Combinatorial analysis of 2D–NOESY spectra in Nuclear Magnetic Resonance spectroscopy of RNA molecules

Summary of Ph.D. thesis

Marta Szachniuk

Poznan University of Technology, Piotrowo 2, 60-965 Poznan, Poland

Introduction

Many phenomena in the animate and inanimate realms of nature can be traced back to molecular processes. Thus, cognition of organism structure as well as exploring functions, dependencies and processes on a molecular level have been, from years, one of the most fundamental tasks in many different research areas. Recognition of biomolecule structures has become possible due to the development of several analytical methods. A quick spread of NMR spectroscopy in the last decades of the 20th century has resulted in a superiority of this method over others, as far as a structure determination of biomolecules in solution is concerned. This method appeared a very good choice for studying tertiary structure of RNA molecules - the least known of the biomolecules and hard to be analyzed with other methods.

NMR procedure of structure determination is composed of two general stages: experimental, where multidimensional correlation spectra are acquired and computational, where spectra are analyzed and spatial structure is determined. In all methods of NMR structure analysis experimental data are processed by the procedures of peak-picking, assignment, restraints determination, structure generation and refinement, respectively. An assignment of the observed NMR signals to the corresponding protons and other nuclei is a bottleneck 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 spectra resulting from NMR experiments. For short DNA and RNA duplexes it is performed manually in accordance with the experimenter's knowledge and intuition. However, for the longer nucleic chains the assignment step becomes troublesome, due to a large number of signals and their overlapping in the spectrum. Therefore, *it has been necessary to facilitate NMR structural analysis of biopolymers by an introduction of automatic procedures at this level*.

Scope of the thesis

The goal of the research, described in the thesis, may be stated as a g*eneration of the new automatic methods of nuclei assignment, dedicated to RNA molecules*. An assignment process starts from the construction of NOE (Nuclear Overhauser Effect) pathway in 2D–NOESY (Nuclear Overhauser Enhancement SpectroscopY) spectrum resulting from NMR experiment. The NOE peaks illustrated in the 2D–NOESY spectrum and representing correlation signals are connected to form the path, called the NOE pathway. The pathway shows a transfer of magnetization between the following pairs of protons within RNA chain: H6 (of pyrimidine residues) and H1' (of ribose), or H8 (of purine residues) and H1' (of ribose). After finding a pathway, each of its cross-peaks is assigned to an appropriate pair of protons, which generated the signal.

In the thesis, *a new combinatorial model* of an automatic generation of pathways between H6/H8 and H1' resonances observed for RNA duplexes in a 2D–NOESY spectra, is proposed. As a result, the NOE pathways analysis is *reduced to a variant of the Hamiltonian path problem*. The 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 proved to be NP–hard in the strong sense*, thus, unlikely to admit a polynomial time algorithm. Hence, *two metaheuristics algorithms, based on tabu search and genetic procedures*, are proposed. Both take into account the combinatorial model and structure-specific aspects of the path generated. Their performance is compared to the third, *enumerative algorithm* also proposed and described in the thesis. A representative set of NMR spectra used for an experimental validation of the algorithms proposed proves high efficiency of the methods.

Theoretical part of the research

One of the first analytical step in the tertiary structure determination using NMR, is an 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. 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 noncomplementary duplexes two NOE paths exist. They will be called the *original paths*.

The NOE interactions between protons are represented as cross-peaks in the 2D–NOESY spectrum generated for the molecule during 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 $2 \cdot M - 1$ (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 centre, 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 lay vertically or horizontally. Fig. 1 demonstrates an exemplary NOE pathway found in the analyzed region of the 2D–NOESY spectrum (b) and the corresponding magnetization transfer pathway drawn in RNA chain (a).

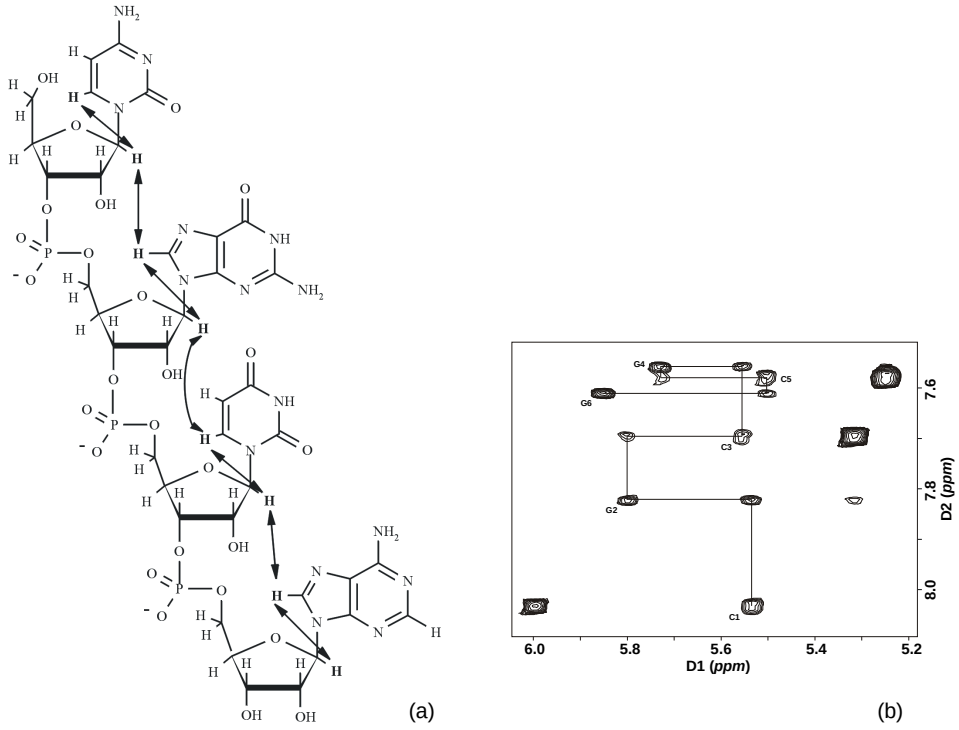

Fig. 1. Main NOE interactions in RNA molecule (a) and tthe corresponding NOE path in the spectrum (b).

Respecting the biochemical description of the problem, its graph-theoretic model, being a background for the complexity analysis and for a construction of the algorithms solving the problem has been proposed. The process of sequential assignments of H6/H8–H1' corresponds to a construction of a path between graph vertices. 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, that can be suggested during NOE pathway reconstruction, define edges of the graph. The following definition characterizes a new type of a graph representing the selected region of 2D–NOESY spectrum 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,

(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.

Let us notice that in general, a NOESY graph may not represent any specific experimental NOESY spectrum. Such a case is called a *theoretical model* of the problem. Otherwise, an *experimental* (or *real*) *model* of the problem (i.e. graph representing the spectrum) is considered. It is important to realize that in the theoretical model, NOE pathway may not exist in the graph, while in the experimental model it always exists.

On the basis of the problem graph-theoretical model, NOE path has been defined in terms of graph theory as well:

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:

(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 nor vertical line,

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

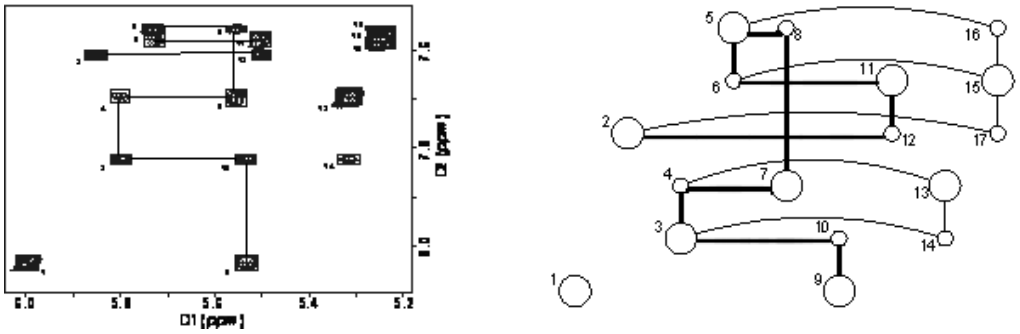

Fig. 2. NOE path found in 2D–NOESY spectrum (a) of r(CGCGCG)2 and - in the corresponding NOESY graph (b).

On the basis of the graph model of the problem, its computational complexity has been analyzed. It has been proved, that the *theoretical problem* of the NOE path construction in the NOESY graph in its search version is *strongly NP–hard*, while the decision version of the problem in question is *strongly NP–complete*. In the proof, it has been shown that the considered problem is in NP and next, the decision version of the strongly NP-complete Hamiltonian path problem has been polynomially transformed to a decision version of the NOE path problem. Polynomial transformation has been defined as a reduction $R:G_H \rightarrow G$, where $G_H=(V_H,E_H)$ is an instance of the Hamiltonian path problem and $G=(V,E)$ denotes an instance of the NOE path problem. Additionally, it was assumed that graph $G_H=(V_H,E_H)$ has no self-loops and no vertex with degree exceeding three. Hamiltonian path problem restricted in such a way remains strongly NP–complete.
The reduction $R:G_H \rightarrow G$ proceeds in the following way:

1. For every vertex $w_i \in V_H$ place the corresponding vertex $v_i \in V$ on a plane at the point of coordinates (i,i) and assign to it the weight equal 1. Consequently, coordinates of vertex $v_i \in V$ satisfy the equation f(x)=x.

2. For every edge $e_j=(w_p,w_k)\in E_H$, construct a square subgraph as shown in Figure 3 and place it in graph G between the appropriate vertices $v_p\in V$ and $v_k\in V$ (corresponding to w_p, $w_k\in V_H$).
3. Assume the following coordinates of the vertices: $v_{jt} = (p,k)$, $v_{jd} = (k,p)$. Let us observe that edges e_{jt}^1 and e_{jt}^2, as well as e_{jd}^1 and e_{jd}^2, respectively, are perpendicular to each other.
4. Assign weights equal 0 to vertices v_{jt} and v_{jd}.

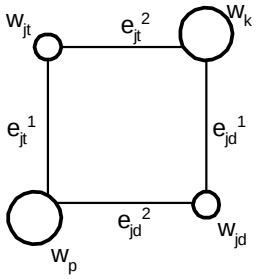

Fig. 3. A square subgraph.

Proving strong NP-completeness of the problem decision version in the theoretical model has not finished the discussion on computational complexity of the whole problem. Since NOE pathway always exists in the NOESY spectrum obtained from NMR experiment, the decision version of NOE path problem in the real model appeared trivial. The question "whether or not NOESY graph representing spectrum *S* contains a NOE pathway?" will be always answered "yes". Thus, the search version of the problem in the real model has been defined as a promise problem (promise: spectrum contains at least one NOE pathway) and it has been proved that such a problem is NP-hard.

Algorithms and Experiments

In the presented thesis, three new algorithms for NOE pathway reconstruction have been proposed and implemented in C programming language: exact, tabu search and evolutionary algorithm. Exact algorithm performs a complete search on the search space of the problem and enumerates all solutions being feasible NOE pathways. Feasible pathway has been characterized on the basis of Definition 2, by removing conditions which are satisfied only in ideal cases and by adding some constraints resulting from supplemental expert knowledge (doublets, overlapping, spectrometer resolution, H5/H6 signals, length of the path, etc.). The enumerative method is based on a Hamiltonian path construction procedure. Using domain expert knowledge it introduces additional constraints that limit the search space to the reasonable proportions.

All the algorithms have been tested on the same set of NMR spectra acquired for the following molecules: I - $r(CGCGCG)_2$, II - 2'–$OMe(CGCGCG)_2$, III, IV - $r(CGCG^FCG)_2$, V - $d(GACTAGTC)_2$, VI - $r(GGCAGGCC)_2$, VII - $r(GAGGUCUC)_2$, VIII - $r(GGCGAGCC)_2$, and IX, X - $r(GGAGUUCC)_2$. Two tests T1, T2 have been performed for each instance. In the first test (T1) algorithms used all available supplemental data, thus, processing much reduced search space. In the second case (T2) a minimum amount of expert supplemental information has been considered. In both heuristic procedures, feasible solution is evaluated by a goal (criterion) function *f* according to a set of criteria, like length of the pathway, edge deviations, alternative appearance of vertices with different weights, etc. Both heuristics tend to maximize a number of cross-peaks in the solution and minimize edge deviations, inconsistency in neighboring cross-peaks alternative appearances as well as cross-peaks incompatibility with such predefined conditions like known positions within the pathway or H5–H6 signals. A feasible solution with the best score has been defined as the optimal solution. The global criterion function *f* has been constructed as a weighted sum combining a set of different criteria:

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

Computational tests have been performed on Indigo 2 Silicon Graphics workstation (1133 MHz, 64 MB) in IRIX 6.5 environment. They have shown that all the algorithms work fast. In many cases the results are obtained after 1–5 seconds. This is very important, especially when we recall that, the optimization version of the problem of the NOE pathways reconstruction is NP–hard. Fortunately, the NOESY graphs created upon the 2D–NOESY spectra belong to the class of sparse graphs, thus, the cardinality of their edge set is rather small, which considerably reduces the time of computations.

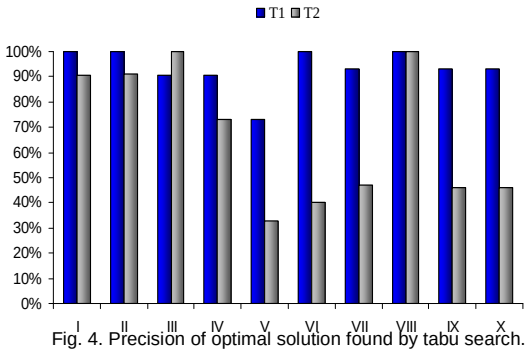

Fig. 4. Precision of optimal solution found by tabu search.

Apart from computation time, quality of optimal solutions constructed by heuristic algorithms has been considered. The value of a solution precision has been calculated as a percentage of the original path covered by the best solution (i.e. percentage of edges from original path which included in the optimal solution). Figure 4 presents the results of solution quality analysis for tabu search.

Computational experiments show that in test T1 both heuristics give very good results, while results of test T2 are worse and depend on the specificity of the input data. In practice however, supplying additional expert data is usually possible. Thus, the proposed methods for NOE pathway reconstruction are much better than the manual procedure, since they give the possibility to analyze NMR spectra obtained for bigger RNA molecules.

Conclusions

In the thesis, the problem of automatic resonance assignment in 2D–NOESY NMR spectra of RNA molecules has been considered. The problem appears at the beginning of the process of molecule tertiary structure determination with NMR techniques. So far, no automatic method for resonance assignment has existed and the assignment of cross-peaks in the 2D–NOESY spectra of nucleic acids was accomplished by hand with a help of interactive graphics. Thus, it was crucial to propose automatic procedures at this level of structural study of RNA molecules.

To solve the problem the theoretical aspect of it has been first analyzed. An examination has resulted in constructing a graph-theoretic model representing 2D–NOESY spectra of RNA (or DNA) molecules and NOE path being the principle of resonance assignment.

Formulation of the assignment problem on the basis of graph theory has been followed by an analysis of the computational complexity of the problem in question. It has been proved that the search version of NOE path assignment belongs to NP–hard class of problems. Thus, a branch-and-cut algorithm for automatic signal assignment, enumerating all feasible NOE paths, has been designed, implemented and applied to a set of real data. Analyzing results of computational experiments performed with this algorithm, one can notice that it constructed surprisingly small number of alternative pathways in tests with supplemental data provided, thus, proving its high accuracy in these cases. Since analysis of bigger molecules implies processing of more crowded spectra, designing the other method that could deal with large instances of the problem in a short time seemed of a great importance. In consequence, two heuristics, tabu search and evolutionary algorithm, generating optimal solution (being an approximation of the original NOE path), have been proposed. Both appeared very useful in practice, when a number of feasible paths returned by enumerative algorithm for the instances without the additional knowledge was large. Even if the heuristic algorithm finds only half of the original path it facilitates the assignment problem to a very large degree. Having the partial assignment, an experimenter is able to complete the NOE pathway in a reasonable time without an extreme effort.

During computational experiments, quality of optimal solutions, generated by heuristic algorithms, has been considered. One can see that both algorithms are quite precise for most instances. Especially tabu search seems to give superior results and for most instances the obtained solutions coincide with the majority of vertices in the original path.

Finally, let us notice that all the algorithms perform quite fast. Detailed analysis of the NOESY graphs makes one 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. Computation time will be crucial in case of an analysis of longer nucleic chains, for which manual assignment is a hard and tiresome work, usually impossible to be done in a period of days or even weeks.

Tabu search is the fastest of all the tested methods designed for an automatic assignment of NOE pathways. Thus, its application to more complicated cases and instances as well as to an analysis of long nucleic chains seems promising. However, any other tool (enumerative or evolutionary algorithm) that can facilitate this analysis is of great importance. The designed algorithms might be also useful when applied to a verification of the assignment correctness.

As a continuation of the research reported in the thesis, one may consider the analysis of spectra which contain a lot of noise signals as well as three dimensional spectra of RNA molecules. Especially 3D NMR spectra deserve a special attention. These spectra represent a wider range of interactions than their 2D equivalents. Thus, they carry more information about the structure and help in more precise determination of input sample characteristics. Furthermore, it seems evident that 3D and finally d-dimensional (d>3) NMR spectra analysis will be considered in the continuation of this research. Solving the problem of finding NOE path on the basis of 2D–NOESY, NMR spectrum appears to be a good platform for this purpose. As it was demonstrated, however, the problem of finding NOE paths in 2D spectra has been already troublesome. Consequently, we should expect that adding one or more dimensions into the search space will complicate the searching algorithms.

Let us stress that the proposed method of NOE pathway reconstruction has been the subject of Polish patent pending application. Moreover, main results of the thesis have been published and presented on many seminars and conferences around the world.