181. A combinatorial analysis of 2D NMR spectra of RNA duplexes [1]

Marta Szachniuk[2], Ryszard W. Adamiak[3], Piotr Formanowicz[23], Zofia Gdaniec[3], Marta Kasprzak[23], Mariusz Popenda[3], Jacek Błażewicz[23]

Keywords: automatic assignments of NMR spectra, 2D-NOESY spectra, RNA duplexes, computational complexity, heuristic algorithms

1 Introduction

Nuclear magnetic resonance (NMR) spectroscopy has been now well established as a method for structure determination of biomolecules in RNA duplexes and protein chain. [4]. The procedure is composed of two general stages, i.e. the experimental one, where multidimentional correlation spectra are acquired, and the computational one, where the spectra are analysed and the structure is determined. Types of NMR experiments to be chosen differ for proteins and nucleic acids. Quality and quantity of the experimental data obtained influence very strongly the computational stage. Nevertheless, in all types of NMR structure analysis the following steps have to be performed on raw experimental data: processing, peak-picking, assignment, restraints determination, structure generation and refinement.

The assignment of the observed signals to corresponding protons and other nuclei is a bottleneck of the structure elucidation process. For non-labeled small proteins and short DNA and RNA duplexes the assignment of NMR signals is usually based on the analysis of 2D spectra like NOESY, TOCSY and COSY. Due to a 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 automation on this level [2], the case of RNA chains being far more complicated.

In this work we propose a new algorithm for an automatic generation of paths between H6/H8 and H1' resonances observed for short RNA duplexes in a 2D-NOESY spectra. It reduces the NOE paths 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 strongly NP-hard. Hence, a heuristic 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 proposed algorithm proves its high efficiency and surprisingly good predictive power highly exceeding the existing approaches [3].

2 The combinatorial model and the algorithm

The problem of NMR spectra analysis (cf. [1]) may be modeled as the following graph-theoretic problem.

Let us consider undirected graph $G = (V, E)$ on a Cartesian plane, where V is a set of vertices

[1]The research has been partially supported by KBN grant 7T11F02621.

[2]Institute of Computing Science, Poznań University of Technology, Piotrowo 3A, 60-965 Poznań Poland. E-mails: Marta.Szachniuk@cs.put.poznan.pl, piotr@cs.put.poznan.pl

[3]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12/14, 61-704 Poznań, Poland.

and E is a set of edges. Furthermore, let us assume that G has the following properties:
1) every vertex $v_i \in V$ corresponds to one cross-peak from the spectrum;
2) vertices are weighted: weight 1 is assigned to every vertex representing intranucleotide NOE, and 0 - to every vertex representing internucleotide NOE;
3) the number of vertices in G is equal to the number of cross-peaks in the spectrum;
4) every edge $e_{ij} \in E$ represents a possible connection between two cross-peaks with different intensities having one coordinate in common (thus G contains only horizontal and vertical edges);
5) the number of edges in G is equal to the number of all possible correct connections (i.e. lines between two cross-peaks of different intensities having one coordinate in common) that may be drawn in the spectrum.
We will call graph G a *NOESY graph*.

The aim of the spectral analysis is finding a H8/H6$_{(i)}$-H1'$_{(i)}$-H8/H6$_{(i+1)}$ path in 2D-NOESY spectra of RNA duplexes. It corresponds to looking for a NOE path in the NOESY graph. The path may be characterized similarly to the magnetization transfer path in the spectrum, i.e.: every vertex end every edge may occur in the path at most once, every two neighboring edges have to be perpendicular, no two edges lie on the same horizontal or vertical line, and the number of edges in the path equals $2|V_1| - 2$, where $|V_1|$ is the number of intranucleotide signals.

The following theorem concerning the computational complexity of the considered problem may be proved [1]:
THEOREM: The problem of finding a NOE path in a NOESY graph is strongly NP-hard. □

Since the problem is computationally intractable a need of finding good suboptimal algorithms has arisen and a heuristic algorithm that automatically groups H6/H8-H1' cross-peaks of the nucleotide residues according to their position in the sequence has been proposed. The algorithm is based on a Hamiltonian path construction procedure and uses domain expert knowledge in order to take into account additional constraints that limit the search space.

The algorithm has been implemented in C language and tested on Silicon Graphics Indigo 2 workstation. As a testing instances a set of experimental and simulated 2D-NOESY spectra has been used. All the instances have been already solved manually, hence the verification of the correctness of the results provided by the algorithm was possible. Moreover, it was also possible to examine the way the expert knowledge influences the algorithm output. Minimal expert knowledge has been used in every example, but it should be noticed that in some cases such knowledge is necessary for an appropriate interpretation of the input data. The analysis of the results provided by the algorithm showed that it gives very good results in the situation of some expert knowledge availability. Let us notice, that even a small amount of information about the analysed chain results is a significant reduction of the final solution set.

References

[1] Adamiak, R. W., Błażewicz, J., Formanowicz, P., Gdaniec, Z., Kasprzak, M., Popenda, M. and Szachniuk, M. 2002. An algorithm for an automatic NOE pathways analysis of 2D NMR spectra of RNA duplexes, *submitted for publication*.

[2] Popenda, M., Biała, 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. *Nucleic Acid Research* 25:4589-4598.

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

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