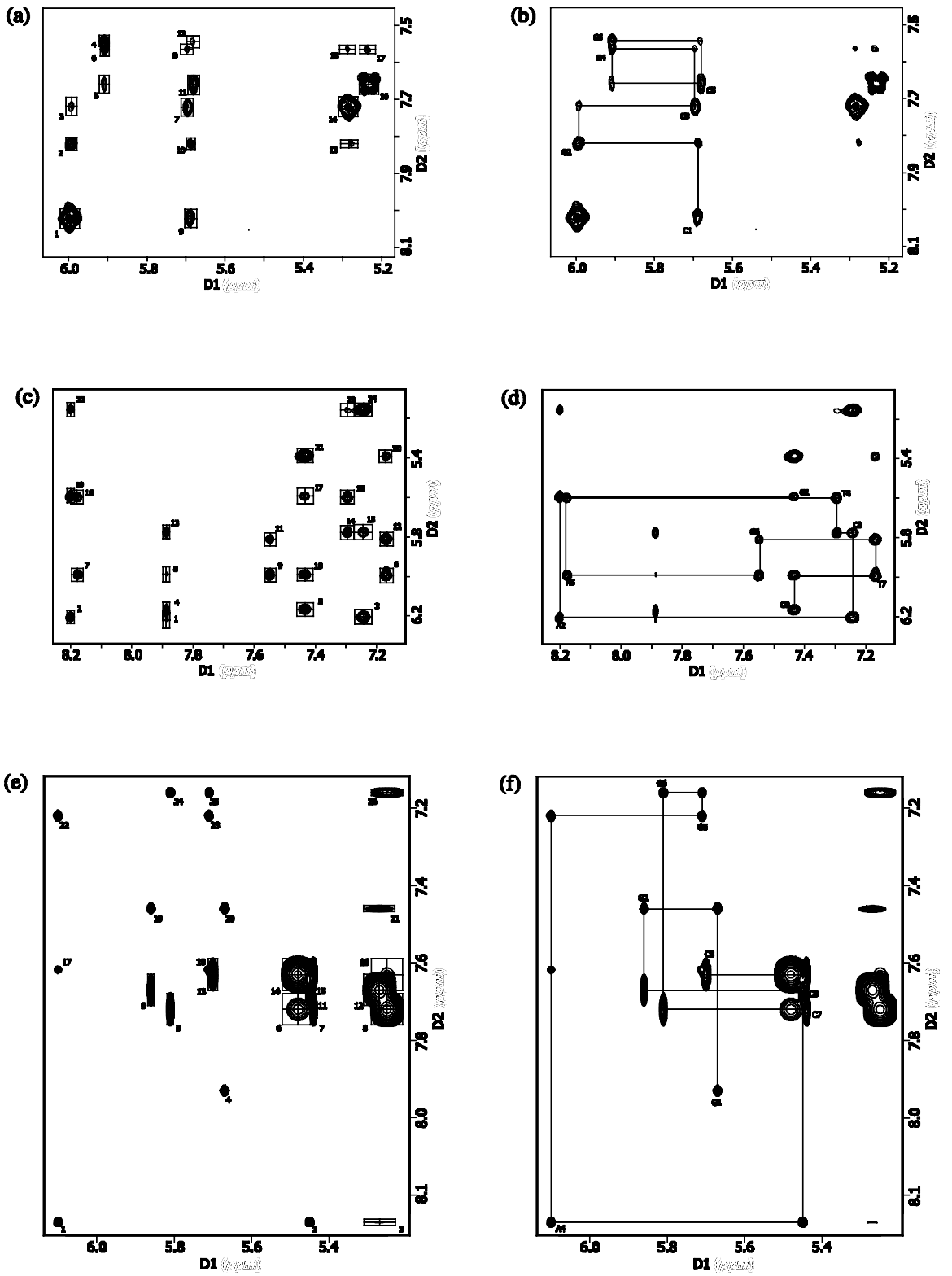

FIG. 9. (**a**) 2′-OMe(CGCGCG)$_2$ spectrum. (**b**) 2′-OMe(CGCGCG)$_2$ NOE pathway. (**c**) d(GACTAGTC)$_2$ spectrum. (**d**) d(GACTAGTC)$_2$-NOE pathway. (**e**) r(GGCAGGCC)$_2$ spectrum. (**f**) r(GGCAGGCC)$_2$-NOE pathway.

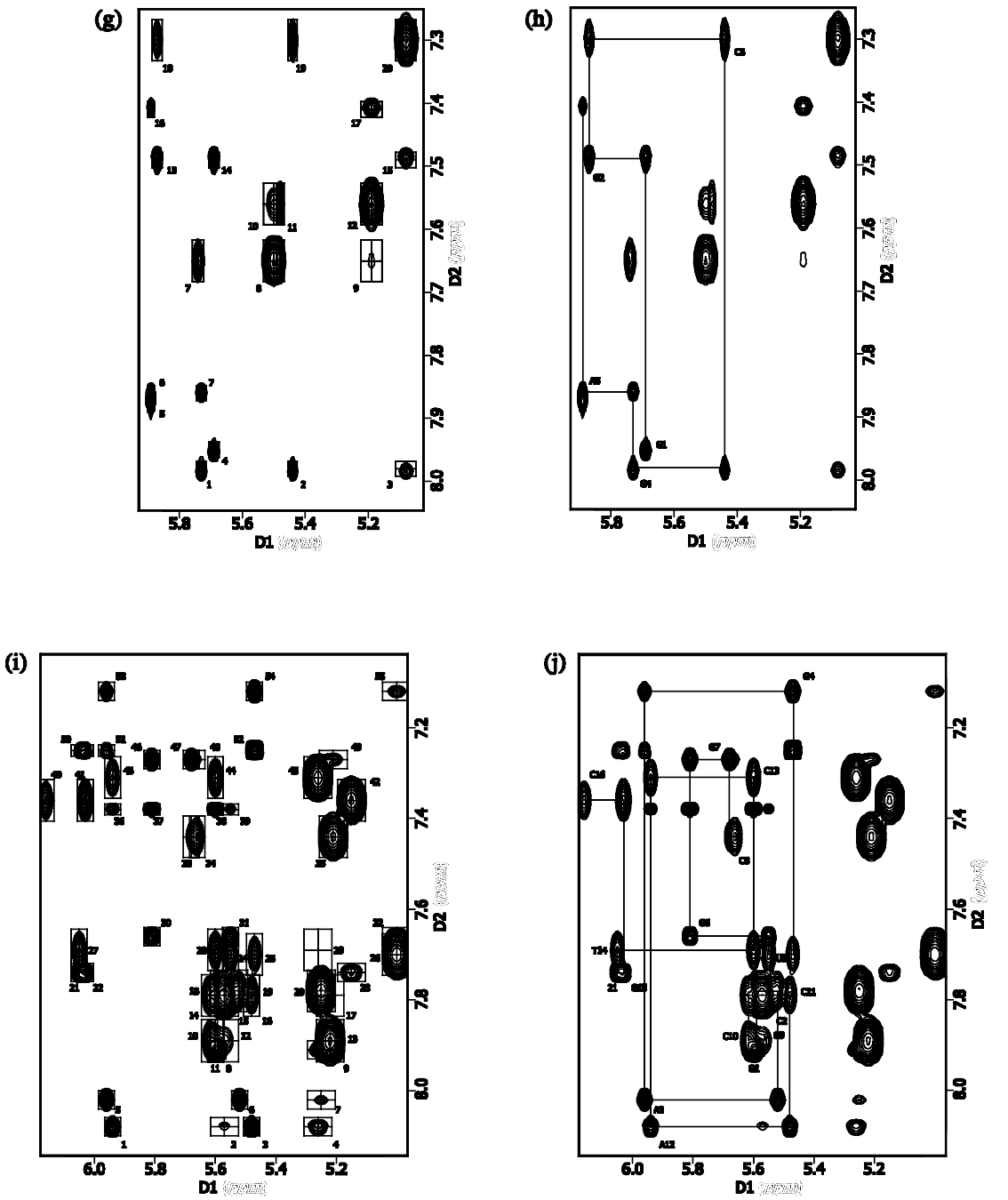

FIG. 9. (*Continued*) (**g**) r(GGCGAGCC)$_2$ spectrum. (**h**) r(GGCGAGCC)$_2$-NOE pathway. (**i**) r(GCAGUGGC). r(GCCA)d(CTGC) spectrum. (**j**) r(GCAGUGGC).r(GCCA)d(CTGC)-NOE pathways.

algorithm quite fast, despite its computational complexity which equals $O(m^m)$, where m is the number of graph edges. Detailed analysis of the NOESY graphs allows us to observe that they belong to the class of sparse graphs. Thus, the cardinality of the edge set is rather small, which considerably reduces the time of computations.

5. CONCLUSIONS

In this paper, the problem of automatic resonance assignment of 2D–NOESY NMR spectra of RNA duplexes has been considered, and its combinatorial model has been proposed. Since the basic problem has been proved to be strongly NP-hard, a branch-and-cut algorithm has been presented. This algorithm gives very good results when some expert knowledge is available. Note that even a small amount of information about the analyzed chain results in a significant reduction of the final solution set.

Thus far, the assignment of cross-peaks in the 2D NOESY spectra of nucleic acids was accomplished by hand with the help of interactive graphics. This manual assignment of NOE resonances is very tedious and time consuming due to the large number of cross-peaks present in the NOESY spectra of biomolecules and a possibly large number of existing alternative pathways. Thus, any tool that can facilitate this analysis is of great importance. On the other hand, the algorithm proposed here might be very useful when applied to a verification of the assignment correctness.

As a continuation of the research reported in this paper, one may consider the 3D NMR spectra analysis. They represent a wider range of interactions than their 2D equivalents. Thus, they carry more information about the structure and allow precise determination of input samples characteristics. Furthermore, it seems evident that 3D and finally XD (X > 3) NMR spectra analysis will be considered in the continuation of our research. Solving the problem of finding a NOE path on the basis of 2D–NOESY, an NMR spectrum appears to be a good platform for this purpose. As it was demonstrated in Section 2, however, the problem of finding NOE paths in 2D spectra already has been troublesome. Consequently, we should expect that adding one or more dimensions into the search space will complicate the searching algorithm.

ACKNOWLEDGMENTS

This research is supported by grants 7T11F02621 and 7T09A09720 from the State Committee for Scientific Research, Poland. The authors are grateful to the anonymous referee for his helpful remarks leading to a better presentation of the paper.

REFERENCES

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

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

Garey, M.R., and Johnson, D.S. 1979. *Computers and Intractability. A Guide to the Theory of NP-Completeness.* W.H. Freeman, San Francisco.

Haasnoot, C.A.G, de Leeuw, F.A.A.M., and Altona, C. 1980. The relationship between proton-proton NMR coupling constants and substituent electronegativities—I. *Tetrahedron Lett.* 36, 2783–2792.

Jeener, J., Meier, B.H., Bachmann, P., and Ernst, R.R. 1979. Investigation of exchange processes by 2-D NMR spectroscopy. *J. Chem. Phys.* 71, 4546–4593.

Kraulis, P.J. 1989. ANSIG: A program for the assignment of protein ^{1}H 2D NMR spectra by interactive graphics. *J. Magn. Reson.* 24, 627–633.

Lankhorst, P.P, Haasnoot, C.A.G., Erkelens, C., and Altona, C. 1984. Carbon-13 NMR in conformational analysis of nucleic acid fragment. *J. Biomol. Struct. Dyn.* 1, 1387–1405.

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

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

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.

Mollova, E.T., and Pardi, A. 2000. NMR solution structure determination of RNAs. *Curr. Opin. Struct. Biol.* 10, 298–302.

Moseley, H.N.B., Monleon, D., and Montelione, G.T. 2001. Automatic determination of protein backbone resonance assignments from triple-resonance NMR data. *Methods Enzymol.* 339, 91–108.

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.

Popenda, M., Biala, E., Milecki, J., and Adamiak, R. 1997. Solution structure of RNA duplexes containing alternating CG base pairs: NMR study of $r(CGCGCG)_2$ and $2'$–O–$Me(CGCGCG)_2$ under low salt conditions. *Nucl. Acids Res.* 25, 4589–4598.

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

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

Sattler, M., Schleucher, J., and Griesinger, C. 1999. Heteronuclear multidimensional NMR experiments for the structure determination of proteins in solution employing pulsed field gradients. *Prog. NMR Spectrosc.* 34, 93–158.

Szyperski, T., Goette, M., Billeter, M., Perola, E., Cellai, L., Heumann, H., and Wüthrich, K. 1999. NMR structure of the chimeric hybrid duplex R(GCAGUGGC).R(GCCA)D(CTGC) comprising the tRNA-DNA junction formed during initiation of HIV-1 reverse transcription. *J. Biomol. NMR* 13, 343–355.

Varani, G., Aboul-ela, F., and Allain, F.H.T. 1996. NMR Investigation of RNA Structure. *Prog. NMR Spectrosc.* 29, 51–127.

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

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

Wu, M., SantaLucia Jr., J., and Turner, D.H. 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 Wiley, New York.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C.-Y., Powers, R., and Montelione, G.T. 1997. Automated Analysis of Protein NMR Assignments Using Methods from Artificial Intelligence. *J. Mol. Biol.* 269, 592–610.

Address correspondence to:
Marta Szachniuk
Institute of Computing Science
Poznan University of Technology
Piotrowo 3a
60-965 Poznan, Poland

E-mail: Marta.Szachniuk@cs.put.poznan.pl